On New Selection Procedures for Unequal Probability Sampling

Muhammad Qaiser Shahbaz, Saman Shahbaz and Muhammad Hanif

Department of Mathematics, COMSATS Institute of IT, Lahore, Pakistan
e-mail: qshahbaz@gmail.com
Department of Mathematics, COMSATS Institute of IT, Lahore, Pakistan
e-mail: samans@ciitlahore.edu.pk
Department of Mathematics, LUMS, Lahore

Abstract

We propose a new selection procedures in unequal probability sampling that can be used with Horvitz and Thompson Estimator. Expression for inclusion probability for $i$th unit ($\pi_i$) and joint inclusion probability of $i$th and $j$th unit ($\pi_{ij}$) has been obtained. Some desirable properties of $\pi_i$ and $\pi_{ij}$ has been verified for proposed procedure.

Keywords: Horvitz and Thompson Estimator, Selection Procedure, Unequal Probability Sampling.

1 Introduction

Unequal probability sampling has its roots in early thirties. The premier idea of unequal probability sampling was introduced by Neyman [7] in his ground breaking article. The first mathematical framework of unequal probability sampling with replacement was given by Hansen and Hurwitz [5]. The technique proposed by Hansen and Hurwitz [5] could not find much applicability due to possibility of selection of a population units more than once. Horvitz and Thompson [6] developed general sampling theory by restricting selection of population units to sample for one time only. The estimator of population total proposed by Horvitz and Thompson [6] is given as:

$$\hat{Y}_{HT} = \sum_{i \in S} \frac{Y_i}{\pi_i}; \quad (1.1)$$
where \( \pi_i \) is probability of selection of \( i \)th unit in the sample. The variance of Horvitz and Thompson [6] estimator has two different forms. The variance of (1.1) proposed by Horvitz and Thompson [6] is given as:

\[
\text{Var}(\hat{y}_{HT}) = \sum_{i=1}^{N} \left( \frac{1-\pi_i}{\pi_i} \right) y_i^2 + \sum_{i=1}^{N} \sum_{j=1}^{N} \left( \frac{\pi_{ij} - \pi_i \pi_j}{\pi_i \pi_j} \right) y_i y_j;
\]  

(1.2)

where \( \pi_{ij} \) is joint probability of inclusion of two units in the sample. Another form of variance of (1.1) proposed by Sen [8] and independently by Yates and Grundy [10] is:

\[
\text{Var}(\hat{y}_{HT}) = \sum_{i=1}^{N} \sum_{j=1}^{N} \left( \pi_i \pi_j - \pi_{ij} \right) \left( \frac{y_i - y_j}{\pi_i} \right)^2.
\]  

(1.3)

The estimator of (1.3), proposed by Sen [8] and independently by Yates and Grundy [10] is:

\[
\text{Var}(\hat{y}_{HT}) = \sum_{i=1}^{N} \sum_{j=1}^{N} \left( \pi_i \pi_j - \pi_{ij} \right) \left( \frac{y_i - y_j}{\pi_i} \right)^2.
\]  

(1.4)

The variance expression given in (1.3) is more popular as compared with (1.2). Both variance expressions given in (1.2) and (1.3) are based upon the quantities \( \pi_i \) and \( \pi_{ij} \). Suitable choice of these quantities can lead to substantial reduction in variance of Horvitz and Thompson [6] estimator. Survey statisticians, from time to time, has proposed number of selection procedures that can be used with Horvitz and Thompson [6] estimator. These selection procedures have been proposed with a view that the variance of (1.1) is minimum and (1.4) remain positive for all possible samples from a population of size \( N \). Brewer [2] proposed a procedure which ensures that \( \pi_i \propto p_i \), where \( p_i \) is probability of selection of \( i \)th unit in the sample. Shahbaz and Hanif [9] have proposed a more simpler selection procedure for use with (1.1). Brewer and Hanif [3] and Hanif and Brewer [4] has given a comprehensive list of selection procedures that can be used with (1.1). Recently Alodat [1] has proposed a simpler selection procedure that can be used with Horvitz and Thompson [6] estimator by following the lines of Shahbaz and Hanif [9]. In the following we proposed an extension of the selection procedure proposed by Alodat [1].

2 The New Selection Procedure

Suppose a population of \( N \) units is available and a sample of size 2 is to be selected. We propose the following selection procedure for selection of the sample:
• Select first unit with probability proportional to \( q_i = \frac{ap_i}{(2-p_i)} \) and without replacement.

• Select second unit with probability proportional to size of remaining units.

The expression for probability of inclusion of \( i \)th unit in the sample is derived below:

\[
\pi_i = q_i + \sum_{j=1}^{N} q_j p(i \mid j)
\]

\[
= \frac{ap_i}{B(2-p_i)} + \sum_{j=1}^{N} \frac{ap_j}{B(2-p_j)} \cdot \frac{p_j}{1-p_j} \text{ with } B = \sum_{j=1}^{N} \frac{ap_j}{2-p_j}
\]

\[
= \frac{ap_i}{B} \left[ \frac{1}{2-p_i} + \sum_{j=1}^{N} \frac{p_j}{(1-p_j)(2-p_j)} - \frac{p_j}{(1-p_j)(2-p_j)} \right]
\]

\[
= \frac{ap_i}{B} \left[ \frac{1}{2-p_i} + \sum_{j=1}^{N} \frac{p_j}{(1-p_j)(2-p_j)} \right]
\]

(2.1)

The expression for joint inclusion probability \( \pi_{ij} \) is:

\[
\pi_{ij} = q_i p(j \mid i) + q_j p(i \mid j)
\]

\[
= \frac{ap_i p_j}{B} \left[ \frac{1}{(1-p_i)(2-p_i)} + \frac{1}{(1-p_j)(2-p_j)} \right]
\]

(2.2)

We now verify some desirable properties for (2.1) and (2.2) in the following.

