J. Statistics
1998(1), 07-18

ON PROBABILITY PROPORTIONAL TO SIZE SAMPLING

By

ABDUL KADIR A. KATTAN

Department of Mathematical Sciences. Ummul Qurrah University
P.O. Box 7303, Makkah Al-Mukarramah, Saudi Arabia

and

MS. IFFAT KHAWAJA

National Fertilizing and Marketing
Jail Road, Lahore, Pakistan

SUMMARY

A property of Raj-Murthy estimator is discussed. Comparisons of different sampling schemes under a super-population model is made and numerical evaluation is also provided.

Keywords: Unbiassedness, Super-Population Model. Expected Variance.

1. INRODUCTION

Numerous sampling schemes with resulting estimates of the population total in unequal probability sampling are discussed in the literature. Some of the schemes and/or estimates which are more frequently referred to in the literature and with which the present paper is concerned occur in Raj (1956). Murthy (1957), Horvitz and Thompson (1952), Lahiri (1951), Sampford (1967), Durbin (1967) and Brewer (1963). One comes across three broad types of estimators here in called (i) Raj-Murthy. (ii) Ratio and (iii) Hovitz-Thompson estimators. As early as 1955 Godambe (1955) proved the nonexistence of a best linear unbiased estimators. This makes the comparison of different sampling schemes with their resulting estimates extremely difficult. However, numerical comparison by Rao and

Bayless (1969) and Bayless and Rao (1970) seem to suggest that Raj-Murthy estimates perform better in an over all sense. Some progress can be made if the general set up considered by Godambe (1955) is restricted through a super-population model. Such models were discussed by Smith (1938) and Cochran (1953). Relevent work in this direction are Godambe (1955) and Rao, T. J (1971). In the following sections, we discuss a new property of Raj-Murthy estimators. The super-population model is then considered in which among a given class of estimators Hovitz-Thompson estimators are found to be superior. Numerical comparisons are also provided.

2. RAJ-MURTHY ESTIMATORS

Let the characteristics of interest for N units in the population be $Y_1, Y_2, \ldots Y_N$ and the population total be $Y = \sum_{i=1}^{N} Y_i$. Let X_r be the size of the rth unit which is known to us and let $P_r = X_r / X$ where $X = \sum X_r$. Raj (1956) suggested a sampling scheme which chooses the i^{th} unit at the first draw with probability P_i. At the next draw j^{th} unit is selected from amongst the remaining units with probability proportional to p_j and so on. If $y_1, y_2, \ldots\ldots y_n$ are the units selected in the sample of size n in the same order then

$$\left.\begin{aligned} t_1 &= \frac{y_1}{p_1} \\ t_1 &= y_1 + y_1 + \frac{y_1}{p_1}(1 - p_1) \\ &\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ t_n &= y_1 + y_2 + \ldots + y_{n-1} + \frac{y_n}{p_n}(1 - p_1 - p_2 - \ldots - p_{n-1}) \end{aligned}\right\} \quad (2.1)$$

are each unbiased estimators of Y and are uncorrelated. Any linear combination $\sum c_i t_i$, where $\sum c_i = 1$ is also an unbiased estimator of Y. Raj (1956) suggested using $y'_R(n) = \frac{1}{n}\sum t_i R(n)$ which for n = 2 is

$$y'_R(2) = \frac{1}{2}\left[\frac{y_1}{p_1}(1+p_1) + \frac{y_2}{p_2}(1-p_1)\right] \tag{2.2}$$

Murthy (1957) considered all possible permutations of $(y_1, y_2, \ldots\ldots y_n)$ which lead to different estimates $y'_R(n)$. He then proved that weighted average of these estimates, with weights proportional to the probability of the sample in that particular order, leads to $y'_M(n)$ Which has smaller variance than $y'_R(n)$. His estimate for n = 2 is

$$y'_M(2) = \frac{\frac{y_1}{p_1}(1-p_2) + \frac{y_2}{p_2}(1-p_1)}{2-p_1-p_2} \tag{2.3}$$

We will call this Murthy's method of symmetrizing $y'_R(n)$. Perhaps it has not been noticed that if we start with any linear combination $dt_1 + (1-d)t_2$ and then symmetrize we get back $y'_M(2)$ i.e. an expression which is free from d. This result extends to the general case of n. Consider $y'_R(n,C) = \sum C_i t_i$ Where C_i=1. The resulting Murthy's estimate is $y'_M(n)$ which does not depend on the set C of ($C_1, C_2, \ldots, C_n$). The result is contained in the following theorem.

