EFFICIENCY OF ESTIMATORS AND VARIANCE ESTIMATORS IN SAMPLING WITH TWO UNITS PER STRATUM FOR SMALL ECOLOGICAL POPULATIONS

A. R. SEN and HALYNA BEZNACZUK
Environment Canada, Ottawa, Ontario

(Received : June, 1981)

SUMMARY

Stabilities of estimators of the population total and of their variance estimators have been compared in single-stage sampling with unequal probability by Rao and Bayless [12] [1] for small agricultural and demographic populations.

This paper examines the techniques for estimating the characteristics of breeding bird populations and investigates the conditions under which the estimators would be more efficient. For simple random sampling, an adaptive estimator which is more reliable than based on mean per unit is recommended for use in populations where the minimum value may be very low and the maximum very high.

Introduction

Rao and Bayless [12] [1] compared (*a*) the efficiencies of the estimators $\hat{Y}$ of the population total Y as judged by the inverse of the actual variances and (*b*) stabilities of the sample estimates of the variances of $\hat{Y}$ as judged by the inverses of the variances of the estimators $\hat{v}(\hat{Y})$ in single-stage sampling. The methods were compared in three situations :

(1) 7 very small ($N = 4, 5, 6$) artificial populations,

(2) 20 natural populations with N ranging from 9 to 35,

(3) the super population model with a linear regression

$$y_i = \beta x_i + e_i, i = 1, \ldots, N$$

$$E(e_i \mid x_i) = 0, E(e_i^2 \mid x_i) = ax_i^g$$

$$E(e_i e_j \mid x_i, x_j) = 0, a > 0, g = 1, 1.5, 1.75, 2$$

The authors presented their results as percent gains in efficiency of the estimators over the Brewer [2], Rao [11] and Durbin [5] estimators as standard. Their main conclusions are (i) Murthy's [9] method is preferable, when a stable estimator of total as well as of variance are required, (ii) the Rao-Hartley-Cochran (RHC) estimator [10] of variance is the most stable, but the RHC estimator of population total might lead to significant loss in efficiency.

Cochran [3] summarized Murthy, RHC, probability proportional to size and with replacement (PPSWR) and Brewer methods for the natural populations and the super-population model of Rao and Bayless with $g = 1$, 1.5 and 2 using median values to study percentage gain in efficiency of the variance estimators owing to the highly skewed nature of the distributions.

For the natural populations, the three "without replacement methods" were very close in efficiency for estimating total, the order of preference being Murthy, RHC and Brewer; for the super-population model the Brewer method improved as g increased, the rank order at $g = 2$ being Brewer, Murthy and RHC. For estimating variance, the order of preference was RHC, Murthy and Brewer.

Rao and Singh [13] considered 14 additional natural populations with N ranging from 8 to 13 and presented results on 34 of them. We have investigated the position for 27 natural populations (N ranging from 8 to 25) based largely on breeding bird surveys conducted in North America. We will estimate the gains for these populations and examine if the gains in efficiency over the Brewer-Rao-Durbin estimator are related to other statistics e.g., correlation coefficient so that further gains could be achieved by stratification with respect to the statistics.

For some small ecological populations, it is possible to determine with a high degree of precision if the sample drawn would contain units having high and/or low values in the population. In such cases the estimator based on simple random sampling has been modified to yield an adaptive estimator which is much more efficient than the mean per unit estimator.

The empirical study is based on a sampling scheme of selecting two units from each of the natural populations.

