Sankhyā : The Indian Journal of Statistics
1988, Volume 50, Series B, Pt. 1, pp. 0-38.

ON THE METHOD OF OVERLAPPING MAPS IN SURVEY SAMPLING*

By SUJIT KUMAR MITRA
Indian Statistical Institute

[Dedicated to Professor Debabrata Lahiri on his seventyfifth birthday]

SUMMARY. Consider a finite population containing N units on which it is proposed to carry out k separate surveys, each with it's own probabilities of selection specified for the N units. This paper reviews the existing literature on the subject of cost optimal integration of the k surveys for various types of cost functions. Some of the approaches are heuristic while others formulate the problem as a linear programming problem and recommend standard LP techniques. This paper also considers sampling with probabilities proportional to the total size of the sample units and integration of k surveys of this type, each survey possibly using a different indicator of size.

1. THE PROBLEM : HISTORY AND THE KEYFITZ AND OTHER SOLUTIONS

1.1. *The problem* : The problem appears to have been originally formulated by Nathan Keyfitz in 1951 and can be stated thus. We have a finite population of N units on which it is proposed to carry out k separate surveys. Each survey assigns distinct sets of probabilities of selection to the N units. For example, the units may be villages and for the first survey a demographic one, the sampling design may dictate selection of villages with probabilities proportional to the population of the village. The second one could be a land utilization survey for which it seems desirable to select the villages with probabilities proportional to the area of the village and so on.

---

*Text of the Presidential address at the 75-th session of the Indian Science Congress Association, (Statistics Section) held at Pune in January 1988.

Editorial Note : Reprinted with the permission of the Deputy Executive Secretary, Indian Science Congress Association.

AMS (1980) *subject classification* : 62D05.

Key words and phrases : Algorithm, Configuration, Cost optimality, Integrated survey, Maximum matching, Transportation problem, Linear programming approach, Midzuno-Sen survey.

When the same agency is carrying out all the surveys simultaneously, there may be some distinct advantage if the sample units for the different surveys overlap as this would reduce the amount of time the investigator would have to spend moving from one unit to another, with a consequent reduction in the cost of the survey.

Even if all the surveys were not held simultaneously and there are short gaps between them, it may be advantageous if the investigator visits a large number of common villages for the separate surveys, as the knowledge and experience gained from his first visit may lead to a better execution of the subsequent surveys if they are carried out in the same village.

The problem can be mathematically formulated as follows in the context of k-surveys. Let X_i denote the serial no. of the unit selected for the i-th survey. The sample design for the i-th survey stipulates that Prob $\{X_i = j\} = P_{ij} (j = 1, 2, ..., N ; i = 1, 2, ..., k)$; we have thus k random variables $X_1, X_2, ..., X_k$. Let $\mathcal{S}$ denote the set of integers $\{1, 2, ..., N\}$. By integration of surveys, we mean defining a joint distribution of the k random variables, $X_1, X_2, ..., X_k$ on $\mathcal{S}^k$ the k-th Cartesian power of $\mathcal{S}$ which realises for each X_i, the same marginal distribution as dictated by the sampling design of the i-th survey. In the integrated survey, using the joint probability distribution so defined, at one stroke, one selects a k-tuple $(i_1, i_2, ..., i_k)$ which will require him to consider the i_1-th unit for survey 1, i_2-th unit for survey 2 and so on. Let $\nu(i)$ denote the number of distinct units appearing as coordinates of $i = (i_1, i_2, ..., i_k)$. Put

$$\mathcal{S}_u = \{i : \nu(i) = u\}, \quad u = 1, 2, ..., k. \qquad ... (1)$$

An integrated survey is called *optimal* if the joint distribution of X is so determined that $E[\nu(X)]$ is a minimum.

Such a survey is cost optimal if the cost function of the survey is of the form

$$a + b\nu, \quad b > 0. \qquad ... (2)$$

The problem is to find an integrated survey scheme which is cost-optimal. Maczynski and Pathak (1980) have shown that this problem always has a solution though not necessarily unique. The problem of optimal integration has a close resemblance to that of controlled selection beyond stratification. A typical example here is the problem studied by Goodman and Kish (1950). These authors considered two strata, one containing three coastal and three inland units while the other contains one coastal and four inland

units. The probabilities of selections of various units within a stratum are specified for both the strata. The object is to select one unit from each stratum by assigning probabilities to the various possible pairs of units so as to maximise the probability of selecting one coastal and one inland unit.

Keyfitz (1951) gave a solution to the problem of optimal integration of two surveys.

1.2. *Keyfitz solution for $k = 2$:* Assume that j-th unit has been selected for survey 1, following a procedure, which ensures probabilities of selection $P_{11}, P_{12}, \ldots, P_{1N}$ to the N units 1, 2, ..., N.

(1) If $P_{2j} > P_{1j}$, select j for survey 2 as well.

(2) Otherwise if $P_{2j} < P_{1j}$, then select the j-th unit for survey 2 with a probability P_{2j}/P_{1j} and reject the j-th unit with a probability $1-P_{2j}/P_{1j}$.

(3) If the j-th unit is so rejected then for survey 2, select a unit from among those units w in the population for which $P_{2w} > P_{1w}$, assigning a probability of selection proportional to $P_{2w}-P_{1w}$.

It is easily seen that this procedure will ensure that the probability that unit j is selected for both survey 1 and survey 2 = min (P_{1j}, P_{2j}).

This establishes the optimality of the Keyfitz solution.

1.3. *The problem of overlapping maps as a special case of the transportation problem :* Des Raj (1957) pointed out that the problem of optimal integration of surveys is a special case of the transportation problem and hence the simplex method or other methods for solving the LP problem could be used to obtain an optimal solution, even for cost functions which are more complicated than the one we have just considered. This approach is described in greater detail in the book by Arthanari and Dodge (1981).

In the context of two surveys the problem can be stated as an LP problem as follows :

Let $p_{jj'}$ denote the probability of selecting unit j for survey 1 and j' for survey 2.

$$\text{Maximise} \sum_{j} p_{jj} \qquad \ldots (3)$$

$$\text{s.t.} \quad \sum_{j'} p_{jj'} = P_{1j}, \ \sum_{j} p_{jj'} = P_{2j'}, \ p_{jj'} \geq 0. \qquad \ldots (4)$$

We shall come back to the LP approach in a later section.

1.4. *Another algorithm for optimal integration of surveys* : Mitra and Pathak (1984) present an alternative algorithm for this problem. This can be described as follows.

We first arrange the k marginal distributions in a two-way table with k rows and N columns, the i-th row describing the probabilities of selection for the i-th survey. Such an arrangement will be called the initial configuration. A configuration in general will denote a two-way arrangement of nonnegative entries with each row adding up to the same number [not necessarily = 1, which is true only at the initial configuration]. Each step of the algorithm transforms one configuration into another with the common row sum of the configuration progressively shrinking to zero. When the final configuration, which has all entries equal to 0, is reached, this is an indication that all the marginal distributions have now been fully accounted for.

We first introduce some notations :

Let $P_{1j}, P_{2j}, \ldots, P_{kj}$, (the entries in the j-th column of the initial configuration) be arranged in increasing order of magnitude and the ordered values be denoted by

$$P_{(1)j} \leqslant P_{(2)j} \leqslant \ldots \leqslant P_{(k)j}. \qquad \ldots (5)$$

Thus $P_{(1)j}$ is the smallest column entry and $P_{(k)j}$ is the largest column entry.

Put

$$0_i = \sum_j P_{(i)j} \qquad \ldots (6)$$

Clearly,

$$0_1+0_2+\ldots+0_k = \sum_{i,j} P_{ij} = k. \qquad \ldots (7)$$

The algorithm can be best described through a numerical example :

Consider the problem of two surveys required to be carried out on four villages (data as in Arthanari and Dodge : Table 5.9.1).

This corresponds to the initial configuration as given below :

TABLE 1 : PROBABILITIES OF SELECTION OF FOUR VILLAGES UNDER TWO SURVEYS

	village			
survey	1	2	3	4
1 (P_{1j})	0.5	0.2	0.1	0.2
2 (P_{2j})	0.3	0.1	0.4	0.2

For $k = 2$, in the j-th step of stage 1 of the algorithm, the smallest entry in the j-th column of the configuration is zeroed out through an assignment

of probability $P_{(1)j}$ to the pair (j, j). The successive configurations would look like the following.

				Total
0.5	0.2	0.1	0.2	1
0.3	0.1	0.4	0.2	1
		$p_{11} = 0.3$		
0.2	0.2	0.1	0.2	0.7
0	0.1	0.4	0.2	0.7
		$p_{22} = 0.1$		
0.2	0.1	0.1	0.2	0.6
0	0	0.4	0.2	0.6
		$p_{33} = 0.1$		
0.2	0.1	0	0.2	0.5
0	0	0.3	0.2	0.5
		$p_{44} = 0.2$		
0.2	0.1	0	0	0.3
0	0	0.3	0	0.3

In stage 2 of the algorithm, we again proceed from column 1 and notice that its non-zero entry appears in the first row. We then scan the successive columns of the configuration and look for a column where, for the first time, the non-zero entry appears in the second row. We thus have a pair of entries one from each row, appearing in columns 1 and j.

