AN EMPIRICAL STUDY ON MAXIMIZING (OR MINIMIZING) THE NUMBER OF RETENTIONS IN UNEQUAL PROBABILITY SAMPLING WITHOUT REPLACEMENT: TWO UNITS PER STRATUM*

How Tsao and Tommy Wright, Oak Ridge National Laboratory

ABSTRACT

Sampling strategies are determined to maximize (or minimize) the expected number of overlaps between two successive samples using the linear programming approach. The abilities to control overlaps between two successive samples are compared for five methods of sampling with two units selected, in terms of the maximum (or minimum) expected number of overlaps achieved. Six different types of artificial populations are considered to support the empirical study and to illustrate the differences among the five sampling procedures.

INTRODUCTION

In the past, various sample selection methods with unequal probabilities without replacement (UP wor sampling methods) were proposed when sampling n units from a stratum of N units. UP wor sampling methods such as Murthy's (1957) method will determine the selection probabilities of getting any sample of size n from a stratum when "measures of stratum unit size" are known for a particular occasion. This means that an UP wor sampling method will induce a discrete probability density function on the sample space Ω, the set of all possible samples of size n from the given stratum, based on given measures of stratum unit size.

Suppose samples of size n are to be drawn in two different sampling occasions using a particular UP wor sampling method, and that the same units are in a given stratum on the two occasions. The change of time as well as other time-associated factors over the two occasions may change the original measures of stratum unit size and these changes in measures of stratum unit size will, in turn, change the selection probabilities of the UP wor sampling method.

In this article, we consider the use of an UP wor sampling method, denoted by A, for two sampling occasions. We will use the term "component" to represent actions taken on one of the two occasions but not both. We will also use the term "population" to replace "stratum" since a stratum will act like a single population under the cases that we studied.

Hence we let

$U = \{1, \ldots, N\}$ represent the N population units,

$M = \binom{N}{n}$ be the number of all possible samples of size n from U, and

$\Omega = \{s_1, \ldots, s_M\}$ be the M possible samples of size n from U.

On the first sampling occasion, A induces a component probability density μ_A characterized by the M numbers $p_1, \ldots, p_M$ with

$$\sum_{i=1}^{M} p_i = 1 \text{ and } p_i > 0 \text{ for all } i,$$

where for each sample $s_i \in \Omega$,

$\mu_A(s_i) = p_i$ = the probability of selecting sample s_i on the first occasion when method A is used.

Similarly, on the second sampling occasion; A induces a component probability density ν_A characterized by $q_1, \ldots, q_M$ with

$$\sum_{i=1}^{M} q_i = 1 \text{ and } q_i > 0 \text{ for all } i,$$

where for each sample $s_i \in \Omega$,

$\nu_A(s_i) = q_i$ = the probability of selecting sample s_i on the second occasion when method A is used.

Note that the functions μ_A and ν_A are uniquely determined by Method A based on the corresponding measures of unit size on the two occasions.

Definition 1. A successive sampling procedure P of A is a joint probability density function defined on the sample space of all ordered pairs

$$\Omega^2 = \Omega \times \Omega = \{(s_i, s_j): s_i, s_j \in \Omega\}$$

such that $P(s_i, s_j)$ represents the joint probability density function of selecting s_i on the first occasion and s_j on the second occasion, and

$$\mu_A(s_i) = p_i = \sum_{j=1}^{M} P(s_i, s_j) \quad (1)$$

$$\nu_A(s_j) = q_j = \sum_{i=1}^{M} P(s_i, s_j) \quad (2)$$

Relationships (1) and (2) state that a successive sampling procedure P of A marginally agrees with the two component probability density functions μ_A and ν_A. One special case of a successive sampling procedure P of Method A is that P is equal to the product of μ_A and ν_A where

$$P(s_i, s_j) = \mu_A(s_i) \times \nu_A(s_j) \quad \text{for all } i, j.$$

This occurs when the sampler chooses to draw a "fresh sample" on the second occasion which is independent of the original sample.

