Statistica Sinica **10**(2000), 1171-1178

ON A SIMPLIFICATION OF THE LINEAR PROGRAMMING APPROACH TO CONTROLLED SAMPLING

P. Lahiri and Rahul Mukerjee

University of Nebraska-Lincoln and Indian Institute of Management

Abstract: With reference to the linear programming approach to controlled sampling, this paper shows how from consideration of symmetry one can achieve a drastic reduction in the dimensionality of the problem and hence reduce the computing time to a great extent.

Key words and phrases: Equivalence class, inclusion probabilities, symmetry, undesirable sample.

1. Introduction

With reference to a finite population of size N, consider the problem of finding a sampling design, with fixed size n $(< N)$, so as to minimize the probability of selecting certain *undesirable* samples while at the same time matching the inclusion probabilities of the first two orders with those under simple random sampling without replacement; see Rao and Nigam (1990) for a discussion on the problem.

To avoid trivialities, suppose $n \geq 3$. Rao and Nigam (1990) proposed an elegant linear programming formulation for the above problem. While this strengthens and unifies some previous results based on block designs, a difficulty with the linear programming approach, in its present form, is that both the number of variables and the number of constraints increase very rapidly with increase in N and n and, despite the computing facilities available nowadays, can be prohibitively large even for moderately large N and n.

In an attempt to make the aforesaid linear programming formulation more user-friendly, the present article shows how from consideration of symmetry it is possible to achieve a drastic reduction in the dimensionality of the problem and hence a substantial saving of computing time in many situations of practical interest. In particular, as illustrated in last section, this can be of considerable help in surveys spread over a geographical area which possibly represent the most common field of application of controlled sampling.

The difficulty with the linear programming formulation for large N and n was noted also by Rao and Nigam (1992) who remarked that the formulation might become more manageable with a stratified population provided the set of undesirable samples is specified separately in each stratum. This implicitly calls for matching the inclusion probabilities separately for each stratum and hence amounts to treating each stratum as a population. Consequently, even when such a stratification is possible, the present approach can entail a significant further simplification when applied separately to each stratum. In this connection, see Examples 1 and 4 which show that even with relatively small populations, that could as well be interpreted as strata, the reduction in the dimensionality through the present procedure can be substantial.

2. Main Result

Let $p(s)$ be the selection probability for a sample s which is an n-subset of $U = \{1, \ldots, N\}$ and S be the set of all the $\binom{N}{n}$ possible samples. If S_1 $(\subset S)$ denotes the set of undesirable samples then following Rao and Nigam (1990), the linear programming formulation of the present controlled sampling problem is as follows:

A. Find the sampling design $\{p(s) : s \in S\}$ so as to minimize

$$\phi = \sum_{s \in S_1} p(s) \tag{1}$$

subject to

$$\sum_{s \ni i,j} p(s) = n(n-1)/\{N(N-1)\}, \quad 1 \le i < j \le N \tag{2}$$

and

$$p(s) \ge 0, \text{ for all } s \in S. \tag{3}$$

The above formulation involves $\binom{N}{n}$ decision variables and $\binom{N}{2}$ equality constraints. We now show how, from consideration of symmetry, the dimensionality of this formulation can be reduced. Two units i and j of the population are said to be associates of each other, written $i \sim j$, if the set S_1 of undesirable samples remains unaltered when the roles of i and j are interchanged. Obviously, $i \sim i$ and the relation $\sim$ partitions U into equivalence classes such that the members within each equivalence class are associates of one another. Let there be t equivalence classes $U_1, \ldots, U_t$ with respective cardinalities $N_1, \ldots, N_t$, where $N_i \ge 1$ for all i and $\sum_{i=1}^{t} N_i = N$. Let V be the set of ordered t-plets $v = v_1 \cdots v_t$ such that $v_1, \ldots, v_t$ are integers satisfying

$$0 \le v_i \le N_i \ (1 \le i \le t) \text{ and } \sum_{i=1}^{t} v_i = n. \tag{4}$$

For any $v\ (= v_1 \cdots v_t) \in V$, let $S(v)$ be a subset of S consisting of those samples which contain exactly v_i units from $U_i,\ 1 \leq i \leq t$. The cardinality of $S(v)$ is then

$$h_v = \prod_{i=1}^{t} \binom{N_i}{v_i} \tag{5}$$

Clearly, the class of sets $\{S(v) : v \in V\}$ represent a disjoint partition of S so that $\sum_{v \in V} h_v = \binom{N}{n}$.

