International Journal of Statistics and Analysis.
ISSN 2248-9959 Volume 6, Number 1 (2016), pp. 9-16

http://www.ripublication.com

Estimation of Finite Population Total under PPS Sampling in Presence of Extra Auxiliary Information

P. A. Patel[1] and Shraddha Bhatt[2]

Department of Statistics, Sardar Patel University,
Vallabh Vidyanagar 388120, India.
[1,2]E-mail:patelpraful_a@yahoo.co.in, shraddha.bhatt83@gmail.com

Abstract

This article deals with probability proportion to size (pps) estimation of the population total incorporating auxiliary information at estimation stage via model-based approach. An optimal estimator is obtained. Motivated from the optimal estimator few estimators are suggested and compared empirically with the conventional estimator.

Keywords: Auxiliary information, Model-based estimation, Probability proportional to size,

Subject Classification: 62D05

INTRODUCTION

It is well-known that when the survey variable and selection probabilities are highly correlated, pps estimators of the population mean leads to considerable gain in efficiency as compared with the customary estimator simple mean for equal probability sampling. The pps estimator depends on multiplicity but not on the order and hence is inadmissible. An improve estimator is then available by applying the Rao-Blackwell theorem. Rao-Blackwellization of this estimator, considered by Pathak (1962), yields a rather complicated estimator which does not admit a simple variance estimator as does the pps estimator. Moreover the gain in efficiency is considered to be small, unless the sampling fraction is large (see, Cassel et al., 1977). Thus, the resulting estimator is less useful in practice than the original pps estimator. In this paper, alternative estimators for estimating population mean under pps sampling are suggested.

Let U be a finite population of size N. Let (y_i, z_i) be pair of values of the study variable y and an auxiliary variable z associated with each unit $i \in U$. Let $s \subset U$ be a sample of size n drawn according to probability proportional to size (pps) z, $p_i \propto z_i$, and with replacement (ppswr) sampling design $p = p(s)$. Suppose that the values y_i, $i \in s$, and z_i, $i \in U$, are known. The problem is how to use this information to make inference about the finite population total $Y = \sum_{i \in U} y_i$. If $A \subseteq U$, we write $\sum_A$ for $\sum_{i \in A}$ and $\sum\sum_A$ for $\sum\sum_{i \neq j \in A}$. The customary design-based unbiased estimator of Y which makes no use of auxiliary information at the estimation stage is the Hansen-Hurwitz (HH) (1943) estimator

$$\hat{Y}_{HH} = \sum_s y_i / n p_i \tag{1}$$

The sampling variance of $\hat{Y}_{HH}$ is given by

$$V(\hat{Y}_{HH}) = \frac{1}{n} \sum_U p_i \left(\frac{y_i}{p_i} - Y \right)^2 \tag{2}$$

A design-unbiased estimator of $V(\hat{Y}_{HH})$ is given by

$$v(\hat{Y}_{HH}) = \frac{1}{n(n-1)} \sum_s \left(\frac{y_i}{p_i} - \hat{Y}_{HH} \right)^2 \tag{3}$$

In survey sampling, auxiliary information about the finite population is often available at the estimation stage. Utilizing this information more efficient estimators may be obtained. There exist several approaches, such as model-based, calibration, etc., each of which provides a practical approach to incorporate auxiliary information at the estimation stage. Here, we will use the following model as working model.

Model GT. Assume that $y_1, \ldots, y_N$ are random variables having joint distribution ξ (see, Cassel et al., 1977, p.102) with

$$E_\xi(y_i) = a_i \mu + b_i$$
$$V_\xi(y_i) = a_i^2 \sigma^2$$
$$C_\xi(y_i, y_j) = a_i a_j \rho \sigma^2 \quad (i \neq j)$$

where μ, $\sigma > 0$ and $\rho \in (-(N-1)^{-1}, 1)$ are the parameters, and $a_i > 0$, b_i are known numbers. Here $E_\xi(\cdot)$, $V_\xi(\cdot)$ and $C_\xi(\cdot, \cdot)$ denote ξ–expectation, ξ-variance and ξ-covariance, respectively. We try to make as efficient use of auxiliary information as possible through model. To find an optimal (in the sense minimum $E_\xi E_p (T - Y)^2$) strategy (a combination of sampling design and estimator), given a model ξ, we minimize the

$$E_\xi V_p(T) = E_\xi E_p(T-Y)^2 = E_p\{V_\xi(T)\} + E_p\{B_\xi(T)\}^2 \quad (4)$$

subject to $E_\xi E_p(T) = E_\xi(Y)$, where $B_\xi(T) = E_\xi(T-Y)$.

