An Efficient and Easy to Carry out Sampling Design in Environmental Studies

M. Moradi [a], **J. Brown** [b], **N. Karimi** [c]

[a]*Department of Statistics, College of Science, Razi University, Kermanshah, Iran.*
[b]*Department of Mathematics and Statistics, University of Canterbury, Private Bag 4800, Christchurch, New Zealand.*
[c]*Department of Biology, College of Science, Razi University, Kermanshah, Iran.*
Email: moradi_m@razi.ac.ir

Abstract:

In survey sampling, finding an efficient sampling design, with both a logistically feasible sampling scheme and an efficient estimator is the ultimate goal. Some efficient estimators when used with complex sampling design have such complicated formula that calculations are impossible even by computers. This can be the case with the Murthy's estimator (1957) when used in designs with unequal selection probabilities and without replacement. This complexity of the estimator often means that the designs are not used to field studies in environmental studies, ecological studies, biological studies, and so on. In this paper, Murthy's estimator is approximated for sampling designs with unequal selection probabilities and without replacement. This approximation produces an estimator that is fairly simple to calculate. We use it for a real population, a contaminated area with arsenic in Kurdistan. The efficiency of the introduced sampling design is compared with other counterpart estimators. The auxiliary variable, arsenic concentration values in the soil are used to improve the efficiency of arsenic concentration estimator in plants.

Keywords: Approximated estimator, Arsenic contamination in Kurdistan, Raj's estimator, Murthy's estimator, Sampling with unequal selection probabilities and without replacement

1 INTRODUCTION

Surveys for rare populations is difficult, and when confronted with difficult conditions in the field study designing a survey scheme that will produce a sample with appropriate information to describe this population is very hard. In such cases, a probability sampling scheme that aims to select a sample with a large rate of rare units is desirable. An example of difficult field conditions is estimating the Arsenic (As) pollution value of soil or plants in a large area, when measuring the pollution for each unit involves an expensive and time consuming test. Moreover, when the polluted area is only a small subpopulation of wide area, the sample selection will be more cumbersome. A post stratification sampling design is a conventional method for estimating subpopulation parameters. This method has problems when applied to studies of rare subpopulation which the number of selected rare units is variable and it may be an insufficient value. To overcome such problem and to avoid selecting predetermined number of rare units, a kind of inverse sampling designs can be used. Haldane (1945) proposed inverse sampling in which one continues sampling until a predetermined number of rare events of interest is observed. In inverse sampling, it is guaranteed that a predetermined number of rare units are selected. Inverse sampling with unequal selection probabilities and with replacement was introduced by Greco and Naddeo (2007). In Greco and Naddeo's sampling design, the sampling is continued until the predetermined numbers of rare units are selected. Since sampling is with replacement the number of distinct rare units may be smaller than the predetermined number of rare units. This is a very real problem and an unfortunate deficiency in studies of rare subpopulations.

In this study we use a simple sampling design and a simple estimator which has an acceptable efficiency. We use the conventional design of sampling with unequal selection probabilities. To overcome the risk of selecting a small number of rare units, we increase the chance of selecting rare units and decrease the chance of selecting none rare ones by appropriate defining selection probabilities. Auxiliary variables which correlated to the response variable can be used to determine these probability selections.

In survey sampling, sampling designs without replacement are usually more efficient than those that are with replacement. On the other hand, finding estimators for the with replacement case is easier. Some unbiased estimators are introduced for sampling with unequal selection probabilities and without replacement by Das(1951), Raj(1956), Murthy (1957), and Rao *et al.* (1962). Murthy's estimator is more efficient than Das's and Raj's estimator. Murthy's estimator was more studied by other authors like Pathak (1967a), Pathak (1967b) , Rao and Bayless (1969), Bayless and Rao (1970), and Samiuddin *et al.* (1992). Salehi and Seber (2001) generalized Murthy's estimator for any sequential sampling design.

The Murthy's estimator in the sampling with unequal selection probabilities and without replacement has a complicated formula and cannot be calculated without computers for medium or large samples. We introduce a method to approximate the Murthy's estimator. The approximated Murthy's estimator is less efficient than Murthy's estimator but it can be calculated quickly and it is easy for use. For example in a sample of size 6, Murthy's estimator is calculated based on $6! = 720$ permutations, where the approximated estimator can be calculated based on 3 pairs of units with 2 permutations for each of them, or based on two sets with 3 units. The efficiency of Raj's estimator and approximated Murthy's estimator are accessed in a simulation study for As pollution in Kurdistan.