2. Empirical Results

As stated earlier we have chosen 27 populations. Table 1 presents the source, description of y and x, population size, coefficients of variation

TABLE 1—STATISTICS AND DESCRIPTIONS OF ECOLOGICAL POPULATIONS

	Source	*y*	*x*	*N*	*CV(y)*	*CV(x)*	ρ
	1	*2*	*3*	*4*	*5*	*6*	*7*
1.	G. Adams' Breeding Bird Study 1976 (Personal communication)	Blue-winged Teal (pop'n) Early June 1976 (*a*)	Number of ponds May 1976 (*a*)	10	0.45	0.38	0.42
2.	G. Adams' Breeding Bird Study 1976 (Personal communication)	Blue-winged Teal (pop'n) Early June 1976 (*b*)	Number of ponds May 1976 (*b*)	10	0.48	0.32	0.70
3.	G. Adams' Breeding Bird Study 1976 (Personal communication)	Number of ponds May 1976 (*a*)	Number of ponds May 1975 (*a*)	10	0.38	0.40	0.95
4.	G. Adams' Breeding Bird Study 1976 (Personal communication)	Number of ponds May 1976 (*b*)	Number of ponds May 1975 (*b*)	10	0.32	0.33	0.98
5.	G. Adams' Breeding Bird Study 1976 (Personal communication)	Dabblers (pop'n) 1976 (*a*)	Number of Type 3 open ponds, 1976 (*a*)	10	0.51	0.81	0.37
6.	G. Adams' Breeding Bird Study 1976 (Personal communication)	Dabblers (pop'n) 1976 (*b*)	Number of Type 3 open ponds, 1976 (*b*)	10	0.50	0.70	0.93
7.	G. Adams' Breeding Bird Study 1976 (Personal communication)	Number of ponds July 1976 (*a*)	Number of ponds July 1975 (*a*)	10	0.40	0.43	0.84
8.	G. Adams' Breeding Bird Study 1976 (Personal communication)	Number of ponds July 1976 (*b*)	Number of ponds July 1975 (*b*)	10	0.30	0.30	0.95
9.	G. Adams' Breeding Bird Study 1976 (Personal communication)	Blue-winged Teal (pop'n) Early June 1976 (*a*)	Number of ponds July 1975 (*a*)	10	0.48	0.43	0.59

Table 1 (*contd. on page* 36)

Table 1 (*contd. from page* 35)

	1	*2*	*3*	*4*	*5*	*6*	*7*
10.	G. Adams' Breeding Bird Study 1976 (Personal communication)	Blue-winged Teal (pop'n) Early June 1976 (*b*)	Number of ponds July 1975 (*b*)	10	0.48	0.30	0.77
11.	G. Adams' Breeding Bird Study 1976 (Personal communication)	Number of Type 4 & 5 ponds, 1975 (*a*)	Number Type 4 & 5 ponds, 1974(*a*)	10	0.36	0.40	0.88
12.	G. Adams' Breeding Bird Study 1976 (Personal communication)	Number of Type 4 & 5 ponds, 1975 (*b*)	Number Type 4 & 5 ponds, 1974 (*b*)	10	0.25	0.25	0.76
13.	Crissey (1969), CWS Report # 6, Saskatoon Wetlands Seminar, p. 162	Breeding duck pop'n following year, 1954-65	Number of ponds July 1954-65	12	0.15	0.66	0.89
14.	Breeding Bird Survey (1979-80) Southern Ontario	Killdeer (pop'n) per route, 1980	Killdeer (pop'n) per route 1979	10	0.49	0.54	0.68
15.	Crissey (1969), CWS Report # 6, Saskatoon Wetlands Seminar, p. 164	Number of mallard young produced in N. America, (millions), 1955-65	Number of ponds, s. Prairie prov., July 1955-65	11	0.41	0.65	0.71
16.	Crissey (1969), CWS Report # 6, Saskatoon Wetlands Seminar, p. 164	Number of mallard young produced in N. America, (millions), 1955-6, 58-65	Number of ponds, S. Prairie prov., July 1955-6, 58-65	10	0.35	0.70	0.92
17.	Stoudt (1969), CWS Report # 6, Saskatoon Wetlands Seminar, p. 123	Total broods observed, Redvers study areas, 1952-66	Number of ponds, Redvers study area, May 10, 1952-66	15	0.76	0.43	0.32

