A Model-based Estimation of Finite population Variance under PPS Sampling

P. A. Patel\(^1\) and Shraddha Bhatt\(^2\)

Department of Statistics, Sardar Patel University, VallabhVidyanagar 388 120

Abstract: This article deals with estimation of the population variance under pps sampling incorporating auxiliary information at estimation stage through model-based approach. An optimal estimator is obtained and compare empirically with the conventional variance estimator.

Key Words: Model-based estimation, Finite population variance, Probability proportional to size, U-statistics

Subject Classification: 62D05

1 Introduction

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. This article deals with estimation of the population variance incorporating auxiliary information at estimation stage through model-based approach.

Let U be a finite population of size N. Let \( S \subset U \) be a sample of size n drawn according to a known probability sampling design \( p(s) \) (non-informative) with positive first and second order inclusion probabilities \( \pi_i \) and \( \pi_{ij} \). Let \( (y_i, x_i) \) be pair of values associated with each unit \( i \in U \). Suppose that the values \( y_i, i \in S \), and \( x_i, i \in U \), are known. The problem is how to use this information to make inference about the finite population total \( Y = \sum_{i \in U} y_i \) (or mean \( \bar{Y} \)) or variance \( \sigma^2_Y = \frac{1}{N} \sum_{i \in U} (y_i - \bar{Y})^2 \). If \( A \subseteq U \), we write \( \sum_A \) for \( \sum_{i \in A} \) and \( \sum_A \) for \( \sum_{i \in A} \sum_{j \in A} \). The customary design-based unbiased estimator of Y which makes no use of auxiliary information at the estimation stage is the Horvitz-Thompson (HT) estimator

\[
\hat{Y}_{HT} = \frac{1}{n} \sum_{i \in S} y_i / \pi_i
\]

with (Horvitz and Thompson, 1952) variance

\[
V_{HT} = \sum_{i \in U} \Delta_i y_i^2 + \sum_{i \neq j} \Delta_{ij} y_i y_j,
\]

for which an unbiased estimator is given by

\[
\hat{V}_{HT} = \sum_{i \in S} \Delta_i y_i^2 / \pi_i + \sum_{i \neq j} \Delta_{ij} y_i y_j / \pi_{ij}
\]

where \( \Delta_{ij} = \pi_i^{-1} - 1 \) if \( i = j \) and \( = \pi_{ij} / \pi_i^{-1} / \pi_j^{-1} - 1 \) if \( i \neq j \).

There are two difficulties associated with HT estimator. (i) When both the pps requirement and the fixed sample size requirement are imposed on the sampling scheme it is tedious and often computationally difficult to calculate \( \pi_{ij} \) satisfying \( \pi_i \pi_j - \pi_{ij} < 0 \). (ii) The design-based variance estimation uses the \( \pi_{ij} \) in a cumbersome double-sum calculation with \( n(n-1)/2 \) terms. This very large number of terms effectively rules out correct variance calculation to many pps surveys.

Unequal probability sampling was first suggested by Hansen and Hurwitz (HH) (1943) in the context of with-replacement (wr) sampling. Variance estimation
for rr sampling requires no heavy calculations. This simplicity speaks in favour of the HH estimator and probability proportional to size (PPS) sampling. For a \( \pi ps \)-sampling the HH and HT estimators are identical, and so, sometimes, one can use HT estimator to estimate the population total or mean and can use variance estimator of the HH estimator to estimate the variance of the HT estimator. In this article we focus on the estimation of population quadratic function under pps sampling which includes the estimation of the variance of the linear estimator and estimation of the finite population variance.

\[
\sigma_y^2 = \frac{1}{N} \sum_{i \in U} (y_i - \bar{Y})^2
\]  

(1)

using pps sampling with selection probabilities \( p_1, ..., p_N \). \( \sigma_y^2 \) can alternatively be written as

\[
\sigma_y^2 = \frac{1}{N} \left( 1 - \frac{1}{N} \right) \sum_{i \in U} y_i^2 - \frac{1}{N^2} \sum_{i \in U} \sum_{j \in U} y_i y_j
\]  

(2)

The usual unbiased estimator of \( \sigma_y^2 \) can by obtained by estimating unbiasedly the two terms in \( \sigma_y^2 \). Thus the unbiased estimator of \( \sigma_y^2 \) is given by

\[
\hat{\sigma}_y^2 = \frac{1}{N} \sum_{i \in s} y_i^2 - \frac{1}{N^2} \sum_{i \in s} \sum_{j \in s} y_i y_j
\]  

(3)

This estimator can be obtained in different ways as follows.

