# 다차원 층화에서 선형계획법을 이용한 표본배정 방법

최재혁[1], 남궁 평[2]


## 요 약

다차원층화에서 선형계획법을 이용한 표본배정 방법은 Winkler(1990, 2001), Sitter와 Skinner(1994, 2002)가 제안하였다. 이 방법들은 표본크기가 층 개수보다 크지 않는 경우에 공통적으로 선형계획법을 이용하여 표본배정을 실시하였다. 반복 비율 적합방법(IPF), 일반화 반복 비율 적합(GIFP), SS 방법을 통해 셀 값을 결정하고 선형계획법을 이용하여 표본의 배정확률을 통해 표본배정을 실시한다. 이 3가지 방법들로 표본을 배정하고 평균 및 분산추정량을 비교한다.


주요용어 : 선형계획법, 반복 비율 적합 방법(IPF), 일반화 반복 비율 적합(GIFP)

## 1. 서 론

모집단을 두 개 이상의 변수를 사용해서 층화하는 경우를 다차원 층화라고 한다. 모집단을 다차원 층화하는 경우에 추정오차를 구하기 위해서는 각 층에서 두 개 이상의 표본을 추출해야 한다. 즉, 모집단이 $R\times C$ 개의 층으로 구성되어 있다면 적어도 표본의 수가 $2\times R\times C$ 개 이상이어야 한다. 보통 다차원 층화 추출에서는 비율배정을 통해 각 층에 표본을 배정한다. 하지만 이것은 표본크기가 층화된 층의 개수보다 크기 않을 경우 기대되는 층의 크기가 매우 작을뿐더러 정수로 마무리되어야 하는 비율적 할당의 특성에 큰 방해가 될 것이다.

다차원 층화에서의 표본배정 방법으로 Winkler(1990, 2001)가 제안한 반복 비율 적합 방법(IPF 방법), 일반화 반복 비율 적합 방법(GIFP방법)과 Sitter와 Skiner(1994, 2002)가 제안한 SS 방법이 있는데 이 방법들은 모두 선형계획법(Linear Programing)을 이용한다. 이 방법들은 표본크기가 층화된 층의 개수보다 크지 않는 경우에 한에서 사용되는 표본 배정 방법이다.

본 논문에서는 이차원 층화에서 선형계획법을 이용하여 IPF 방법 및 GIFP 방법과 SS 방법을 알아보고 각 추정량들을 비교하고자 한다.

## 2. 선형계획법을 이용한 표본배정

선형계획법을 통해 표본을 배정하는 방법은 표본크기를 선형계획법을 통해 해로 구하는 것이 아니라 표본배정확률 $p(x)$ 를 해로 구하는 방법이다. 따라서 표본배정에 사용되는 목적함수 및 제약식은 다음과 같다.

① 목적함수 : $\min\{c_1p(x_1)+c_2p(x_2)+\cdots+c_np(x_n)\}$

1) 성균관대학교 경제학부 통계학과 박사과정, E-mail:leonash@skku.edu
2) 성균관대학교 경제학부 통계학과 교수, E-mail:namkung@skku.ac.kr

$$② \text{ 제약조건 : } \begin{cases} \sum_{x=1}^{n} p(x_i) = 1 \\ \sum_{x=1}^{n} m_i p(x_i) = n_i \\ p(x_i) \geq 0 \\ i = 1, 2, \cdots, n \end{cases}$$

여기서 $p(x_i)$는 $i$번째 표본배열이 선택될 확률이고 $c_i$는 주변합의 손실값이 된다. 주변합의 손실값이란 주변합이 반올림하면서 생기는 손실값을 의미한다. $m_i$는 표본배정을 위해 모든 배열들이고 $n_i$는 비율배정이나 우리가 적합하여 얻어진 배열이다. 위의 목적함수와 제약식을 통해 표본배열 확률이 결정되면 그 값을 통해 표본배정을 결정하는 것이 다차원 층화에서 선형계획법을 이용한 표본배정방법이 된다.