**Result 1:** The quantity \( \pi_i \) satisfies \( \sum_{i=1}^{N} \pi_i = n \).

**Proof:** Consider (2.1) as:

\[
\pi_i = \frac{ap_i}{B} \left[ \frac{1-2p_i}{(1-p_i)(2-p_i)} + \sum_{j=1}^{N} \frac{p_j}{(1-p_j)(2-p_j)} \right]
\]

Applying summation on both sides we have:
\[
\sum_{i=1}^{N} \pi_i = \sum_{i=1}^{N} \left[ \frac{ap_i}{B} \left( \frac{1-2p_i}{(1-p_i)(1-2p_i)} + \sum_{j=1}^{N} \frac{p_j}{(1-p_j)(2-p_j)} \right) \right]
\]
\[
= \frac{a}{B} \left[ \sum_{i=1}^{N} p_i \frac{(1-2p_i)}{(1-p_i)(2-p_i)} + \sum_{i=1}^{N} p_i \sum_{j=1}^{N} \frac{p_j}{(1-p_j)(2-p_j)} \right]
\]
\[
= \frac{a}{B} \sum_{i=1}^{N} \frac{p_i}{(1-p_i)(2-p_i)} \left( 1-2p_i + 1 \right)
\]
\[
= \frac{2a}{B} \sum_{i=1}^{N} \frac{p_i}{2-p_i} = \frac{2}{B} \cdot B = 2 \quad (2.3)
\]

Since \( n = 2 \), so from (2.3) we can readily see that \( \sum_{i=1}^{N} \pi_i = n \).

**Result 2:** The quantity \( \pi_j \) satisfies \( \sum_{j=1}^{N} \pi_j = (n-1)\pi_i \).

**Proof:** Consider quantity \( \pi_j \) from (2.2.) as:

\[
\pi_j = \frac{ap_i p_j}{B} \left[ \frac{1}{(1-p_i)(2-p_i)} + \frac{1}{(1-p_j)(2-p_j)} \right]
\]

Applying conditional summation, we have:

\[
\sum_{j=1}^{N} \pi_j = \sum_{j=1}^{N} \left[ \frac{ap_i p_j}{B} \left( \frac{1}{(1-p_i)(2-p_i)} + \frac{1}{(1-p_j)(2-p_j)} \right) \right]
\]
\[
= \frac{ap_i}{B(1-p_i)(2-p_i)} \sum_{j=1}^{N} p_j + \frac{ap_i}{B} \sum_{j=1}^{N} \left( \frac{p_j}{(1-p_j)(2-p_j)} \right)
\]
\[
= \frac{ap_i}{B(2-p_i)} + \frac{ap_i}{B} \left[ \sum_{j=1}^{N} \frac{p_j}{(1-p_j)(2-p_j)} - \frac{p_i}{(1-p_i)(2-p_i)} \right]
\]
\[
= \frac{ap_i}{B} \left[ \frac{1}{2-p_i} + \sum_{j=1}^{N} \frac{p_j}{(1-p_j)(2-p_j)} - \frac{p_i}{(1-p_i)(2-p_i)} \right]
\]
\[
= \frac{ap_i}{B} \left[ \frac{1-2p_i}{(1-p_i)(2-p_i)} + \sum_{j=1}^{N} \frac{p_j}{(1-p_j)(2-p_j)} \right] = \pi_i \quad (2.4)
\]

Since \( n = 2 \) so \( \sum_{j=1}^{N} \pi_j = (n-1)\pi_i \).
**Result 3:** The quantity $\pi_{ij}$ satisfies $\sum_{i=1}^{N} \sum_{j=i+1}^{N} \pi_{ij} = n(n - 1)$.

**Proof:** The proof is straightforward from proves of results 1 and 2.

**Result 4:** The Sen–Yates–Grundy variance estimator is always non-negative under this selection procedure.

**Proof:** For non-negativity of Sen–Yates–Grundy variance estimator we must have $\pi_i \pi_j - \pi_{ij} \geq 0$. Using (2.1) and (2.2) we have:

$$
\pi_i \pi_j - \pi_{ij} = \frac{ap_i p_j}{B} \left[ a \left( D^2 + \frac{(1-2p_i)(1-2p_j)}{(1-p_i)(1-p_j)(2-p_i)(2-p_j)} \right) \times 

\left( D(1-p_i)(2-p_i) + D(1-p_j)(2-p_j) + 1 \right) \times

\frac{(1-p_i)(2-p_i) + (1-p_j)(2-p_j)}{(1-p_i)(1-p_j)(2-p_i)(2-p_j)} \right]

(2.5)

where $D = \sum_{j=1}^{N} p_j \left[ (1-p_j)(2-p_j) \right]^{-1}$. The expression (2.5) is always non-negative and hence Sen–Yates–Grundy variance estimator is always non-negative under this procedure.

**Result 5:** The quantities $\pi_i$ and $\pi_j$ reduces to classical results of simple random sampling for $p_i = p_j = N^{-1}$.

**Proof:** The proof is straightforward.

### 3 Open Problem

We have proposed a new selection procedure for use with Horvitz and Thompson estimator. The procedure can be extended to sample of any size. The optimum value of constant $a$ can be located by conducting empirical study.

### References