A design-model-based (i.e. $p\xi$-) estimation of the population total *Y* under ppswr is considered in Section 2. The optimal $p\xi$-unbiased estimator depends on unknown model parameters. Motivated by this in Section 3 we have suggested a few pps-type estimators for *Y* and obtained their approximate bias and variance. In Section 4 we demonstrated through a limited simulation study the improved performance of the proposed estimators over their convention counterpart.

THE OPTIMAL ESTIMATOR

Consider a linear predictor T of *Y* as

$$T = w_{0s} + \sum_s w_{si} y_i$$

where w_{0s} and w_i are constants free from y-values. The predictor T is p-unbiased if $E_p(T) = Y$ for all $\underline{y} = \{y_1, \dots, y_N\} \in \mathbb{R}^N$ and $p\xi$-unbiased if, for every *s*, $E_\xi E_p(T) = E_\xi(Y)$.

Under Model G_T, $p\xi$-unbiasedness implies that

$$\sum_U p_i w_i a_i = \sum_U a_i/n \text{ and } E_p(w_{0s}) = \sum_U b_i - \sum_U p_i w_i b_i$$

Denoting $y_i' = y_i - b_i, i = 1, \dots, N, Y' = \sum_U y_i'$ and$T' = \sum_s w_{si} y_i'$, under Model G_T, we obtain

$$E_\xi V_p(T') = n[(\sigma^2+\mu^2)\sum_U p_i(1-p_i)w_i^2 a_i^2 + (\rho\sigma^2 + \mu^2)\{\sum_U p_i^2 w_i^2 a_i^2 - n(\sum_U a_i)^2\}]$$

Minimization of (4) (equivalently minimization of $E_p\{V_\xi(T)\}$ since $B_\xi(T) = 0$) subject to the conditions $p\xi$-unbiasedness yields the optimal value of w_{si} as

$$w_{si}^* = \frac{1}{nQ} \cdot \frac{1}{\{(\sigma^2+\mu^2)(1-p_i) + (\rho\sigma^2+\mu^2)p_i\}} \cdot \frac{1}{a_i/\sum_U a_i}$$

Therefore, the optimal predictor of Y' in the class of $p\xi$-unbiased linear predictor is obtained as

$$T'^* = \sum_s w_{si}^* y_i' = \sum_s y_i'/np_i^*$$

where

$$Q = \sum_U [p_i/\{(\sigma^2+\mu^2)(1-p_i) + (\rho\sigma^2+\mu^2)p_i\}]$$

and

$$p_i^* = Q[\{(\sigma^2 + \mu^2)(1 - p_i) + (\rho\sigma^2 + \mu^2)p_i\}]\frac{a_i}{\sum_U a_i}$$

Finally, the optimal predictor of $Y = Y' + \sum_U b_i$ is obtained as

$$\begin{aligned} T^* &= \sum_s w_{si}^*(y_i - b_i) + \sum_U b_i \\ &= \sum_s \frac{y_i}{np_i^*} + \left(\sum_U b_i - \sum_s \frac{b_i}{np_i^*}\right) \end{aligned} \tag{5}$$

In particularly, if $\rho = 0, b_i = 0\ \forall\ i,\ a_i = p_i,$ then (5) reduces to the estimator suggested by Arnab (2004)

$$T_1^* = \sum_s \frac{y_i}{np_{1i}^*}$$

where $p_{1i}^* = \{\delta(1 - p_i) + 1\}p_i \cdot \sum_U [p_i/\{\delta(1 - p_i) + 1\}]$ with $\delta = \sigma^2/\mu^2$.

