Scand J Statist 11: 257–263, 1984

Algorithms for Optimal Integration of Two or Three Surveys

S. K. MITRA[1] and P. K. PATHAK[1,2]

[1]*Indian Statistical Institute, Delhi Centre and the* [2]*University of New Mexico*

ABSTRACT. Let $\mathscr{S}=\{1,2,\ldots,N\}$ and for each i, $1\leq i\leq k$, let $\mathscr{P}_i=\{P_{ij}, 1\leq j\leq N\}$ denote a probability distribution on $\mathscr{S}$. Let X_i denote a random variable with the distribution $\mathscr{P}_i$. For $k=2$ and 3, we present here simple algorithms which yield a joint probability distribution for random variables $(X_1,\ldots,X_k)$ with the prescribed marginal distributions such that the expectation of the number of distinct values among $(X_1,X_2,\ldots,X_k)$ is minimized.

Key words: multiple surveys, joint probability distribution, algorithm, optimal integrated surveys

1. Introduction

The problem of optimal integration of surveys has its origin in multiple surveys, sampling over successive occasions and to a lesser extent in certain problems of controlled selection in stratified sampling. For two surveys, this problem has been studied in some detail by Keyfitz (1951), Lahiri (1954), Raj (1957) and others, while Pathak & Maczynski (1980) have recently studied the more general problem of integration of k surveys with $k\geq 2$. The object of this paper is to furnish an algorithm which provides a complete solution of the problem of optimal integration of $k=2$ or 3 surveys.

Briefly the problem of optimal integration of surveys can be described as follows. At the outset, we have a population consisting of N units serially numbered $1,2,\ldots,N$. Let $\mathscr{S}$ denote the set of the first N integers. It is proposed to carry out k separate surveys on this population. The ith survey corresponds to a random variable X_i having a preassigned probability distribution $\mathscr{P}_i=\{P_{ij}, 1\leq j\leq N\}$ on $\mathscr{S}$, $1\leq i\leq k$. (Thus for each i, $1\leq i\leq k$, $P(X_i=j)=P_{ij}$, $1\leq j\leq N$, with $\Sigma_j P_{ij}=1$.) An *integrated survey* with marginals $\mathscr{P}_1,\ldots,\mathscr{P}_k$ is a joint probability distribution $\mathscr{P}$ on the kth cartesian power of $\mathscr{S}$ which realizes for each i the marginal distribution $\mathscr{P}_i$. Let $\mathbf{x}=(x_1,x_2,\ldots,x_k)$ be the observed sample in the integrated survey and $\nu=\nu(\mathbf{x})$ denote the number of distinct units represented in the sample $\mathbf{x}$. An integrated survey is called *optimal* (for the given marginals) if it minimizes $E[\nu(\mathbf{X})]$. For any given marginal surveys $\mathscr{P}_i$, $1\leq i\leq k$, an optimal integrated survey always exists, it is, however, not unique (cf. Maczynski & Pathak (1980)). We shall describe here algorithms for deriving optimal integrated surveys for $k=2$ and 3.

To begin with consider the $k\times N$ array of the P_{ij}'s, which will be called a configuration. A *configuration* in general is an arrangement of kN nonnegative numbers in k rows and N columns such that the row totals are all equal. Consider the initial configuration of the P_{ij}'s and let the smallest entry in Column j be denoted by $P_{(1)j}$, the next smallest by $P_{(2)j}$ and so on. We let

$$\theta_i=\Sigma_j P_{(i)j}. \tag{1.1}$$

We consider a partition of the N^k points in $\mathscr{S}^k$ as follows. Let

$$\mathscr{S}_u=\{\mathbf{x}: \nu(\mathbf{x})=u\}, \quad 1\leq u\leq k. \tag{1.2}$$

Clearly the $\mathcal{S}_u$'s are disjoint sets and

$$\mathcal{S}_1 \cup \mathcal{S}_2 \ldots \cup \mathcal{S}_k = \mathcal{S}^k. \tag{1.3}$$

2. Algorithm for $k=2$

Stage 1. Assign a probability of $P_{(1)j}$ to the pair (j,j) in $\mathcal{S}_1$, $1 \leq j \leq N$, and replace P_{ij} by $\bar{P}_{ij}=P_{ij}-P_{(1)j}$. In the resulting configuration of the $\bar{P}_{ij}$'s, at least one entry in each column is equal to zero, and the two row sums are equal to $1-\theta_1$, where $\theta_1=\Sigma_j P_{(1)j}$.

