The Korean Communications
in Statistics Vol. 8, No. 3, 2001
pp. 719-729

Maximizing the Overlap of Sample Units for Two Stratified Designs by Linear Programming

Jea-Bok Ryu[1], Sun-Woong Kim[2]

Abstract

Overlap Maximization is a sampling technique to reduce survey costs and costs associated with the survey. It was first studied by Keyfitz(1951). Ernst(1998) presented a remarkable procedure for maximizing the overlap when the sampling units can be selected for two identical stratified designs simultaneously. But the approach involves mimicking the behaviour of nonlinear function by linear function and so it is less direct, even though the stratification problem for the overlap corresponds directly to the linear programming problem. Furthermore, it uses the controlled selection algorithm that repeatedly needs zero-restricted controlled roundings, which are solutions of capacitated transportation problems. In this paper we suggest a comparatively simple procedure to use linear programming in order to maximize the overlap. We show how this procedure can be implemented practically.

Keywords : Overlap maximization, Controlled selection algorithm, Zero-restricted controlled rounding, Method of cumulative sums, Linear programming.

1. 서 론

서로 다른 두 개의 층화설계에서 조사관련 비용을 줄이기 위해 추출단위의 중복을 최대로 하는 방법이 널리 사용되어 왔다. 그런데 추출단위의 중복최대화(overlap maximization) 방법에 관한 연구는 초기 층화설계와 이 설계에 대한 재층화설계인 경우에 주로 다루어진다.

Keyfitz(1951)는 초기 층과 새로운 층은 서로 일치하고 추출단위들의 추출확률이 다를 때 층당 1개의 추출단위를 중복 추출하는 방법을 제시하였다. Kish와 Scott(1971)는 두 개 층이 서로 다르고 추출단위들의 추출확률도 다른 경우에 층당 1개의 추출단위를 중복 추출하는 방법을 개발하였다. Causey, Cox 그리고 Ernst(1985)와 Ernst와 Ikeda(1995)는 두 개 층과 추출확률이 다를 때 수송이론(transportation theory)을 사용하여 중복 추출을 최대화시키는 방법을 제안하였다. Causey, Cox 그리고 Ernst의 방법은 Ernst와 Ikeda의 방법에 비해 추출단위의 중복 추출 확률을 높일 수

1) Professor, Statistics, Division of Natural Science, Chongju University, Chongju, Korea (360-764)
E-mail : jbryu@chongju.ac.kr
2) Lecturer, Department of Statistics, Dongguk University, Seoul, Korea (100-715)
E-mail : kswyh@chollian.net

있다. 반면에 Ernst와 Ikeda의 방법은 수송이론을 적용할 때 변수의 개수 즉 모든 가능한 표본의 개수가 매우 큰 경우에도 쉽게 표본추출계획을 얻을 수 있다.

한편 Ernst(1996)는 추출단위들과 추출확률이 다른 2개의 층화표본설계에서 동시에 표본을 중복추출하는 방법을 제시하였다. Kish와 Scott(1971), Causey, Cox 그리고 Ernst(1985)와 Ernst와 Ikeda(1995) 등이 제시한 방법과 Ernst(1996)가 제안한 방법은 추출단위와 추출확률이 다른 경우에 사용된다는 점은 같지만 전자는 초기 설계와 재설계 간의 중복 최대화를 위해 사용되는 반면 후자는 중복을 최대화시키고자 하는 서로 다른 층화설계 간에 사용된다. 그러나 Ernst(1996)의 방법은 실행 절차가 복잡할 뿐만 아니라 중복단위의 추출확률을 최대화시키는 표본추출계획을 세우기 어렵다. 김종호, 류제복, 김선웅(1999)은 Ernst(1996)의 방법에 비해 표본추출계획을 세우는 절차를 보다 단순화시키면서 중복 추출을 최대화할 수 있는 방법을 제시하였다. 그런데 이 2가지 방법은 모두 층당 1개의 추출단위만을 중복 추출하는 경우에만 사용할 수 있으므로 실제 조사시에 사용하는 데에는 한계가 있다.

Ernst(1998)는 Ernst(1996)의 방법에서 사용되는 것보다 제한된 표본설계 즉, 추출단위는 동일하되 추출확률은 서로 다른 2개의 층화표본설계에서 1개 이상의 추출단위들을 중복 추출할 수 있는 방법을 제안하였다. 류제복과 김선웅(2000)은 Ernst(1998)의 방법에서 사용되는 동일한 표본설계 하에서 선형계획법(linear programming)을 사용하여 추출단위의 중복추출을 최대화시킬 수 있는 표본추출계획을 세우는 방법을 제시하였다.