The smallest of the two entries is then zeroed out by assigning a probability, equal to the smallest entry, to the pair $(1, j)$ or $(j, 1)$ according as the non-null entry in column 1 appears in row 1 or row 2. If the entry in column

1 is not zeroed out in this process we scan for an appropriate column beyond column j and repeat this step as many times as may be necessary. In the given example we have $p_{13} = 0.2$.

				total
0	0.1	0	0	0.1
0	0	0.1	0	0.1

In this example, we assign further $p_{23} = 0.1$ and then both the marginal distributions will be fully explained. It may be noted that in this write-up we have made some small changes in the steps of stage 2 described in Mitra and Pathak (1984). This is in the hope of reducing the total number of computational steps involved in the execution of the algorithm. In a later section we shall have an opportunity to compare the computational complexities of this algorithm vis-a-vis other algorithms.

1.5. *Solution for* $k = 3$: Stage 1 here is parallel to that of the case $k = 2$, in that the probability of $P_{(1)j}$ is assigned to the point (j, j, j) in $\mathcal{S}_1$ and the entries in the j-th column of the configuration reduced accordingly. At the end of stage 1, each column of the configuration has one zero entry and in the j-th column the other entries (possibly nonzero) are $P_{(2)j} - P_{(1)j}$ and $P_{(3)j} - P_{(1)j}$.

In stage 2, the initial attempt is to zero out the second minimum entry in each column assigning maximum possible probability to points in $\mathcal{S}_2$. If for example the j-th column has its minimum in the first row, this is done by assigning a probability $P_{(2)j} - P_{(1)j}$ to the points of the type (x, j, j) where $x \neq j$. The strategy is to choose x in such a manner that the second minimum in no column is reduced in the process. It is thus clear that for a particular column x to qualify for inclusion in (x, j, j) it is necessary that $P_{1x} = P_{(3)x}$. Further the maximum probability that could be assigned to (x, j, j) is the minimum of $P_{(2)j} - P_{(1)j}$ and $P_{(3)x} - P_{(2)x}$. If $P_{(3)x} - P_{(2)x} < P_{(2)j} - P_{(1)j}$, then we look out for other columns with a similar property to meet the deficiency.

If the strategy succeeds then in stages 1 and 2 we would have assigned a mass of $\sum_j P_{(1)j} = \theta_1$ to $\mathcal{S}_1$ and a mass of $\sum_j (P_{(2)j} - P_{(1)j}) = \theta_2 - \theta_1$ to $\mathcal{S}_2$, leaving a mass of $1 - \theta_2$ to be distributed to $\mathcal{S}_3$. For the success of this strategy it is therefore necessary that $\theta_2 \leqslant 1$. It was shown in Krishnamoorthy and Mitra (1980) that the condition $\theta_2 \lesssim 1$ is sufficient as well.

In the beginning of stage 3, each column of the configuration has at least two zero entries and at most one non-zero entry. We start from column 1 and scan the columns until we reach a column with a non-zero entry, say column c. If e.g. this nonzero entry occurs in the first row we scan the subsequent columns until we reach a nonzero entry in a different row. We continue further scanning until we find a column with a nonzero entry in yet another row. The minimum of the three nonzero entries is now zeroed out by assigning a probability equal to this minimum to (j_1, j_2, j_3), where j_1, j_2, j_3 represent the columns with the non-zero entry in the first row, second row and third row respectively. If the non-zero entry in column c, which for the sake of illustration we have assumed to be j_1, is not zeroed out in the process, then scan the subsequent columns again until we find two columns with nonzero entries occurring in two distinct rows other than the first. The entire process is repeated until the nonzero entries in each column are completely zeroed out.

When $\theta_2 > 1$, the strategy in stage 2, will eventually fail at some step. Assume that it fails for the first time at column j. Without loss of generality assume further that $P_{1j} = P_{(1)j}$. At this stage the first row contains either a zero entry or the second minimum entry of each column. In case it does not, it will do so, once the required mass is removed from the first row of column x to be assigned to the point (x, j, j) in $\mathcal{S}_2$. In any case this step is executed completely in column j so that the configuration has the above mentioned property, call it property π. In columns 1 to j, the first row has a zero entry and the single nonzero entry occurs, if at all, either in row 2 or in row 3. We now show that the balance of the three marginal distributions, as described in the configuration, can be fully explained by distributing the probability mass only to points in $\mathcal{S}_2$. We first zero out the first nonzero entry in row 1 which occurs in some column following column j, say column k. Assume further that the second row of column k contains a zero entry. This is done by assigning appropriate probability masses to the point (k, x, k) in $\mathcal{S}_2$ where x is either a column preceding column j with a nonzero entry in the second row or a subsequent column. If a subsequent column has to be used and it has a zero entry in the first row, then there is no restriction on the probability mass that can be removed from the nonzero entry in the second row position, otherwise just enough mass can be removed from the second row of column x, so that property π is not disturbed. This process is repeated to zero out successively the nonzero entries in the first row for subsequent columns. Since this way a probability mass of θ_1 is assigned to $\mathcal{S}_1$ (which is the maximum mass that could be so assigned) and the rest of the mass is distributed to

$\mathcal{S}_3$, the suggested integration plan is clearly optimal. When $\theta_2 < 1$ for the integration plan

$$\text{Prob}\,(\mathcal{S}_1) = \theta_1,\ \text{Prob}\,(\mathcal{S}_2) = \theta_2 - \theta_1,\ \text{Prob}\,(\mathcal{S}_3) = 1 - \theta_2$$

and $$E(\nu) = 1\times\theta_1 + 2\times(\theta_2-\theta_1) + 3\times(1-\theta_2) = 3-\theta_1-\theta_2 = \theta_3.$$

We note that if δ_j denotes the variable which assumes the value 1, if the j-th population unit is included in the survey and assumes the value 0 otherwise, then

$$\nu = \delta_1+\delta_2+\ldots+\delta_N. \qquad \ldots\ (8)$$

Hence

$$E(\nu) = \sum_j \text{Prob}\,(\delta_j = 1) \geqslant \sum_j P_{(3)j} = \theta_3. \qquad \ldots\ (9)$$

Since the suggested integration plan attains the lower bound θ_3, it is again clearly optimal.

For the case $\theta_2 > 1$, we shall now illustrate the algorithm through a numerical example.

Consider the initial configuration given by Table 2.

TABLE 2: PROBABILITIES OF SELECTION OF THREE VILLAGES UNDER THREE SURVEYS

	village		
survey	1	2	3
1 (P_{1j})	0.1	0.5	0.4
2 (P_{2j})	0.4	0.1	0.5
3 (P_{3j})	0.5	0.4	0.1

Setting $p_{111} = p_{222} = p_{333} = 0.1$, at the end of stage 1, we are left with the configuration,

0	0.4	0.3
0.3	0	0.4
0.4	0.3	0

We put $p_{211} = 0.3$. This leads to the following configuration with the property π.

0	0.1	0.3
0	0	0.4
0.1	0.3	0

The choice of $p_{232} = 0.1$, $p_{331} = 0.1$ and $p_{332} = 0.2$ leads to the final configuration with all zero entries. The case $\theta_2 > 1$ displays certain interesting peculiarities which we describe below.

1.6. *Induced design, not necessarily optimally integrated for a subset of surveys* : Consider the plan for optimum integration in the numerical example. Suppressing the 2nd coordinate, we have the following joint distribution for X_1 and X_3 representing the first and third surveys

$$p_{1\times1} = p_{3\times3} = 0.1,\ p_{2\times2} = 0.2,\ p_{2\times1} = 0.3,\ p_{3\times1} = 0.1,\ p_{3\times2} = 0.2.$$

This plan assigns a probability mass of 0.4 to $\mathcal{S}_1$ and a probability 0.6 to $\mathcal{S}_2$ and is therefore not optimal, the value of θ_1 being 0.6. However, from the algorithm described for the case $\theta_2 > 1$, it should be clear that if it is so desired, one could always ensure that the second minimum entry in each column is zeroed out provided this entry as well as the maximum column entry are confined to the second or third row. Upto column j this process works smoothly. At column j, it was ensured that the configuration has property π. In subsequent columns the 2nd minimum entry could be zeroed out without destroying property π as long as it and the maximum entry were present among the 2nd and 3rd rows. This would imply optimality of the induced design for 2nd and 3rd surveys. A little thought will reveal that there is nothing sacrosant about the 2nd and 3rd rows. One could in fact enforce the optimality of the induced design for any two predetermined surveys, of course at the cost of losing optimality for certain other subsets. Thus in the numerical example, if it is desired to preserve the optimality of the induced design for the first and third surveys, starting from the configuration at the end of stage 1, one could take an alternative route to the final configuration setting in succession $p_{232} = 0.3$, $p_{211} = 0.1$, $p_{311} = 0.2$, and $p_{331} = 0.1$. We conclude this subsection by raising the following converse problem, yet awaiting a solution. Given an arbitrary optimally integrated plan, for any two surveys is it possible to extend this plan to obtain an optimally integrated

plan for the 3rd survey as well ? If the answer turns out to be in the affirmative, a Keyfitz-like stepwise algorithm would be available for optimally integrating 3 surveys. The following counter-example shows that a stepwise algorithm may not exist beyond 3 surveys.