Stage 2. Assume without any loss of generality that $\bar{P}_{ij}$'s have the following configuration

$$\begin{matrix} 0 & 0 & \ldots & 0 & \bar{P}_{1,M+1} & \ldots & \bar{P}_{1N} \\ \bar{P}_{21} & \bar{P}_{22} & \ldots & \bar{P}_{2M} & 0 & \ldots & 0. \end{matrix} \tag{2.1}$$

We also assume without any loss of generality that $\bar{P}_{1,M+1}$ is the smallest positive entry in the configuration of the $\bar{P}_{ij}$'s. (If necessary this can be achieved by permuting the rows and the columns of the configuration in (2.1).) Clearly $\bar{P}_{1,M+1} \leq \bar{P}_{21}$. Assign a probability of $\bar{P}_{1,M+1}$ to the pair $(M+1, 1)$ in $\mathcal{S}_2$ and replace $\bar{P}_{1,M+1}$ by zero and $\bar{P}_{21}$ by $\bar{\bar{P}}_{21}=\bar{P}_{21}-\bar{P}_{1,M+1}$. The resulting configuration has one less positive entry and at most $(N-1)$ repetitions of this step may be necessary to zero out all the entries in (2.1) and thus reach the complete specification of an integrated survey, which in fact is optimal.

Note that for the suggested survey a probability of θ_1 is assigned to $\mathcal{S}_1$ and $(1-\theta_1)$ to $\mathcal{S}_2$. Hence

$$Ev = 1 \cdot \theta_1 + 2(1-\theta_1) = \theta_2, \tag{2.2}$$

since in this case $\theta_1+\theta_2=2$.

Proof of optimality: Let I_j be the indicator variable which assumes the value 1 if unit j is included in the survey and the value zero otherwise. Then $v=I_1+\ldots+I_N$, so that

$$\begin{aligned} Ev &= \Sigma_j \mathcal{P}(I_j = 1) = \Sigma_j \mathcal{P}\left(\bigcup_{i=1}^{k} \{X_i = j\} \right) \\ &\geq \Sigma_j \max P_{ij} = \Sigma_j P_{(k)j} = \theta_k. \end{aligned} \tag{2.3}$$

The optimality of our algorithm clearly follows from (2.3). Observe that in our case $k=2$.

3. Algorithm for $k=3$

Stage 1. This is parallel to that for the case $k=2$ in that a probability of $P_{(1)j}=\min(P_{1j}, P_{2j}, P_{3j})$ is assigned to the point (j,j,j) in $\mathcal{S}_1$ and P_{ij} is replaced by $\bar{P}_{ij}=P_{ij}-P_{(1)j}$ for each j, $1 \leq j \leq N$.

Stage 2. Assume without any loss of generality that $P_{11}=\min(P_{11}, P_{21}, P_{31})$, so that $\bar{P}_{11}=0$, and let $\bar{P}_{(2)j}=P_{(2)j}-P_{(1)j}$ be the second smallest entry in Column j. Assume further that it is possible to remove nonnegative reals $\delta_1, \delta_2, \ldots, \delta_N$ from $\bar{P}_{11}, \bar{P}_{12}, \ldots, \bar{P}_{1N}$ respectively without affecting the numerical values of the second smallest entries in the respective columns and that the δ's add up to $\bar{P}_{(2)1}$. (Note that in this case $\delta_1=0$.) Assign a probability δ_j to the point $(j, 1, 1)$, $j=1, 2, \ldots, N$, and replace $\bar{P}_{i1}$ by $\bar{\bar{P}}_{i1}=\bar{P}_{i1}-\bar{P}_{(2)1}$, $i=2, 3$, $\bar{P}_{1j}$ by $\bar{\bar{P}}_{1j}=\bar{P}_{1j}-\delta_j$, $1 \leq j \leq N$, and $\bar{\bar{P}}_{ij}=\bar{P}_{ij}$ for all other pairs (i,j). The object of this operation is to "zero out" the second smallest entry, $\bar{P}_{(2)1}$, from

Column 1 without affecting the second smallest entries in other columns. Thus after this operation Column 1 has at most one nonzero entry. Carry out analogous operations of zeroing out second smallest entries as far as possible for all the N columns. A systematic way of doing this would be to first zero out all columns of the form $(0, b, c)$, then all columns of the form $(a, 0, c)$, and finally all columns of the form $(a, b, 0)$. For example if Column 2 has the configuration $(a, b, 0)$, then this operation would involve assigning probabilities δ_k^* to points of the form $(2, 2, k)$, where $k=1, 3, \ldots, N$.