본 논문에서는 Ernst(1998) 방법을 사용할 때 발생하는 몇 가지 문제점들을 개선시키면서 2개의 층화 설계간에 추출단위들이 동일하나 그 추출확률이 다른 경우와 보다 일반적인 경우로서 추출단위와 추출확률이 모두 다른 경우에도 1개 이상의 추출단위들의 중복 추출을 최대화할 수 있는 선형계획법을 이용한 방법을 제안한다.

2. 층화설계간 동시중복표본추출계획

Ernst(1998)가 제안한 것으로서 추출단위는 동일하나 추출확률이 다른 경우 실제 조사에 사용될 수 있는 새로운 표본설계방법으로서 다음과 같은 표본추출문제를 고려한다.

동일한 층들로 구성되는 표본설계 D_1과 D_2로부터 표본단위들을 추출한다고 하자. 층은 I로 표기하자. 또한 두 표본설계의 층 I에 있는 단위들은 동일하나 각 단위들의 추출확률은 서로 다르다고 하자. 이때 아래의 조건들을 만족시키도록 두 표본설계의 층 I로부터 동시에 표본단위들을 추출한다. 이를 위한 표본추출계획을 동시중복표본추출계획이라 하며 표본단위추출은 층마다 독립적으로 이루어진다.

(i) 표본설계 $D_j,\ j=1,2$ 각각에 대해 층 I로부터 추출되는 표본단위들의 개수 n_j는 미리 결정된다.

(ii) 표본설계 D_j의 층 I의 i번째 표본단위는 정해진 확률 π_{ij}로 추출된다.

(iii) 표본설계 $D_j,\ j=1,2$에 대해 공통인 표본단위들의 개수의 기대값은 최대화된다.

동시중복표본추출계획을 세우기 위해서는 실수값들로 구성되는 $N\times 4$ 배열 $S=(s_{ij})$을 일차적으로 얻어야 하는데 이는 식(2.1)-(2.3)으로부터 구해진다. 여기서 N은 층 I에 있는 추출단위들의 개수이고 i는 층 I의 i번째 추출단위에 대응되는 행을 의미한다. 그리고 열 $j=1,2,3,4$

에서 1과 2는 각각 표본설계 D_1과 표본설계 D_2, 3은 D_1과 D_2의 중복표본설계, 4는 이들을 제외한 나머지 표본설계에 대응하는 열을 의미한다. 또한 각 행과 열에 대한 주변을 얻을 수 있으므로 행과 열이 각각 1개씩 추가되어 $(N+1)\times 5$ 배열 $S=(s_{ij})$을 얻는다.

$$s_{i3} = \min\{\pi_{i1}, \pi_{i2}\} \tag{2.1}$$

$$s_{ij} = \pi_{ij} - s_{i3}, \quad j=1, 2 \tag{2.2}$$

$$s_{i4} = 1 - \sum_{j=1}^{3} s_{ij} \tag{2.3}$$

위의 식들에서 s_{i3}는 i번째 추출단위가 표본설계 D_1과 D_2의 공통인 표본단위가 될 확률, s_{ij}는 i번째 추출단위가 표본설계 D_j의 표본단위가 될 확률, s_{i4}는 i번째 추출단위가 두 표본설계 D_1와 D_2에서 모두 표본단위로 뽑히지 않을 확률을 각각 나타낸다.

이렇게 얻어진 $S=(s_{ij})$에 대한 동시중복표본추출계획은 Causey, Cox 그리고 Ernst(1985)가 제안한 관리적선정(controlled selection) 알고리즘을 이용하여 얻을 수 있다.

알고리즘 상에서 두 종류의 배열과 추출확률이 얻어진다. 먼저 실수값들의 배열 $A_{k+1}=(a_{ij(k+1)})$, $k=1,\cdots,l$는 $(N+1)\times 5$ 행렬로 각 셀의 값 $a_{ij(k+1)}$는 식(2.4)로부터 얻어진다. 단 $A_1=(a_{ij1})=S=(s_{ij})$이다.