1.7. *The case of four surveys* : *R. Chandrasekaran's counter-example* : The following example shows that the simple algorithm proposed by Mitra and Pathak cannot be extended in a routine fashion to 4 surveys.

TABLE 3 : PROBABILITIES OF SELECTION OF FOUR VILLAGES UNDER FOUR SURVEYS

survey	village			
	1	2	3	4
1 (P_{1j})	1/3	0	1/3	1/3
2 (P_{2j})	1	0	0	0
3 (P_{3j})	1/3	1/3	0	1/3
4 (P_{4j})	1/3	1/3	1/3	0

A routine extension of the algorithm of Mitra and Pathak leads to the following plan for integrating the 4 surveys : $p_{1111} = 1/3$, $p_{3122} = 1/3$, $p_{4143} = 1/3$ with an expected number of distinct units $E(\nu) = 7/3$. The following alternative integration plan $p_{1122} = 1/3$, $p_{3113} = 1/3$, $p_{4141} = 1/3$ has a lower expected value of 2 for the number of distinct units and is therefore a superior plan. Since the value of 0_4 is also equal to 2, the expected value of the number of distinct units in the alternative plan attains the lower bound. The alternative plan is thus an optimally integrated plan. One also faces a difficulty of another type. Sample points in $\mathcal{S}_2$ are seen to have two different structures of the type (1, 1, 2, 2) or (1, 2, 2, 2). In stage 2 one is thus occasionally unable to decide which path to proceed along to eliminate the 2nd smallest entry in each column.

2. Other cost functions

2.1. *Cost depends on ν in a non-linear fashion* : We first consider the case where the cost of the survey depends exclusively on ν, the number of distinct units, increases monotonically with ν but the amount of increase is progressively smaller as ν increases. Let $C(\nu)$ denote the cost of a survey involving ν distinct units. Our assumption implies $C(1) < C(2) < C(3)$, and $C(3)-C(2) < C(2)-C(1)$. It was shown by Krishnamoorthy and Mitra (1986) that the optimal integration plan of Mitra and Pathak retains the optimality under such a cost function.

Consider a point in $\mathcal{S}_1$ and a point in $\mathcal{S}_3$. We say that these 2 points are matched if they agree in one coordinate. Since the coordinates of a point in $\mathcal{S}_1$ are identical and those of a point in $\mathcal{S}_3$ are necessarily distinct, these two points could agree at most in one coordinate. As for example, the point (1, 1, 1) and (2, 3, 1). Consider an integration plan which assigns a mass of δ to (1, 1, 1) and a mass of δ' to (2, 3, 1). A mass equal to $\min(\delta, \delta')$ could be removed from the point (1, 1, 1) in $\mathcal{S}_1$ and (2, 3, 1) in $\mathcal{S}_3$ and redistributed equally to the points (2, 1, 1) and (1, 3, 1) both in $\mathcal{S}_2$. If the cost increases linearly with ν, the resulting plan retains its optimality as these manoeuvres do not affect $E(\nu)$. However, alternative plans desired through the algorithm of Mitra and Pathak may allow for different degrees of matching. Krishnamoorthy and Mitra (1986) have therefore suggested a strategy for maximum matching. Consider a plan for a maximally matched optimally integrated survey and a plan for optimally integrated survey derived from the same by transferring equal masses to the maximum extent possible from $\mathcal{S}_1$ and $\mathcal{S}_3$ to points in $\mathcal{S}_2$. Krishnamoorthy and Mitra (1986) have shown that the resulting plan is cost optimal if

$$1 < [C(3)-C(2)]/[C(2)-C(1)] \leqslant 2. \qquad \ldots \quad (10)$$

Observe that there is zero matching now between points in $\mathcal{S}_1$ and points in $\mathcal{S}_3$. However, transfers to $\mathcal{S}_2$ can still take place under slightly more unfavourable condition. Thus, given a mass 2δ attached to the point $(u, u, u) \in \mathcal{S}_1$ and a mass δ at the point $(j, k, l) \in \mathcal{S}_3$ with u, j, k, l all distinct, a mass of δ could be transferred to each of the points (j, u, u), (u, k, u), (u, u, l) in $\mathcal{S}_2$. These transfers though not profitable under (10) will turn out to be profitable if

$$2 < [C(3)-C(2)]/[C(2)-C(1)]. \qquad \ldots \quad (11)$$

We keep on making these transfers until zero mass is left in $\mathcal{S}_1$ or in $\mathcal{S}_3$ or in both. Krishnamoorthy and Mitra (1986) show that the resulting plan is cost-optimal if $[C(3)-C(2)]/[C(2)-C(1)] > 2$.

2.2. *Lahiri's serpentine arrangement of contiguous geographical units*: Lahiri (1954) proposed a serpentine arrangement of the population units which approximately ensures that the distance between the geometric centres of the j-th and j'-th units is roughly proportional to $|j-j'|$. The following diagram and table which we reproduce from Des Raj (1957) shows to what extent this model helps in a particular case. When the cost of an integrated survey which uses unit j for survey 1 and unit j' for survey 2 is proportional to $|j-j'|$,

Lahiri (1954) proposes the following plan for optimal integration of surveys 1 and 2.

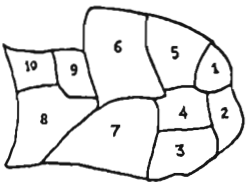

Figure 1 : Map showing boundaries of ten villages and Lahiri's serpentine arrangement

Let

$$q_{1j} = \sum_{u=1}^{j} P_{1u} \quad \{j = 1, 2, \ldots, N\} \qquad \ldots \quad (12)$$

and

$$q_{2j'} = \sum_{u=1}^{j'} P_{2u} \quad \{j' = 1, 2, \ldots, N\} \qquad \ldots \quad (13)$$

TABLE 4 : COST MATRIX c_{ij}

village no. survey 1	survey 2									
	1	2	3	4	5	6	7	8	9	10
1	0	5	10	5	5	11	14	20	15	21
2	5	0	7	5	8	13	13	21	18	23
3	10	7	0	6	11	14	8	17	16	20
4	5	5	6	0	6	9	9	16	13	18
5	5	8	11	6	0	6	11	16	12	17
6	11	13	14	9	6	0	10	11	5	10
7	14	13	8	9	11	10	0	9	9	12
8	20	21	17	16	16	11	9	0	6	5
9	15	18	16	13	12	5	9	6	0	5
10	21	23	20	18	17	10	12	5	5	0

Choose a number X at random in the interval [0,1]. Let j and j' be respectively the largest integers satisfying the inequalities $X \leqslant q_{1j}$, $X \leqslant q_{2j'}$. We select accordingly unit j for survey 1 and unit j' for survey 2. Des Raj (1957) shows that Lahiri's solution is indeed an optimal solution of the corresponding

LP problem, thus establishing the optimality of Lahiri's integration plan. Before we conclude this section, we wish to make the following remarks. It can be seen that the arguments given by Des Raj can be suitably modified so that the optimality of Lahiri's solution can be established for a wider class of cost functions e.g., cost proportional to $(j-j')^2$ or even for cost proportional to $f(j-j')$, where $f(x)$ is any convex function of x. In fact the optimality holds if cost is proportional to a convex function of $x_j-y_{j'}$, where $x_1, x_2, \ldots, x_N$ and $y_1, y_2, \ldots, y_N$, are arbitrary sets of numbers. Here the units have to be first arranged in increasing order of the values of x_j and also in increasing order of the values of $y_{j'}$ before the probabilities can be cumulated. It may be of interest to note that Lahiri's solution minimises $E(X_1-X_2)^2$ where X_1 and X_2 are discrete random variables supported respectively on the set $\{x_1, x_2, \ldots, x_N\}$ and $\{y_1, y_2, \ldots, y_N\}$ with the corresponding probability vectors $\{p_{11}, p_{12}, \ldots, p_{1N}\}$ and $\{p_{21}, p_{22}, \ldots, p_{2N}\}$. We thus have a joint distribution of the pair of random variables X_1 and X_2 which maximizes the covariance between X_1 and X_2, given the two marginal distributions (Whitt, 1976; Mitra and Mohan, 1987). In a given concrete case, if the marginal distributions of X_1 and X_2 are given, we are thus able to provide a lower and an upper bound for cov (X_1, X_2) which is an improvement over the Cauchy-Schwartz inequality,

$$-\{\text{var}\,(X_1).\ \text{var}(X_2)\}^{\frac{1}{2}} \leqslant \text{cov}(X_1, X_2) \leqslant +\{\text{var}(X_1).\ \text{var}(X_2)\}^{\frac{1}{2}}. \quad \ldots \quad (14)$$