Dashkasan antimony - arsenic - gold deposit located in Kurdistan province, western Iran, is one of the most antimony producing areas in Iran; and is associated with elevated environmental levels of As, gold and antimony (Moritz *et al.*, 2006). This contamination originated in As-rich ores that were liberated both by the mining and smelting operations carried out in this area (Lescuyer *et al.*, 2003). Arsenic in soils, water and food is a global health concern due to its toxicity, even at low concentrations. Arsenic-contaminated soil is one of the major sources of arsenic in drinking water (Polya *et al.*, 2008). The concentration of arsenic in cereals, vegetables and fruits is directly related to the level of arsenic in the soil. Severe arsenic contamination of soils may cause arsenic toxicity in plants, animals, and humans (Warren *et al.*, 2003). Remediation of arsenic contaminated soils has thus become a major environmental issue. The promising technology available for the removal of As from contaminated soils is the phytoremediation technology, in which living plants are used to remove As from impacted soil (karimi *et al.*, 2010). Therefore the relationship of As concentration in soil and plants is good guide for knowing the fate of plants for remediation of As contaminated area.

2 Sampling with unequal selection probabilities and without replacement.

Let $U = \{1, 2, ..., N\}$ be the under studying population and the set $\{p_1, p_2, ..., p_N\}$ be the selection probabilities for units in the population, respectively. In sampling with unequal selection probabilities and without replacement, units are selected sequentially correspond to the set selection probability in each selection until to get the predetermined sample size n.

The probability of obtaining the ordered sample set $s_O = (i_1, i_2, ..., i_n)$, and the unordered sample set $s = \{i_1, i_2, ..., i_n\}$ are given as follows. The sets are shown by $s_O = (1, 2, ..., n)$ and $s = \{1, 2, ..., n\}$, for simplicity.

$$P(s_O) \quad = \quad p_1 \times \frac{p_2}{1-p1} \times ... \times \frac{p_n}{1-p_1-...p_{n-1}} = \frac{\prod_{i=1}^{n} p_i}{\prod_{t=1}^{n-1}(1-\sum_{i=1}^{t} p_i)}$$

and,

$$P(s) \quad = \quad \sum_{g=1}^{n!} P(s_{O_g}) = (\prod_{i=1}^{n} p_i) \sum_{g=1}^{n!} \frac{1}{\prod_{t=1}^{n-1}(1-\sum_{i=1}^{t} p_{i_g})} \tag{1}$$

where s_{O_g} is the gth permutation and $g = 1, 2, ..., n!$.

3 Conventional estimators

For this sampling design usually some estimators like, Das (1951), Raj (1956), Murthy (1957), and Rao *et al.* (1962) are used. In this paper we introduce Das's estimator, Raj's estimator and Murthy's estimator, then we introduce an approximating method to calculating Murthy's estimator.

3.1 Das's estimator

Das (1951) introduced the following estimator for sampling with unequal selection probabilities and without replacement:

$$\hat{\tau} = \sum_{r=1}^{n} c_r t_r$$

where c_r are such that $\sum_{r=1}^{n} c_r = 1$, and for simplicity Das choose $c_r = n^{-1}$ and

$$t_r = \frac{(1-p_1)(1-p_1-p_2)...(1-p_1-p_2-...-p_{r-1})}{(N-1)...(N-r+1)p_1...p_r} y_r$$

In such an estimator the order of selection is taken into account, means the estimator is calculated based on s_o.

3.2 Raj's estimator

Raj (1956) introduced a series of estimators based on the order of selection given by Das (1951), where the general form of the introduced estimator is as follows:

$$\hat{\tau}_1 \quad = \quad \frac{y_1}{p_1}$$

and $\hat{\tau}_i$ for $i = 2, ..., n$ is given by:

$$\hat{\tau}_i = \sum_{j=1}^{i-1} y_j + \frac{y_i}{p_i} \times (1 - \sum_{j=1}^{i-1} p_j)$$

The final estimator is calculated as follows:

$$\hat{\tau}_R = \sum_{i=1}^{n} a_i \hat{\tau}_i$$

where $\sum_{i=1}^{n} a_i = 1$. Raj (1956) let $a_i = 1/n$, therefore the estimator is given by:

$$\hat{\tau}_R = \frac{1}{n} \sum_{i=1}^{n} \hat{\tau}_i \tag{2}$$

3.3 Murthy's estimator

Murthy (1957) improved Raj's estimator by Rao-Blackwel theorem, such that the estimator is given by calculating the expectation of Raj's estimator conditioned on minimal sufficient statistics s. Murthy's estimator which is a function of unordered sample set has the following form. Rao-Blackwel estimator can be calculated as following

$$\hat{\tau}_M = \sum_{i=1}^{n} \frac{P(s|i)}{P(s)} y_i$$

In sampling with unequal selection probabilities and without replacement, $P(s)$ is calculated from (1) and $P(s|i)$ is given by:

$$P(s|i) \quad = \quad (\frac{\prod_{i=1}^{n} p_i}{p_i}) \sum_{s_{og} \in B_i} \frac{1}{\prod_{t=1}^{n-1} (1 - \sum_{i=1}^{t} p_{i_g})}$$

where B_i is the set of all permutations that ith unit is the first.

