Journal of Mathematics and Statistics

Original Research Paper

Modification on PPS Sample Scheme with Replacement

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

[1]Department of Mathematics, College of Science and Human Studies at Hotat Sudair, Majmaah University, Majmaah 11952, Saudi Arabia
[2]Department of Mathematics, Jadara University, P.O. Box (733), postal code 21111, Irbid-Jordan
[3]Department 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 $Y = \Sigma y_i$ from the y values of the units drawn in a sample (u_1, $u_2,\ldots, u_n$) with maximum precision. The easiest of the probability sampling scheme for drawing a sample u is the Simple Random Sampling with Replacement (SRSWR) scheme for which an unbiased estimator of y is given by:

$$\hat{T}_{wr} = \frac{N}{n}\sum_{i=1}^{n} y_i \tag{1}$$

With variance is given by:

$$V\left(\hat{T}wr\right) = \frac{N}{n}\left[\sum_{i=1}^{N} y_i^2 - \frac{Y^2}{N}\right]. \tag{2}$$

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 ith unit selected from the population, a selection probability is given by:

$$p_i = \frac{x_i}{x}$$

where, x_i is the measure for ith unit and:

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

They gave the following estimator of population total Yi as:

$$\hat{T}_{HH} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i} \tag{3}$$

With variance is given by:

$$v\left(\hat{T}_{HH}\right) = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i^2}{p_i} - Y^2 \tag{4}$$

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:

$$X^* = X_i + \left(\frac{1}{L} - 1\right)\bar{X} \qquad 0 < L < 1$$

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 highly 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 x to x^* such that $x^* = (x+x_i)$.

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 y_i and pi denote the value of characteristics y and the relative measure of size p for the ith (i = 1,2,…,N) unit in the population, respectively. A general superpopulation model suitable for our case is:

$$y_i = Bp_i + e_i,\ i = 1,2,\ldots,N, \tag{5}$$

where, e_i are the errors such that:

$$\begin{aligned} &E\left(e_i / p_i\right) = 0,\\ &E\left(e_i^2 / p_i\right) = \sigma^2 p_i^g,\\ &\sigma^2 > 0, g \geq 0,\\ &E\left(e_i e_j / p_i p_j\right) = 0, \end{aligned}$$

where, $E(.)$ 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 sample 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 y on x 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 $x>0$ has a positive correlation with study variable y. Then we suggest the following transformation on x to x^* such that $x^* = (x+x_i), i = 1,2,\ldots,N$. Naturally x^* is greater than zero. Further, we can easily see that correlation between y and x^* is also positive. Hence the modified probabilities of selection become:

$$p_i^* = \frac{1+p_i}{N+1}\ i, = 1,2,\ldots,N \tag{6}$$

Then the estimator of the population total Y is give by:

$$\widehat{Y}_p = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i^*}$$

The Variance and its Expected Value of the Suggested Estimator

It is well known that the variance of the usual estimator $\widehat{T}_{HH}$ is given by:

$$v\left(\widehat{T}_{HH}\right) = \frac{1}{n}\left[\sum_{i=1}^{N}\frac{y_i^2}{p_i} - \left(\sum_{i=1}^{n} y_i\right)^2\right] \tag{7}$$

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

$$v\left(\widehat{T}_R\right) = \frac{N^2}{n}\left[\sum_{i=1}^{N} y_i^2 p_i - \left(\sum_{i=1}^{N} y_i\, p_i\right)^2\right] \tag{8}$$

The variance of proposed estimator is obtain by replacing p_i by p_i^* in (7) and is given by:

$$v\left(\widehat{Y}_p\right) = \frac{1}{n}\left[\sum_{i=1}^{N}\frac{y_i^2}{p_i^*} - \left(\sum_{i=1}^{N} y_i\right)^2\right] \tag{9}$$

Robustness Estimator

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

Lemma 1

Royall (1970) Let $0 \leq b_1 \leq b_2 \leq \ldots \leq b_m$ and $c_1 \leq c_2 \leq \ldots \leq c_m$ satisfying:

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

then:

$$\sum_{i=1}^{m} b_i\, c_i \geq 0.$$

Lemma 2

Let $b_1 \geq b_2 \geq ... \geq b_m \geq 0$ and $c_1 \geq c_2 \geq ... \geq c_m$ 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 $\hat{T}_{HH}$ has smaller expected variance than $\hat{Y}_p$ is:

$$g \geq 1 + \frac{p_i}{1 + p_i}.$$

Proof

Under the superpopulation model the expected variance of $\hat{T}_{HH}$ and $\hat{Y}_p$ are respectively given by:

$$nE\left(v\left(\hat{T}_{HH}\right)\right) = \sigma^2 \sum_{i=1}^{N} p_i^g \left(1 - p_i\right),$$

and:

$$nE\left(v\left(\hat{Y}_p\right)\right) = B^2 \left[\sum_{i=1}^{N} \frac{p_i^2}{p_i^*} - \left(\sum_{i=1}^{N} p_i\right)^2\right] + \sigma^2 \sum_{i=1}^{N} p_i^g \left(\frac{1}{p_i^*} - 1\right).$$

The difference between them can be written as:

$$n\left[E\left(v\left(\hat{Y}_p\right)\right) - E\left(v\left(\hat{T}_{HH}\right)\right)\right] = B^2 \left[\sum_{i=1}^{N} \frac{p_i^2}{p_i^*} - \left(\sum_{i=1}^{N} p_i\right)^2\right]$$

$$+\sigma^2 \sum_{i=1}^{N} p_i^{g-1} \left(\frac{p_i - p_i^*}{p_i^*}\right) = B^2 \left[\sum_{i=1}^{N} \frac{p_i^2}{p_i^*} - \left(\sum_{i=1}^{N} p_i\right)^2\right]$$

$$+\sigma^2 \sum_{i=1}^{N} p_i^{g-1} \frac{Np_i - 1}{(N+1)p_i^*} = B^2 \left[\sum_{i=1}^{N} \frac{p_i^2}{p_i^*} - \left(\sum_{i=1}^{N} p_i\right)^2\right]$$

$$+\sigma^2 \sum_{i=1}^{N} p_i^{g-1} \frac{Np_i - 1}{(1+p_i)} = B^2 \left[\sum_{i=1}^{N} \frac{p_i^2}{p_i^*} - \left(\sum_{i=1}^{N} p_i\right)^2\right] + \sigma^2 \sum_{i=1}^{N} b_i c_i,$$

where, $c_i = (Np_i - 1)$ and $b_i = \frac{p_i^{g-1}}{1 + p_i}$. 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 T_{HH} has smaller variance than $\hat{Y}_p$ is:

$$g \geq 1 - \frac{p_i}{1 + p'_1}$$

Hence the theorem is proved.

Empirical Study

To study the behavior of the estimator $\hat{Y}_p$ with the conventional estimator $\hat{T}_{HH}$, 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

	A		*B*	*C*				
Unit No	*x*	*y*	*y*	*y*				
1	0.1	0.8	0.8	0.2				
2	0.2	1.2	1.4	0.6				
3	0.3	2.1	1.8	0.9				
4	0.4	3.2	2.0	0.8				
5								
	D		*E*		*F*			
Unit No	*x*	*y*	*x*	*y*	*x*	*y*	*x*	*y*
1	0.01	4	0.1	4	0.07	11	0.06	6
2	0.09	9	0.1	9	0.09	7	0.09	13
3	0.16	16	0.2	16	0.04	5	0.12	9
4	0.25	25	0.2	25	0.20	27	0.14	14
5	0.49	36	0.4	36	0.07	30	0.12	18

Table 2: Variance of the estimator

POP	$\hat{T}_{WR}$	$\hat{T}_{HH}$	$\hat{T}_{R}$	$\hat{Y}_{p}$
A	6.815	0.305	6.384	4.576
B	1.680	0.500	1.104	0.911
C	0.575	0.125	0.337	0.416
D	498.000	1842.000	890.902	569.108
E	498.000	1229.200	765.200	597.871
F	3350.000	3708.440	3221.620	3151.820