Writing

\[
\sigma_y^2 = \frac{1}{N} \sum_{i \in U} y_i^2 - \bar{Y}^2
\]

an unbiased estimator of \( \sigma_y^2 \) can be written as

\[
\hat{\sigma}_y^2 = \text{Est} \left[ \frac{1}{N} \sum_{i \in U} y_i^2 \right] - \left[ \text{Est} \bar{Y} \right]^2 + \text{Est}[V(\text{Est} \bar{Y})]
\]  

(4)

Noting that \( \bar{Y}^2 = E \left( \bar{Y}^2 \right) - V(\bar{Y}) \) for unbiased estimator, under pps sampling

\[
\text{Est} \left[ \frac{1}{N} \sum_{i \in U} Y_i^2 \right] = \frac{1}{N} \sum_{i \in s} y_i^2
\]

\[
\text{Est} \bar{Y} = \bar{Y}_{HH} = \frac{1}{N} \sum_{i \in s} y_i / np_i
\]

\[
\text{Est} V(\bar{Y}_{HH}) = \text{Est} \left[ \frac{1}{n} \sum_{i \in s} p_i \left( \frac{y_i}{np_i} - \bar{Y} \right)^2 \right] = \frac{1}{n(n - 1)} \sum_{i \in s} \left( \frac{y_i}{np_i} - \bar{Y}_{HH} \right)^2
\]
Inserting all these in (4) we obtain the unbiased estimator of \( \sigma^2 \) as

\[
\hat{\sigma}^2 = \frac{1}{N} \sum_{i \in s} y_i^2 n p_i - \left[ \frac{1}{N} \sum_{i \in s} y_i n p_i \right]^2 + \frac{1}{n(n-1)} \sum_{i \in s} \left( \frac{y_i}{N p_i} - \bar{Y}_{HH} \right)^2
\]

which reduces on simplification to \( \hat{\sigma}^2_y \), given at (3).

Again, the variance expression of \( \bar{Y}_{HH} \) is alternatively written as

\[
V(\bar{Y}_{HH}) = \frac{1}{2nN^2} \sum_{i \not= j} \sum_{i \in U} p_i p_j \left( \frac{y_i}{p_i} - \frac{y_j}{p_j} \right)^2
\]

Its unbiased estimator can easily been seen as

\[
\hat{V}(\bar{Y}_{HH}) = \frac{1}{2n^2(n-1)N^2} \sum_{i \not= j} \sum_{i \in U} \left( \frac{y_i}{p_i} - \frac{y_j}{p_j} \right)^2
\]

Substituting this expression in (4) we obtain another estimator of \( \sigma^2 \) as

\[
\hat{\sigma}^2_y = \frac{1}{N} \sum_{i \in s} y_i^2 n p_i - \left[ \frac{1}{N} \sum_{i \in s} y_i n p_i \right]^2 + \frac{1}{2n^2(n-1)N^2} \sum_{i \not= j} \sum_{i \in U} \left( \frac{y_i}{p_i} - \frac{y_j}{p_j} \right)^2
\]

Surprisingly, this estimator is same as \( \hat{\sigma}^2_y \), given at (3). This is not the case in unequal probability sampling without replacement (see, Swain and Mishra, 1994).

2. Model-based Estimation of Quadratic Form

In this section we first consider estimation of a quadratic form

\[
Q = \sum_{i \in U} \beta_{ii} y_i^2 + \sum_{i \not= j} \beta_{ij} y_i y_j
\]

(5)

where \( \beta_i \)'s are known numbers independent of \( y \)-values and \( \beta_{ij} = \beta_{ji} \forall i, j \), using model-based approach.