*Research partially supported by the Energy Information Administration and the Office of Energy Research, U.S. Department of Energy, under contract DE-AC05-840R21400 with Martin Marietta Energy Systems, Inc.

To investigate the expected number of retained original sample units using a successive sampling procedure P of method A, we define a function X on Ω^2 such that

$X(s_i, s_j)$ = the number of population units in $s_i \cap s_j$.

For each pair of samples (s_i, s_j) observed on the two occasions, X measures the number of originally selected units in s_i that are retained in the second occasion sample s_j. The expected value of X under P is

$$E_p X = \sum_{(s_i, s_j)} X(s_i, s_j)\, P(s_i, s_j) \qquad (3)$$

which measures the expected number of retained first occasion sample units.
For n = 2, this reduces to

$$E_p X = \sum_{\{X=1\}} P(s_i, s_j) + \sum_{\{X=2\}} 2P(s_i, s_j) \qquad (4)$$

For n=1, it further reduces to

$$E_p X = \sum_{i=1}^{N} P(s_i, s_i) \qquad (5)$$

where $P(s_i, s_i)$ is the probability of selecting the ith sample on the first occasion and on the second occasion.

It is often desirable that the expected value of X be maximized (or minimized) when sampling on two occasions. Keyfitz (1951) and Des Raj (1956) considered the problem of maximizing (5) when n = 1 on both occasions. Fellegi (1966) considered the problem of maximizing (4) when n=2 and using Fellegi's (1963) sampling method. Causey, Cox, and Ernst (1983) considered the more general problem of maximizing (or minimizing) (3) for $n \leq N$. The need for maximizing (or minimizing) the expected number of overlaps helps to establish the following selection criterion among possible UP wor sample selection methods.

Let $\underline{P}_A$ be the family of all successive sampling procedures of Method A satisfying Definition 1. For a given A, the selection criterion is to select a P^* in $\underline{P}_A$ or a P° in $\underline{P}_A$ such that

a. $E_{p^*}X = \max \{E_p X : P \in \underline{P}_A\}$ when it is desirable to maximize the expected number of retained first occasion sample units, or

b. $E_{p^\circ}X = \min \{E_p X : P \in \underline{P}_A\}$ when there is a need to minimize the expected number of retained first occasion sample units.

Definition 2. A successive sampling procedure P of A is called an optimal successive sampling procedure of A if P satisfies either condition a or b above.

Lemma: For each UP wor method, A, there exist a P^* and a P° of $\underline{P}_A$ such that

a. $E_{p^*}X = \max \{E_p X : P \in \underline{P}_A\}$
= maximum expected number of overlaps when method A is being employed, and

b. $E_{p^\circ}X = \min \{E_p X : P \in \underline{P}_A\}$
= minimum expected number of overlaps when method A is being employed.

Proof of the Lemma is straightforward due to the fact that the set $\{E_p X : P \in \underline{P}_A\}$ is a compact subset of R^1 for any A. This is true because (i) $\underline{P}_A$ is non-empty, (ii) $\underline{P}_A$ can be considered as a compact subset of R^t where $t = M^2$, and (iii) the functions described in a and b above are real-valued, continuous functions on R^t.

The lemma states that each UP wor method, A, uniquely determines a maximum value of the expected number of overlaps, and a minimum value of the expected number of overlaps under A. Each of the two optimal values is attainable by some optimal successive sampling procedure of A. The two optimal values of an UP wor method reflect its abilities to maximize (or minimize) the expected number of retained first occasion sample units.

THE EMPIRICAL STUDY

The objective of this study is to compare UP wor methods in terms of their ability to maximize (or minimize) the expected number of overlaps when sampling two units on each of the two occasions.

We investigated five UP wor methods in combination with six artificial test populations. The six test populations are given in Table 1 where each population is described by the change in measures of unit size over the two occasions. Population 1 has four units and was taken from Keyfitz (1951). Populations 2 — 5 have six units each and were taken from Fellegi (1966). Population 6 is the case where relative measures of size were not changed in the two occasions.