## 3. 반복비율적합 방법(IPF)과 일반화 반복 비율 적합 방법(GIFP)

### 3.1 칸 도수 추정

반복비율적합(Iterative Proportional Fitting)은 주변합을 통해 셀 크기를 추정한다. 먼저 $x_{ij}^s$를 $s$번째 반복의 표준화 칸 값이라고 하면 $x_{ij}^0 = x_{ij}$은 비율배정을 통한 셀 크기인 관측값이 된다. 그리고 $x_{i+}^*$와 $x_{+j}^*$는 표준화 주변합이다. 다음 식를 통해 반복적합하게 된다.

$$x_{ij}^s = \left(\frac{x_{ij}^{s-1}}{x_{i+}^{s-1}}\right) x_{i+}^* , \quad x_{ij}^{s+1} = \left(\frac{x_{ij}^s}{x_{+j}^s}\right) x_{+j}^*$$

$$| x_{ij}^{2k} - x_{ij}^{2k-2} | < \delta$$

여기서 $k$는 반복주기이다. 위 두 개의 식을 반복의 첫 번째 주기로 하여 허용오차 $\delta$에 이를 때까지 반복하게 된다. 이러한 주변합에 의한 반복계산법은 정확도 $\delta$에 이를 때까지 정확한 값에 수렴하게 된다. 그 수렴된 값을 통해 선형계획법을 이용, 표본배정을 실시하게 된다.

GIFP(Generalized Iterative fitting Procedure) 방법은 분류적 적합으로 Dykstra'의 일반화 반복 비율 적합(GIFP)에 의해 실현된다. 그 적합 과정은 다음과 같다.(L.Dykstra, 1985)

1. $s_{11} = r$, $p_{11} = \pi_1(s_{11})$ 라고 하고 $s_{12} = p_{11} = r(p_{11}/s_{11})$ 라고 생각하자. $s_{11}(k) = 0$ 이면 $p_{11}(k) = 0$ 이다. 0/0은 1이 된다.
2. $p_{12} = \pi_2(s_{12})$ 이고 $s_{13} = p_{12} = r(p_{11}/s_{11})(p_{12}/s_{12})$ 이다.
3. 계속해서 반복하면 $s_{1t} = p_{1(t-1)} = r(p_{11}/s_{11}) \cdots (p_{1(t-1)}/s_{1(t-1)})$ 이므로 $p_{1t} = \pi_t(s_{1t})$ 이고 $s_{21} = r(p_{12}/s_{12}) \cdots (p_{1t}/s_{1t})$ 이므로 $p_{2t} = \pi_{1t}/(p_{11}/s_{11})$ 이다.
4. $p_{21} = \pi_1(s_{21})$ 이라고 하면 $s_{22} = r(p_{21}/s_{21})(p_{13}/s_{13}) \cdots (p_{1t}/s_{1t})$ 이고 이 값은 $p_{22} = \pi_{21}/(p_{12}/s_{12})$ 의 값과 같다.
5. 계속해서 반복하면 일반적인 다음의 식을 얻을 수 있다.

$$s_{ni} = r\frac{p_{n1}}{s_{n1}} \cdots \frac{p_{n(i-1)}}{s_{n(1-1)}} \frac{p_{(n-1)(i+1)}}{s_{(n-1)(i+1)}} \cdots \frac{p_{(n-1)t}}{s_{(n-1)t}} = \begin{cases} p_{(n-1)t}\left(\frac{p_{(n-1)1}}{s_{(n-1)1}}\right)^{-1} & \text{if } i=1 \\ p_{n(i-1)}\left(\frac{p_{(n-1)i}}{s_{(n-1)i}}\right)^{-1} & \text{if } 2 \leq i \leq t \end{cases}$$

위의 방법을 통해 반복 계산하면 GIFP 방법에 의한 셀 값을 추정할 수 있다. 이렇게 얻어진 셀 값을 통해 선형계획법을 이용하여 표본크기를 결정할 수 있다. GIFP 방법은 각 셀 간의 변동량을 최소로 하는 방법이므로 상호 관련 있는 모집단 수에 의해 칸내(그룹내) 표본 배분을 실현할 수 있다. 추가적으로 층화를 필요로 하는 주변값이 선형관계인 경우에는 GIFP 방법은 고전적 IPF 방법과 동등한 결과를 얻는다.

### 3.2 표본배정 및 추정량

다변량 표본추출 방법에 정의되는 주변값 $m_j\,(j=1,\cdots,s)$ 의 집합이 있다고 하자. 각 $m_j$는 하나의 층화변수로 층화된 표본크기들 이다. 하나 이상의 층화변수로 정의되는 모집단 크기를 $N_i\,(i\in I)$ 라고 하면 IPF 방법과 GFIP 방법에 의해 이것은 $g_i\,(i\in I)$ 로 수렴하게 된다. 배열 $g_i\,(i\in I)$ 은 반드시 다음을 만족해야 한다.

$$\sum_{ij\in I_j} g_{ij} = m_j,\quad j=1,\cdots,s \quad \text{and} \quad g_i \le N_i,\quad i\in I$$

여기서 $I_j$는 주변값 $m_j\,(j=1,\cdots,s)$ 에 대한 $I$의 부분집합이다.