3. Sample size $n > 1$

3.1. *Linear programming approach* : So far, we have considered integrating surveys each of sample size 1. Arthanari and Dodge (1981) show how the case of a general n as common sample size can be treated in the framework of an LP problem. They restrict their attention to sampling without replacement. We shall illustrate this method using the data given in Table 1 and a sample of size 2 drawn with replacement from this population. From standard results in the classical occupancy problems (Feller, vol. I, 1950, p. 52) it follows that the total no. of distinct samples that have to be considered (ignoring order) is $s = \binom{N+n-1}{n} = \binom{4+2-1}{2} = 10$. The ten possible samples for survey 1 are listed along the rows of Table 5 while those for survey 2 are listed along the columns. The probabilities of selection for the various types of samples worked out from the original probabilities of selection assigned to the four villages under the assumption of independent draws are given on the respective marginals.

TABLE 5. THE COST MATRIX FOR INTEGRATING TWO SAMPLES OF SIZE TWO EACH

	survey 2										
survey 1	1,1	2,2	3,3	4,4	1,2	1,3	1,4	2,3	2,4	3,4	Prob
1,1	1 .09	2	2	2	2	2	2	3	3	3	0.09
2,2	2	1	2	2	2 .01	3	3	2	2	3	0.01
3,3	2 .15	2	1 .01	2	3	2	3	2	3	2	0.16
4,4	2	2	2	1 .04	3	3	2	3	2	2	0.04
1,2	2	2	3	3	2 .06	3	3	3	3	4	0.06
1,3	2 .01	3	2	3	3 .13	2 .10	3	3	4	3	0.24
1,4	2	3	3	2	3	3	2 .12	4	3	3	0.12
2,3	3	2 .04	2	3	3	3	4	2 .04	3	3	0.08
2,4	3	2	3	2	3	4	3	3	2 .04	3	0.04
3,4	3	3	2	2	4	3	3 .08	3	3 .04	2 .04	0.16
Prob	0.25	0.04	0.01	0.04	0.20	0.10	0.20	0.04	0.08	0.04	

In the top left hand corner of the (j, j') cell of the table, we record the number of distinct units in the combined sample, one of type j for survey 1 with one of type j' for survey 2. This will provide the cost matrix for the LP problem. To guard against mistake in writing out the cost matrix, one can apply the following simple checks.

(1) For the j-th row, the row total of these entries in the table must be equal to

$$m\binom{N+n-2}{n}+N\binom{N+n-2}{n-1} \qquad \ldots (15)$$

where m denotes the number of distinct units appearing in the j-th sample.

(2) The grand total of all the s^2 entries is equal to

$$N\left[\binom{N+n-1}{n}^2-\binom{N+n-2}{n}^2\right]. \qquad \ldots (16)$$

Both these formulae are due to Balasubramanian (see Appendix).

The table also shows an optimum solution to the LP problem. For the cells which receive a non-zero probability mass, the probability values are noted in the bottom right hand corner of the respective cells.

It is seen that the size of the LP problem is fairly large even for moderate values of N and n. Krishnamoorthy and Mitra (1987) therefore looked for an optimum solution in a narrow class of integration plans.

3.2. *An alternative approach*: Krishnamoorthy and Mitra (1987) consider a plan $\mathcal{P}$ for integration of k surveys for the special case of a sample size one for each survey and n independent repetitions of $\mathcal{P}$ so as to ensure a sample size n for each survey. They restrict their attention only to the plans of this type which they denote by $\mathcal{P}^n$. They have shown that if $k=2$ and $\mathcal{P}$ is obtained through the Mitra-Pathak algorithm, then $\mathcal{P}^n$ is indeed optimal in the sense that it minimizes the expected number of distinct units in the integrated survey. The same is also true for $k = 3$ and $\theta_2 \leqslant 1$. Let p_j be probability of inclusion of the j-th unit under the integration plan $\mathcal{P}$. If the plan $\mathcal{P}$ is independently repeated n times then the probability of inclusion of the j-th unit is given by $1-[1-p_j]^n$. Hence

$$E(\nu) = \sum_j [1-(1-p_j)^n]. \qquad \ldots \quad (17)$$

Under the Mitra-Pathak algorithm if either $k = 2$ or $k = 3$ and $\theta_2 \leqslant 1$ for each j, p_j attains its lower bound. Therefore for any such plan $\mathcal{P}$, $\mathcal{P}^n$ inherits the optimality property. When $\theta_2 > 1$ however, the same is no longer true for an arbitrary plan $\mathcal{P}$ derived through the Mitra-Pathak algorithm and the optimal plan of the type $\mathcal{P}^n$ has to be separately worked out.

To minimize the expression (17) one has to make the p_j's as small as possible. However this cannot be arbitrarily done. We have seen that $p_j \geqslant P_{(3)j}$ for each j. Further once the assignment is made as in stage 1 of Mitra-Pathak algorithm, we have $p_j \leqslant P_{(3)j}+P_{(2)j}-P_{(1)j}$. If one restricts one's attention to optimal integration plans $\mathcal{P}$, the inequality can be sharpened further to

$$P_{(3)j} \leqslant p_j \leqslant P_{(3)j}+\Delta_j \qquad \ldots \quad (18)$$

where $\Delta_j = \min\{P_{(2)j}-P_{(1)j}, \theta_2-1\}$. Consider the plan $\mathcal{P}_b$ which is defined as follows. Let

$$P_{(3)b_1} = \min_j \{P_{(3)j}\}. \qquad \ldots \quad (19)$$

We first fix

$$p_{b_1} = P_{(3)b_1}. \qquad \ldots \quad (20)$$

This will require assigning a mass $P_{(2)b_1}-P_{(1)b_1}$ to points in $\mathcal{S}_2$ with two coordinates equal to b_1. One would like to scan the remaining columns for the next smallest value of $P_{(3)j}$ and fix the p_j value accordingly. This however may not be possible if the corresponding value of $P_{(2)j}-P_{(1)j}$ is large. Since the maximum mass that could be assigned to $\mathcal{S}_2$ is $1-\theta_1$ the possibility of the following is ruled out,

$$P_{(2)b_1}-P_{(1)b_1}+P_{(2)j}-P_{(1)j} > 1-\theta_1.$$

Accordingly, let

$$P_{(3)b_2}+P_{(2)b_2}-P_{(1)b_2}-\min\{P_{(2)b_2}-P_{(1)b_2},\ 1-\theta_1-(P_{(2)b_1}-P_{(1)b_1})\}$$

$$=\min_{j\neq b_1}[P_{(3)j}+P_{(2)j}-P_{(1)j}-\min\{P_{(2)j}-P_{(1)j},\ 1-\theta_1-(P_{(2)b_1}-P_{(1)b_1})\}].$$

We set

$$p_{b_2}=P_{(3)b_2}+P_{(2)b_2}-P_{(1)b_2}-\min\{P_{(2)b_2}-P_{(1)b_2},\ 1-\theta_1-(P_{(2)b_1}-P_{(1)b_1})\} \quad \ldots (21)$$

This process is continued until p_j is fixed for all j.

Consider another plan $\mathcal{P}_t$ where we try to preserve the values of p_j from the top. Thus let $P_{(3)t_n}+\Delta_{t_n}=\max_j\{P_{(3)j}+\Delta_j\}$. We fix

$$p_{t_n}=P_{(3)t_n}+\Delta_{t_n}. \quad \ldots (22)$$

Let

$$P_{(3)t_{n-1}}+\min\{P_{(2)t_{n-1}}-P_{(1)t_{n-1}},\ \theta_2-1-(p_{t_n}-P_{(3)t_n})\}$$

$$=\max_{j\neq t_n}[P_{(3)j}+\min\{P_{(2)j}-P_{(1)j},\ \theta_2-1-(p_{t_n}-P_{(3)t_n})\}].$$

Set

$$p_{t_{n-1}}=P_{(3)t_{n-1}}+\min\{P_{(2)t_{n-1}}-P_{(1)t_{n-1}},\ \theta_2-1-(p_{t_n}-P_{(3)t_n})\} \quad \ldots (23)$$

This process is continued till p_j is fixed for all j.