3.4 An approximating method for calculating Murthy's estimator

We approximate Murthy's estimator by computing Murthy's estimators for sequential paired units in the ordered set then combining the estimator values, by appropriate linear combination, into a single estimator. Let n be odd and the ordered set is partitioned into $n/2$ pairs $(1, 2), (3, 4), ..., (n - 1, n)$. The first Murthy's estimator (1957) is calculated based on the first paired units, as:

$$\hat{\tau}_{M_1} = \sum_{i=1}^{2} \frac{P(s|i)}{P(s)} y_i = \frac{1}{2 - (p_1 + p_2)} [\frac{y_1}{p_1}(1 - p_2) + \frac{y_2}{p_2}(1 - p_1)]$$

An unbiased variance estimator of $\hat{\tau}_{M_1}$ is

$$\hat{V}_1 = V\hat{a}r(\hat{\tau}_{M_1}) = \frac{(1 - p_1)(1 - p_2)}{4} (\frac{y_1}{p_1} - \frac{y_2}{p_2})^2.$$

The second Murthy's estimator is calculated based on the second paired units and conditioned on selecting the first pair, as:

$$\hat{\tau}_{M_2} = \sum_{i=1}^{2} y_i + \frac{1}{2 - \frac{p_3 + p_4}{1 - \sum_{i=1}^{2} p_i}} [\frac{y_3}{p_3}(1 - \sum_{i=1}^{2} p_i - p_4) + \frac{y_4}{p_4}(1 - \sum_{i=1}^{2} p_i - p_3)]$$

An unbiased variance estimator of $\hat{\tau}_{M_2}$ is

$$\hat{V}_2 = V\hat{a}r(\hat{\tau}_{M_2}) = \frac{(1 - \sum_{i=1}^{2} p_i - p_3)(1 - \sum_{i=1}^{2} p_i - p_4)}{4} (\frac{y_3}{p_3} - \frac{y_4}{p_4})^2.$$

...

The $n/2$th Murthy's estimator is calculated based on the last paired units and conditioned on selecting the previous pairs, as:

$$\hat{\tau}_{M_{n/2}} = \sum_{i=1}^{n-2} y_i + \frac{1}{2 - \frac{p_{n-1}+p_n}{1-\sum_{i=1}^{n-2} p_i}} [\frac{y_{n-1}}{p_{n-1}}(1 - \sum_{i=1}^{n-2} p_i - p_n) + \frac{y_n}{p_n}(1 - \sum_{i=1}^{n-2} p_i - p_{n-1})]$$

An unbiased variance estimator of $\hat{\tau}_{M_2}$ is

$$\hat{V}_{n/2} = V\hat{a}r(\hat{\tau}_{M_{n/2}}) = \frac{(1 - \sum_{i=1}^{n-2} p_i - p_{n-1})(1 - \sum_{i=1}^{n-2} p_i - p_n)}{4} (\frac{y_{n-1}}{p_{n-1}} - \frac{y_n}{p_n})^2.$$

Now, the final estimator is calculated by linear combination $\hat{\tau}_A = \sum_{i=1}^{n/2} a_i \hat{\tau}_{M_i}$. If $\sum_{i=1}^{n/2} a_i = 1$, then $\hat{\tau}_A$ is unbiased. In order to find the optimum values of a_i such that $\hat{\tau}_A$ be unbiased with minimum variance, assume the population size N is so large and the pairs are approximately independent. Therefore the optimum values for a_i that guarantees unbiasedness and minimum variance of $\hat{\tau}_A$ are given by:

$$a_i = \frac{1/Var(\hat{\tau}_{M_i})}{\sum_{j=1}^{n/2} 1/Var(\hat{\tau}_{M_j})}$$

It is clear that the optimum a_is are dependent on the parameters and cannot be given from the sample. We use the estimator $\hat{a}_i$ in the linear combination as:

$$\hat{a}_i = \frac{1/V\hat{a}r(\hat{\tau}_{M_i})}{\sum_{j=1}^{n/2} 1/V\hat{a}r(\hat{\tau}_{M_j})} = \frac{1/\hat{V}_i}{\sum_{j=1}^{n/2} 1/\hat{V}_j}$$