다음과 같이 양의 정수로 주변값 $m_j\,(j=1,\cdots,s)$ 을 가지고 양의 정수값 $p_k\,(k=1,\cdots,t)$ 가 존재하는 배열 $M_{ik}\,(i\in I,\ k=1,\cdots,t)$ 을 찾는 것이 목적이다.

$$\sum_{k=1}^{t} p_k = 1 \quad \text{and} \quad \sum_{k=1}^{t} p_k M_{ik} = g_i,\ \ i\in I$$

위의 식은 기대값 $g_i\,(i\in I)$ 를 가지는 양의 정수 행렬을 만들어내는 가능성 있는 구조를 산출한다. 이런 배열을 찾았다면 단지 확률적으로 비율크기 $p_1$를 가지는 배열 $M_{i1}\,(i\in I)$ 를 선택하고 나서 모든 $i\in I$인 칸 $i$에서 크기 $M_{i1}$인 표본을 선택한다.

IPF 방법 및 GIFP 방법으로 적합시킨 결과를 가지고 선형계획법을 이용하여 표본을 배정한다. 첫 번째 $M_0$는 IPF 방법 및 GIFP 방법에 의한 배열과 같고 계속해서 $M_k\,(k=1,\cdots,t)$ 을 계산하면 선형계획법 과정에 의해 양의 정수로만 이루어진 배열을 찾을 수 있다

IPF 방법 및 GIFP 방법에 대한 평균과 분산추정량을 구하는 방법은 포함확률을 이용하는 호르비츠와 톰슨(Horvitz-Thompson) 추정량을 이용한다(Winkler, 2001). IPF 방법 및 GIFP 방법에서 평균과 분산추정량은 다음과 같이 계산된다.

$$\bar{y} = \frac{1}{N}\sum_{i}^{I}\left(\frac{N_i}{M_0}\right)\sum_{j}^{n_i} y_{ij}$$

$$\widehat{Var}(\bar{y}) = \frac{1}{N^2}\sum_{i}^{I}\left(\frac{n_i}{M_0}-1\right)\left(\sum_{j}^{n_i} y_{ij}\right)^2 + \frac{1}{N^2}\sum_{i\neq i}^{I}\left(\frac{n_i}{M_0}-1\right)\left(\frac{n_{i'}}{M_0}-1\right)\left(\sum_{j}^{n_i} y_{ij}\right)\left(\sum_{j}^{n_{i'}} y_{i'j}\right)$$

여기서 $N$은 모집단 크기, $i$는 각 셀, $n_i$은 각 셀의 표본크기, $M_0$은 IPF방법 및 GIFP방법으로 적합시킨 셀 크기, $y_{ij}$는 $i$번째 셀에서 선택된 표본값이다.

따라서 모집단 총합 추정량은 $\hat{\tau} = \sum_{i}^{I}\left(\frac{N_i}{M_0}\right)\sum_{j}^{n_i} y_{ij}$이다.

## 4. Sitter와 Skinner 방법 : SS방법

### 4.1 칸 도수 추정

먼저 2차원 층화를 보면 $N$개의 모집단은 $R\times C$의 분할표로 분류되어 있다. 2단 추출 과정을 고려해 볼 때 첫 번째로 표본크기 $n_{ij}$를 특별한 랜덤 과정을 따르는 셀이라 하고 $s$를 $R\times C$ 배열( $n_{ij}$, $i=1,\cdots,R$, $j=1,\cdots,C$ )로 정의하고 가능한 배열 $S$ 집합에서 각 $s$의 확률 $p(s)$을 배정한다. $s$에 대한 $n_{ij}$의 독립을 강조하기 위해 $n_{ij}(s)$ 라고 쓴다. 두 번째로 $n_{ij}(s)$ 값들의 단순임의표본은 셀 $ij$으로부터 추출된다. 모든 배열의 집합 $S_n$가 되는 $S$를 제한한다. 여기서 $P_{ij}=N_{ij}/N$는 각 셀의 반응비율이다. 아래의 식이 제약조건이 된다.