Krishnamoorthy and Mitra show that the plans $\mathcal{P}_b$ and $\mathcal{P}_t$ can be derived using the Mitra-Pathak algorithm. They further show that if these two plans had identical vectors of inclusion probabilities, that is, if $(p_{b_1}, p_{b_2}, \ldots, p_{b_n})=(p_{t_1}, p_{t_2}, \ldots, p_{t_n})$, then $\mathcal{P}_b^n$ is an optimal integration plan irrespective of the value of n and so is $\mathcal{P}_t^n$. It is easily seen that otherwise $\mathcal{P}_b^n$ is cost-optimal for sufficiently large sample size n. They show through an example that occasionally independent planning of the two surveys could show better results at least for some sample size n compared to an arbitrary Mitra-Pathak plan. We reproduce the following example from Krishnamoorthy and Mitra (1987) to illustrate a situation where $\mathcal{P}_b$ and $\mathcal{P}_t$ are different.

Consider the data in Table 6.

TABLE 6. VALUES OF P_{ij}

i \ j	1	2	3	4	5	6	7	8	9	
1	.10	0.0	.15	0.0	.20	.20	0.0	.05	.30	$\theta_1 = 0.0$
2	.10	.10	0.0	.20	.15	0.0	.15	.30	0.0	$\theta_2 = 1.15$
3	0.0	.12	.15	.15	0.0	.21	.27	0.0	.10	$\theta_3 = 1.85$

Plan $\mathcal{P}_b$:

$$\mathcal{P}_{b_1} = .10,\ \mathcal{P}_{b_2} = .12,\ \mathcal{P}_{b_3} = .15,\ \mathcal{P}_{b_4} = .20,$$

$$\mathcal{P}_{b_5} = .20,\ \mathcal{P}_{b_6} = .21,\ \mathcal{P}_{b_7} = .26,\ \mathcal{P}_{b_8} = .35,\ \mathcal{P}_{b_9} = .40$$

where $b_j = j,\ \ j = 1, 2, \ldots, 9.$

Plan $\mathcal{P}_t$:

$$\mathcal{P}_{t_1} = .10,\ \mathcal{P}_{t_2} = .12,\ \mathcal{P}_{t_3} = .15,\ \mathcal{P}_{t_4} = .20,\ \mathcal{P}_{t_5} = .20,$$

$$\mathcal{P}_{t_6} = .21,\ \mathcal{P}_{t_7} = .30,\ \mathcal{P}_{t_8} = .30,\ \mathcal{P}_{t_9} = .42.$$

where $t_j = j$, for $j = 1, 2, \ldots, 6$, $t_7 = 8$, $t_8 = 9$, and $t_9 = 7$.

Let ν_n denote the number of distinct units in the combined sample of size $3n$. We compute the value $E(\nu_n)$ for the plans $\mathcal{P}_b$ and $\mathcal{P}_t$ for $n = 2, 3, \ldots, 10$ and present in the following table.

TABLE 7. VALUES OF $E\nu_n$

n	Plan $\mathcal{P}_b$	Plan $\mathcal{P}_t$
2	3.4736	3.4726
3	4.5787	4.5773
4	5.4214	5.4201
5	6.0739	6.0732
6	6.5862	6.5863
7	6.9935	6.9942
8	7.3208	7.3221
9	7.5865	7.5881
10	7.8040	7.8057

The above table values shows that $E(\nu_n)$ of the plan $\mathcal{P}_b$ is greater than that of the plan $\mathcal{P}_t$ for $2 \leqslant n \leqslant 5$ and less than that of the plan $\mathcal{P}_b$ for $n \geqslant 6$. Also note that the absolute difference between them is numerically insignificant for all $n \geqslant 2$. One should not rush to the conclusion that if $\mathcal{P}_b$ and $\mathcal{P}_t$

are different, $\mathscr{V}_1^*$ is better than $\mathscr{V}_0^*$ for small sample size. Another numerical example in the same paper shows that this may not always be the case.

3.3. *Sampling with probability proportional to total size* : Roychoudhury (1956) proposes the following plan for selection of samples of size n, one for survey 1 and another for survey 2, which ensures that at least $n-1$ units overlap in the two samples and the probability of selection of any one particular sample is proportional to sum of the probabilities of the constituent units.

Step 1 : Select one unit at random for survey 1 assigning a probability P_{1j} to unit j, $j = 1, \ldots, N$. Let the unit u be selected this way.

Step 2 : Select one unit at random for survey 2 assigning a probability P_{2j} to unit j, $j = 1, \ldots, N$. Let the unit u' be selected this way.

Step 3 : If $u = u'$, from the remaining $N-1$ units select $n-1$ units at random with equal probability and without replacement. These $n-1$ units along with unit $u(=u')$ are common for both the surveys.

Step 4 : If $u \neq u'$ form a composite unit (u, u') which along with the remaining $N-2$ units constitute a collection of $N-1$ units. From these $N-1$ units draw a sample of $n-1$ units at random with equal probability and without replacement.

Step 5 : If the composite unit gets selected in this process, then unit u' is also included for survey 1 and u for survey 2. Units u, u' along with the other selected $n-2$ units are thus common for both the surveys.

Step 6 : If the composite unit is rejected, then unit u along with the selected $n-1$ units is a sample for survey 1 and unit u' along with the selected $n-1$ units is a sample for survey 2.

One notices a certain lack of clarity in the description of the plan as given in Roychoudhury (1956) which also persists in Murthy's presentation (Murthy, 1967). This may have led Arthanari and Dodge (1981) to interpret the plan in their own way and conclude that the Roychoudhury plan ensures the correct probability of selection for survey 1 but not for survey 2.

With the Roychoudhury plan the number of common units between the two samples is $n-1$ or n and the expected number of common units is equal to $n-1+\pi$ where π is the probability of having all the n units common. Here

$$\pi = \frac{n}{N}\left[\frac{N^2(N-n)}{n(N-1)}\sigma_{12}+1\right] \qquad \ldots (23)$$

and $$\sigma_{12} = \frac{1}{N} \sum_j \left(P_{1j} - \frac{1}{N}\right)\left(P_{2j} - \frac{1}{N}\right).$$

Roychoudhury (1956) states that the above algorithm for integrating two surveys could be extended to provide integration of k surveys, requiring no more than $(n+k-1)$ distinct units and ensuring that the individual probability requirements of each survey be satisfied. No further details are given. We present below one such extension.

The first k steps are similar to steps 1 and 2 of the previous algorithm. Thus in step i $[1 \leqslant i \leqslant k]$ we select a unit from the population assigning probabilities P_{i1}, P_{i2}, ..., P_{iN} to the N population units,

$$1, 2, \ldots, N.$$

Let the unit $u^{(i)}$ be so selected for survey i. Let the units $u^{(1)}, u^{(2)}, \ldots, u^{(k)}$ consist of precisely d distinct units which we denote by $u^{[1]}, u^{[2]}, \ldots, u^{[d]}$ where $u^{[1]} = u^{(1)}$.

There are now $N-d$ units left in the population. We add to these $N-d$ units $d-1$ artificial units $\alpha_1, \alpha_2, \ldots, \alpha_{d-1}$.

From the collection of $(N-d)+(d-1) = N-1$ units so constructed, we draw a sample of $n-1$ units at random with equal probability and without replacement.

Let the sample so selected include e artificial units and $n-e-1$ original units.

These $n-e-1$ original units are common units for the k surveys. Other units are chosen for the various surveys using the following Latin square.

TABLE 8. $d \times d$ LATIN SQUARE AND ITS USE IN INTEGRATING k MIDZUNO-SEN TYPE SURVEYS

	α_1 ↓	α_2 ↓	...	α_{d-1} ↓
$u^{[1]}$	$u^{[2]}$	$u^{[3]}$		$u^{[d]}$
$u^{[2]}$	$u^{[3]}$	$u^{[4]}$		$u^{[1]}$
$u^{[d]}$	$u^{[1]}$	$u^{[2]}$		$u^{[d-1]}$

Each artificial unit corresponds to one particular column of this Latin squares. Thus α_1 corresponds to the second column, α_2 to the third column and so on.

We strike out from this Latin square all columns that correspond to artificial units which are absent in the selected sample.

If $u^{(i)}=u^{(h)}$ the units to be included for the i-th survey are now read out from the h-th row of the Latin square, after deleting the delinquent columns.

Thus if $e = 2$ and α_1 and α_2 are the only artificial units to appear in the selected sample, then survey 1 will cover the units $u^{[1]}$, $u^{[2]}$, $u^{[3]}$ in addition to the $n-3$ original units, already selected which are common to all the k surveys. We note that in Table 8 one could have used any $d \times d$ Latin square in whose first column $u^{[1]}$, $u^{[2]}$ ... appear in the natural order.