If in the configuration $\{P_{ij}^*\}$ that will finally emerge, each column has at most one nonzero entry then go to Stage 3, otherwise we are in Stage 2* described below.

Stage 3. The nonzero entries in $\{P_{ij}^*\}$ can be zeroed out through steps similar to Stage 2 for the case $k=2$ by putting the remaining probability on $\mathscr{S}_3$.

The constructed survey assigns a probability of θ_1 to $\mathscr{S}_1$, a probability of $\theta_2-\theta_1$ to $\mathscr{S}_2$ and thus a probability of $(1-\theta_2)$ to $\mathscr{S}_3$. Hence

$$Ev = 1\cdot\theta_1+2(\theta_2-\theta_1)+3(1-\theta_2) = 3-\theta_1-\theta_2 = \theta_3. \tag{3.1}$$

From (2.3) it follows that the suggested algorithm then does yield an optimal integrated survey.

Note that the possibility of carrying out the above construction entails that $\theta_2 \leq 1$.

Stage 2*. If the required operations get blocked during Stage 2, it will be because either all the nonzero entries in one of the rows are also the second smallest column entries in their respective columns, or that it is not possible to remove the stipulated amount of probability from this row without lowering the magnitude of the second smallest column entries. For example, if it is the first row which reaches this configuration during Stage 2, then either each nonzero entry in the first row is the second smallest column entry in the respective column, or zeroing out of the second smallest entry from any column of the form $(0, b, c)$ cannot be accomplished without affecting the magnitude of the second smallest entries in other columns; if so, then we can and do zero out a column of the form $(0, b, c)$ in such a way that all nonzero entries in the first row turn into second smallest column entries. From this stage go to Stage 3*.

Stage 3*. In Stage 3*, the positive entries in the configuration are zeroed out by distributing the probability mass suitably in $\mathscr{S}_2$. We assume without any loss of generality that the configuration at this stage has the appearance of Table 1, in which $a_{t+1} \leq b_{t+1}, \ldots, a_u \leq b_u$, $a_{u+1} \leq c_{u+1}, \ldots, a_N \leq c_N$.

It is important to note here that for reasons of notational simplicity we now denote the entries in any configuration by generic symbols a, b and c in the three rows respectively. Thus entries in a new configuration are not necessarily the same as that of the preceding tables from which they may have been derived, and their meanings depend very much on the context in which they are being used; the only exception to this rule is when a given entry has been zeroed out.

Table 1

0	0	...	0	0	0	...	0	0	...	0	a_{t+1}	...	a_u	a_{u+1}	...	a_N
b_1	b_2	...	b_r	0	0	...	0	b_{s+1}	...	b_t	b_{t+1}	...	b_u	0	...	0
0	0	...	0	c_{r+1}	c_{r+2}	...	c_s	c_{s+1}	...	c_t	0	...	0	c_{u+1}	...	c_N

Since $\Sigma_{t+1}^{N} a_j = \Sigma_1^t b_j + \Sigma_{t+1}^{u} b_j$ and $a_j \leqslant b_j$ for $j=t+1, \ldots, u$, we have

$$\sum_{u+1}^{N} a_j \geqslant \sum_{1}^{t} b_j. \tag{3.2}$$

Consequently we can choose nonnegative numbers $\delta_{u+1} \leqslant a_{u+1}, \ldots, \delta_N \leqslant a_N$ such that $\Sigma_{u+1}^{N} \delta_j = b_1$. Since $a_j \leqslant c_j$ for $j=u+1, \ldots, N$, it follows that $\delta_{u+1} \leqslant c_{u+1}, \ldots, \delta_N \leqslant c_N$. Now define $\mathcal{P}(X_1=j, X_2=1, X_3=j)=\delta_j$ for $j=u+1, \ldots, N$. Removing the probability mass $b_1 = \Sigma\, \delta_j$ from this assignment, zeroes out the first cell in the second row, the entries a_j in the first row are replaced by $\bar{a}_j = a_j - \delta_j$ and the entries in the third row are replaced by $\bar{c}_j = c_j - \delta_j$ for $j=u+1, \ldots, N$. Note that this operation leaves the structure of the configuration intact in that the new a_j's continue to be the second smallest entries in their respective columns. The new configuration now has the appearance of Table 2. The entries $b_2, \ldots, b_t$ in Table 2 are zeroed out in a similar fashion by assigning suitable probabilities to the cells (j, k, j) for $k=2, \ldots, r, s+1, \ldots, t$ and $j=u+1, \ldots, N$ and keeping the structure of the first row intact. After these operations, the configuration assumes the appearance of Table 3. The entries $c_{r+1}, \ldots, c_t$ in Table 3 are now zeroed out one-by-one in a similar fashion by assigning suitable probabilities to the cells (j, j, k) for $j=t+1, \ldots, u$ and $k=r+1, \ldots, t$. As a result of this operation, the configuration that emerges has the appearance of Table 4.

