Modification on PPS Sample Scheme with Replacement

1Ayed R.A. Alanzi, 2Naser A. Alodat and 3Ahmad M. Qazza

1Department of Mathematics, College of Science and Human Studies at Hotat Sudair, Majmaah University, Majmaah 11952, Saudi Arabia
2Department of Mathematics, Jadara University, P.O. Box (733), postal code 21111, Irbid-Jordan
3Department of Mathematics, Zarqa University, P.O. Box (132222), postal code 13132, Zarqa-Jordan

Article history
Received: 30-03-2019
Revised: 30-05-2019
Accepted: 18-06-2019

Corresponding Author:
Ayed R.A. Alanzi
Department of Mathematics, College of Science and Human Studies at Hotat Sudair, Majmaah University, Majmaah 11952, Saudi Arabia
Email: a.alanzi@mu.edu.sa, auid403@hotmail.com

Abstract: In this paper we have developed an alternative estimator for the Probability Proportional to Size (PPS) with replacement sampling scheme when certain characteristics under study are positively correlated with the selection probability. An analogue to the well-known superpopulation model for finite population is also suggested, using which, we compare the proposed estimator with Hansen and Hurwitz estimator. Finally, an empirical investigation of the performance of the propose estimator has also been made.

Keywords: Correlation Coefficients, Probability Proportional to Size PPS with Replacement, Superpopulation Model

Introduction

Probability Proportional to Size (PPS) sampling is a method of sampling from finite population in which a size measure is available for each population unit before sampling and where the probability of selecting a unit is proportional to size.

Consider a finite population \( U = (U_1, U_2, \ldots, U_N) \) consisting of \( N \) distinct and identifiable units. Let \( Y_i \) be the value of the study variable \( y \) on the unit \( U_i, i = 1, \ldots, N \). In practice we wish to estimate the population total\( \sum_{i=1}^{N} y_i \) from the sample \( (u_1, u_2, \ldots, u_n) \) with maximum precision. The easiest of the probability sampling scheme for drawing a sample is the Simple Random Sampling with Replacement (SRSWR) scheme for which an unbiased estimator of \( y \) is given by:

\[
\hat{Y}_{SRSWR} = \frac{1}{n} \sum_{i=1}^{n} y_i
\]

With variance is given by:

\[
V(\hat{Y}_{SRSWR}) = \frac{N}{n^2} \sum_{i=1}^{n} y_i^2 - \frac{Y^2}{N}.
\]

Hansen and Hurwitz (1943) proposed the idea of sampling with Probability Proportional to Size (PPS) with replacement for positive correlated characteristics. This scheme was carried out as follows: One unit is selected at each of the \( n \) draws. For each \( t \)th unit selected from the population, a selection probability is given by:

\[
p_i = \frac{x_i}{x},
\]

where, \( x_i \) is the measure for \( i \)th unit and:

\[
x = \sum_{i=1}^{n} x_i.
\]

They gave the following estimator of population total \( Y_t \) as:

\[
\hat{Y}_{PPS} = \frac{1}{n} \sum_{i=1}^{n} \frac{y_i}{p_i}
\]

With variance is given by:

\[
V(\hat{Y}_{PPS}) = \frac{1}{n^2} \sum_{i=1}^{n} y_i^2 - \frac{Y^2}{n}.
\]

PPS sampling is expected to be more efficient than SRS sampling if the regression line of \( y \) on \( x \) passes through the origin. When it is not so, a transformation on the auxiliary variable can be made so that the PPS sampling with modified sizes becomes more efficient. Reddy and Rao (1977) suggested that the sample be selected by probability proportional to revised sizes scheme and with replacement, the revised sizes are obtained through a location shift in the auxiliary variable as:

\[
x^* = x_i + \left( \frac{1}{L} - 1 \right) \bar{x}, \quad 0 < L < 1
\]
However, only one measure of size is usually used in selecting primary sampling units in PPS scheme. In contrast, it may sometimes happen that some of these study variables are poorly but positively correlated with selecting probabilities, thereby reducing the existing estimator inadequate. An alternative estimator was proposed by Rao (1966). Bansal and Singh (1985), Amahia et al. (1989), Enang and Amahia (2012) and others have proposed an estimator for characteristics that are poorly correlated with selecting probabilities.