It is interesting to note that if we let $\rho = 0, b_i = cx_i\ \forall\ i,\ a_i = p_i$ and c is known scalar, in (5) the resulting estimator can be viewed as a generalized difference estimator

$$T_2^* = \sum_s \frac{y_i}{np_{2i}^*} + c\left(X - \sum_s \frac{x_i}{np_{2i}^*}\right)$$

where $p_{1i}^* = p_{2i}^*$. In this article we shall not focus on T_2^*.

Remark 1. The optimal estimator T^* involves model parameters and hence is useful only when these parameters are known. In such situation auxiliary information can be effectively used through the fitted values.

THE PROPOSED ESTIMATORS

We assume further that the data $\{x_i, i \in s\}$ are observed. Here x_i is the value for unit i of an extra auxiliary variable x whose total, $X = \sum_U x_i$, and coefficient of variation (cv), C_x, are assumed to be known from a reliable source.

In Model G_T, inserting $\rho = 0,\ a_i = x_i$ and $b_i = 0\ \forall\ i,$ we obtain from T^* the model-assisted optimal estimator of Y as

$$\hat{Y}_1 = \sum_s \frac{y_i}{n\{\delta(1-p_i)+1\}(x_i/X)\cdot\sum_U[p_i/\{\delta(1-p_i)+1\}]}$$

where

$$\delta = \sigma^2/\mu^2 = \sigma^2 a_i^2/\mu^2 a_i^2 = V_\xi(y_i)/\left(E_\xi(y_i)\right)^2 = \left(CV_\xi(y_i)\right)^2$$

is assumed to be known. Often this is difficult. This motivates to suggest the following estimators.

$$\hat{Y}_2 = \sum_s \frac{y_i}{n\{C_x^2(1-p_i)+1\}(x_i/X)\cdot\sum_U[p_i/\{C_x^2(1-p_i)+1\}]} \quad (6)$$

$$\hat{Y}_3 = \sum_s \frac{y_i}{n\{c_y^2(1-p_i)+1\}(x_i/X)\cdot\sum_U[p_i/\{c_y^2(1-p_i)+1\}]} \quad (7)$$

$$\hat{Y}_4 = \sum_s \frac{y_i}{n\left\{\left(c_y C_x/c_x\right)^2(1-p_i)+1\right\}\frac{x_i}{X}\cdot\sum_U\left[p_i/\left\{\left(c_y C_x/c_x\right)^2(1-p_i)+1\right\}\right]} \quad (8)$$

$$\hat{Y}_5 = \sum_s \frac{y_i}{n\left\{\left(c_y+b(C_x-c_x)\right)^2(1-p_i)+1\right\}\frac{x_i}{X}\sum_U\left[p_i/\left(c_y+b(C_x-c_x)\right)^2\{(1-p_i)+1\}\right]}$$

where c_y and c_x denote coefficient of variations of *y* and *x* variables and *b* denotes regression coefficient between c_y and c_x based on pps sample. Obviously, the exact bias and exact mean square error (MSE) of $\hat{Y}_i, i = 3, 4, 5$, are hard to obtain.

Approximate Bias and MSE of the Proposed Estimators

Suppose that $p_i^*, i \in U$ are the revised probabilities of selection and consider

$$\hat{Y}^* = \sum_s y_i/np_i^*$$

as a pps estimator of the population total *Y*. The exact bias and exact variance of $\hat{Y}^*$ are obtained as

$$B\left(\hat{Y}^*\right) = \sum_U \frac{y_i}{p_i^*}p_i - Y \quad (9)$$

and

$$V\left(\hat{Y}^*\right) = \frac{1}{n}\left[\sum_U \frac{y_i^2}{p_i^{*2}}p_i - \left(\sum_U \frac{y_i}{p_i^*}p_i\right)^2\right] \quad (10)$$