We now zero out the entry a_{t+1} in the first row of Table 4. We note that $\Sigma_{t+1}^{u} a_j = \Sigma_{u+1}^{N} (c_j - a_j)$. Therefore

$$a_{t+1} \leqslant \sum_{u+1}^{N} (c_j - a_j). \tag{3.3}$$

Consequently there exist nonnegative numbers $\delta_j \leqslant (c_j - a_j)$, $j=u+1, \ldots, N$, such that

$$a_{t+1} = \sum_{u+1}^{N} \delta_j. \tag{3.4}$$

Table 2

0	0	…	0	0	…	0	0	…	0	a_{t+1}	…	a_u	$\bar{a}_{u+1}$	…	$\bar{a}_N$
0	b_2	…	b_r	0	…	0	b_{s+1}	…	b_t	b_{t+1}	…	b_u	0	…	0
0	0	…	0	c_{r+1}	…	c_s	c_{s+1}	…	c_t	0	…	0	$\bar{c}_{u+1}$	…	$\bar{c}_N$

Table 3

0	0	…	0	0	…	0	a_{t+1}	…	a_u	a_{u+1}	…	a_N
0	0	…	0	0	…	0	b_{t+1}	…	b_u	0	…	0
0	0	…	0	c_{r+1}	…	c_t	0	…	0	c_{u+1}	…	c_N

Table 4

0	…	0	a_{t+1}	…	a_u	a_{u+1}	…	a_N
0	…	0	b_{t+1}	…	b_u	0	…	0
0	…	0	0	…	0	c_{u+1}	…	c_N

Now define $\mathscr{P}(X_1=t+1,\ X_2=t+1,\ X_3=j)=\delta_j$ for $j=u+1, \ldots, N$. Removing these probabilities leads to the configuration given by Table 5 in which $\bar{b}_{t+1}=b_{t+1}-a_{t+1}$ and $\bar{c}_j=c_j-\delta_j$ for $j=u+1, \ldots, N$.

Table 5 is easily seen to have the structure of Table 3. The entry $\bar{b}_{t+1}$ of Table 5 can be zeroed in a manner similar to that of Table 3. The configuration that emerges now is like that of Table 4 in which the $(t+1)$st column has now been zeroed out. We repeat this procedure until all the entries in Columns $(t+1)$ through u have been zeroed out. The final configuration that now emerges must necessarily be of the form of Table 6. By assumption the three row sums of Table 6 are all equal and so all the a's and c's in the final configuration must be zero.

It is easily seen that in this case, our algorithm assigns a probability of θ_1 to $\mathscr{S}_1$, a probability of $1-\theta_1$ to $\mathscr{S}_2$ and zero probability to $\mathscr{S}_3$. Hence

$$Ev = 1 \cdot \theta_1 + 2(1-\theta_1) = 2-\theta_1. \tag{3.5}$$

Proof of optimality. To prove the optimality in this last case, we observe that for any integrated survey

$$Ev \geqslant \mathscr{P}(\mathscr{S}_1) + 2[1-\mathscr{P}(\mathscr{S}_1)] = 2-\mathscr{P}(\mathscr{S}_1). \tag{3.6}$$

Since $\mathscr{P}(\mathscr{S}_1) \leqslant \theta_1$, it follows that we must always have

$$Ev \geqslant 2-\theta_1. \tag{3.7}$$

Since for our algorithm the equality in (3.7) is attained, it is necessarily optimal.

It is worth noting that the quantity θ_2 plays a crucial role in the preceding algorithm. The algorithm shows that if $\theta_2>1$ then the given algorithm must get blocked at Stage 2 so that the achieved optimal solution must have its support in $\mathscr{S}_1 \cup \mathscr{S}_2$.

4. Remark

It is perhaps natural to speculate that the optimality of the integrated surveys worked out in Section 3 would go through if apart from possibly a constant term, the cost instead of being proportional to the number of distinct units, is actually a monotonic increasing function thereof. The following counterexample for $N=3$ sets at rest any such speculation.