18.	Stoudt (1969), CWS Report # 6, Saskatoon Wetlands Seminar, p. 123	Total broods observed, Redvers study area, 1952-66	Number of ponds Redvers study area, July 10, 1952-66	15	0.76	0.58	0.66
19.	Dzubin (1969), CWS Report # 6, Saskatoon Wetlands Seminar, p. 206	Mallard breeding pairs, Kindersley study area, 1956-67	Number of ponds, Kindersley study area, May 1956-67	12	0.87	0.51	0.40
20.	Dzubin (1969), CWS Report # 6, Saskatoon Wetlands Seminar, p. 206	Total breeding pairs, Kindersley study area, 1956-67	Number of ponds, Kindersley study area, May 1956-67	12	0.76	0.51	0.51
21.	K. Ross' Snow Goose Study 1976 (Personal communication)	Number of snow geese per plot (*a*)	Number of ponds per plot (*a*)	14	0.57	0.41	—0.02
22.	K. Ross' Snow Goose Study 1976 (Personal communication)	Number of snow geese per plot (*b*)	Number of ponds per plot (*b*)	14	0.84	—0.44	—0.07
23.	Filion (1974), CWS Biometrics Section Manuscript Report Number 13	Total duck kill by age category, Nova Scotia & New Brunswick	Number of hunters by age category, Nova Scotia & New Brunswick	12	0.93	0.74	0.96
24.	Breeding Bird Survey (1979-80) Quebec	Killdeer (pop'n) per route, 1980	Killdeer (pop'n) per route, 1979	10	0.66	0.91	0.92
25.	Breeding Bird Survey (1979-80) Quebec	Eastern Wood Pewee (pop'n) per route, 1980	Eastern Wood Pewee (pop'n) per route, 1979	10	0.95	1.09	0.87
26.	Pilou, "Population and Community Ecology", p. 114	Insects per quadrant	Flowers per quadrant	8	0.57	0.39	0.90
27.	Pilou, "Population and Community Ecology", p. 119	Number of beetles	Weight of fungus (grams)	25	0.80	0.50	0.85

(c. v.) of y and x and correlation ρ between y and x. It would be seen that 25 out of the 27 populations relate to migratory game birds; of these, most of the cases relate to breeding bird populations.

Among the estimators of total we will consider those due to Des Raj [4], Murthy, RHC, Lahiri [7] and PPSWR and among estimators of variance those due to Des Raj, Murthy, RHC, PPSWR and by Rao-Vijayan [14]. The Lahiri estimator was also obtained independently by Hájek [6], Midzuno [8] and Sen [16], and in the sequel we shall, for convenience, denote all these by Lahiri estimator.

A word may be said about the use of the probability proportional to size (PPS) method in selecting the sampling units for the ecological populations. Consider for example, the populations (1 to 4, 7, 8, 9, 10, 11, 12, 14, 24, 25) for which the variable (x) used to determine the selection probabilities is the number of ponds (or birds) in a time-period preceding the period for which the populations totals (number of waterfowl or ponds) are to be estimated. For the small populations this is feasible for selecting the units with PPS in time for measurement of y population; for large populations which can be stratified into several small populations, the selection procedure should not pose any operational problem; often measurement on units of x-population can be made by photographic methods. For the populations of waterfowl (5, 6, 13, 15, 16, 17, 18, 19, 20, 21, 22) for which the variable (x) is the number of ponds, measurement of x variable, which merely amounts to counting of ponds, is a simpler and quicker process than counting of birds and can be done ahead of time for each unit in the population for selection of units with PPS for making count of waterfowl (y).

Similarly, for populations 23, 26 and 27 measurement of the x-population for each unit is relatively much easier to make for PPS selection in time of the sampling units for estimation of the y-population, particularly for the small populations considered in the study.

2.1 *Stabilities of the Estimators and of Variance Estimators*

In presenting results, the Brewer-Rao-Durbin methods will be taken as standard, the figures given being the percent gains (+) or losses (—) in efficiency of the other methods with respect to this method. For obtaining more reliable information a comparison will be made with the relative gains from the 34 agricultural and demographic populations of Rao and Bayless [12] and Rao and Singh [13]. The relative efficiencies have slightly skewed distributions so that their arithmetic means together with the lowest and the highest extreme values will serve as summary statistics which are presented in Table 2.