$$a_{ij(k+1)} = m_{ijk} + (a_{ijk} - m_{ijk})/d_k \tag{2.4}$$

여기서 m_{ijk}는 정수이고, 배열 $M_k=(m_{ijk})$는 알고리즘 상에서 생성되는 각 셀의 값으로 Cox와 Ernst(1982)가 제시한 다소 복잡한 수송문제에 관한 이론에 적용시켜 얻어지는 배열 $A_1, A_2, \cdots, A_l$에 관한 각각의 해이다. 배열 M_k를 배열 A_k에 대한 영점제한컨트롤라운딩(zero-restricted controlled rounding)이라 한다. 이때 d_k는 식(2.5)와 같으며 l은 알고리즘 상에서 얻어지는 배열 $A_k=(a_{ijk})$ 또는 $M_k=(m_{ijk})$의 총 개수이고 $d_k=0$일 때 알고리즘이 종결된다.

$$d_k = \max\{\,|m_{ijk} - a_{ijk}| : i=1,\cdots,N+1,\ j=1,\cdots,5\} \tag{2.5}$$

그리고 배열 $M_k=(m_{ijk})$의 추출확률 p_k는 알고리즘 상에서 식(2.6)으로부터 얻어진다.

$$\begin{aligned} p_k &= 1-d_k, \quad k=1 \\ &= \left(1-\sum_{i=1}^{k-1} p_i\right)(1-d_k), \quad k>1 \end{aligned} \tag{2.6}$$

동시중복표본추출계획은 배열 $M_k=(m_{ijk})$와 p_k를 사용하며 식(2.7)과 (2.8)를 만족해야 한다.

$$\sum_{k=1}^{l} p_k = 1 \tag{2.7}$$

$$\sum_{k=1}^{l} m_{ijk}\, p_k = s_{ij}, \quad i=1,\cdots,N+1, \quad j=1,\cdots,5 \tag{2.8}$$

동시중복표본추출계획은 관리적선정 알고리즘을 이용하여 얻은 확률추출법으로서 식(2.7)과

(2.8)를 만족시키므로 원래의 배열 $S=(s_{ij})$에 있는 각 셀의 값을 유지할 수 있다. 그러나 이 방법은 실행 과정에 있어 다음과 같은 문제점들을 가지고 있다.

첫째, Ernst는 배열 $S=(s_{ij})$를 얻기 위해서 다음과 같은 방법을 이용하였다. 먼저 식(2.1)의 s_{i3}는 표본단위 i가 두 표본설계 D_1과 D_2로부터 공통으로 추출될 확률이 π_{i1} 또는 π_{i2}을 초과할 수 없다는 것으로부터 단순히 얻어진다. 그리고 표본단위 i가 표본설계 D_j에서만 표본으로 추출될 확률인 식(2.2)는 원래의 추출확률인 π_{ij}로부터 표본단위 i가 D_1과 D_2의 공통인 표본단위가 될 확률인 s_{i3}을 빼줌으로서 얻어진다. 또한 표본단위 i가 D_1과 D_2에서 모두 표본으로 뽑히지 않을 확률인 식(2.3)은 앞의 두 가지 경우를 제외한 나머지 경우의 확률이므로 모든 경우의 확률의 합인 1로부터 다른 경우의 확률을 빼줌으로서 얻어진다.

그런데 한 가지 유의해야 할 것은 배열 $S=(s_{ij})$에 관하여 최종적으로 얻어지는 동시중복표본추출계획은 $M_k=(m_{ijk})$와 p_k로부터 각 표본설계 D_1과 D_2에 대한 표본추출단위들을 동시에 결정하기 위해서 누적합방법(method of cumulative sums)을 사용한다. 따라서 표본추출계획을 세우기 위한 1차 과정의 배열 $S=(s_{ij})$도 앞서 설명한 방법에 의해 얻어지기보다는 누적합방법으로부터 얻어져야 할 것이다. Ernst의 방법은 두 표본설계의 층 I에 있는 단위들이 반드시 동일할 때에만 사용할 수 있는 반면 누적합방법은 층 I에 단위들이 동일하지 않은 경우에도 배열 $S=(s_{ij})$를 얻을 수 있다.