Let $\hat{\mathcal{S}} = (i_1, i_2, \ldots, i_n)$ be a subset of $\mathcal{S}$. We shall now calculate the probability that units with serial numbers as in $\hat{\mathcal{S}}$ constitute the chosen sample for survey 1. Let $\mathcal{J}$ be a subset of $\mathcal{S}^k$ with the first coordinate, restricted to $\hat{\mathcal{S}}$. Other coordinates are unrestricted in $\mathcal{S}$.

Clearly for $\hat{\mathcal{S}}$ to be a chosen sample for survey 1, it is necessary that vector $\{u^{(1)}, u^{(2)}, \ldots, u^{(k)}\}$ assume values only in $\mathcal{J}$. Further any such choice of the vector $\{u^{(1)}, u^{(2)}, \ldots, u^{(k)}\}$ corresponds uniquely to the set of distinct units $u^{[1]}, u^{[2]}, \ldots, u^{[d]}$ and in turn to artificial units $\alpha_1, \alpha_2, \ldots \alpha_{d-1}$. They also correspond to a unique sample of size $n-1$ containing possibly some artificial units so as to lead to $\hat{\mathcal{S}}$ as the chosen sample for survey 1, via the Latin square in Table 8. The probability that units in $\hat{\mathcal{S}}$ constitute a sample for survey 1 is therefore given by

$$(P_{1i_1}+P_{1i_2}+\ldots+P_{1i_n}) \prod_{u=2}^{k} \left(\sum_{j=1}^{N} P_{uj} \right) \Big/ \binom{N-1}{n-1}$$

$$= (P_{1i_1}+P_{1i_2}+\ldots+P_{1i_n}) \Big/ \binom{N-1}{n-1}.$$

The argument is similar for other surveys.

The probability of all the n units being common to all the k surveys is similarly seen to be equal to

$$\sum \prod_{u=1}^{k} (P_{ui_1}+P_{ui_2}+\ldots+P_{ui_n}) \Big/ \binom{N-1}{n-1}$$

where the summation extends over all the $\binom{N}{n}$ ways of choosing $(i_1, i_2, \ldots, i_n)$ out of $\mathcal{S}$. The probability is thus $\binom{N-1}{n-1}^{k-1}$ times the corresponding expression when the Midzuno-Sen procedure (Murthy, 1967, p. 218) is independently applied to all the k surveys.

The Roychoudhury plan is not necessarily cost optimal. This is best seen in the special case $k = 2$, $n = N-1$. Here if $(i_1, i_2, \ldots, i_n)$, $i_1 < i_2 < \ldots < i_n$ be the sample selected for survey 1 and $(j_1, j_2, \ldots, j_n)$, $j_1 < j_2 < \ldots < j_n$ the one selected for survey 2 and if $(i_1, i_2, \ldots, i_n) = (j_1, j_2, \ldots, j_n)$ then the number of distinct units in the combined sample is n. If $(i_1, i_2, \ldots, i_n) \neq$

$(j_1, j_2, \ldots, j_n)$ it is equal to $n+1$. The cost function is thus, apart from a shift, same as that considered by Keyfitz (1951) and by Mitra and Pathak (1984). The Roychoudhuri plan assigns a probability

$$\frac{(P_{1i_1}+P_{1i_2}+\ldots+P_{1i_n})\,(P_{2i_1}+P_{2i_2}+\ldots+P_{2i_n})}{\binom{N-1}{n-1}}$$

to the event that $(i_1, i_2, \ldots, i_n)$ is the common sample for both surveys 1 and 2 and this is less than

$$\min\left\{\frac{(P_{1i_1}+P_{1i_2}+\ldots+P_{1i_n})}{\binom{N-1}{n-1}},\ \frac{(P_{2i_1}+P_{2i_2}+\ldots+P_{2i_n})}{\binom{N-1}{n-1}}\right\}$$

unless either $\max\{P_{1i_1}+P_{1i_2}+\ldots+P_{1i_n},\ (P_{2i_1}+P_{2i_2}+\ldots+P_{2i_n}\} = 1$, or $\min\{(P_{1i_1}+P_{1i_2}+\ldots+P_{1i_n}),\ (P_{2i_1}+P_{2i_2}+\ldots+P_{2i_n})\} = 0$, when the equality holds. If the equality is to hold for all possible samples $(i_1, i_2, \ldots, i_n)$ as is necessary for the cost optimality of the Roychoudhury plan, it is seen that only one of the following two conditions are possible. Either

(i) atleast one of two probability distributions on the set of integers $1, 2, \ldots, N$ as specified for surveys 1 or 2 is a degenerate distribution, or

(ii) the two probability distributions have disjoint supports.

Barring these situations the Roychoudhury plan is not cost optimal in the special case.

However atleast for large n the Roychoudhury plan is nearly cost optimal in the sense that

$$\lim_{n\to\infty} \frac{E(\nu_n)}{n} = 1$$

while the comparable expressions for the case of k surveys, each survey independently planned are

$$\overline{\lim_{n\to\infty}} \frac{E(\nu_n)}{n} = \frac{1-(1-\underline{f})^k}{\underline{f}}$$

where $\underline{f} = \underline{\lim}_{n\to\infty} n/N$ and

$$\underline{\lim_{n\to\infty}} \frac{E(\nu_n)}{n} = \frac{1-(1-\bar{f})^k}{\bar{f}} \text{ where } \bar{f} = \overline{\lim_{n\to\infty}}\, n/N.$$

These expressions are equal to k if $\underline{f}$ or $\bar{f}$ are respectively equal to 0 (see Appendix II).

4. Other strongly polynomial algorithms

A 'strongly polynomial algorithm', according to Grotschel, Lovasz and Schrijver (1981), is one which (1) involves only the four basic arithmetic operations : addition, comparison, multiplication and division and (2) the number of such steps is polynomial bounded in the dimension of the input, that is in the number of data items in the input and (3) when the algorithm is applied to rational inputs, then the size of the numbers occuring during the algorithm is polynomially bounded in the dimension of the input and the size of intput numbers. Kabadi, Chandrasekaran and Nair (1987) point out that for the problem of optimal integration of surveys, if the number of surveys, k, is held fixed and N is regarded as a variable parameter then a strongly polynomial algorithm can be constructed for its solution following the approach suggested by Tardos (1986). However when k is also regarded as a variable parameter the problem is NP-hard.

Tardos considers a linear programme

$$\max\ cx$$
$$\text{s.t.}\quad Ax = b,\quad x \geqslant 0. \qquad \ldots\ (25)$$

where A is a $m \times n$ integer matrix. Let $\Delta(A)$ denote an integer greater than or equal to

$$\max\{\det B \mid B \text{ is a submatrix of } A\}. \qquad \ldots\ (26)$$

Also let $S(A, b)$ denote the number of arithmetic steps used by the subroutine of the basic algorithm when used to find

$$\max\ \bar{c}x$$
$$\text{s.t.} Ax = b,\ x \geqslant 0,\ x_i = 0 \text{ for } i \in K \qquad \ldots\ (27)$$

for an integer vector $\bar{c}$ with $\|\bar{c}\|_\infty \leqslant n^2\,\Delta$ and a subset K of the indices.

She proves that the number of arithmetic operations in the basic algorithms is to $o(n^4+n^2 \log \Delta+nS(A, b))$ and that all numbers by which the algorithm divides or multiplies have size polynomial in the size of matrix A. In our context m is k. N and n is N^k. We emphasize the fact that the Mitra-Pathak algorithm is not only a strongly polynomial algorithm but also a considerably simpler algorithm requiring a number of steps of arithmetic operations (subtraction and comparison) which is essentially linear in N. Kabadi, Chandrasekaran and Nair (1987) point out that the fact that the variable k problem is NP-hard, may be an indication that the Mitra-Pathak type simple algorithm may not exist for larger values of k.

Appendix I

DERIVATION OF EXPRESSIONS (15) AND (16)

By K. BALASUBRAMANIAN
Indian Statistical Institute

Lemma 1 : (a) *Number of non-decreasing n-sequences* $S(n, r)$ *over* $\{1, 2, \ldots, r\}$ *using each of these atleast once is* $|S(n, r)| = \binom{n-1}{n-r}$.

(b) *Number of non-decreasing n-sequences* $S(n, \leqslant r)$ *over* $\{1, 2, \ldots, r\}$ *without restriction on number of times any symbol is to be used is* $\binom{n+r-1}{n}$.

Proof : (a) Let $n_1, n_2, \ldots, n_r$ be the number of times $1, 2, \ldots, r$ respectively occurs in an element of $S(n, r)$. Then $n_1+n_2 \ldots +n_r = n$ and $n_i \geqslant 1$ for $i = 1, 2, \ldots, r$. Hence $|S(n, r)|$ is the coefficient of x^n in $(x+x^2+\ldots)^r = x^r\,(1-x)^{-r} = x^r \sum_{s=0}^{\infty} \binom{r+s-1}{s} x^s$. Thus

$$|S(n, r)| = \binom{n-1}{n-r}.$$

(b) Clearly $|S(n, \leqslant r)|$ is the coefficient of x^n in $(1+x+x^2+\ldots)^r$.