TABLE 2—PERCENT GAINS IN EFFICIENCY OF THE ESTIMATORS OVER THE BREWER-RAO-DURBIN ESTIMATOR

Natural Population		*Method*				
		Des Raj	*Murthy*	*RHC*	*Lahiri*	*PPSWR*
Table 1	Mean	−0.33	1.19	−0.70	−3.74	−10.37
(#27)	Extremes	(−4,7)	(−2,9)	(−8,4)	(−27,22)	(−18,−4)
Rao and Bayless and Rao and	Mean	0.18	1.94	−0.38	13.38	−9.88
Singh (#34)	Extremes	(−6,12)	(−3,18)	(−8,7)	(−31,511)	(−21,−1)
(#61)	Mean	−0.05	1.61	−0.52	6.08	−10.09
(Total)	Extremes	(−6,12)	(−3,18)	(−8,7)	(−31,511)	(−21,−1)

It would be seen that of the 5 estimators without replacement, the Lahiri estimator performed erratically as regards efficiency of the estimator although the mean gain based on the 61 populations was high compared to others: The gain was more apparent than real being based on one of the 34 populations for which the percentage gain was as high as 511, the population proving most suitable to the Lahiri method because one unit in the population had unusually high value of both y_i and x_i. Omitting this population, the average gain was negative being −2.33. The remaining four without replacement schemes due to Des Raj, Murthy, RHC and Brewer showed little difference amongst themselves though the Murthy estimator proved slightly better than others.

Both the coefficients of skewness and kurtosis were rather high for the percentage gains in efficiency of the variance estimators. We have, therefore, used the median, quantiles and the extreme values as summary statistics which are presented in Table 3. The order of preference for the natural populations is Murthy, Des Raj, Brewer-Rao-Durbin and Rao-Vijayan. There appears to be substantial gain in efficiency in about $\frac{1}{4}$ of the populations for Des Raj, Murthy and Rao-Vijayan estimators, the gains being highly pronounced for the Rao-Vijayan estimators.

3. Relation Between Gains and Known Statistics

A multiple correlation between the gains in efficiency (y) of the estimators due to Des Raj, Murthy, RHC, Lahiri and PPSWR (over the Brewer-

TABLE 3—PERCENT GAINS IN EFFICIENCY OF THE VARIANCE ESTIMATOR OVER THE BREWER-RAO-DURBIN VARIANCE ESTIMATOR

		Method				
Natural Populations		*Des Raj*	*Murthy*	*RHC*	*Rao-Vijayan*	*PPSWR*
	Median	*6*	*6*	*−11*	*−1*	*−3*
Table 1 (#27)	Quartiles	(3.25,10)	(3.25,9.75)	(−54.5,4.5)	(−22.75,42.75)	(−11,8)
	Extremes	(−1,46)	(−1,44)	(−100,63)	(−42,84)	(−28,82)
Rao and Bayless and Rao and Singh (#34)	Median	5.5	5.5	−46	−3.5	−4.5
	Quartiles	(2,21)	(2,19)	(−89,4)	(−36,33)	(−13,8)
	Extremes	(−5,332)	(−5,301)	(−100,132)	(−60,642)	(−25,322)
(#61) Total	Median	6	6	−28	−2	−5
	Quartiles	(3,15)	(3,13 25)	(−81,4)	(−26,34.5)	(−12,0.5)
	Extremes	(−5,332)	(−5,301)	(—100,132)	(−60,642)	(−25,332)

Rao-Durbin estimator) and the variables ρ and λ (= c.v. (y/x)/c.v. (y)) was worked out to find the relative importance of the variables and hence to suggest a basis for stratification. It is known that λ represents approximately the relative variances of the methods to variance of the estimator based on simple random sampling. None of the regression lines was significant though it appeared that higher gains in Murthy estimator were associated with higher (five) values of ρ and of Lahiri estimator with positive and higher values of λ.