Table 5

0	…	0	0	a_{t+2}	…	a_u	a_{u+1}	…	a_N
0	…	0	$\bar{b}_{t+1}$	b_{t+2}	…	b_u	0	…	0
0	…	0	0	0	…	0	$\bar{c}_{u+1}$	…	$\bar{c}_N$

Table 6

0	…	0	a_{u+1}	…	a_N
0	…	0	0	…	0
0	…	0	c_{u+1}	…	c_N

Example 4.1. Let $N=3$ and consider the problem of integrating three surveys. Let $C(\nu)$ denote the cost of selecting an integrated sample with ν distinct units and suppose that $C(1)=1$, $C(2)=2$ and $C(3)=10$. Further suppose that the marginal probabilities of selection for the three surveys are given by the entries in Table 7.

An optimally integrated survey based on the algorithm of Section 3 is given in Table 8. The integrated survey of Table 8 yields the following distribution for the number of distinct units in the sample: $\mathcal{P}(\nu=1)=0.6$, $\mathcal{P}(\nu=2)=0.3$, $\mathcal{P}(\nu=3)=0.1$ so that $EC(\nu)=2.2$. On the other hand the integrated survey given in Table 9 assigns all the probability exclusively to $\mathcal{S}_2$ and therefore has a lower expected cost $EC(\nu)=2.0$.

5. Integrating four or more surveys

The algorithm described in this paper cannot be extended to four or more surveys in a routine manner. Consider the marginal probabilities of selection for four units in four surveys as given by the entries in Table 10. Applying a routine extension of the algorithm in Section 3, one arrives at the following plan for an integrated survey

Table 7. *Stochastic matrix for three surveys*

i	P_{i1}	P_{i2}	P_{i3}
1	.2	.3	.5
2	.5	.2	.3
3	.3	.5	.2

Table 8. *Plan for an optimally integrated survey (Section 3): values of* P_{ijk}

	P_{1jk}			P_{2jk}			P_{3jk}		
(j, k)	1	2	3	1	2	3	1	2	3
1	.2	0	0	.1	0	0	0	0	0
2	0	0	0	0	.2	0	0	0	0
3	0	0	0	.1	.1	0	.1	0	.2

Table 9. *Plan for an alternative integrated survey: values of* P_{ijk}

	$150P_{1jk}$			$150P_{2jk}$			$150P_{3jk}$		
(j, k)	1	2	3	1	2	3	1	2	3
1	0	9	0	17	5	0	5	0	9
2	2	0	0	17	0	2	0	9	0
3	17	0	2	0	17	17	17	5	0

Table 10. *Stochastic matrix for four surveys*

i	P_{i1}	P_{i2}	P_{i3}	P_{i4}
1	1/3	0	1/3	1/3
2	1	0	0	0
3	1/3	1/3	0	1/3
4	1/3	1/3	1/3	0

$$p_{1111} = 1/3, \quad p_{3122} = 1/3, \quad p_{4143} = 1/3$$

with an expected number of distinct units $E(v) = 1/3$. The following alternative plan however assigns probability 1 to sample points with two distinct units

$$p_{1122} = 1/3, \quad p_{3113} = 1/3, \quad p_{4141} = 7/3$$

which points out the nonoptimality of the plan arrived via Section 3. One also faces 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 to eliminate the second largest entry in each column.

Acknowledgements

The authors would like to thank a referee for a very careful reading of our paper which led to considerable improvements. The authors are grateful to Professor R. Chandrasekaran of the University of Texas at Dallas for providing the example discussed in Table 10, and for the corresponding observations made in this section.

This research was partially supported by the NSF grant INT-8020450 during P. K. Pathak's sabbatical at the Indian Statistical Institute, New Delhi.

Dedication

This paper is dedicated to Professor D. Basu on the occasion of his sixtieth birthday.

References

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

Lahiri, D. B. (1954). Technical paper on some aspects of the development of the sample design. *Sankhya* **14**, 264–316.

Maczynski, M. J. & Pathak, P. K. (1980). Integration of surveys. *Scand. J. Statist.* **7**, 130–138.

Raj, D. (1957). On the method of overlapping maps in sample surveys. *Sankhya* **17**, 89–98.

Received January 1982, in final form May 1984

Pramod Pathak
Department of Mathematics and Statistics
The University of New Mexico
Albuquerque, New Mexico 87131, USA