Theorem 1: $y'_M(n,C)$ when symmetrized does not depend on C.

Proof: The result can be easily proved by the method of induction and the fact that the result is clearly true for n = 2.

A consequence of theorem 1 is that while symmetrizing to get $y'_M(n)$ one can consider only t_1 (when $C_1 = 1, C_i = 0, i \neq 1$). This directly leads to the often quoted result that $y'_M(n) = \dfrac{\sum y_i P(s\mid i)}{P(s)}$, where P(s|i) is the probability of the sample with y_i as the first unit selected, where P(s) is the probability of sample. Replacement of $t_1, t_2, \ldots\ldots, t_n$ by t_1 only in symmetrizing to get $y'_M(n)$ may result in some loss of information. Das (1951) has considered a general set of estimators like $(t_1, t_2, \ldots\ldots, t_n)$ which for n = 2 gives

$$t_1' = y_1 / p_1 \ and \qquad t_2' = \frac{1-p_1}{p_1 p_2} \quad \frac{y_2}{N-1} \tag{2.4}$$

Yet an other set of estimators may be

$$\left.\begin{aligned} T_1 &= y_1 + \frac{y_2}{p_2}(1-p_1) \\ T_1 &= y_2 + \frac{y_1}{p_1}(1-p_1)\left(K - \frac{p_2}{1-p_2}\right) \end{aligned}\right\} \tag{2.5}$$

Where $K = \sum_{i=1}^{N} \frac{p_i}{1-p_i}$

Both these can be symmetrized bythe method of Murthy. We denote the symmetrized estimators obtained by (2.4) and (2.5) by t` and T. Numerical calculations for the cases considered here shows that T performs well. However, there is one thing very awkward about both (2.4) and (2.5). Since X`s are known to us (and therefore X is also known) and the estimators in (2.4) and (2.5) are unbiased for Y (for any Y_i`s). One would expect that if Y`s are replaced in (2.4) and (2.5) by X's, the result should be X. We shall call this property A which runs as follows.

Property A: Correspondence to any sampling procedure let y' be the unbiased estimate of Y. If when Y_i-s in y' is replaced by X_i, y' = X the estimator y' is said to enjoy property A.

Estimators in (2.4) and (2.5) do not enjoy property A and as such are seriously defective. Some estimators in Rao, T. J (1971) do not enjoy this property.

We do not recommend use of T. The point of including this here is just to indicate that numerical evidence (even if it spreads over 30 or 40 isolated cases) in favour of an estimator is not a sufficient justification for recommending it.

3. THE LINEAR STOCHASTIC MODEL

For any effective comparison of the estimates, the message from Godambe (1955) is to restrict the generality of the situation. If for example some other characteristics Z_i is known for the ith unit and Y_i is known to depend on Z_i in some stochastic way the generality of the situation can be restricted in a meaningful way. For example, we may have on the lines of Smith (1938) and Cochran (1953).

$$Y_i = f(Z_i) + \in_i \tag{3.1}$$

For its simplicity, Smith (1938) and Cochran (1953) choose $f(Z_i) = Bx_i$ so that (3.1) becomes

$$Y_i = \beta X_i + \in_i \tag{3.2}$$

(3.2) is the usual super-population model with the extra assumptions that $E(\in_i) = 0, E(\in_i \in_j) = 0$ for $i \neq j$ and $Var(\in_i) = \sigma^2 X_i^{2\gamma}$, *where* $\frac{1}{2} \leq \gamma \leq 1$. We assume first a random sample of N units is selected from (3.2) from which a subsequent sample of n units is selected. Given the validity of the model in (3.2) estimation of $\sum_{i=1}^{N} Y_i = Y$ is equivalent to estimating βX. The best linear unbiased estimator of X (in the context of model (3.2)) is

$$\frac{\sum_{i=1}^{n} \frac{y_i}{P_i} p_i^{2(1-\gamma)}}{\sum_{i=1}^{n} p_i^{2(1-\gamma)}} \tag{3.3}$$