A multiple regression of the gains in efficiency of the variance estimator (over the Brewer-Rao-Durbin estimator) over the variables ρ and λ was worked out for the Murthy, RHC, Rao-Vijayan and PPSWR estimators. The regression equations along with the F values (where significant) on percent gains are:

$$\text{RHC} \qquad y = -7.05 - \underset{\substack{\pm\\ 24.67}}{56.20^{*}\rho} + \underset{\substack{\pm\\ 11.50}}{20.24\lambda} \; (F_{2,24} = 5.51,\; P < 0.05)$$

$$\text{Rao-Vijayan} \quad y = -30.90 + \underset{\substack{\pm\\ 26.64}}{27.46\rho} + \underset{\substack{\pm\\ 12.42}}{27.93^{*}\lambda} \; (F_{2,24} \doteq 2.64,\; P < 0.10)$$

It would be seen that gains in efficiency of the variance estimator for RHC were negatively associated with increase in values of ρ but positively associated (regression of y on λ was on the verge of significance) with increase in values of λ. Similarly for the Rao-Vijayan estimator, the gains are positively and significantly associated with increase in λ; the association with ρ is positive but not significant. Thus, stratification of the populations on the basis of ρ and λ is likely to result in overall gains in efficiency of the variance estimators for RHC and Rao-Vijayan. For the Murthy estimator of variance, values of ρ were positively associated with gains though the regression coefficient was not significant. Assuming that the y's and x's are correlated and the regression function involving y, ρ and λ fairly stable over years, the values of the unknowns y, ρ and λ can be replaced by their corresponding values from the previous seasons to provide a basis for stratification during the current season.

4. Efficient Estimators Based on Simple Random Sampling

In many biological populations simple random sampling (*SRS*) is found most convenient. Where it is realized after a decision to take a simple random sample of size n has been made that Y_1 would be unusually low and Y_N unusually high, Särndal [15] has suggested the following adaptative estimator of total which is unbiased and more efficient than the unbiased estimator $\hat{Y}$ based on *SRS*.

$\hat{Y}_s = \hat{Y} + c$ if the sample contains at least one element $\leqslant Y_1$ but no element $\geqslant Y_1$

$= \hat{Y} - c$ if the sample contains at least one element $\geqslant Y_N$ but no element $\leqslant Y_1$

$= \hat{Y}$ for all other samples

where c is a constant.

It can be shown that $\hat{Y}_s$ is unbiased with

$$V(\hat{Y}_s) = N^2(1-f)\left[\frac{S^2}{n} - \frac{2c}{(N-1)}(Y_N - Y_1 - nc)\right]$$

so that $V(\hat{Y}_s) < V(\hat{Y})$ if $0 < c < (Y_N - Y_1)/n$; also $V(\hat{Y}_s)$ will decrease (and hence the precision of $\hat{Y}_s$ will increase) with increasing values of $Y_N - Y_1$ where S^2 is the population variance of Y. In particular Y_1 may be zero kill by a hunter which is rather of frequent occurrence, and Y_N a reasonably high kill by a hunter determined apriori on the basis of past performance of the population of hunters whose kill is to be estimated during the current season.

In general, Y_N and Y_1 of a population will not be known and hence $\hat{Y}_s$ cannot be obtained. However, in some situations it may be possible to know if the sample would contain Y_N and Y_1 with a high degree of precision without knowing the true values. Thus, for example, y may be highly correlated with an auxiliary variable $x(x_1, \ldots, x_N$ being known) and Y_N, Y_1 can be estimated from the relation between y and x in the sample and using values of X_N, X_1 in the population. In particular, x may be the value of y on a previous occasion. For this, using optimum values of c, expression for $V(\hat{Y}_s)$ reduces to

$$V_1(\hat{Y}_s) = N^2(1-f)\left[\frac{S^2}{n} - \frac{b^2(X_{Max} - X_{Min})^2}{2n(N-1)}\right]$$

where b is the sample regression coefficient of y on x.

The expected percentage gain in efficiency of $\hat{Y}_s$ over $\hat{Y}$ for $n = 2$ based on an enumeration of all possible samples in the populations 3, 4, 6, 8, 13, 16, 23, 24, and 26 from Table 1 are given in Table 4 which shows that the gains are appreciably high in all the cases. In practice, the gains may be somewhat lower since the regression coefficient b will be subject to error.