둘째, 관리적선정 알고리즘 상에서 식(2.4)에 의해 얻어지는 배열 A_k에 대해서 식(2.9)가 항상 성립한다. 또한 수송이론을 적용시켜 얻은 배열 A_k의 해인 영점제한컨트롤라운딩 배열 M_k에 대해서도 식(2.10)이 성립한다.

$$a_{(N+1)jk}+a_{(N+1)3k}=n_j,\ \ j=1,2 \tag{2.9}$$

$$m_{(N+1)jk}+m_{(N+1)3k}=n_j,\ \ j=1,2 \tag{2.10}$$

그런데 두 표본설계에 공통인 표본단위 수의 기대값인 $a_{(N+1)3k}$는 알고리즘이 반복되는 과정에서 식(2.4)에 의한 고의적인 조정으로 그 값이 점점 작아지므로 $a_{(N+1)3k}$의 컨트롤라운딩인 $m_{(N+1)3k}$도 작아지게 된다. 따라서 이 조정에 의해 원래의 $a_{(N+1)3k}$ 값이 왜곡될 수 있으므로 $m_{(N+1)3k}$의 값도 크게 영향을 받을 수 있다. 뿐만 아니라 식(2.4)에 의해 다른 m_{ijk}값들도 바뀌게 되므로 배열 $M_k=(m_{ijk})$의 추출확률인 p_k도 마찬가지로 영향을 받게 된다. 이러한 일련의 과정에 의해 동시중복표본추출계획이 얻어지는 것이다. 예를 들어, 배열 $S=(s_{ij})$에 대해서 식(2.11)과 같은 척도 $D(m_{ijk},s_{ij})$를 최소화시키는 컨트롤라운딩 즉, 최적컨트롤라운딩(optimal controlled rounding)이 여러 개 존재할 경우에 이 중 일부가 알고리즘이 반복되는 과정에서 전혀 고려되지 않을 수 있다.

$$D(m_{ijk},s_{ij})=\left[\sum_{i=1}^{N}\sum_{j=1}^{4}(m_{ijk}-s_{ij})^{2p}\right]^{1/2p},\ \ k=1,\cdots,l \tag{2.11}$$

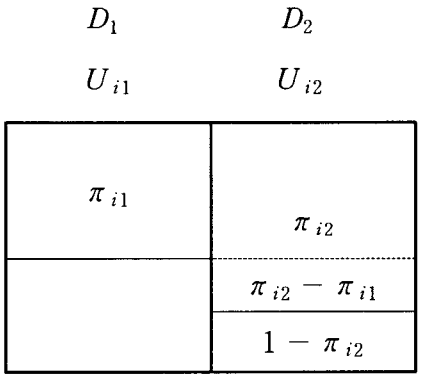
여기서 p는 정수이고 $1\le p<\infty$이다.

셋째, Cox와 Ernst(1982)가 제안한 수송이론을 적용시켜 얻은 배열 A_k에 대한 해인 영점제한 컨트롤라운딩 배열 M_k를 얻는 과정에서 비선형 함수인 식(2.11)을 선형함수로 나타내는 간접적인 방법을 사용하였다. 또한 알고리즘 상에서 배열 A_k는 1개의 수송문제이므로 수송계획법을 사용하여 A_k, $k = 1, \cdots, l$에 대한 해를 구해야한다. 즉, l개의 수송문제에 대한 해를 얻기 위하여 수송계획법을 l번 사용해야 한다. 이때 l의 상한은 $3N$이다.

3. 선형계획법을 이용한 동시중복표본추출 절차

2절에서 다룬 Ernst(1998) 방법의 문제점들을 개선할 수 있는 다음과 같은 절차를 제시한다.

우선 첫 번째로 언급한 문제를 해결하기 위하여 <그림 1>을 이용한 누적합방법을 이용할 수 있다. U_{i1}과 U_{i2}는 각각 층화설계 D_1과 D_2의 층 I에 있는 추출단위로서 서로 동일한 단위이다.