Since, for sufficiently large n, c_y , $c_y C_x/c_x$ and $c_y + b(C_x - c_x)$ are asymptotically consistent and asymptotically unbiased estimators for C_y (Patel & Shah, 2009) the approximate bias and variance of each of $\hat{Y}_i, i = 2,3,4,5$, can be obtained using (9) and (10) with

$$p_i^* = \{C_y^2(1-p_i)+1\}(x_i/X) \cdot \sum_U [p_i/\{C_y^2(1-p_i)+1\}]$$

AN EMPIRICAL STUDY

The estimators $\hat{Y}_{HH}, \hat{Y}_2, \hat{Y}_3$ and $\hat{Y}_4$ given at (1), (6), (7) and (8) were compared empirically on 3 natural populations given below. For comparison of the estimators, a sample was drawn using pps sampling from each of the populations and these estimators were computed. These procedure was repeated $M = 5000$ times. For an estimator $\hat{Y}$, its relative percentage bias (RB%) was calculated as

$$RB(\hat{Y}) = 100 * \left(\overline{\hat{Y}} - Y\right)/Y$$

and the relative efficiency in percentage (RE%) as

$$RE(\hat{Y}) = 100 * MSE_{sim}(\hat{Y}_{HH})/MSE_{sim}(\hat{Y})$$

where

$$\overline{\hat{Y}} = \sum_{j=1}^{M} \hat{Y}_j/M \text{ and } MSE_{sim}(\hat{Y}) = \sum_{j=1}^{M} (\hat{Y}_j - Y)^2/(M-1)$$

Data set I: Murthy (1967)

y : Output for factories, x : Number of workers, z : Fixed capital

$\rho_{yz} = .9149,\ \rho_{yx} = .9413$

Data set II: Murthy (1967)

y : Number of cultivators, x : Number of persons, z : area in sq. miles

$\rho_{yz} = .6611,\ \rho_{yx} = .8311$

Data set III: Fisher (1936) (combined all three data sets)

y : Patel width x : Sepal length z : Patel length

$\rho_{yz} = 0.8179$, $\rho_{yx} = .9628$

The simulated results are presented in the following table.

Table 1: Relative bias and Mean Square Error

Data Set	Population Total	N	n	Estimator	Estimate	RB%	MSE_{sim}	RE%
I	414611	80	10	$\hat{Y}_{HH}$	466632.5	12.54	1.1E+10	-
				$\hat{Y}_2$	413330.7	- 0.31	2.75E+09	400
				$\hat{Y}_3$	415116.5	0.12	2.94E+09	374
				$\hat{Y}_4$	411429.9	- 0.76	2.94E+09	374
II	109248	128	20	$\hat{Y}_{HH}$	167792	53.59	5.6E+09	-
				$\hat{Y}_2$	109099	-0.14	79893656	7011
				$\hat{Y}_3$	109141	-0.10	79829722	7017
				$\hat{Y}_4$	109119	-0.12	80058668	6997
III	179.9	150	20	$\hat{Y}_{HH}$	213.307	18.57	1437.936	-
				$\hat{Y}_2$	180.126	0.125609	96.8654	1484
				$\hat{Y}_3$	180.127	0.126236	96.8312	1485
				$\hat{Y}_4$	180.117	0.120456	96.8243	1485

The above simulation reveals that (1) the absolute RBs % of the suggested estimators are in reasonable range and (2) the efficiency of the suggested estimator is substantial as compared to the conventional estimator.

REFERENCES

[1] Arnab R. (2004). Optimum estimation of a finite population total in PPS sampling with replacement for multi-character surveys, Jour. of Indian Soc. Agri. Statist., 58 (2), 231-43

[2] Cassel, C. M., Särndal, C. E., and Wretman, J. H. (1977). Foundation of Inferencein Survey Sampling., New York, John Wiley.

[3] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problem, Annals of Eugenics, 7, 179-188.

[4] Hansen, M. H. and Hurwitz, W. N. (1943). On the thory of sampling from finitepopulations. Ann. Math. Statist. 14, 333–362.

[5] Murthy, M. N. (1967). Sampling Theory and Methods, Statistical Publishing House:Culcutta.
[6] Patel, P. A. and Shah, R. M. (2009). A Monte Carlo comparison of some suggested estimators of coefficient of variation in finite population, Journal of Statistics Sciences, 1 (2), 137-48.
[7] Pathak P. K. (1962). On sampling with unequal probabilities. Sankhya A 24, 315-326.