Sahoo et al. (1994) suggested a simple transformation of the auxiliary variable where the correlation between study variable and auxiliary variable is negative.

Bedi and Rao (1997) gave a new direction in determining estimator of population total under the PPSWR sampling scheme when the correlation between the auxiliary variable and study variable is negative.

In this paper we suggested a simple transformation on  to such that  where  and .

We have also obtained the condition under which the proposed estimator will be more efficient than Hansen and Hurwitz (1943) estimator. The condition has been derived under the superpopulation model given below.

**The Superpopulation Model**

Let  and  denote the value of characteristics  and the relative measure of size  for the  unit in the population, respectively. A general superpopulation model suitable for our case is:

\[
y_i = Bp_i + e_i, \quad i = 1, 2, \ldots, N, \quad (5)
\]

where,  are the errors such that:

\[
E\left( e_i | p_i \right) = 0, \\
E\left( e_i^2 | p_i \right) = \sigma_e^2 p_i, \\
\sigma_e^2 > 0, \quad g \geq 0, \\
E\left( e_i | p, p_i \right) = 0,
\]

where,  denote the average overall finite population that can be drawn from the super population. There are many papers in which the supper population model is successfully used for the purpose of comparing the different size strategies, see, Godambe (1955), Brewer (1963), Rao (1966), Hanurav (1976) and many others.

PPS sampling is considered to be more efficient than SRS sampling if the regression line of  on  passes through the origin Raj (1954). When it is not so a transformation on the auxiliary variable can be made so that the PPS sampling with modified sizes become more precise.

**Suggested Estimator**

Suppose that the auxiliary variable  has a positive correlation with study variable . Then we suggest the following transformation on  such that  for . Naturally  is greater than zero. Further, we can easily see that correlation between  and  is also positive. Hence the modified probabilities of selection become:

\[
p_i^* = \frac{1 + p_i}{N + 1}, \quad i = 1, 2, \ldots, N \quad (6)
\]

Then the estimator of the population total  is given by:

\[
\hat{Y}_p = \frac{1}{n} \sum_{i=1}^{N} \frac{y_i}{p_i^*} \quad (7)
\]

**The Variance and its Expected Value of the Suggested Estimator**

It is well known that the variance of the usual estimator  is given by:

\[
v(\hat{T}_{sw}) = \frac{1}{n} \left[ \sum_{i=1}^{N} y_i^2 \left( \frac{\sum_{i=1}^{N} y_i}{n} \right)^2 \right] \quad (7)
\]

The corresponding variance of the estimator due to Rao is obtained by:

\[
v(\hat{T}_y) = \frac{N^2}{n} \left[ \sum_{i=1}^{N} y_i^2 p_i - \sum_{i=1}^{N} y_i p_i \right] \quad (8)
\]

The variance of the proposed estimator is obtained by replacing  by  in (7) and is given by:

\[
v(\hat{Y}_p) = \frac{1}{n} \left[ \sum_{i=1}^{N} y_i^2 - \frac{\sum_{i=1}^{N} y_i}{n} \right] \quad (9)
\]

**Robustness Estimator**

Now, we state two lemmas, which are useful for estimator's comparisons.