Table 1. Measures of Unit Sizes for the Five Artificial Test Populations - Sampling on Two Occasions

Population	Sampling Occasion	Unit Relative Measure of Size U_1	U_2	U_3	U_4	U_5	U_6
1[a]	first	0.07281	0.32310	0.29267	0.31142		
	second	0.08202	0.33509	0.27980	0.30309		
2[b]	first	.10	.14	.17	.18	.19	.22
	second	.22	.19	.18	.17	.14	.10
3[b]	first	.10	.14	.17	.18	.19	.22
	second	.14	.10	.18	.17	.22	.19
4[b]	first	.10	.14	.17	.18	.19	.22
	second	.10	.14	.17	.19	.18	.22
5[b]	first	.10	.14	.17	.18	.19	.22
	second	.0820	.1148	.1393	.1475	.1557	.3607
6	first	.10	.14	.17	.18	.19	.22
	second	.10	.14	.17	.18	.19	.22

[a]Taken from Keyfitz (1951).
[b]Taken from Fellegi (1966).

For each given method A of UP wor and a given test population, FORTRAN programs were developed to compute the maximum and the minimum expected number of overlaps attainable under A. The following steps were taken to achieve this goal:

STEP I

For n = 2, the first step is to determine μ_A based on the original measures of size and to determine ν_A based on the measures of size for the second occasion. Without loss of generality, we describe the computations of μ_A using relative measures of size $Z_1, \ldots, Z_N$ on the first occasion and the method A. The computation algorithms for ν_A is the same as μ_A, except that the relative measures of size will usually change on the second occasion.

To determine μ_A, we need to determine the selection probabilities of each sample $s = \{i, j\}$ with $i \neq j$ and $i, j = 1, \ldots, N$, that is, the probability of s being selected under the selection rule of method A:

Method 1

If A is the Murthy's (1957) method, compute

$$\mu_A(s) = \mu_A(\{i, j\}) = \frac{Z_i Z_j (2 - Z_i - Z_j)}{(1 - Z_i)(1 - Z_j)}$$

for all samples $\{i, j\}$.

Method 2

If A is the Brewer's (1963) method, compute

$$\mu_A(s) = \mu_A(\{i, j\}) = \frac{2 Z_i Z_j (1 - Z_i - Z_j)}{D \cdot (1 - 2Z_i)(1 - 2Z_j)}$$

for each sample $s = \{i, j\}$ where

$$D = \sum_{i=1}^{N} \frac{Z_i (1 - Z_i)}{1 - 2Z_i} .$$

Method 3

If A is the Lahiri's (1951) method, compute

$$\mu_A(s) = \mu_A(\{i,j\}) = (Z_i + Z_j)/(N-1)$$

for each sample $s = \{i, j\}$.

Method 4

If A is the Fellegi's (1966) method, compute

$$\mu_A(s) = \mu_A(\{i,j\}) = 2Z_i \left(\frac{b_j}{1 - b_i} \right)$$

where

(i) $\sum_{\substack{j=1 \\ j \neq i}}^{N} \mu_A(\{i,j\}) = Z_i$

(ii) $\sum_{\substack{i=1 \\ i \neq j}}^{N} \mu_A(\{i,j\}) = Z_j$ and

(iii) $\sum_{i=1}^{N} b_i = 1$, and $b_i > 0$; $i = 1, \ldots, N$.

Brewer (1967) showed that $\{b_i\}_{i=1}^{N}$ is uniquely determined if $\max_i \{Z_i\} < 1/2$.

A successive approximation algorithm suggested in Fellegi (1966) was used to compute $b_1, \ldots, b_N$. A bound for testing convergence to each b_i was set at 0.000001 so that $| b_i^{(m)} - b_i^{(m-1)} | < 0.000001$ for all $i = 1, \ldots, N$ were $b_i^{(m)}$ is the approximated value of b_i after the mth iteration.

