ISSN 1533-9211

ON SAMPLING WITH PROBABILITY PROPORTIONAL TO SIZE WITH REPLACEMENT

NASER A. ALODA[1] , AYED RHEAL A ALANZI[2] AND AYMAN A . HAZAYMEH[3]

[1]Department of Mathematics, Jadara University, P.O. Box (733), postal code 21111, Irbid-Jordan

[2]Department of Mathematics, College of Science and Human Studies at Hotat Sudair, Majmaah University, Majmaah 11952, Saudi Arabia.

[3]Department of Mathematics, Jadara University, P.O. Box (733), postal code 21111, Irbid-Jordan

DOI: 10.5281/zenodo.6552294

Abstract

In this paper we suggested a new transformation for the selection probability under positive correlation coefficient between study variable (y) and measure of size variable (x). The relative efficiency of the proposed estimator has been studied under a superpopulation model. A numerical investigation into the performance of the estimator has been made.

Keywords: Hansen Hurwitz, Probability Proportional to size, Estimator, sampling with replacement.

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 units 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 $Y = \sum y_i$ 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:

$$\hat{T}_{srs} = \frac{N}{n} \sum_{i=1}^{n} y_i \qquad (1)$$

With variance is given by:

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

Hansen & Hurwitz (1943) proposed the idea of sampling with probability proportional to size and with replacement (PPSWR).Under the scheme, one unit to be selected at each of the n draw. For each of the i^{th} unit selected from population, at selection probability is given by

The Seybold REPORT

ISSN 1533-9211

$$p_i = \frac{x_i}{X}, \quad \text{where} \quad X = \sum_{i=1}^{N} x_i$$

Hansen & Hurwitz (1943) give the estimator of the population total , as

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

with variance

$$v(\hat{T}_{HH}) = \frac{1}{n}\left[\sum_{i=1}^{n}\frac{y_i^2}{p_i} - Y^2\right] \qquad (3)$$

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 & Rao (1977) suggested that the sample by Rao, Hartley & Cochran (1962) proposed a method for estimation of variance that always have smaller variance than the standard in sample with unequal probability with replacement.

Amahia, Chaubey & Rao (1989) provide simple alternative estimator of the population total when is positive correlation between the study and auxiliary variable, the estimator is

$$\hat{T} = \sum_{i=1}^{N}\frac{y_i}{p_i^*}, p_i^* = \frac{1-\rho}{N} + \rho p_i, p_i = \frac{x_i}{\sum x_i}$$

Singh & Horn (1998) proposed an alternative estimator for estimating a population total when the certain variable have poor positive correlation with selection probabilities. Singh & Tailor (2003) suggested the following estimator of population total

$$p_i^* = \frac{(1-\rho)(1+\rho)}{N} + \frac{1}{2}[\rho(1+\rho)p_i^+ - \rho(1-\rho)p_i^-] \qquad (4)$$

where $p_i^+ = \frac{x_i}{X}, X = \sum_{i=1}^{N} x_i, p_i^- = \frac{z_i}{\sum x}$, with $z_i = \frac{X - nx_i}{N-n}$

Bansal & Singh (1985), noticed that the Rao (1966a) model deal with zero correlation and so developed a new transformed estimator of population total when the characteristics under study are poorly correlated with selected probability. Amahia, Chaubey & Rao (1989), suggested simple alternatives to the transformations in Bansal & Singh (1985) procedure. Kumar bedi (1995), Bedi & Rao (2001), Singh & Horn (1998), Sahoo, Mishra & Senapati (2005), Sahoo, Singh & Das (2006), and Sahoo, SC. & AK. (2010) worked in negatively correlation characteristics.

The Seybold REPORT

ISSN 1533-9211

The super population model

Let y_i and p_i denote the value of characteristics y and the relative measure of size p for the i^{th}, $(i = 1, 2, \ldots, N)$ unit in the population, respectively. A general superpopulation model suitable for our case is

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

where e_i are the errors such that

$$E(e_i/p_i) = 0, E(e_i^2/p_i) = \sigma^2 p_i^g, \sigma^2 > 0, g \geq 0, E(e_i e_j / p_i p_j) = 0$$

where $E(.)$ denote the average overall finite population that can be drawn from the superpopulation. 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 (1966b), Hanurav (1967) and many others.