However, there are several ways in which the model can go wrong. First $f(Z_i)$ may not provide an adequate description of the deterministic part of the relation between Y_i and Z_i. Secondly assumptions about $\in_i$ may not be valid and finally both these may go wrong. We can take care of the first defect by demanding that the sampling scheme should be such that (3.3) becomes unbiased for Y (for any set of $Y_i^{'}s$). For fixed n there are $\binom{N}{n}$ distinct samples. A sampling scheme assigns $\binom{N}{n}$ probabilities adding upto one to the distinct samples. We do not know whether for a general γ there will exist a sampling procedure which renders (3.3) unbiased for Y but it looks probable because there are $\binom{N}{n}$ adjustable parameters on satisfy N equations for unbiasedness. Note that the estimate in (3.3) enjoys property A.

4. HORVITZ-THOMPSON AND RATIO ESTIMATES

For $\gamma = \frac{1}{2}$ (3.3) becomes

$$y'_{(1/2)} = y'_r = \frac{\sum^{n} y_i}{\sum p_i} \tag{4.1}$$

and for $\gamma = 1$ (3.3) becomes

$$y'_{(1)} = y'_{HT} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{p_i} \tag{4.2}$$

(4.1) is the usual ratio estimate and (4.2) is the Horvitz-Thompson estimates when probability of inclusion of the ith unit $\pi_i = nP_i$. Both the ratio estimator and Horvitz-Thompson estimator enjoy property A. Lahiri's (1951) and Midzuno (1951) sampling schemes make (4.1) unbiased for Y and likewise. Brewer's (1963), Durbin's (1967) and Sampford's (1967) schemes make (4.2) unbiased for Y.

5. EXPECTED VARIANCES OF DIFFERENT SCHEMES

Let s denote the sample ($y_1, y_2, \ldots, y_n$) and P(s) the probability of the sample s (regardless of order). P(s) defines a sampling scheme and we assume that the sampling scheme is such that (3.3) becomes unbiased for a general γ. The expected variance of y'_γ is defined as

$$E\left[Var\left(y'_\gamma\right)\right] = E\sum_{s}\left\{y'_{(\gamma)} - Y\right\}^2 P(s) \tag{5.1}$$

Since y'_γ is unbiased (5.1) reduces to

$$= E\left[\sum_{s}\left\{y'_\gamma\right\}^2 P(s) - Y^2\right]$$

$$= E\left[\sum_s \left\{\frac{\sum \frac{\in_i P_i^{2(1-\gamma)}}{P_i}}{\sum P_i^{2(1-\gamma)}}\right\}^2 p(s) - (\sum \in_i)^2\right]$$

$$= \sigma^2 (\sum X_i)^{2\gamma}\left[E\left\{\sum_s \frac{1}{\sum P_i^{2(1-\gamma)}} P(s)\right\} - \sum P_i^{2\gamma}\right] \qquad (5.2)$$

Now since y'_γ is unbiased for Y for all $y'_i(s)$ we put $Y_i = \frac{1}{n} P_i^{2(\gamma-1)+1}$ which gives $\sum_s \frac{1}{P_i^{2(1-\gamma)}} P(S) = \frac{1}{n}\sum_I^N P_i^{2(1-\gamma)}$ (5.2) then becomes

$$\sigma^2 (\sum X_i)^{2\gamma}\left[\sum P_i^{2\gamma}\left(\frac{1}{nP_i} - 1\right)\right] \qquad (5.3)$$

At $\gamma = 1/2$ and 1 (5.3) reduce to

$\sigma^2 (\sum X_i)\left[\frac{N}{n} - 1\right] and \sigma^2 (\sum X_i)^2\left[\frac{1}{n} - \sum P_i^2\right]$ An implication of (5.3) is that if there are more than one sampling schemes which make y'_γ unbiased there is little to choose between them because the expected variance is the same for all.