Method 5

If A is the Rao-Hartley-Cochran (1962) method, compute

$$\mu_A(s) = \mu_A(\{i,j\}) = \begin{cases} \sum \dfrac{1}{\binom{N}{N/2}} \cdot \dfrac{Z_i Z_j}{Z_{g_1} Z_{g_2}} & \text{if } N \text{ is even} \\ \sum \dfrac{1}{\binom{N}{\frac{N+1}{2}}} \cdot \dfrac{Z_i Z_j}{Z_{g_1} Z_{g_2}} & \text{if } N \text{ is odd} \end{cases}$$

where the summation is over all possible subgroups g_1 and g_2 such that

(i) $g_1 \cup g_2 = \{1, \ldots, N\}$, $g_1 \cap g_2 = \phi$, and

(ii) $i \varepsilon g_1$, $j \varepsilon g_2$ or $i \varepsilon g_2$, $j \varepsilon g_1$.

When N is even, each group has N/2 units. When N is odd, g_1 (group 1) has (N+1)/2 units and g_2 (group 2) has (N-1)/2 units. Also,

$$Z_{g_1} = \sum_{i \varepsilon g_1} Z_i \quad \text{and}$$

$$Z_{g_2} = \sum_{i \varepsilon g_2} Z_i .$$

Thus for each method A, μ_A determines $\{p_1, \ldots, p_M\}$, the selection probabilities of all possible samples of size n = 2 on the first occasion, where

$$\mu_A(s_i) = p_i; \; i = 1, \ldots, M.$$

Similarly, we can determine ν_A based on the given relative measures of size on the second sampling occasion. The selection probabilities of all possible samples of size n = 2 are $q_1, \ldots, q_M$, where

$$\nu_A(s_i) = q_i; \; i = 1, \ldots, M.$$

STEP II

After μ_A and ν_A are calculated, the next step is to select a $P^* \varepsilon \underline{P}_A$ and a $P^\circ \varepsilon \underline{P}_A$ which will attain the two optimal values:

$$E_{P^*} X = \max \{E_P X : P \varepsilon \underline{P}_A\} \quad \text{and} \qquad (2)$$

$$E_{P^\circ} X = \min \{E_P X : P \varepsilon \underline{P}_A\} \qquad (3)$$

subject to the constraints on $P \in \underline{P}_A$ which are given in Definition 1.

Thus, the optimization problem is presented as a simple linear programming problem with the objective function (2) or (3), and the constraints are given in Definition 1. A FORTRAN coded transportation subroutine obtained from B. Holcomb (1983) was employed to compute P^* and P°. The subroutine was originally designed to solve transportation problems based on the algorithms given by Hillier and Lieberman (1980), and it has been converted to fit the needs of our empirical study. Finally, values of $E_{P^*}X$ and $E_{P^\circ}X$ are obtained for each method and each given population.

A simple linear programming approach to this type of optimization problem was discussed in Des Raj (1956), and Wright and Tsao (1984) for $n = 1$, and in Causey, Cox, and Ernst (1983) for $n > 1$. The previous work did not consider the changes in the ability of optimizing expected number of overlaps when different component UP wor sample selection methods are used.

For each test population and for each sample selection method, the algorithms for computing μ_A and ν_A and the developed FORTRAN programs allowed us to compute the desired optimal values. Table 2 lists the computation results for $\max \{E_P X : P \in \underline{P}_A\}$ and $\min \{E_P X : P \in \underline{P}_A\}$ under different method-population combinations.

Table 2. Comparisons of the Five UP wor Sample Selection Methods Based on the Maximum (Minimum) Expected Number of Overlaps Attained

UP wor method used on both occasions	Populations tested[a]					
	1	2	3	4	5	6
Murthy's Method	1.9654 (0.6388)	1.6786 (0.0)	1.8568 (0.0)	1.9829 (0.0)	1.7983 (0.0521)	2.0 (0.0)
Brewer's Method	1.9550 (0.6901)	1.6400 (0.0)	1.8398 (0.0)	1.9801 (0.0)	1.7188 (0.1612)	2.0 (0.0)
Lahiri's Method	1.9850 (0.2301)	1.8560 (0.0)	1.9360 (0.0)	1.9920 (0.0)	1.8875 (0.0)	2.0 (0.0)
Fellegi's Method	1.9551 (0.6901)	1.6378 (0.0)	1.8400 (0.0)	1.9799 (0.0)	1.7188 (0.1614)	2.0 (0.0)
Rao-Hartley-Cochran's Method	1.9699 (0.5967)	1.7062 (0.0)	1.8678 (0.0)	1.9848 (0.0)	1.8290 (0.0028)	2.0 (0.0)

[a]Populations tested are described in Table 1.