Where wide variation in the high and low values is envisaged before selection of the sample, an alternative plan for $n > 2$ would be to use stratified sampling by including X_N and X_1 in every sample, and drawing a simple random sample of n-2 members from $x_2, \ldots, x_{N-1}$ (excluding

TABLE 4—PERCENT GAIN IN EFFICIENCY OF $\hat{Y}_s$ OVER *SRS* ESTIMATOR $\hat{Y}$ FOR SPECIFIED POPULATIONS FROM TABLE 1

Populations #	*Percent gain* $\hat{Y}_s$
3	77
4	50
6	54
8	78
13	54
16	94
23	75
24	65
26	95

X_N and X_1) and using the corresponding y's to obtain $\bar{y}_{n-2}$. An unbiased estimate of total would be

$$Y_N + (N-2)\,\bar{y}_{n-2} + Y_1$$

where Y_N and Y_1 can be estimated from the regression of y on x.

However, when data are available on an auxiliary variable highly correlated with the variable under study, it would be preferable to adopt a PPS design for selection of sampling units and the Murthy method for estimation of population total and its variance.

REFERENCES

[1] Bayless, D. L. and Rao, J. N. K. (1970): An empirical study of stabilities of estimators and variance estimators in equal probability sampling ($n = 3$ or 4), *J. Amer. Statist. Ass.*, **65**; 1645-1667.

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

[3] Cochran, W. G. (1974): Two Recent Areas of Sample Survey Research. In: J. N. Srivastava (ed.), *A Survey of Statistical Design and Linear Models*. North-Holland Publishing Company, 1975, 101-115.

[4] Des Raj (1956): Some estimators in sampling with varying probabilites without replacement, *J. Amer. Statist. Ass.*, **51**; 269-284.

[5] Durbin, J. (1967): Estimation of sampling errors in multi-stage samples, *Appl. Statistics*, **16**; 152-164.

[6] Hájek, J. (1949): A two-phase method for cluster Sampling, *Stat. Obzov.*, **29**(4); 384-394.

[7] Lahiri, D. B. (1951): A method for sample selection providing unbiased ratio estimates, *Bull. Int. Stat. Inst.*, 33(2); 133-140.

[8] Midzuno, H. (1952): On the sampling system with probability proportionate to sum of sizes, *Ann. Inst. Stat. Math.*, 2; 99-108.

[9] Murthy, M. N. (1957): Ordered and unordered estimators in sampling without replacement. *Sankhya*, 18, 379-390.

[10] Rao, J. N. K., Harthy, H. O. and Cochran, W. G. (1962): On a simple procedure of unequal probability sampling without replacement, *J. Roy. Statist. Soc. Ser.* 13, 24; 482-491.

[11] Rao, J. N. K. (1965): On two simple schemes of unequal probability sampling without replacement, *J. Indian Statist. Ass.*, 3; 173-180.

[12] Rao, J. N. K., and Bayless, David L. (1969): An empirical study of the Stabilites of Estimators and Variance Estimators in Unequal Probability Sampling of Two Units per Stratum, *J. Amer. Statist. Ass.*, 64; 540-59.

[13] Rao, J. N. K., and Singh, M. P. (1973): On the choice of estimator in survey sampling, *Australian Jour. Stat.*, 15; 95-104.

[14] Rao, J. N. K. and Vijayan (1977). On estimating the variance in sampling with probability proportional to Aggregate Size, *J. Amer. Statist. Ass.*, 72; 579-584.

[15] Särndal, Carl-Erik (1972): Sample Survey Theory vs. General Statistical Theory: Estimation of the Population Mean, *Rev. Int. Stat. Inst.*, 40; 1-12.

[16] Sen, A. R. (1952): Further developments of the theory and application of the selection of primary sampling units with special reference to N. C. Agricultural Population. *Ph.D. Thesis*, Library, N. C. State College.