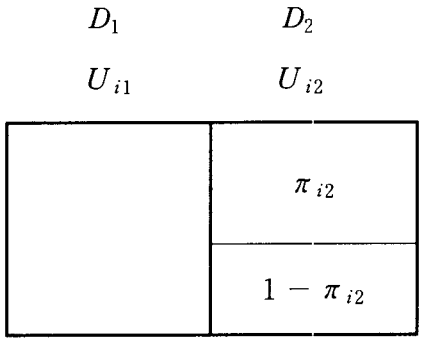
<그림 1>

동시중복표본추출계획을 세우기 위한 첫 번째 과정으로서 $(N+1) \times 5$ 배열인 $S = (s_{ij})$를 세우기 위해 다음과 같은 사항들을 고려한다. 단 π_{i1}과 π_{i2}가 소수 첫째 자리의 수인 경우 1~10의 난수를 사용하며 소수 둘째 자리의 수인 경우 1~100의 난수를 사용한다. 만약 π_{i1}과 π_{i2}가 소수 첫째 자리의 수라면 1~10의 난수를 발생시켜 $\pi_{i1} \times 10$보다 작거나 같은 난수가 얻어지면 층 S에 있는 추출단위 U_{i1}과 U_{i2}는 표본설계 D_1과 D_2에서 공통인 표본단위로 사용한다. 마찬가지로 난수를 발생시켜 $\pi_{i1} \times 10$보다 크고 $\pi_{i2} \times 10$보다 작거나 같은 난수를 얻으면 표본설계 D_2에서만 추출단위 U_{i2}가 표본단위로 사용된다. 마지막으로 $\pi_{i2} \times 10$보다 크고 10이하인 난수가 발생되면 두 표본설계 D_1과 D_2에서 모두 추출단위 U_{i1}과 U_{i2}를 표본단위로 사용하지 않는다.

이러한 절차에 의해서 추출단위 U_{i1}과 U_{i2}가 D_1과 D_2의 공통 표본단위가 될 확률은 π_{i1}, D_2에서만 추출단위 U_{i2}가 표본단위로 사용될 확률은 $\pi_{i2} - \pi_{i1}$, 두 표본설계에서 모두 추출단위 U_{i1}과 U_{i2}가 표본단위로 사용되지 않을 확률은 $1 - \pi_{i2}$이 된다. <그림 1>은

$\pi_{i1} \le \pi_{i2}$ 인 경우만을 다루고 있지만 $\pi_{i1} > \pi_{i2}$ 인 경우도 동일한 방법으로 사용될 수 있다.

또한 D_1 과 D_2 에서 대응되는 층이 서로 동일하지 않은 경우, 예를 들어 D_2 의 층에는 있으나 D_1 의 층에는 없는 단위의 경우 U_{i1} 에 관한 추출확률은 주어지지 않으므로 <그림 2>와 같이 나타낼 수 있다. 따라서 추출단위 D_1 과 D_2 의 공통 표본단위가 얻어질 확률은 0이고, D_2 에서만 U_{i2} 가 표본단위로 사용될 확률은 π_{i2}, D_1 에서는 물론이고 D_2 에서도 U_{i2} 가 표본단위로 사용되지 않을 확률은 $1 - \pi_{i2}$ 가 된다.



<그림 2>

다음 두 번째, 세 번째로 언급한 문제를 개선시키면서 앞의 누적합방법으로부터 얻은 배열 $S = (s_{ij})$ 에 대한 동시중복표본추출계획을 세우기 위하여 선형계획법을 이용할 수 있으며 이 방법은 다음과 같은 장점들을 갖는다.

우선 층화와 관련된 문제는 선형계획문제와 직결되므로 Ernst(1998)의 방법에서와 같이 다소 복잡한 수송문제 이론을 적용한 간접적인 방법을 사용하지 않고 선형계획법을 직접적으로 사용하여 해를 얻을 수 있다.

다음으로 선형계획법을 사용함으로써 각 셀의 값 a_{ijk} 에 대한 정수값 m_{ijk} 을 얻기 위해서 식 (2.4)에 의한 고의적인 조정을 피하면서 동시중복표본추출계획을 세울 수 있다.

선형계획법을 이용하기 위한 절차는 다음과 같다.

먼저 $S = (s_{ij})$ 에 대한 모든 가능한 영점제한컨트롤라운딩을 구한다. 이것을 구하는 과정이 복잡하지 않으므로 어떠한 프로그램 언어를 사용해서도 쉽게 결과를 얻을 수 있다. 이때 모든 가능한 컨트롤라운딩 집합을 R 이라 하고 각 컨트롤라운딩은 $N \times 4$ 배열 $R_v = (r_{ijv})$, $i = 1, \cdots, N$, $j = 1, \cdots, 4$, $v = 1, \cdots, q$ 로 표시하자. 여기서 R_v 는 두 층화설계에서 각 층의 표본배분을 나타내고 주변인 $i = N+1$, $j = 5$ 는 제외되며 q 는 모든 가능한 컨트롤라운딩의 개수이다. .