**Lemma 1**

Royall (1970) Let  and  and  satisfying:

\[
\sum_{i=1}^{N} c_i \geq 0
\]

then:

\[
\sum_{i=1}^{N} b_i c_i \geq 0.
\]
Lemma 2

Let \( b_1 \geq b_2 \geq \ldots \geq b_m \geq 0 \) and \( c_1 \geq c_2 \geq \ldots \geq c_m \) satisfy:

\[
\sum_{i=1}^{m} c_i \geq 0
\]

then:

\[
\sum_{i=1}^{m} b_i c_i \geq 0.
\]

Theorem

Under the superpopulation model, the sufficient condition that \( \hat{T}_{ini} \) has smaller expected variance than \( \hat{Y}_r \) is:

\[
g \geq 1 + \frac{p_i}{1 + p_i}.
\]

Proof

Under the superpopulation model the expected variance of \( \hat{T}_{ini} \) and \( \hat{Y}_r \) are respectively given by:

\[
nE\left(\hat{Y}_{r}\right) = \sigma^2 \sum_{i=1}^{N} \frac{p_i}{1 + p_i}.
\]

and:

\[
nE\left(\hat{T}_{ini}\right) = B^2 \left[ \sum_{i=1}^{N} \frac{p_i^2}{1 + p_i} - \left( \sum_{i=1}^{N} p_i \right)^2 \right] + \sigma^2 \sum_{i=1}^{N} b_i c_i.
\]

where, \( c_i = (N p_i - 1) \) and \( b_i = \frac{p_i^{* - 1}}{1 + p_i} \). Note that, the above first term of the above expression is always positive. For the second term we observe that \( \sum c_i = 0 \) and \( c_i \) is an increasing function of \( i \). So in view Royall's lemma 1 it can be shown that \( \sum b_i c_i > 0 \) provided \( b_i \) is also increasing function of \( p_i \). By deriving \( b_i \) with respect to \( p_i \) we get that the sufficient condition that makes \( \hat{T}_{ini} \) has smaller variance than \( \hat{Y}_r \) is:

\[
g \geq 1 - \frac{p_i}{1 + p_i^*}.
\]

Hence the theorem is proved.

Empirical Study

To study the behavior of the estimator \( \hat{Y}_r \) with the conventional estimator \( \hat{T}_{ini} \), we consider the five population \( A, B, C, D \) and \( E \), details of which are given in Table 1. The population \( A, B \) and \( C \) are the same as the three population of the Yates and Grundy (1953).

Table 1: Populations under

<table>
<thead>
<tr>
<th>Unit No</th>
<th>A</th>
<th>B</th>
<th>C</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( x )</td>
<td>( y )</td>
<td>( y )</td>
</tr>
<tr>
<td>1</td>
<td>0.1</td>
<td>0.8</td>
<td>0.8</td>
</tr>
<tr>
<td>2</td>
<td>0.2</td>
<td>1.2</td>
<td>1.4</td>
</tr>
<tr>
<td>3</td>
<td>0.3</td>
<td>2.1</td>
<td>1.8</td>
</tr>
<tr>
<td>4</td>
<td>0.4</td>
<td>3.2</td>
<td>2.0</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Unit No</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>( x )</td>
<td>( y )</td>
<td>( x )</td>
</tr>
<tr>
<td>1</td>
<td>0.01</td>
<td>4</td>
<td>0.1</td>
</tr>
<tr>
<td>2</td>
<td>0.09</td>
<td>9</td>
<td>0.1</td>
</tr>
<tr>
<td>3</td>
<td>0.16</td>
<td>16</td>
<td>0.2</td>
</tr>
<tr>
<td>4</td>
<td>0.25</td>
<td>25</td>
<td>0.2</td>
</tr>
<tr>
<td>5</td>
<td>0.49</td>
<td>36</td>
<td>0.4</td>
</tr>
</tbody>
</table>
Table 2: Variance of the estimator