Example 1. Let $N = 10, n = 3$ and $S_1 = \{124, 125, 134, 135\}$ where, for notational simplicity, we write $i_1 \cdots i_n$ to denote a sample $\{i_1, \ldots, i_n\}$. Clearly, S_1 does not change if, say, the roles of the units 2 and 3 are interchanged. Hence $2 \sim 3$. From similar considerations, it is seen that here $t = 4,\ U_1 = \{1\},\ U_2 = \{2, 3\},\ U_3 = \{4, 5\},\ U_4 = \{6, 7, 8, 9, 10\},\ N_1 = 1,\ N_2 = N_3 = 2,\ N_4 = 5$. Hence by (4), $V = \{0003, 1002, 0102, 0012, 1020, 0120, 0021, 1200, 0210, 0201, 1110, 1101, 1011, 0111\}$. In particular, $S(0210) = \{234, 235\},\ S(1110) = \{124, 125, 134, 135\}$, $h_{0210} = 2$, $h_{1110} = 4$, and so on. Note that $S_1 = S(1110)$.

Lemma 1. *For each $v \in V$, either $S(v) \subset S_1$ or $S(v)$ and S_1 are disjoint.*

Proof. If possible suppose the result is not true. Then there exist samples s_1 and s_2 such that

$$s_1 \in S_1, \quad s_2 \notin S_1 \tag{6}$$

$$\text{and } s_1 \in S(v), s_2 \in S(v), \tag{7}$$

for some $v = v_1 \cdots v_t \in V$. Then by (7), $s_i = s_{i1} \cup \cdots \cup s_{it}\ (i = 1, 2)$ where, for $1 \leq j \leq t$, both s_{1j} and s_{2j} are v_j-subsets of U_j. But then by the definition of $U_1, \ldots, U_t$, one can obtain s_2 from s_1 in a finite number of steps where in each step the roles of two units that are associates of each other are interchanged. Since, by the definition of associates, any such interchange leaves S_1 unaffected, it follows that (6) is impossible.

In view of Lemma 1, we have

$$S_1 = \cup_{v \in V_0} S(v), \tag{8}$$

where V_0 is a nonempty subset of V. Consider now the following linear programming problem where $\{x_v : v \in V\}$ are real numbers and $Q = \{r : 1 \leq r \leq t, N_r \geq 2\}$.

B. Find $\{x_v : v \in V\}$ so as to minimize

$$\psi = \sum_{v \in V_0} x_v \tag{9}$$

subject to

$$(N_rN_{r'})^{-1}\sum_{v\in V} v_rv_{r'}x_v = n(n-1)/\{N(N-1)\},\quad 1\le r<r'\le t \tag{10}$$

$$\{N_r(N_r-1)\}^{-1}\sum_{v\in V} v_r(v_r-1)x_v = n(n-1)/\{N(N-1)\},\quad r\in Q \tag{11}$$

$$x_v\ge 0,\ \text{for all } v\in V. \tag{12}$$

Lemma 2. *The linear programming problem* B *has a feasible solution. Moreover, if $\phi^\star$ and $\psi^\star$ are the minimum possible values of ϕ and ψ under the problems* A *and* B *respectively, then $\phi^\star \ge \psi^\star$.*

Proof. First note that one feasible solution of the problem A is $p(s)=\binom{N}{n}^{-1}$ for all $s\in S$. Consider now any feasible solution, say, $\{\hat{p}(s), s\in S\}$ of the problem A and define

$$\hat{x}_v = \sum_{s\in S(v)} \hat{p}(s),\quad v\in V. \tag{13}$$

We shall show that $\{\hat{x}_v : v\in V\}$ is a feasible solution of the problem B. Since the quantities $\hat{p}(s)$ satisfy (2), for any r, r' $(1\le r<r'\le t)$, summing these identities over $i\in U_r$ and $j\in U_{r'}$,

$$\sum_{i\in U_r}\sum_{j\in U_{r'}}\sum_{s\ni i,j}\hat{p}(s) = N_rN_{r'}n(n-1)/\{N(N-1)\}. \tag{14}$$