다음 각 컨트롤라운딩 $R_v = (r_{ijv})$ 의 가중치로서 식(3.1)의 $w(d_v)$ 를 사용한다.

$$w(d_v) = \max\{ | r_{ijv} - s_{ij} | : i = 1, \cdots, N,\ j = 1, \cdots, 4 \},\ v = 1, \cdots, q \tag{3.1}$$

여기서 $w(d_v)$ 는 식(3.2)의 $p = \infty$ 에 대하여 얻어지는 값이다.

$$D(r_{ijv}, s_{ij}) = \left[\sum_{i=1}^{N} \sum_{j=1}^{4} (r_{ijv} - s_{ij})^{2p} \right]^{1/2p} \tag{3.2}$$

이때 선형계획법을 사용하기 위하여 다음과 같은 목적함수 식(3.3)과 제약조건 식(3.4)를 둔다.

$$\phi = \sum_{R_v \in R} w(d_v)\, p(R_v) \tag{3.3}$$

$$\sum_{ij \in R_v,\, R_v \in R} p(R_v) = s_{ij}, \quad i = 1, \cdots, N, \quad j = 1, \cdots, 4, \quad v = 1, \cdots, q \tag{3.4}$$

식(3.3)과 식(3.4)를 살펴보면 모두 각 컨트롤라운딩 R_v의 추출확률인 $p(R_v)$에 관한 함수 형태임을 알 수 있다. 이때 각 R_v에 대한 가중치를 부여한 목적함수 ϕ를 최소화시키는 해인 $p(R_v)$는 정해진 제약조건인 식(3.4) 하에서 선형계획법을 이용하여 얻을 수 있다. 여기서 가중치 $w(d_v)$는 r_{ijv}와 s_{ij}에 대한 거리함수이므로 목적함수는 최소화되어야 한다. R_v와 $p(R_v)$는 층화표본설계 D_1과 D_2의 표본을 결정할 수 있는 동시중복표본추출계획이 된다.

그런데 각 컨트롤 라운딩 $R_v = (r_{ijv})$는 선형계획문제의 변수이므로 모든 가능한 컨트롤라운딩의 개수가 증가할수록 해를 얻는데 어려움이 있다. 그러나 실제로는 $S = (s_{ij})$에 대한 모든 가능한 영점제한컨트롤라운딩은 식(3.5)가 만족되어야 하므로 수송계획법을 사용하는 Ernst(1998) 방법에서 변수의 개수인 $4N$을 크게 넘지 않는다.

$$r_{(N+1)jk} + r_{(N+1)3k} = n_j, \quad j = 1, 2 \tag{3.5}$$

Ernst방법은 해를 얻기 위하여 수송계획법을 l번 사용하였으나 이 방법은 선형계획법을 1회만 사용하므로 반복절차의 회수를 크게 줄일 수 있다. 또한 각 컨트롤라운딩의 가중치 $w(d_v)$를 반영한 목적함수 식(3.3)을 사용하여 해를 얻음으로서 Ernst방법과 같이 해를 얻기 위하여 알고리즘 상에서 두 표본설계에 공통 표본단위들의 개수의 기대값인 $a_{(N+1)3k}$을 고의적으로 조정하지 않아도 된다.

4. 예 제

본 절에서는 2개의 층화 설계간에 추출단위들은 동일하나 그 추출확률이 다른 경우인 [예제1]과 추출단위와 추출확률이 모두 다른 경우인 [예제2]를 통하여 본 논문에서 제안한 선형계획법을 이용하여 동시중복표본추출계획을 얻는 절차를 설명한다.

[예제 1](Ernst(1998))

다음 <표 1>은 층화표본설계 D_1과 D_2의 층 I의 5개 추출단위는 동일하며 각 단위의 추출확률은 서로 다르다. <표 1>의 추출확률을 갖는 표본설계 D_1과 D_2에 대해 <그림 1>을 이용한 누적합방법을 사용하여 <표 2>와 같은 6×5 배열인 $S = (s_{ij})$을 얻는다. 이때 식(3.5)를 만족시키는 모든 가능한 컨트롤라운딩인 5×4 배열의 개수는 24개로서 이들이 선형계획문제의 변수가 된다. 수송계획법을 사용한 Ernst(1998)의 방법을 사용할 때는 변수의 개수가 $4N = 20$이므로 변수의 개수는 크게 차이가 나지는 않는다. Ernst(1998)의 방법은 해를 얻기 위해 수송계획법

을 4회 사용하였지만 본 연구에서 제안한 방법에서는 선형계획법을 단지 1회 사용하여 해를 얻을 수 있다.

<표 1> 각 추출단위의 추출확률

구 분	I				
	1	2	3	4	5
π_{I1}	.6	.4	.8	.6	.6
π_{I2}	.8	.4	.2	.4	.2

<표 2> $S = (s_{ij})$

0	.2	.6	.2	1
0	0	.4	.6	1
.6	0	.2	.2	1
.2	0	.4	.4	1
.4	0	.2	.4	1
1.2	.2	1.8	1.8	5

선형계획법을 이용하여 제약조건 식(3.4) 하에서 목적함수인 식(3.3)을 최소화시키는 해를 얻으면 <표 3>과 같은 동시중복표본추출계획이 된다. 본 논문에서는 SAS/OR의 선형계획절차(LP procedure)를 사용하였다.