An design-based unbiased quadratic estimator of \( Q \) is given by

\[
\hat{Q} = \sum_{i \in s} b_{ii} y_i^2 + \sum_{i \not= j} \sum_{i \in s} b_{ij} y_i y_j
\]

(6)

where \( b_{ij} (i, j \in s) \) are such that \( E(\hat{Q}) = Q \).

2.1 Model GT.

Assume that \( y_1, ..., y_N \) are random variables having joint distribution \( \xi \) (see, Cassel et al., 1977, p.102) with

\[
\begin{align*}
E_i(y_i) &= a_i \mu + b_i \\
V_i(y_i) &= a_i^2 \sigma^2 \\
C_i(y_i, y_j) &= a_i a_j \rho \sigma^2 \quad (i \neq j)
\end{align*}
\]

(7)
where \( \mu, \sigma > 0 \) and \( \rho \in (-N^{-1}, 1) \) are the parameters, and \( a_i > 0, b_i \) are known numbers. Here \( E(\cdot) \), \( V(\cdot) \) and \( C(\cdot, \cdot) \) denote \( \xi \)–expectation, \( \xi \)–variance and \( \xi \)–covariance, respectively. To find an optimal (in the sense minimum \( E_p(Q - \hat{Q})^2 \)) strategy (a combination of sampling design and estimator), given a model \( \xi \), we minimize

\[
E_p(Q - \hat{Q})^2 = E_p[V(\hat{Q})] + E_p[B(\hat{Q})]^2
\]

subject to \( E_p(\hat{Q}) = E(\hat{Q}), \) i.e., \( \hat{Q} \) is unbiased for \( Q \), where \( B(\hat{Q}) = E(\hat{Q} - Q) \).

Denoting \( \sum_{i \in \Omega} \beta_{ii} a_i^2 = C_1 \) and \( \sum_{i \neq j} \sum_{s \in \Omega} \beta_{ij} a_i a_j = C_2 \), \( p\xi \) – unbiasedness of \( \hat{Q} \) implies that

\[
\sum_{s \in \Omega} p(s) \sum_{i \in s} b_i a_i^2 = C_1 \quad \text{and} \quad \sum_{s \in \Omega} p(s) \sum_{i \neq j} \sum_{s \in \Omega} b_{ij} a_i a_j = C_2
\]

or equivalently

\[
\sum_{s \in \Omega} p(s) \sum_{i \in s} h_{ii} = C_1 \quad \text{and} \quad \sum_{s \in \Omega} p(s) \sum_{i \neq j} \sum_{s \in \Omega} h_{ij} = C_2 \quad (8)
\]

where \( h_{ii} = b_{ii} a_i^2 \) and \( h_{ij} = b_{ij} a_i a_j \).

\[
EV(\hat{Q}) = EV\left(\sum_{i \in s} b_i y_i^2 + \sum_{i \neq j} b_{ij} y_i y_j\right)
\]

\[
= E \left[ \sum_{i \in s} b_i^2 V(y_i^2) + \sum_{i \neq j} b_i b_{ij} C(y_i^2, y_j^2) + \sum_{i \neq j} \sum_{s \in \Omega} b_{ij}^2 V(y_i, y_j) + \sum_{i \neq j} \sum_{s \in \Omega} b_{ij} b_{kl} C(y_i^2, y_j y_k) + \sum_{i \neq j \neq k} \sum_{s \in \Omega} b_{ij} b_{kl} C(y_i y_j, y_k y_l) \right]
\]

To simplify the derivation we use (10.12) and (10.13) of Kendall et al. (1987) to obtain the approximate variances and covariances. Using (7) we find

\[
EV(\hat{Q}) = 4\mu^2 \sigma^2 \sum_{s \in \Omega} p(s) \left[ (1 - \rho) \sum_{i \in s} h_{ii}^2 + \rho \left( \sum_{i \in s} h_{ii} \right)^2 + \frac{1 - \rho}{2} \sum_{i \neq j} h_{ij}^2 + \rho \left( \sum_{i \neq j} h_{ij} \right)^2 + 2(1 - \rho) \sum_{i \neq j} h_{ij} h_{ij} + \text{constant} \right]
\]