<table>
<thead>
<tr>
<th>POP</th>
<th>$\bar{\bar{T}}_{w_k}$</th>
<th>$\bar{\bar{T}}_{m_l}$</th>
<th>$\bar{T}_R$</th>
<th>$\bar{\bar{Y}}_p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>6.815</td>
<td>0.305</td>
<td>6.384</td>
<td>4.576</td>
</tr>
<tr>
<td>B</td>
<td>1.680</td>
<td>0.500</td>
<td>1.104</td>
<td>0.911</td>
</tr>
<tr>
<td>C</td>
<td>0.575</td>
<td>0.125</td>
<td>0.337</td>
<td>0.416</td>
</tr>
<tr>
<td>D</td>
<td>498.000</td>
<td>1842.000</td>
<td>890.902</td>
<td>569.108</td>
</tr>
<tr>
<td>E</td>
<td>498.000</td>
<td>1229.200</td>
<td>765.200</td>
<td>597.871</td>
</tr>
<tr>
<td>F</td>
<td>3350.000</td>
<td>3708.440</td>
<td>3221.620</td>
<td>3151.820</td>
</tr>
</tbody>
</table>

Table 3: Percentage variances relative to the suggested estimator

<table>
<thead>
<tr>
<th>POP</th>
<th>$\bar{\bar{T}}_{w_k}$</th>
<th>$\bar{\bar{T}}_{m_l}$</th>
<th>$\bar{T}_R$</th>
<th>$\bar{\bar{Y}}_p$</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>149.12</td>
<td>6.67</td>
<td>139.5</td>
<td>1</td>
</tr>
<tr>
<td>B</td>
<td>184.41</td>
<td>54.88</td>
<td>121.18</td>
<td>1</td>
</tr>
<tr>
<td>C</td>
<td>138.22</td>
<td>30.05</td>
<td>81.01</td>
<td>1</td>
</tr>
<tr>
<td>D</td>
<td>87.51</td>
<td>323.66</td>
<td>156.5</td>
<td>1</td>
</tr>
<tr>
<td>E</td>
<td>83.30</td>
<td>205.61</td>
<td>129.99</td>
<td>1</td>
</tr>
<tr>
<td>F</td>
<td>106.23</td>
<td>117.65</td>
<td>102.21</td>
<td>1</td>
</tr>
</tbody>
</table>

Whereas population $D$ is of Stuart (1986). The population $E$ is of Stuart (1986) and population $F$ is of Amahia et al. (1989).

Table 2 gives the variances of the proposed estimator $\bar{\bar{Y}}_p$ with the conventional estimators $\bar{\bar{T}}_{w_k}$, $\bar{\bar{T}}_{m_l}$, $\bar{T}_R$ for $n = 2$.

Table 3 gives the percentages efficiency of the proposed estimator $\bar{\bar{Y}}_p$ with the conventional estimators $\bar{\bar{T}}_{w_k}$, $\bar{\bar{T}}_{m_l}$ and $\bar{T}_R$.

**Conclusion**

Table 3 give the percentage efficiency of the proposed estimators $\bar{\bar{Y}}_p$ with the conventional estimator $\bar{\bar{T}}_{w_k}$, $\bar{\bar{T}}_{m_l}$ and $\bar{T}_R$ for $n = 2$.

It is clear from Table 3 that the proposed estimator $\bar{\bar{Y}}_p$ performed better in populations $A$ and $B$ than $\bar{\bar{T}}_{w_k}$, $\bar{\bar{T}}_{m_l}$ and $\bar{T}_R$. In population $C$ it is clear that the proposed estimator $\bar{\bar{Y}}_p$ performed better than $\bar{\bar{T}}_{w_k}$. But in population $D$ and $E$ the proposed estimator performed better than $\bar{\bar{T}}_{m_l}$ and $\bar{T}_R$.

We can see that in population $F$ the proposed estimator $\bar{\bar{Y}}_p$ performed better than the $\bar{\bar{T}}_{w_k}$, $\bar{\bar{T}}_{m_l}$ and $\bar{T}_R$.

**Acknowledgement**

The authors gratefully acknowledge with thanks the very thoughtful and constructive comments and suggestions of the Editor-in-Chief and the reviewers which resulted in much improved paper.

**Author’s Contributions**

Authors contributed to the same extent to all the process of preparing and developing the manuscript since we operate as a group.

**Ethics**

This article is original and contains unpublished material. The corresponding author confirms that all of the other authors have read and approved the manuscript and there are no ethical issues involved.

**References**