Our empirical study showed consistent results over the six different populations for all five sampling methods that we studied. Observe that the Lahiri's method performed best, Rao-Hartley-Cochran method ranked second, and Murthy's method ranked third. Brewer's and Fellegi's methods have negligible differences among them, and they were ranked last as a group. A method that is best for maximizing the overlaps is also the best for minimizing overlaps regardless of our choice of population in the experiments we conducted.

The study is preliminary and is at an exploratory stage. Much more computation results can be generated for other possible situations to expand the current level of work, and to gain more background knowledge on how the ability of a component sample selection method can optimize the number of expected overlaps when successive sampling is required.

Brewer and Hanif (1983) listed 50 sample selection methods. One classification criterion they used is "classification by equivalence class," where two methods are considered to be in the same equivalence class whenever they have the same selection probabilities on all possible samples. For a fixed sample of size n, methods that fall into the same equivalence class will have the same maximum (or minimum) number of expected overlaps. Methods in different equivalence classes can be compared with different types of test populations and evaluated for their ability to optimize the expected number of overlaps when sampling on two occasions.

REFERENCES

Brewer, K. R. W. (1963), "A Method of Systematic Sampling with Unequal Probabilities," Australian Journal of Statistics, 5: 5-13.

Brewer, K. R. W. (1967), "A Note on Fellegi's Method of Sampling Without Replacement with Probability Proportional to Size," Journal of the American Statistical Association, 62: 79-85.

Brewer, K. R. W. and Hanif, M. (1983), Lecture Notes in Statistics: Sampling with Unequal Probabilities, Springer-Verlag, New York.

Causey, B. D., Cox, L. H., and Ernst, L. R. (1983), "Applications of Transportation Theory to Statistical Problems," Proceedings of the American Statistical Association, Section on Survey Research Methods, pp. 112-117.

Fellegi, I. (1963), "Sampling with Varying Probabilities Without Replacement: Rotating and Non-Rotating Samples," Journal of the American Statistical Association, 58: 183-201.

Fellegi, I. (1966), "Changing the Probabilities of Selection When Two Units are Selected with PPS Without Replacement," American Statistical Association, Proceedings of Social Statistics Section, pp. 434-442.

Hillier, F. S. and Lieberman, G. J. (1980), Introduction to Operations Research, Holden Day, Inc., San Francisco.

Holcomb, B. D. (1983), "Transportation Algorithm," Computing Technology Center, Oak Ridge National Laboratory, Oak Ridge, Tennessee.

Keyfitz, N. (1951), "Sampling with Probabilities Proportional to Size: Adjustment for Changes in the Probabilities," Journal of the American Statistical Association, 46: 104-109

Lahiri, D. (1951), "A Method of Sample Selection Providing Unbiased Ratio Estimates," International Statistical Institute Bulletin, 33 (Part 2): 133-140.

Murthy, M. N. (1957), "Ordered and Unordered Estimators in Sampling Without Replacement," Sankhyá, 18: 379-390.

Raj, D. (1956), "On the Method of the Overlapping Maps in Sample Surveys," Sankhyá, 17 (No. 1): 89-98.

Rao, J. N. K., Hartley, H. O., and Cochran, W. G. (1962), "On a Simple Procedure of Unequal Probability Sampling Without Replacement," Journal of Royal Statistical Society, Series B24: 482-491.

Wright, T. and Tsao, H. (1984), "On an Optimal Solution for Maximizing the Probability of Retention in PPS Sampling," Journal of Linear Algebra and Its Applications (to appear).