<표 3> 동시중복표본추출계획

R_v	0 0 1 0	0 0 0 1	0 0 1 0	0 1 0 0
	0 0 0 1	0 0 1 0	0 0 0 1	0 0 1 0
	1 0 0 0	0 0 0 1	0 0 1 0	1 0 0 0
	0 0 1 0	1 0 0 0	0 0 0 1	0 0 0 1
	0 0 0 1	0 0 1 0	1 0 0 0	1 0 0 0
$P(R_v)$	.4	.2	.2	.2

각 R_v에서 i번째 행은 i번째 추출단위를 나타내며 1열은 D_1에서만 단위들이 추출되는 경우, 2열은 D_2에서만 단위들이 추출되는 경우, 3열은 D_1과 D_2에서 동시에 단위들이 추출되는 경우, 4열은 D_1과 D_2에서 모두 단위들이 추출되지 않는 경우를 의미한다.

<표 3>의 동시중복표본추출계획으로부터 누적합방법을 사용하여 추출된 R_v는 층화표본설계 D_1과 D_2의 층 S에 대한 표본 배분이 결정된다. 예를 들어 R_1이 추출될 때 D_1의 층 I에서는 첫 번째, 세 번째, 네 번째 추출단위가 표본으로 배정되고 D_2의 층 I에서는 첫 번째, 네 번째 추출단위가 표본으로 배정된다. 두 번째와 다섯 번째 추출단위는 D_1과 D_2 모두에서 표본단위로 사용되지 않는다. 따라서 D_1과 D_2의 층 I의 공통표본은 첫 번째, 네 번째 추출단위이다.

참고적으로 <표 4>는 앞서 설명한 수송이론을 간접적으로 적용시키는 Ernst(1998)의 방법에

의해 얻어진 동시표본추출계획이다.

< 표 4 > Ernst(1998) 방법에 의한 동시중복표본추출계획

M_k	0 0 1 0	0 0 1 0	0 1 0 0	0 0 0 1
	0 0 0 1	0 0 1 0	0 0 1 0	0 0 0 1
	1 0 0 0	0 0 0 1	1 0 0 0	0 0 1 0
	0 0 1 0	0 0 0 1	0 0 0 1	1 0 0 0
	0 0 0 1	1 0 0 0	1 0 0 0	0 0 1 0
p_k	.4	.2	.2	.2

< 표 3 >과 비교해보면 표본의 추출확률은 동일하나 첫 번째 표본만 동일할 뿐 다른 표본들은 전혀 다르다는 것을 알 수 있다. 또한 단순히 < 표 3 >과 < 표 4 >를 보고 결과를 비교하기보다는 어떠한 방법이 보다 더 대표성이 있는 표본을 추출할 수 있는가에 관심을 두어야 할 것이다. 이런 점에서 선형계획법을 이용한 방법은 Ernst(1998)의 방법에서 사용된 식(2.4)에 의한 조정 방법에 의해 각 표본의 추출확률을 얻지 않고 각 표본의 추출확률에 보다 객관적인 가중치를 부여함으로써 정해진 조건하에서 각 표본에 대한 최적 추출확률을 얻을 수 있다.

[예제 2]

층화표본설계 D_1과 D_2에서 동시 중복 추출하고자하는 두 개 층의 일부 추출단위가 다르고 각 단위의 추출확률이 서로 다른 < 표5 >와 같은 예제를 다루어보자.

< 표 5 > 각 추출단위의 추출확률

구 분	I				
	1	2	3	4	5
π_{I1}	.7	.1	.8	.4	
π_{I2}	.6	.4	.6	.8	.6

< 그림 2 >를 이용한 누적합방법을 사용하여 < 표 6 >와 같은 6×5 배열인 $S = (s_{ij})$을 얻는다.

< 표 6 > $S = (s_{ij})$

.1	.0	.6	.3	1
0	.3	.1	.6	1
.2	0	.6	.2	1
0	.4	.4	.2	1
0	.6	0	.4	1
.3	1.3	1.7	1.7	5

모든 가능한 컨트롤라운딩의 개수는 22개이며 앞의 예제와 마찬가지로 본 논문에서 제안된 선형계획법을 이용하여 < 표 7 >과 같은 동시중복표본추출계획을 얻을 수 있다.