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 $x^* = \frac{x_i + nX}{N - n}, i = 1,2, \ldots, N$. 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

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

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

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

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

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

The corresponding variance of the estimator due to Rao (1966b) is obtained by

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

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

ISSN 1533-9211

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

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

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

Lemma 2: Let $b_1 \geq b_2 \geq \ldots \geq b_m \geq 0$ and $c_1 \geq c_2 \geq \ldots \geq c_m$ satisfying

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

Then

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

Theorem 1: Under the superpopulation model, the sufficient condition that $\hat{T}_{HH}$ has smaller expected variance than $\hat{Y}_p$ is

$$g \geq 1 + \frac{np_i}{1 + np_i}$$

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

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

and

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

The difference between them can be written as

The Seybold REPORT

ISSN 1533-9211

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

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

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

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

where $c_i = (Np_i - 1)$ and $b_i = \frac{p_i^{g-1}}{1+np_i}$. 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 p_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 bi with respect to p_i we get that the sufficient condition that makes $\hat{T}_{HH}$ has smaller variance than $\hat{Y}_p$ is

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

Hence the theorem is proved.

Theorem 2: Under the superpopulation model the sufficient-condition that the proposed estimator $\hat{Y}$ has smaller expected variance than the estimator $\hat{T}_{srs}$ is

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

Proof: under the superpopulation model the expected variance of the estimator $\hat{T}_{srs}$ and $\hat{Y}_p$ are

$$nEv(\hat{T}_{srs}) = B^2\left[\sum_{i=1} P_i^2 - 1\right] + \sigma^2(N-1)\sum_{i=1} p_i^g$$

and

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

Then

ISSN 1533-9211

$$nEv(\hat{T}_{srs}) - nEv(\hat{Y}_p) = B^2\left[\sum_{i=1}^{N}\frac{P_i^2}{p_i^*}(Np_i - 1)\right] + \sigma^2\left[\frac{p_i^g}{p_i^*}(Np_i^* - 1)\right]$$

$$= B^2\sum_{i=1}^{N} b_i\, c_i + \sigma^2\sum_{i=1}^{N} b_i\, c_i$$

Now because of $\sum c_i = 0$ and c_i is an increasing function of p_i and so b_i.Then the sufficient condition that b_i should also be an increasing function of p_i is

$$g \geq \frac{p_i}{n + p_i}$$

Thus, in view of Roayaii's lemma 1 both part of 2.2 are positive Hence the theorem is prove.

Empirical study:

To study the behavior of the estimator $\hat{Y}_p$ with the conventional estimator $\hat{T}_{srs}$, we will consider the three population, which are given in table 1.

Table 1. Population Under Study.

Unit No	Population 1		Population 2		Population 3	
	x	y	x	y	x	y
1	41	36	3	11	25	11
2	43	47	4	7	32	7
3	54	41	5	9	14	5
4	39	47	8	8	70	27
5	49	47	12	8	24	30
6	45	45	11	9	20	6
7	41	32	8	8	32	13
8	33	37	9	12	44	9
9	37	40	11	10	50	14
10	41	41	10	9	44	18
11	47	37	8	3		
12	39	48	9	14		
13			7	12		
14			8	10		
15			8	10		
16			5	10		
17			6	9		
18			3	5		
19			3	7		
20			9	9		

21		6	6	
22		7	12	
23		8	9	
24		8	6	
25		9	9	
26		11	11	
27		11	10	
28		10	14	
29		5	8	
30		3	7	

Table 2. Result of selection probability and generalized selection probability.

	Population 1			
	X	Y	P_i	P_i^*
	41	36	0.08055	0.082432
	43	47	0.084479	0.083705
	54	41	0.10609	0.090707
	39	47	0.076621	0.081158
	49	47	0.096267	0.087524
	45	45	0.088409	0.084978
	41	32	0.08055	0.082432
	33	37	0.064833	0.077339
	37	40	0.072692	0.079885
	41	41	0.08055	0.082432
	47	37	0.092338	0.086251
	39	48	0.076621	0.081158
Sum	509	498	1	1

Table 3. Result of selection probability and generalized selection probability

The Seybold Report
ISSN 1533-9211

	Population 2			
	X	Y	P_i	P_i^*
	3	11	0.013333	0.033005
	4	7	0.017778	0.033078
	5	9	0.022222	0.033151
	8	8	0.035556	0.03337
	12	8	0.053333	0.033661
	11	9	0.048889	0.033588
	8	8	0.035556	0.03337
	9	12	0.04	0.033443
	11	10	0.048889	0.033588
	10	9	0.044444	0.033515
	8	3	0.035556	0.03337
	9	14	0.04	0.033443
	7	12	0.031111	0.033297
	8	10	0.035556	0.03337
	8	10	0.035556	0.03337
	5	10	0.022222	0.033151
	6	9	0.026667	0.033224
	3	5	0.013333	0.033005
	3	7	0.013333	0.033005
	9	9	0.04	0.033443
	6	6	0.026667	0.033224
	7	12	0.031111	0.033297
	8	9	0.035556	0.03337
	8	6	0.035556	0.03337
	9	9	0.04	0.033443
	11	11	0.048889	0.033588
	11	10	0.048889	0.033588
	10	14	0.044444	0.033515
	5	8	0.022222	0.033151
	3	7	0.013333	0.033005
Sum	225	272	1	1

Table 4. Result of selection probability and generalized selection probability.