But since the sets $\{S(v): v\in V\}$ represent a disjoint partition of S, the left hand side of (14) equals

$$\begin{aligned}\sum_{v\in V}\sum_{i\in U_r}\sum_{j\in U_{r'}}\sum_{\substack{s\ni i,j\\ s\in S(v)}}\hat{p}(s) &= \sum_{v\in V}\sum_{s\in S(v)}\sum_{\substack{i\in U_r\\ i,j\in s}}\sum_{j\in U_{r'}}\hat{p}(s)\\ &=\sum_{v\in V}\sum_{s\in S(v)} v_rv_{r'}\hat{p}(s) = \sum_{v\in V} v_rv_{r'}\hat{x}_v,\end{aligned} \tag{15}$$

using (13). By (14) and (15), the quantities $\hat{x}_v, v\in V$, satisfy (10). Similarly, for any $r\in Q$, summing the identities (2) over $i,j\in U_r, i<j$, it can be seen that these quantities satisfy (11) as well. Also, by (3) and (13), trivially the $\hat{x}_v, v\in V$, satisfy (12). Thus, $\{\hat{x}_v : v\in V\}$ is a feasible solution of the problem B.

Next observe that by (1) and (8), the value of ϕ under $\{\hat{p}(s): s\in S\}$ is given by $\sum_{s\in S_1}\hat{p}(s) = \sum_{v\in V_0}\sum_{s\in S(v)}\hat{p}(s)$, which, by (9) and (13), equals the value of ψ under $\{\hat{x}_v : v\in V\}$. Thus given any feasible solution of the problem A we can find a feasible solution of the problem B such that the value of ϕ under the former equals the value of ψ under the latter. Hence evidently $\phi^\star\ge\psi^\star$.

Theorem 1. *Let $\{x_v^\star : v \in V\}$ be an optimal solution of the linear programming problem* B. *Define*

$$p^\star(s) = x_v^\star/h_v, \text{ for every } s \in S(v) \text{ and every } v \in V. \tag{16}$$

Then$\{p^\star(s) : s \in S\}$ is an optimal solution of the linear programming problem A. *Furthermore,*

$$\sum_{v \in V} x_v^\star = 1. \tag{17}$$

Proof. We first show that $\{p^\star(s) : s \in S\}$ is a feasible solution to the problem A. Consider any i, j, where $1 \le i < j \le N$. If $i, j \in U_r$ for some $r \in Q$ then by (5), (11), and (16),

$$\begin{aligned}\sum_{s \ni i,j} p^\star(s) &= \sum_{v \in V} \sum_{\substack{s \in S(v) \\ s \ni i,j}} x_v^\star/h_v \\ &= \sum_{v \in V} (x_v^\star/h_v)\{\prod_{\substack{r'=1 \\ r' \ne r}}^{t} \binom{N_{r'}}{v_{r'}}\}\binom{N_r - 2}{v_r - 2} \\ &= \sum_{v \in V} x_v^\star v_r(v_r - 1)/\{N_r(N_r - 1)\} = n(n-1)/\{N(N-1)\}\end{aligned}$$

Similarly, by (5), (10) and (16), the above holds also when $i \in U_r, j \in U_{r'}$ for some $r \ne r'$. Hence the quantities $p^\star(s), s \in S$, satisfy (2). By (12) and (16), these quantities satisfy (3) as well. Thus $\{p^\star(s) : s \in S\}$ is a feasible solution of the problem A.

Now for this solution, by (8), (9), (16) and Lemma 2,

$$\sum_{s \in S_1} p^\star(s) = \sum_{v \in V_0} \sum_{s \in S(v)} x_v^\star/h_v = \sum_{v \in V_0} x_v^\star = \psi^\star \le \phi^\star, \tag{18}$$

where $\phi^\star$ and $\psi^\star$ are as defined in Lemma 2. On the other hand, as $\{p^\star(s) : s \in S\}$ is a feasible solution of the problem A, $\sum_{s \in S_1} p^\star(s) \ge \phi^\star$. Hence the equality holds in (18) and optimality of this solution follows.