Thus
$$|S(n, \leqslant r)| = \binom{n+r-1}{n}$$

$S(n, \leqslant N)$ is precisely the sample space in sampling with replacement for a sample of size n from a population of size N, if order is ignored, and the number of sample points, hence, is $\binom{N+n-1}{n}$.

This type of sampling is considered in Section 3.1.

Notation : If $x \in S(n, r)$, let $d(x)$ represent the number of distinct elements in x considered as a set. $d(x \bigcup y)$, $d(x \bigcap y)$ have obvious interpretations when $x, y \in S(n, r)$.

Lemma 2 : *Number of* $x \in S(n, \leqslant N)$ *with* $d(x) = m$ *is*

$$\binom{N}{m}\binom{n-1}{n-m} \text{ and } \sum_{x \in S(n, \leqslant N)} d(x) = N\binom{N+n-2}{n-2}.$$

Proof : Suppose $x \in S(n, \leqslant N)$ and $d(x) = m$. Clearly such an element can occur $\binom{N}{m}\,|S(n, m)|$ times i.e. $\binom{N}{m}\binom{n-1}{n-m}$ times.

$$\sum_{x \in S(n, \leqslant N)} d(x) = \sum_{m=1}^{n} m \binom{N}{m} \binom{n-1}{n-m} = N \sum_{m=1}^{n} \binom{N-1}{n-m} \equiv N \binom{N+n-2}{n-1}$$

[Note that $\sum_{r=0}^{n} \binom{a}{r} \binom{b}{n-r} = \binom{a+b}{n}$ by Vandermonde convolution formula VCF].

Main result: Suppose $x \in S(n, \leqslant N)$ and $d(x) = m$. If $y \in S(n, \leqslant N)$ with $d(x \bigcap y) = r$, then, the number of such y's is

$$\sum_{d(y)=r}^{m} \binom{m}{r} \binom{N-m}{d(y)-r} |S(n, d(y))|.$$

$$= \binom{m}{r} \sum_{t=0}^{n-r} \binom{N-m}{t} \binom{n-1}{n-t-r}$$

$$= \binom{m}{r} \binom{N+n-m-1}{n-r} \text{ by VCF.}$$

Thus

$$\sum_{y \in S(n, \leqslant N)} d(x \bigcap y) = \sum_{r=0}^{n} r \binom{m}{r} \binom{N+n-m-1}{n-r}$$

$$= m \sum_{r=1}^{n} \binom{m-1}{r-1} \binom{N+n-m-1}{n-r}$$

$$= m \binom{N+n-2}{n-1} \text{ by VCF.}$$

Hence

$$\sum_{y \in S(n, \leqslant N)} d(x \bigcup y) = \sum_{y \in S(n, \leqslant N)} (d(x)+d(y)-d(x \bigcap y))$$

$$= m |S(n, \leqslant N)| + N \binom{N+n-2}{n-1} - m \binom{N+n-2}{n-1}$$

$$= m \binom{N+n-1}{n} + N \binom{N+n-2}{n-1} - m \binom{N+n-2}{n-1}$$

$$= N \binom{N+n-2}{n-1} + m \binom{N+n-2}{n}$$

$$\left[\text{using } \binom{n+1}{r} - \binom{n}{r-1} = \binom{n}{r} \right].$$

This proves (15).

$$\sum_{s,\,p(s)(n,\leqslant N)} \sum d(x \bigcup y) = \sum_{x\,\epsilon\,S(n,\leqslant N)} \left\{ N \binom{N+n-2}{n-1} + d(x) \binom{N+n-2}{n} \right\}$$

$$= N \binom{N+n-2}{n-1} \binom{N+n-1}{n}$$

$$+ \binom{N+n-2}{n} . N \binom{N+n-2}{n-1}$$

$$= N \left\{ \binom{N+n-1}{n} \left[\binom{N+n-1}{n} - \binom{N+n-2}{n} \right] \right.$$

$$\left. + \binom{N+n-2}{n} \left[\binom{N+n-1}{n} - \binom{N+n-2}{n} \right] \right\}$$

$$= N \left\{ \binom{N+n-1}{n}^2 - \binom{N+n-2}{n}^2 \right\}.$$

This proves (16).

Appendix II

ON THE EXPECTATION OF THE NUMBER OF DISTINCT UNITS IN k INDEPENDENT REPETITIONS OF MIDZUNO-SEN SAMPLING SCHEME

By S. K. Mitra and K. Balasubramanian

Indian Statistical Institute

Suppose $Z_j = 1$ if j-th population unit is in one of the k samples and 0 otherwise. Then

$$P(Z_j = 0) = \prod_{i=1}^{k} (1 - P_{ij}) \frac{\binom{N-2}{n-1}}{\binom{N-1}{n-1}} = \left(\frac{N-n}{N-1} \right)^k \prod_{i=1}^{k} (1 - P_{ij})$$

$\nu_n = \sum_{j=1}^{N} Z_j$ is clearly the number of distinct units in the pooled sample.

$$E(\nu_n) = \sum_{j=1}^{N} E(Z_j) = \sum_{j=1}^{N} [1 - P(Z_j = 0)] = N - \left(\frac{N-n}{N-1} \right)^k \sum_{j=1}^{N} \prod_{i=1}^{k} (1 - P_{ij})$$

where $$= N - \left(\frac{N-n}{N-1}\right)^k (N+\varepsilon_k)$$

$$N+\varepsilon_k = \sum_{j=1}^{N} \prod_{i=1}^{k} (1-P_{ij}).$$

Clearly $|\varepsilon_k| \leqslant \sum_{i=1}^{k} \sum_{j=1}^{N} P_{ij} + \sum_{i_1<i_2}^{k} \sum_{j=1}^{N} P_{i_1 j} P_{i_2 j} + \ldots \leqslant k + \binom{k}{2} + \ldots = 2^k - 1.$

Writing $\frac{n}{N} = f$, we get

$$\frac{E(\nu_n)}{n} = \frac{1}{f} - \left(\frac{f^{-1}-1}{f^{-1}-\frac{1}{n}}\right)^k \left(\frac{1}{f} + \frac{\varepsilon_k}{n}\right)$$

and for large n this behaves like

$$\frac{1}{f} - \frac{(1-f)^k}{f} = \frac{1-(1-f)^k}{f} = \sum_{r=0}^{k-1} (1-f)^r.$$

R.H.S. is a monotonic decreasing function of f in $[0, 1]$. Thus

$$\lim_{n\to\infty} \frac{E(\nu_n)}{n} = \frac{1-(1-\bar{f})^k}{\bar{f}}$$

and

$$\overline{\lim_{n\to\infty}} \frac{E(\nu_n)}{n} = \frac{1-(1-\underline{f})^k}{\underline{f}}$$

where $$\bar{f} = \overline{\lim_{n\to\infty}} f \text{ and } \underline{f} = \underline{\lim_{n\to\infty}} f.$$

A partition of a positive integer k is a representation of k as the sum of positive integer each one of which is called a part or a summand. The order of the summands is unimportant.

The number of each partitions of k is given by the well known-Euler's partition function $p(k)$. An explicit expression for $p(k)$ in terms of k was given by Hardy and Ramanujan (see G. H. Hardy, 1940 : *Ramanujan*, Cambridge University Press, Cambridge).

Consider the units selected as the first sample unit in the k independent repetitions of the Midzuno-Sen sampling scheme, and let f_j denote the frequency with which the j-th population unit is so selected. If one ignores the zero frequencies the rest of the frequencies constitute a partition of k. Let $\mathcal{C}_\varphi$ denote the event (collection of sample points) which corresponds to the

v-th partition of k, $v = 1, 2, \ldots, p(k)$. Customarily the partitions are listed as follows.

(1) $\{k\}$, (2) $\{k-1, 1\}$, (3) $\{k-2, 2\} \ldots (p(k))\ \{1, 1, \ldots, 1\}$.

We denote by β_v the probability of observing $\mathcal{C}_v$. β_v depends on the initial probabilities of selection of the N population units as specified by the k separate surveys. If ν_n denotes the number of distinct units appearing in the combined sample, it is seen that for an integer r, $n \leqslant r \leqslant \min\{nk, N\}$

$$P(\nu_n = r) = \sum_{v=1}^{p(k)} \beta_v g_v(n, N, k, r)$$

where $g_v(n, N, k, r)$ is a function depending on the partition v as well as on n, N, k and r. Note that $g_v(n, N, k, r)$ is the conditional probability of $\nu_n = r$ given that the first stage sample units for the k separate surveys belong to $\mathcal{C}_v$. Since the second step selection for each survey is done at random with equal probability and without replacement, $g_v(n, N, k, r)$ does not depend on the initial probabilities of selection specified for the k surveys.