Let us minimize
\[
\phi = EV(\hat{Q}) - 8\mu^2\sigma^2\lambda_1 \left( \sum_{s \in S} p(s) \sum_{i \in S} h_{ii} - C_1 \right) - 8\mu^2\sigma^2\lambda_2 \left( \sum_{s \in S} p(s) \sum_{i \neq j} h_{ij} - C_2 \right)
\]

with respect to \(h_{ii}\) and \(h_{ij}\), where \(\lambda_1\) and \(\lambda_2\) are Lagrangian multipliers. By equating the partial derivative of \(\phi\) with respect to \(h_{ii}\) to zero, we find that

\[
\lambda_1 = \sum_{s \in S} p(s) \left( (1 - \rho)h_{ii} + \rho \sum_{i \in S} h_{ii} + (1 - \rho) \sum_{j \neq i} h_{ij} \right)
\]

Summing over \(i \in S\) we have using (8)

\[
\lambda_1 = \frac{1 - \rho}{n} C_1 + \rho C_1 + \frac{1 - \rho}{n} C_2
\]

Inserting this \(\lambda_1\) in (10) we find with help of (8) that

\[
h_{ii} + \sum_{j \neq i \in S} h_{ij} = \frac{1}{n} [C_1 + C_2]
\]

Likewise, setting \(\partial \phi / \partial h_{ij}\) equals to zero and summing over \(i \neq j \in S\) we obtain

\[
h_{ii} + h_{ij} = \frac{1}{n} C_1 - \frac{1}{n(n-1)} C_2
\]

Solving (11) and (12) for \(h_{ii}\) and \(h_{ij}\) under the conditions (8) the unique solution is obtained as

\[
h_{ii0} = \frac{1}{n} C_1 \quad \text{and} \quad h_{ij0} = \frac{1}{n(n-1)} C_2
\]

Consequently,

\[
h_{ii0} = \frac{C_1}{na_i^2} \quad \text{and} \quad h_{ij0} = \frac{C_2}{n(n-1)a_ia_j}
\]

Substituting (12) in (6) the optimal \(p \xi -\) unbiased predictor of \(Q\) is obtained as

\[
\hat{Q}_{opt} = C_1 \sum_{i \in S} \frac{Y_i^2}{na_i^2} + C_2 \sum_{i \neq j \in S} \frac{Y_iY_j}{n(n-1)a_ia_j}
\]

The optimum variance is then obtained by substituting (13) in (9) as

\[
\min EV(\hat{Q}) = \frac{\rho^*}{n} C_1^2 + \frac{\rho^{**}}{n(n-1)} C_2^2 + 2 \frac{\rho^*}{n} C_1 C_2 - Q^2
\]

where \(\rho^* = 1 + (n-1)\rho \) and \(\rho^{**} = 1 + [1 + 2(n-2)(n-3)]\rho\)
2.2 Model \( E_T \)

In Model \( G_T \), further assume that the \( y_1, \ldots, y_N \) are exchangeable variables. Let \( S = \{ s: \text{fixed effective size of } s = n \} \), i.e. \( s \) is a collection of \( n \) distinct units. It has been shown by Basu (1958) that the order statistic \( T = \{ y_{(1)}, y_{(2)}, \ldots, y_{(n)} \} \) is sufficient, where \( y_{(1)}, y_{(2)}, \ldots, y_{(n)} \) are distinct values of \( y \) variable in sample, arranged in ascending order of their unit-indices.

Under Model \( E_T \) \( \xi \)-expectation of \( Q \) is readily found to be

\[
\mathbb{E}(Q) = C_2 (\sigma^2 + \mu^2) + C_2 (\rho \sigma^2 + \mu^2)
\]

(17)

where \( \sigma^2 + \mu^2 \) and \( \rho \sigma^2 + \mu^2 \) are the parameters to be estimated. To obtain the best estimators of these parameters and consequently the optimal predictor of \( Q \) we shall use theory on one-sample U-statistics (see, e.g., Randles and Wolfe, 1979; Rohatgi, 1988, pp. 532-534). Note that \( \sigma^2 + \mu^2 \) and \( \rho \sigma^2 + \mu^2 \) both are \( \xi \)-estimable of degree 2 with the kernels \( y_{(i)}^2/a_i^2 \) and \( y_{(i)}y_{(j)}/a_ia_j \), respectively. Moreover \( y_{(i)}^2/a_i^2 \) and \( y_{(i)}y_{(j)}/a_ia_j \) are symmetric kernels and so the U-statistic estimators of \( \sigma^2 + \mu^2 \) and \( \rho \sigma^2 + \mu^2 \) are, respectively, given by