Then so:

$$\hat{\tau}_A = \frac{\sum_{i=1}^{n/2} \hat{\tau}_{M_i}/\hat{V}_i}{\sum_{j=1}^{n/2} 1/\hat{V}_j}$$

Calculating the variance for the introduced estimator is difficult, for simplicity assume $Var(\hat{\tau}_{M_i}) \simeq \hat{V}_i$ in all pairs, therefor the variance is given by:

$$Var(\hat{\tau}_A) = \frac{\sum_{i=1}^{n/2} Var(\hat{\tau}_{M_i})/\hat{V}_i^2}{(\sum_{j=1}^{n/2} 1/\hat{V}_j)^2} = (\sum_{j=1}^{n/2} \hat{V}_j^{-1})^{-1}$$

For simplicity of introduced approximated estimator, the sample is partitioned into some pairs units. Where the sample can be partitioned into some sets of 3, 4 or larger units. For example when the sample size is 20, we can partition the sample into 4 sets containing 5 units.

4 SIMULATION

We conduct a simulation study to estimate the efficiency of the approximated Murthy's estimator and Raj's estimator in the As contamination of Kurdistan population. Karimi *et al.* (unpublished data) determined the As concentration of soil and plants in an contaminant region of Dashkasan. A total of 50 plant samples belonging to 49 different species were collected from the different locations of study sites. The soil samples were taken near the roots of the plants (0 - 10 cm depth). The process of measuring the As value in soil is easier than plants. In this simulation study, the As values in soil shown by x_i for $i = 1, 2, ..., 50$ is assumed as an auxiliary variable and As values in plants shown by y_i for

Table 1. Relative efficiency of Raj's estimator and Approximated estimator in Populations a and b.

Population	n	$e(\hat{\tau}_R)$	$e(\hat{\tau}_A)$
a	4	8.1	10.7
a	6	8.2	8.9
a	8	8.6	7.5
b	4	0.47	4.2
b	6	0.48	3.4
b	8	0.47	2.8

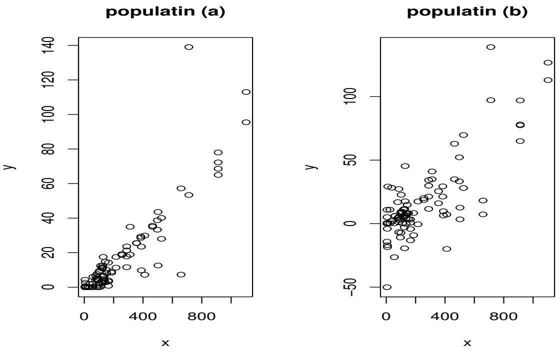

Figure 1. Two simulated populations. Variable x is Arsenic value in soil and variable y is Arsenic value in plant.

$i = 1, 2, ..., 50$ as response variable. In order to construct larger populations of size 100, we simulate two sets containing 50 units based on the initial 50 units. Based on the regression model of the initial 50 units, $\hat{y}_i = -3.947 + 0.0874x_i$, we have simulated two populations of size 100. The first population contains the first 50 units and 50 simulated units by $y_i = -3.947 + 0.0874x_i + e_i$ where $e_i \sim N(0, 4)$. The second population is the first 50 units and 50 simulated units as $y_i = -3.947 + 0.0874x_i + e_i$ where $e_i \sim N(0, 20)$. We show such populations by a and b, respectively and they are shown in Figure (1). We determine the efficiency of estimators $\hat{\tau}_R$, and $\hat{\tau}_A$. For each estimator correspond to $n = 4, 6, 8$, we calculate

$$MSE(\hat{\tau}_\star) = \frac{1}{39999} \sum_{i=1}^{40000} (\hat{\tau}_\star - \bar{\hat{\tau}}_\star)^2 + (\bar{\hat{\tau}}_\star - \tau_y)^2$$

where $\bar{\hat{\tau}}_\star = \sum_{i=1}^{40000} \hat{\tau}_\star / 40000$ and $\star$ stands for R and A. Correspond to each sample size, we have calculated the efficiency as:

$$e(\hat{\tau}_\star) = \frac{N^2(1/n - 1/N)S^2}{MSE(\hat{\tau}_\star)}.$$

The results summarized in Table (1) show that in the first population both of Raj's estimator and Approximated estimator are efficient, and for $n = 4, 6$ Approximated estimator is more efficient than Raj's estimator. In population a, $\rho(x, y) = 0.89$ and in population b, $\rho(x, y) = 0.80$. In population b which variables are less correlated, Raj's estimator is not efficient but Approximated estimator is efficient yet.