<표 7> 동시중복표본추출계획

R_v	0	0	0	1	0	0	1	0	0	0	1	0	0	0	1	0	1	0	0	0
	0	0	0	1	0	0	0	1	0	1	0	0	0	1	0	0	0	0	1	0
	0	0	1	0	0	0	1	0	1	0	0	0	0	0	0	1	0	0	0	1
	0	0	1	0	0	1	0	0	0	0	0	1	0	0	1	0	0	1	0	0
	0	1	0	0	0	0	0	1	0	1	0	0	0	0	0	1	0	1	0	0
$P(R_v)$	.3				.3				.2				.1				.1			

누적합방법을 사용하여 R_3가 추출된다고 할 때 D_1에서는 첫 번째, 세 번째 추출단위가 표본이 되며 D_2에서는 첫 번째, 두 번째, 다섯 번째 추출단위가 표본이 된다. 따라서 첫 번째 추출단위는 D_1과 D_2의 공통표본이 된다. 또한 네 번째 추출단위는 D_1과 D_2 모두에서 표본단위로 사용되지 않는다.

5. 결 론

본 논문에서는 선형계획법을 이용하여 두 층화표본설계에서 추출단위가 동일하거나 동일하지 않은 경우 중복 표본추출계획을 세울 수 있는 방법을 제안하였다. 이러한 동시중복표본추출계획은 초기 설계와 재 설계에서의 중복 추출단위의 사용을 최대화하기 위해서 뿐만 아니라 서로 다른 조사를 목적으로 하는 두 층화표본설계간에도 추출단위의 중복을 최대화함으로서 조사 관련 비용을 줄일 수 있다. 또한 조사원의 편의를 제공할 수 있으므로 실사시 효과적으로 사용할 수 있다.

특히 본 논문에서 중점적으로 다룬 선형계획법을 이용한 동시중복표본추출계획은 수송문제 알고리즘을 사용한 Ernst(1998)의 방법 보다 간단하고 쉽게 해를 얻을 수 있으며 모든 가능한 컨트롤라운딩들에 대한 각 가중치를 사용함으로서 보다 더 대표성있는 표본을 얻을 수 있다.

그러나 선형계획법을 이용한 방법은 표본추출계획을 세우는데 있어 변수의 수가 크게 증가하는 경우 실행에 어려움이 따른다. Ernst(1998)는 Ernst와 Ikeda(1995)의 크기축소수송계획알고리즘(reduced-size transportation algorithm)을 사용하여 변수의 개수를 조정하는 방법을 제시하고는 있으나 원래의 변수에 관한 최적 해를 얻을 수는 없다.

참고문헌

[1] 김종호, 류제복, 김선웅(1999). 추출단위의 중복 사용을 최대화하기 위한 표본추출계획, 한국통계학회 추계학술발표회논문집, 49-54.

[2] 류제복, 김선웅(2000). 층화표본설계에서의 중복 추출 방법, *한국통계학회 춘계학술발표회논문집*, 7-12.

[3] Causey, B. D., Cox, L. H., and Ernst, L. R. (1985). Applications of transportation theory to statistical problems, *Journal of the American Statistical Association*, Vol. 80, 903-909.

[4] Cox, L. H. and Ernst, L. R. (1982). Controlled Rounding, *INFOR*, Vol. 20, 423-432.

[5] Keyfitz, N. (1951). Sampling with probabilities proportionate to size : Adjustment for changes in probabilities, *Journal of the American Statistical Association*, Vol. 46, 105-109.

[6] Kish, L. and Scott, A. (1971). Retaining units after changing strata and probabilities, *Journal of the American Statistical Association*, Vol. 66, 461-470.

[7] Ernst, L. R. (1986). Maximizing the overlap between surveys when information is incomplete, *European Journal of Operational Research*, Vol 27, 192-200.

[8] Ernst, L. R. (1996). Maximizing the overlap of sample units for two designs with simultaneous selection, *Journal of Official Statistics*, Vol. 12, No. 1, 33-45.

[9] Ernst, L. R. (1998). Maximizing and minimizing overlap when selecting a large number of units per stratum simultaneously for two designs, *Journal of Official Statistics*, Vol. 14, No. 3, 297-314.

[10] Ernst, L. R. and Ikeda, M. (1995). A reduced-size transportation algorithm for maximizing the overlap between surveys, *Survey Methodology*, Vol. 21, 147-157.

[11] SAS/OR User's Guide (1989), *SAS Institute Inc.*