The expected variance of the Horvitz-Thompson estimator can be obtained as

$$E\left[Var(y'_{HT})\right] = E\sum_s \left\{\sum \frac{Y_i}{\pi_i} - Y\right\}^2 P(s)$$

$$= \sum_s \left\{\sum \frac{Y_i}{\pi_i}\right\}^2 P(s) - E(Y^2)$$

$$= \beta^2\left[\sum_s \left\{\sum \frac{Y_i}{\pi_i}\right\}^2 P(s) - X^2\right] + \sigma^2\left[\sum X_i^{2\gamma}\left(\frac{1}{\pi_i} - 1\right)\right]$$

When $\pi_i = nP_i$ the first part is zero. Also only in this case property "A" will be enjoyed by the estimator. In this case

$$E[Var(y'_{HT})] = \sigma^2 (\Sigma X_i)^{2\gamma} \left[\Sigma P_i^{2\gamma} \left(\frac{1}{nP_i} - 1 \right) \right]$$

Which is the same as (5.3). We interpret (5.4) as saying that even for general γ, y'_γ has no apparent advantage over y'_{HT} which $\pi_i = nP_i$. This is in contrast to the finding of Rao, T. J (1971) who considers choice of π_i proportional to X_i^γ. However, Rao, T. J's context is more general than ours in that n is not fixed in his study. We also note in passing that the resulting estimator will not enjoy property A. Thus for fixed sample y'_{HT} with $\pi_i = nP_i$ provides a general solution for all values of γ, when assessed on the expected variance of the estimators. In particular $E[\text{var}(y'_\gamma)]$ is equal to $E[\text{var}(y'_{HT})]$ for $\gamma = \frac{1}{2}$. It may be noted that the sampling schemes of Brewer (1963), Durbin (1967) and Sampford (1967) satisfy $\pi_i = nP_i$.

6. NUMERICAL COMPARISON

Six population were generated from the model with $\gamma = \frac{1}{2}$ and 1 they are given in Table 1.

Table 1. Values of X and Y for six generated populations with N = 10.

Population	γ											
1		X	59	47	52	60	67	48	44	58	76	58
	1	Y	124	84	90	110	142	82	101	146	176	106
2	1	X	59	47	52	60	67	48	44	58	76	58
	2	Y	92	63	69	84	105	62	75	107	127	80
3		X	60	52	58	56	62	51	72	48	71	58
	1	Y	76	65	64	72	89	67	101	71	119	107
4	1	X	60	52	58	56	62	51	72	48	71	58
	2	Y	67	57	58	63	76	58	86	60	97	87
5		X	76	138	67	29	381	23	37	120	61	38
	1	Y	79	177	79	36	563	32	50	172	84	47
6	1	X	76	138	67	29	381	23	37	120	61	38
	2	Y	121	338	59	65	1056	73	104	345	171	89

Six other artificial populations with N = 4 were considered in which the ratio Y_i / X_i is D (decreasing), D.F (decreasing fast), I (increasing), I. F (increasing fast) and F.L (fluctuating) with X_i. These are given in table.2

Table 2. Values of X and Y and Y/X for Six artificial population with N=4

Population	7	8	9	10	11	12
X	Y					
0.1	0.7	0.4	0.3	0.6	0.1	0.4
0.2	1.2	1.0	1.0	1.0	0.4	0.8
0.3	1.5	1.8	1.5	1.2	1.2	0.9
0.4	1.6	2.8	1.2	1.2	3.2	1.6
Y_i/X_i	DF	I	FL	D	IF	FL

For n = 2 the variance of y'_M, T, y'_{HT}, y'_r and $y_{HT}(L)$ are provided in Table 3. y'_{HT} is obtained by using Lahiri's scheme but the Horvitz-Thompson estimator.

Table 3. Variances for the different estimators for n = 2 for the 12 populations.