The Seybold REPORT

ISSN 1533-9211

	Population 2			
	X	Y	P_i	P_i^*
	25	11	0.070423	0.058824
	32	7	0.090141	0.103922
	14	5	0.039437	0.101401
	70	27	0.197183	0.109244
	24	30	0.067606	0.102801
	20	6	0.056338	0.102241
	32	13	0.090141	0.103922
	44	9	0.123944	0.105602
	50	14	0.140845	0.106443
	44	18	0.123944	0.105602
Sum	355	140	1	1

From table(2, 3, 4) above, we observed that the linear transformation p_i and hence, the generalized transformation p_i^* satisfied the regularity condition of probability normed size measure

1. $0 < p_i < 1$
2. $\sum_{i=1}^{N} p_i = 1$
3. $0 < p_i^* < 1$
4. $\sum_{i=1}^{N} p_i^* = 1$

Also we observed that the correlation coefficient for population 1,2,3 are 0.162, 0.338, and 0.487 respectively.

Table 5. The Variance of the Estimators for Sample Size = 2.

Population	$\hat{T}_{srs}$	$\hat{T}_{HH}$	$\hat{Y}_p$
I	3708	6364.892	3667.204
II	5276	12715.85	5201.921
III	6700	7478	6478.15

Table 6. Percentage Variance relative for the Suggested Estimator $\hat{Y}_p$.

The Seybold Report
ISSN 1533-9211

Population	$\hat{T}_{srs}$	$\hat{T}_{HH}$	$\hat{Y}_p$
I	98.90	57.62	100
II	98.59	40.91	100
III	96.69	86.63	100

Conclusion

It is clear from table 6 that the estimator $\hat{Y}_p$ is the most efficient than the estimators $\hat{T}_{srs}$ and $\hat{T}_{HH}$ in population I, II, and III.

References

AMAHIA, G., CHAUBEY, Y. & RAO, T. (1989). Efficiency of a new pps sampling for multiple characteristics. Journal of Statistical Planning and Inference 21, 75–84.

BANSAL, M. & SINGH, R. (1985). An alternative estimator for multiple characteristics in pps sampling. Journal of Statistical Planning and Inference 21, 75–84.

BEDI, P. & RAO, T.J. (2001). Pps method of estimation under a transformation. Journal of the Indian Society of Agricultural Statistics 54, 184–195.

BREWER, K. (1963). A method of systematic sampling with unequal probabilities. Aust. J. Stat. 5, 5–13.

GODAMBE, V. (1955). A unified theory of sampling from finite populations. Journal of the Royal Statistical Society: Series B (Methodological) 17, 269–278.

HANSEN, M.H. & HURWITZ, W.N. (1943). On the theory of sampling from finite populations. The Annals of Mathematical Statistics 14, 333–362.

HANURAV, T. (1967). Optimum utilization of auxiliary information: πps sampling of two units from a stratum. Journal of the Royal Statistical Society: Series B (Methodological) 29, 374–391.

KUMAR BEDI, P. (1995). An alternative estimator in midzuno scheme for multiple characteristics. Communications in Statistics-Simulation and Computation 24, 17–30.

RAO, J.N., HARTLEY, H. & COCHRAN, W. (1962). On a simple procedure of unequal probability sampling without replacement. Journal of the Royal Statistical Society: Series B (Methodological) 24, 482–491.

RAO, J.N.K. (1966a). Alternative estimators in pps sampling for multiple characteristics. Sankhya, A 28, 47–60.

RAO, J.N.K. (1966b). On the relative efficiency of some estimators in pps sampling for multiple characteristics. Sankhya, A 28, 61–70.

REDDY, V. & RAO, T. (1977). Modified pps method of estimation. Sankhya C 39, 185–197.

ROYALL, R.M. (1970). On finite population sampling theory under certain linear regression models. Biometrika 57, 377–387.

SAHOO, L., MISHRA, G. & SENAPATI, S. (2005). A new sampling scheme with inclusion probability proportional to size. Journal of Statistical Theory and Applications 4, 361–369.

SAHOO, L., SC., S. & AK., M. (2010). A class of ipps sampling schemes. Revista Investigacion Operacional 31, 217–224.

SAHOO, L., SINGH, G. & DAS, B. (2006). A note on an ipps sampling scheme. Allgemeines Statistisches Archiv 90, 385–393.

SINGH, H. & TAILOR, R. (2003). Use of known correlation coefficient in estimating the finite population mean. Statistics in transition 6, 555–560.

SINGH, S. & HORN, S. (1998). An alternative estimator for multi-character surveys. Metrika 48, 99–107.