Finally, we note that $\sum_{s \in S} p^\star(s) = 1$, since $\{p^\star(s) : s \in S\}$ is a feasible solution of the problem A. But by (16), $\sum_{s \in S} p^\star(s) = \sum_{v \in V} \sum_{s \in S(v)} x_v^\star/h_v = \sum_{v \in V} x_v^\star$, whence (17) follows.

Thus $\{p^\star(s) : s \in S\}$, as defined in Theorem 1, gives an optimal sampling design in the sense described in the Introduction. In view of (4), (5), (16) and (17), this optimal design can be implemented as follows.

Step 1. Find an optimal solution, say $\{x_v^\star : v \in V\}$, of the linear programming problem B.

Step 2. Select a member of V such that any $v \in V$ has a chance $x_v^\star$ of being selected.
Step 3. Let $v^\star = v_1^\star \cdots v_t^\star$ $(\in V)$ be selected in step 2. For $1 \le i \le t$, draw a simple random sample of size $v_i^\star$ from U_i without replacement. Combine these t samples to get a sample of size n.

While the implementation of steps 2 and 3 is straightforward, the linear programming problem considered in step 1 can have a substantially lower dimensionality than the original formulation A in most applications. Note that the problem B involves f decision variables and $\binom{t}{2} + q$ equality constraints where f and q are the cardinalities of V and Q respectively. Thus in the setup of Example 1, the problem B involves 14 decision variables and 9 equality constraints while for the problem A, these numbers are as large as 120 and 45 respectively. More examples are given in the next section to demonstrate how the present approach can entail substantial saving in computing time.

While a compact expression for f is hard to find for general n, before concluding this section we indicate a simple upper bound on f. Note that $V \subset \tilde{V}$, where $\tilde{V}$ is the set of ordered t-plets $v_1 \cdots v_t$ such that $v_1, \ldots, v_t$ are nonnegative integers satisfying $\sum_{i=1}^t v_i = n$. Hence

$$f \le \binom{t+n-1}{n}, \tag{19}$$

the upper bound being attainable when $n \le N_i,\ 1 \le i \le t$; compare Parzen (1960, p.70).

3. Examples

In surveys spread over a geographical area, often there are natural clusters of units based on geographical contiguity and, from consideration of cost or convenience, these clusters determine the desirability or otherwise of any sample. In such a situation, units within each cluster are associates and the clusters correspond to equivalence classes $U_1, \ldots, U_t$. To take a concrete example, consider a population of 10 villages in a partially hilly area where village 1 is located on a hill, villages 2 and 3 are located on another hill, villages 4 and 5 are located on yet another hill while villages $6, \ldots, 10$ are in a plateau surrounded by these hills. The sample size is 3 and, from practical consideration, suppose it is inconvenient to have a sample which contains units from two hills. Then the set of undesirable samples is $\{124, 125, 134, 135\}$ which is dictated by the natural clusters $\{1\}, \{23\}, \{4, 5\}$, and $\{6, \ldots, 10\}$ arising from geographical contiguity. Observe that Example 1 deals precisely with this setup. Some more examples follow. Incidentally the clusters considered here or, for that matter, the equivalence classes of Section 2 are different from strata in the usual sense; for example, the number of units from any particular cluster may vary from one possible sample to another and is randomly determined via Steps 1 and 2 of Section 2.

Example 2. Consider a population of $N = 18$ villages clustered on a stretch of road as follows:

$$1\ 2\ 3\ *\ *\ *\ 4\ 5\ 6\ 7\ 8\ 9\ *\ *\ *\ 10\ 11\ 12\ 13\ *\ *\ *\ 14\ 15\ 16\ 17\ 18$$

Let $n = 4$ and suppose samples other than those which contain units only from two neighboring clusters are considered undesirable. Here $t = 4,\ N_1 = 3,\ N_2 = 6,\ N_3 = 4,\ N_4 = 5$ and by (4), the cardinality of V equals $f = 34$. Thus the linear programming problem B involves 34 decision variables and 10 equality constraints, significantly lower than the corresponding numbers 3060 and 153 arising for the problem A.