Table 3: Percentage variances relative to the suggested estimator

POP	$\hat{T}_{WR}$	$\hat{T}_{HH}$	$\hat{T}_{R}$	$\hat{Y}_{p}$
A	149.12	6.67	139.5	1
B	184.41	54.88	121.18	1
C	138.22	30.05	81.01	1
D	87.51	323.66	156.5	1
E	83.30	205.61	129.99	1
F	106.23	117.65	102.21	1

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 $\hat{Y}_{p}$ with the conventional estimators $\hat{T}_{WR}$, $\hat{T}_{HH}$ and $\hat{T}_{R}$ for $n = 2$.

Table 3 gives the percentages efficiency of the proposed estimator $\hat{Y}_{p}$ with the conventional estimators $\hat{T}_{WR}$, $\hat{T}_{HH}$ and $\hat{T}_{R}$.

Conclusion

Table 3 give the percentage efficiency of the proposed estimators $\hat{Y}_{p}$ with the conventional estimator $\hat{T}_{WR}$, $\hat{T}_{HH}$ and T_{R} for $n = 2$.

It is clear from Table 3 that the proposed estimator $\hat{Y}_{p}$ performed better in populations *A* and *B* than $\hat{T}_{WR}$, $\hat{T}_{R}$. In population *C* it is clear that the proposed estimator $\hat{Y}_{p}$ performed better than $\hat{T}_{WR}$. But in population *D* and *E* the proposed estimator performed than $\hat{T}_{HH}$ and $\hat{T}_{R}$.

We can see that in population *F* the proposed estimator $\hat{Y}_{p}$ performed better than the $\hat{T}_{WR}$, $\hat{T}_{HH}$ and $\hat{T}_{R}$.

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

Amahia, G.N., Y.P. Chaubey and T.J. Rao, 1989. Efficiency of a new estimator in PPS sampling for multiple characteristics. J. Stat. Plann. Inference, 21: 75-84.

Bansal, M.L. and R. Singh, 1985. An alternative estimator for multiple characteristics in PPS sampling. J. Statist. Plann. Inference, 11: 313-320.

Bedi, P.K. and T.J. Rao, 1997. PPS method of estimation under a transformation. J. Indian Soc. Agar. Stat.

Brewer, K.R.W., 1963. A method of systematic sampling with unequal probabilities. Aust. J. Stat., 5: 5-13.

Enang, E. I. and G.N. Amahia, 2012. A class of alternative estimators in probability proportional to size sampling with replacement for multiple characteristics. J. Math. Res., 4: 66-77.

Godambe, V.P. 1955. A unified theory of sampling from finite populations. J. Roy. Stat. Soc., 17: 269-278.

Hansen, M.H. and W.N. Hurwitz, 1943. On the theory of sampling from finite populations. Ann. Math. Stat., 14: 333-362.

Hanurav, T.V., 1976. Optimum utilization of auxiliary information: πps sampling of two units from a straturn. Jout. Roy. Stat. Soc., 29: 374-391.

Raj, D., 1954. On the sampling with probabilities proportional to size. Ganits, 5: 175-182.

Rao, J.N.K., 1966. On the relative efficiency of some estimators in PPS sampling for multiple characteristics. Sankhya A, 28: 61-70.

Reddy, V.N and T.J. Rao, 1977. Modified PPS method of estimation. Sankhya, 39: 185-197.

Royall, R.M., 1970. On finite population sampling theory undercertain linear regression models. Biometrika, 57: 377-387.

Sahoo, J., L.N. Sahoo and Mohanty, 1994. Unequal probability sampling using a transformed variable. Metron, 52: 71-83.

Stuart, A., 1986. Location shifts in sampling with unequal probabilities. J. Roy, Stat. Soc., 149: 169-174.

Yates, F. and P.M. Grundy, 1953. Selection without replacement from within strata with probability proportional to size. J. Roy. Stat. Soc. Ser., 5: 253-261.