Let d be the number of distinct units in the first selection of k units. For the second stage we have to select $(n-1)$ more units from $(N-1)$ independently for k selections. If $\nu_n = r$, then we need $r-d$ more units from $N-d$ units and this can be done in $\binom{N-d}{r-d}$ ways. For any one such choice we consider the r units of the population as columns and the k selections as rows of an incidence matrix of order $k \times r$ whose (i, j)-th entry is 1 if j-th populations unit appears in the sample for the i-th survey and 0 otherwise. The first d columns of such a matrix correspond to the units selected in the first stage. Then it is clear that $P(\nu_n = r) = \binom{N-d}{r-d} \dfrac{W(n, r, k, d)}{\binom{N-1}{n-1}^k}$ where $W(n, r, k, d)$ is the number of ways of filling $(n-1)$ more 1's in each row (note that in the first d columns each row has one 1 corresponding to the first stage selection initially) in such a way that no column is empty (without a 1)—clearly, number of ways in which some particular t columns among the last $r-d$ columns will be empty is $\binom{r-1-t}{n-1}^k$ and t columns can be chosen in $\binom{r-d}{t}$ ways. Hence, using the principle of inclusion—exclusion we get

$$W(n, r, k, d) = \binom{r-1}{n-1}^k - \binom{r-d}{1}\binom{r-2}{n-1}^k + \binom{r-d}{2}\binom{r-3}{n-1} \ldots$$

[Note that we define $\binom{n}{r}$ as zero if $n < r$ or $r < 0$].

Consider $\binom{x-1}{n-1} - \binom{r-d}{1}\binom{x-2}{n-1}^k + \binom{r-d}{2}\binom{x-3}{n-1}^k \cdots$

Using shift operator E and the difference operator $\Delta = E-1$ we can write the expression as $\left\{1 - \binom{r-d}{1}E^{-1} + \binom{r-d}{2}E^{-2} \cdots\right\}\binom{x-1}{n-1}^k$

$$= (1-E^{-1})^{r-d}\binom{x-1}{n-1}^k = \Delta^{r-d}\binom{x-r+d-1}{n-1}^k.$$

Hence $\Pi'(n, r, k, d) = \Delta^{r-d}\binom{x-r+d-1}{n-1}^k\Big|_{x=r} = \Delta^{r-d}\binom{x+d-1}{n-1}^k_{x=0}$

Thus

$$P(\nu_n = r) = \sum_{v=1}^{p(k)} \binom{N-d_v}{r-d_v} \frac{\Delta^{r-d_v}\binom{x+d_v-1}{n-1}^k\Big|_{x=0}}{\binom{N-1}{n-1}^k}\beta_v$$

for $n \leqslant r \leqslant \min\{kn, N\}$ and d_v = number of summands in the partition corresponding to β_v. Thus $P(\nu_n = r)$ depends on β_v only through d_v and not on the actual partitions. A special case of this formula namely $n = 1$ and $P_{ij} = 1/N \;\forall\; i, j$ has been extensively studied in literature. See for example, [Feller, 1950, page 92] and [D. Basu, 1958 : On sampling with and without replacement, *Sankhyā*, 20, 287–294]. Note that n in expressions given by Basu would correspond to k in our formulae. Explicit expressions for the case $k = 2$ are given below: We have here

$$P(\nu_n = r) = g_1(n, N, 2, r)\beta_1 + g_2(n, N, 2, r)\beta_2$$

where $\beta_1 = \sum_{j=1}^{N}\prod_{i=1}^{2} P_{ij}$ and $\beta_2 = 1-\beta_1$.

Clearly we can write $P(\nu_n = 2n-r) = a(n, N, r) + b(n, N, r)\beta_1$ where $a(n, N, r)$ and $b(n, N, r)$ are independent of initial probabilities. Suppose $P_{ij} = \frac{1}{N}$ $\forall\; i, j$. Then $\beta_1 = \frac{1}{N}$ and Midzuno-Sen scheme reduces to SRSWOR.

Under SRSWOR, $P(\nu_n = 2n-r) = \binom{N}{r}\binom{N-r}{n-r}\binom{N-n}{n-r}\Big/\binom{N}{n}^2$

$$= \binom{n}{r}\binom{N-n}{n-r}\Big/\binom{N}{n},$$

a hypergeometric probability $= h(N, n, n\,;\, r)$.

$$\left[\text{ We write } \frac{\binom{a}{x}\binom{N-a}{n-x}}{\binom{N}{n}} \text{ as } h(N, a, n\,;\, x)\right].$$

Thus

$$h(N, n, n; r) = a(n, N, r) + b(n, N, r)\,\frac{1}{N} \qquad \ldots \text{ (I)}$$

Suppose $P_{11} = P_{21} = 1$ and other P_{ij}'s are zero. Clearly in this case $\beta_1 = 1$ and the samples contain unit 1 and those coming from two SRSWOR samples of size $n-1$ from $N-1$ units. Hence

$$P(\nu_n = 2n-r) = h(N-1, n-1, n-1\,;\, r-1) = a(n, N, r)\; rb(n, N, r). \quad \ldots \text{ (II)}$$

Solving (I) and (II) we get

$$P(\nu_n = 2n-r) = \frac{N(1-\beta_1)}{N-1}\, h(N, n, n\,;\, r) + \frac{N\beta_1 - 1}{N-1}\, h(N-1, n-1, n-1\,;\, r-1)$$

or

$$P(\nu_n = r) = \frac{N(1-\beta_1)}{N-1}\, h(N, n, n\,;\, 2n-r) + \frac{N\beta_1 - 1}{N-1}\, h(N-1, n-1, n-1\,;\, 2n-r-1).$$

Note that $\dfrac{N(1-\beta_1)}{N-1} + \dfrac{N\beta_1 - 1}{N-1} = 1$ and hence this expression looks like a mixture of two hypergeometric probabilities. But it is not quite so as $N\beta_1 - 1$ can be negative for small values of β_1. Nevertheless it is easy to verify that the entire expression is non-negative for $\beta_1 \in [0, 1]$.

References

Arthanari, T. S. and Dodge, Y. (1981): *Mathematical Programming in Statistics*, Wiley, New York.

Causey, B. D., Cox, L. H. and Ernst, L. R. (1985): Applications of transportation theory to statistical problems. *J. Amer. Statist. Assoc*, 80, 903-909.

Feller, W. (1950): *An Introduction to Probability Theory and Its Applications*, Wiley, New York.

Goodman, R. and Kish, L. (1950): Controlled selection—a technique in probability sampling. *J. Amer. Statist. Assoc.*, 45, 350—372.

Grotschel, M., Lovasz, J. and Schrijver, A. (1981): The ellipsoid method and its consequences in combinatorial optimization. *Combinatorica*, 1 (2), 169-197.

Kabadi, S. N., Chandrasekaran, R. and Nair, K. P. K. (1987): Optimal integration of sevora surveys. *Sankhyā* Ser. B., To Appear in 50 B(1).

Keyfitz, N. (1951): Sampling with probability proportional to size: adjustment for changes in probabilities. *J. Amer. Statist. Assoc.*, 46, 105-109.

KRISHNAMOORTHY, K. and MITRA, S. K. (1986): Cost robustness of an algorithom for optimal integration of surveys. *Sankhyā*, Ser. B. 48, 233-245.

——— (1987): Optimal integration of two or three PPS surveys with common sample size $n > 1$. *Sankhyā* Ser. B, 49, 283-306.

LAHIRI, D. B. (1954): Technical paper on some aspects of the development of the sample design. *Sankhyā*, 14, 264-316.

MADOWSKI, M. J. and PATHAK, P. K. (1980): Integration of surveys. *Scand. J. Statist.*, 7, 130-138.

MITRA, S. K. and PATHAK, P. K. (1984): Algorithms for optimal integration of two or three surveys. *Scand. J. Statist.*, 11, 257-263.

MITRA, S. K. and MOHAN, S. R. (1987): On the optimality of the northwest corner solution in some applications of the transportation theory. *Technical Report No.* 8712, Indian Statistical Institute, Delhi Centre.

MURTHY, M. N. (1967): *Sampling Theory and Methods*, Statistical Publishing Society, Calcutta.

RAJ, D. (1957): On the method of overlapping maps in sample survey. *Sankhyā*, 17, 89-98.

ROYCHOUDHURY, D. K. (1956): Integration of several PPS surveys. *Science and Culture*, 22, 119-120.

TARDOS, E. (1986): A strongly polynomial algorithm to solve combinatorial linear programmes. *Operations Research*, 34, 250-256.

WHITT, W (1976): Bivariate distributions with given marginals. *Ann. Statist.*, 4, 1280-1289.