Example 3. The setup is as in the last example with the change that now $N = 29$ $t = 5,\ N_1 = N_2 = N_3 = N_4 = 6,\ N_5 = 5$ and $n = 5$. Then the equality holds in (19) so that the problem B involves 126 decision variables and 15 equality constraints. On the other hand, these numbers are as high as 118755 and 406 for problem A.

Even when the undesirable samples are not dictated by a natural clustering of the population, our results can be useful. This happens, for example, when the population contains a fair number of units that are normal in the sense of not being included in any undesirable sample. Then these normal units obviously constitute an equivalence class and, as illustrated below, facilitate a reduction of the linear programming problem A.

Example 4. Let $N = 12,\ n = 3$ and $S_1 = \{123,\ 234,\ 135\}$. Then the units $6, \ldots, 12$ are normal in the above sense and we get $t = 6,\ U_i = \{i\}\ (1 \le i \le 5),\ U_6 = \{6, \ldots, 12\}$. Hence it can be seen that the problem B involves 26 decision variables and 16 equality constraints whereas the corresponding numbers equal 220 and 66, respectively, for the problem A.

We now comment on the amount of effort needed in the identification of the equivalence classes. As noted above, the identification is straightforward when the equivalence classes correspond to natural clusters of units or when an appreciable number of units do not belong to any undesirable sample. Even otherwise, the following considerations help. Among the undesirable samples, let there be a_i which contain unit i, and λ_{ij} which contain units i and j. Then, for units i and j to belong to the same equivalence class, it is necessary that

$$a_i = a_j \tag{20}$$

and

$$\lambda_{ir} = \lambda_{jr}, \text{for every } r \neq i, j. \tag{21}$$

One can easily partition the population into subsets such that units i and j are in the same subset if and only if $a_i = a_j$. By (20), each equivalence class is contained in one of these subsets. This greatly facilitates the task of finding the equivalence classes since, instead of considering the entire population at a time, one needs to consider only the subsets. In the same spirit, further reduction is possible by checking (21) only for pairs of units (i, j) belonging to the same subset. Based on these considerations, our experience suggests that quite commonly the equivalence classes can be identified even by hand computation or just by inspection for moderately large populations, say for N not exceeding 100, a range that covers many practical situations under appropriate stratification.

Before concluding, we remark that the present simplification can be suitably adapted to situations where the right hand side of (2) is replaced by more general expressions. Equivalence classes are then based on interchangeability of units with reference to such modification of (2) as well as the set of undesirable samples. The resulting algebra will be similar to what has been presented here. However, because of possible diversity in the right hand side of (2), the symmetry may hold to a lesser extent and simplification, to the scale observed here, may not be possible. More work is needed in this direction.

Acknowledgement

Thanks are due to Professor Arijit Chaudhuri for suggesting the problem. This work was supported by the NSF grants SES-9511202, INT-9505197, SBR-9705574 and a grant from the Centre for Management and Development Studies, Indian Institute of Management, Calcutta. We thank a referee, an Associate Editor and the Chair Editor for very constructive suggestions.

References

Parzen, E. (1960). *Modern Probability Theory and Its Application.* John Wiley, New York.

Rao, J. N. K. and Nigam, A. K. (1990). Optimal controlled sampling designs. *Biometrika* **77**, 807-814.

Rao, J. N. K. and Nigam, A. K. (1992). "Optimal" controlled sampling: a unified approach. *Internat. Statist. Rev.* **60**, 89-98.

Department of Mathematics and Statistics, 922 Oldfather Hall, University of Nebraska-Lincoln, Lincoln, NE 68588-0323, U.S.A.

E-mail: plahiri@math.unl.edu

Indian Institute of Management, Post Box No. 16757, Calcutta 700 027, India.

E-mail: rahulmuk@hotmail.com

(Received April 1999; accepted December 1999)