Population	y'_M	T	y'_{HT}	y'_r	$y'_{HT}(L)$
	10333	9744	10348	10388	20752
	4438	4185	4453	4400	9511
	6676	6491	6672	6797	10138
	2904	2811	2904	2945	5015
	6508	4665	5606	10792	167462
	119634	104506	127733	149169	686578
	0.3124	0.6075	0.2822	0.3629	0.2178
	0.3124	0.1172	0.2822	0.3629	2.3294
	0.2801	0.4039	0.2376	0.4048	0.7187
	0.3124	0.5466	0.2822	0.3629	0.1071
	2.1087	1.7695	1.5964	3.0618	0.4833
	0.0589	0.0785	0.0600	0.0879	0.4414

Comparing y'_{HT} and y'_r for population 1 to 6, it is clear that y'_{HT} erforms better not only when γ = 1 but also in cases when γ = 1/2. This s to be expexted because for γ =1, y'_{HT} is better than y'_r where as for =1/2 these are equally good as regards their expected variance. In these ases y'_{HT} performs nearly as well as y'_M. It is surprising that the erformance of T in these six cases is excellent. In cases 7 to 12 the performance of y'_M is bad where y_i / X_i is either I or I. F in which cases T performs well. The reason presumably is that larger values of Y_i / P_i get greater weights in y'_M in this case which is the other way round. y'_M performs well in cases Y_i / X_i is F. $y'_{HT}(L)$ performs extremely well in cases where Y_i / X_i is either D or D.F. The general conclusions are

a. No single estimator performs well in all cases. Note that T is best in 7 out of 12 cases but at the same time one notices its erratic behavior for cases 9 to 10.

b. y'_{HT} is reasonably stable in all cases. We do not find its performance particularly bad in any one of 12 cases. Besides on the basis of expected variance, this is not inferior to any other estimator of the class (3.3).

Thus our finding both on theoretical and empirical evidence goes in favour of y'_{HT}.

ACKNOWLEDGEMENT

We are thankful to Dr. Mohammad Hanif, Professor, King Faisal University, Dammam for his guidance and suggestions. We are also thankful to Mr. Mohammad Junaid, King Fahd University of Petroleum and Minerals, Dhahran for his helpful comments in computation

REFERENCES

[1] **Bayless, D. L. and Rao, J. N. K (1970).** Estimators and variance estimators in unequal probability sampling. *J. Am. Statist. Assoc., 65, 1645-1647.*

[2] **Brewer, K. R. W (1963).** A model of systematic sampling with unequal probabilities. *Aust. J. Statist, 5, 5-13.*

[3] **Cochran, W. G (1953).** Sampling Techniques. *John Wiley and Sons.*

[4] **Das, A. C (1951).** On two sampling and sampling with varying probabilities. *Bull. Int. Statist. Inst., 33. 11, 105-112.*

[5] **Durbin, J (1967).** Design of muti-stage surveys for estimation of sampling error. *Appl. Statist., 16. 152-164.*

[6] **Godambe, V. P (1955).** A unified theory of sampling from finite populations. *J. R. Statist, B, 269-278.*

[7] **Horvitz, D. G. and thompson, D. J (1952).** A generalization of sampling without replacement from a finite universe. *J. Am. Statist. Assoc., 47, 663-685.*

[8] **Lahiri, D. B (1951).** A method for sample selection providing unbiased ratio estimates. *Bull. Int. Statist. Math., 33. 11, 133-10.*

[9] **Midzuno, H (1951).** On the sampling system with probabilities proportionate to sum of sizes. *Ann. Inst. Statist. Math., 3, 99-107.*

[10] **Murthy, M. N (1957).** Ordered and unordered estimators in sampling without replacement. *Sankhya, 18, 379-90.*

[11] **Raj, D (1956).** Some estimators in sampling with varying probabilities without replacement. *J. Am. Statist. Assoc., 17, 197-200.*

[12] **Rao, J. N. K. and Bayless, D. L (1969).** An empirical study of the stabilities of estimator and variance estimators in unequal probability of two units per stratum. *J. Am. Statist. Assoc., 64. 540-559.*

[13] **Rao, T. J (1971).** πPS sampling designs and the Horvitz-thompson estimator. *J. Am. Statist. Assoc., 66, 872-875.*

[14] **Sampford, M. R (1967).** *On sampling without replacement with unequal probabilites.* *Biometrika. 54. 499-513.*

[15] **Smith, H. F (1938).** *An empirical law describing heterogeneity in the yields of agricultural crops.* *J. Agri. Soc.. 28. 1-23.*