\[
\sum_i y_{(i)}^2/n a_i^2 = \sum_i y_i^2/n a_i^2 \quad \text{and} \quad \sum_i y_{(i)}y_{(j)}/n(n-1)a_ia_j = \sum_i y_i y_j/n(n-1)a_ia_j
\]

which are \( \xi \)-unbiased and consequently \( p\xi \)-unbiased. Since \( y_i \)'s are exchangeable variables, it follows that such U-statistics are the unique minimum \( \xi \)-variance \( \xi \)-unbiased predictors of \( \sigma^2 + \mu^2 \) and \( \rho \sigma^2 + \mu^2 \). Noting that \( \xi \)-unbiasedness implies \( p\xi \)-unbiasedness and inserting \( p\xi \)-unbiased predictors of \( \sigma^2 + \mu^2 \) and \( \rho \sigma^2 + \mu^2 \) in (17) the optimal \( p\xi \)-unbiased predictor of \( Q \), after is found to be \( \hat{Q}_{opt} \) given at (15).

3. Variance of \( \widehat{\sigma}_y^2 \)

The variance of \( \hat{Q} \) is obtained as

\[
\text{V}(\hat{Q}) = E(\hat{Q}^2) - Q^2
\]

\[
= E \left[ \sum_{i \in S} b_i^2 y_i'^4 + 4 \sum_{i \neq j} \sum_{k \in S} b_{i} b_{j} y_i'^3 y_j' + \sum_{i \neq j} \sum_{k \in S} (b_i b_j + 2b_i^2 b_2) y_i^2 y_j^2 + \sum_{i \neq j} \sum_{k \neq l} \sum_{k \in S} b_{i} b_{k} y_i y_j y_k y_l - Q^2 \right] \quad (18)
\]

Comparison of (3) and (6) gives, \( b_i = 1/nNP_i \) and \( b_j = -1/n(n-1)N^2P_iP_j \) under pps sampling. Consequently we obtain the variance of \( \widehat{\sigma}_y^2 \), given at (3), as

\[
\text{V}(\widehat{\sigma}_y^2) = \frac{1}{nN^2} \sum_{i \in U} y_i'^4 + \frac{2}{n(n-1)N^4} \left( \sum_{i \in U} y_i'^2 \right)^2 + \frac{4(n-2)}{n(n-1)N^2} \left( \sum_{i \in U} y_i'^2 \right) \left( \sum_{i \in U} y_i \right)^2 - \frac{2(n-2)^2}{nN^3} \left( \sum_{i \in U} y_i'^2 \right) \left( \sum_{i \in U} y_i \right)^2 - \frac{(n-2)(n-3)}{n(n-1)N^4} \left( \sum_{i \in U} y_i \right)^4 - \sigma_y^4
\]

(19)

Since \( \beta_i = \frac{1}{N}(1 - \frac{1}{N}) \) and \( \beta_{ij} = -\frac{1}{N^2} \), \( C_1 = \frac{(N-1)}{N^2} \sum_{i \in U} a_i^2 \) and \( C_2 = -\frac{1}{N^2} \sum_{i \neq j \in U} a_i a_j \).
the optimal estimator for the population variance $\sigma_y^2$ under pps sampling is written as

$$\sigma_{yopt}^2 = \frac{1}{N} \sum_{i \in s} \frac{y_i^2}{n_{i}} - \frac{1}{N^2} \sum_{i \in s} \sum_{j \in s} \frac{y_i y_j}{n(n-1)p_i p_j}$$  \hspace{1cm} (20)$$

where

$$p_i = \frac{N a_i^2}{(N-1) \sum_{i \in U} a_i^2} \quad \text{and} \quad p_i p_j = \frac{a_i a_j}{\sum_{i \in U} \sum_{j \in U} a_i a_j}$$

The design variance of $\sigma_{yopt}^2$ can easily be obtained using (18).

References