5 CONCLUSIONS

In environmental studies especially in estimating pollution, expensive and time consuming tests are needed. Therefore researchers are obligated to investigate an only part of the population instead of all the population. In order to generalize the results from these samples to all population, sampling designs based on probability statistical methods are needed. Efficient sampling designs are desirable, because they help in inferring the sample results to the population with a small cost and with accuracy. Information that is related to the target population, especially where the related information is relatively cheap to acquire, can be used in sampling designs to increase the efficiency. The introduced sampling design in this paper is more efficient than other counterpart sampling designs even when the available auxiliary information has only a weak association with the target population.

In surveys of Arsenic pollution, relatively low-cost information is usually available before the study, for instance: 1) the rate of dermatological lesion, hypertension and other symptoms of As toxicity in cattle or human beings, 2) previous studies regarding As content of water, soil and cattle, 3) environmental and geological information from previous studies, and 4) existence of gold and antimony mines or some contaminator manufacturing industries in the area. Such information can be used in the introduced sampling design which has an appropriate efficiency and can be calculated simply.

ACKNOWLEDGEMENT

The first author would like to thank the financial supports from Razi University of Kermanshah.

REFERENCES

Bayless, D. L. and J. N. K. Rao, (1970). An empirical study of stabilities of estimators and variance estimators in unequal probability sampling for n=3 and n=4. J. Amer. Stat. Assoc. *65*, 1645–1667.

Das, A, C. (1951). On the phasmpling and sampling with varying probabilities. Bull. Inter. Stat. Inst *33*(2), 105–112.

Haldane, J, B, S. (1945). On a method of estimating frequencies. , *Biometrika.*, *33*, 222.

Greco, L. and Naddeo. (1945). Inverse sampling with unequal selection probabilities. , *Communication in Statistics Theory and Methods. 36*(5), 1039.

Karimi, N., S. M. Ghaderian, H. Maroofi, H. Schat. (2010). Analysis of arsenic in soil and vegetation of a contaminated area in Zarshuran, Iran. Int. J. Phytoremediat. *12*, 159–173.

Lescuyer, J. L., Z. A. Hushmand, F. Daliran. (2003). Gold metallogeny in Iran: a preliminary review, in Eliopoulos, D.G. et al.., (Eds.) *Mineral exploration and sustainable developement: Rotterdam, Netherlands, Millpress 2*, 1158–1188.

Moritz, R., F. Ghazban, B. S. Singer. (2006). Eocene Gold Ore Formation at Muteh, Sanandaj-Sirjan Tectonic Zone, Western Iran: A result of Late-Stage Extension and Exhumation of Metamorphic Basement Rocks within the Zagros Orogen. *Economical Geology 101*, (8) 1497–1524.

Murthy, M. N. (1957). Ordered and unordered estimators in sampling without replacement. *Sankhya 18*, 379–390.

Pathak, P, K. (1967a). Asymptotic efficiency of Des Raj's strategy-I. *Sankhya A*, (29) 283–298.

Pathak, P, K. (1967b). Asymptotic efficiency of Des Raj's strategy-II. *Sankhya A*, (29) 299–304.

Polya, D. A., M. Berg, A. G. Gault, and Y. Takahashi. (2008). Arsenic in groundwaters of South-East Asia: with emphasis on Cambodia and Vietnam. Appl. Geochem.. *23*, 2968–2976.

Raj, D. (1956). Some estimators in sampling with varying probabilities without replacement. J. Amer. Stat. Assoc. *60*, 278–284.

Rao, J. N. K, H. O. Hartley and W. G. Cochran. (1962). On a simple procedure of unequal probability sampling without replacement. J. Roy. Stat. Soc *B*, (24) 482–491.

Rao, J. N. K. and D. L. Bayless (1969). An empirical study of the stabilities of estimators and variance estimators in unequal probability sampling of two units per stratum. J. Amer. Stat. Assoc. *64*, 540–549.

Salehi, M. M. and G. A. F. Seber (2001). A new proof of Murthy's estimator which applies to sequential sampling. *Australian and New Zealand Journal of Statistics 43*, 3 901–906.

Samiuddin, M., A. K. A. Kattan, M. Hanif, and H, Asad (1992). Some remarks on models, sampling schemes and estimators in unequal probability sampling. Pak, J. Stat. *8(1)*, A 1–18.

Warren, G. P., B. J. Alloway., N. W. Lepp, B. Singh, F. J. M. Bocherau, C. Penny. (2003). Field trials to assess the uptake of arsenic by vegetables from contaminated soils and soil remediation with iron oxides. Sci. Total Environ., *311*, 19–33.