$$\sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij}(s) = n$$

$$\sum_{s\in S_n} n_{ij}(s)p(s) = nP_{ij} \quad \text{for } i=1,\cdots,R,\ j=1,\cdots,C$$

따라서 다음과 같은 문제를 해결함으로써 표본 $s$의 '바람직한 설계'의 기대되는 부족을 최소화하는 표본설계 $p(s)$을 선택하는 것을 가정한다.

$$\underset{p\in P}{\text{mimimize}} \sum_{s\in S_n} w(s)p(s), \qquad 0\le p(s)\le 1 \quad \text{for all } s\in S_n$$

$w(s)$는 특정화된 표본 $s$에 대한 손실함수이고 $P$는 다음을 조건으로 하는 $S_n$에 대한 가능한 표본설계의 클래스인 일정한 제약조건에 종속되는 문제점이 있다.

위의 제한된 식은 $\sum_{s\in S_n} p(s)=1$를 뜻한다. 목적함수와 일정한 제약조건과 일정하지 않은 제약조건 모두 $p(s)$에 대해 선형관계이다. 반면 이러한 문제점은 미지의 $p(s)$, $s\in S_n$에 대해 선형계획법을 통해 직접 해결될 수 있다. 이 해결방법의 어려움은 $S_n$에서 원소의 수가 종종 매우 크게 되거나 또는 미지의 수가 매우 클 때 선형계획법에 의해 결과가 도출되기 어렵게 되는 계산상의 문제점이 있다는 것이다. 그러므로 $S_n$의 부분집합에 관심을 두고 제한하는 것은 바람직한 방법이다. 하나의 자연스러운 제약조건은 $n_{ij}(s)$가 $I_{ij}=[nP_{ij}]$와 같거나 가장 큰 정수가 $nP_{ij}$나 $I_{ij}+1$보다 적은 배열 $s$대해서만 오직 고려한다는 것이다.

$\tilde{n}_{ij}(s)=n_{ij}(s)-I_{ij}$, $r_{ij}=nP_{ij}-I_{ij}$라고 하면 제약조건은 다음과 같다.

$$\underset{p\in P}{\text{mimimize}} \sum_{s\in S_n} w(s)p(s)$$

$$\sum_{s\in S_{\tilde{n}}} \tilde{n}_{ij}(s)p(s) = r_{ij}$$

$$\sum_{s\in S_{\tilde{n}}} p(s)=1,\ 0\le p(s)\le 1 \quad \text{for all } s\in S_{\tilde{n}}$$

여기서 $\mathfrak{S}_{\tilde{n}}$는 모든 원소가 0 또는 1이고 원소의 합이 $\tilde{n} = n - \sum_{ij} I_{ij}$인 $R \times C$ 배열의 집합이다. 물론 모든 $I_{ij}$가 0이면 이것은 이전과 같은 문제점이 발생한다.

선형계획법의 컴퓨터 계산상 과정에서 가장 중요한 $\mathfrak{S}_{\tilde{n}}$의 원소의 수는 이제 $\left(\frac{RC}{\tilde{n}}\right)$이다. 이 숫자는 여전히 매우 클 수 있다. 그러나 조금 축소하면 손실함수 $w(s)$의 알맞은 선택으로 나타날 수 있다. 이러한 접근법을 통해 손실함수 $w(s)$을 선택하는 방법으로 선택된 표본 $s$가 다음과 같은 주변합 제한이 요구된다.

$$| n_{i\cdot}(s) - nP_{i\cdot} | < 1 \quad i = 1, \cdots, R$$

$$| n_{\cdot j}(s) - nP_{\cdot j} | < 1 \quad j = 1, \cdots, C$$

$$n_{i\cdot}(s) = \sum_j n_{ij}(s), \quad n_{\cdot j}(s) = \sum_i n_{ij}(s), \quad P_{i\cdot} = \sum_j P_{ij}, \quad P_{\cdot j} = \sum_i P_{ij}$$

이러한 제한조건은 $S_n$ 집합으로부터의 선택된 표본 $s$가 위와 같은 조건을 만족하지 못하면 배제시키거나 더 간단히 $w(s)$을 효과적으로 유한하게 적용함으로써 접근방법은 조절될 수 있다. 이러한 전통적인 접근이 가지고 있는 문제점은 최적화를 해결할 방법이 존재하지 않는다는 것이다. 그러나 다음과 같은 손실함수를 사용한다면 문제점이 다소 해결될 것이다.

$$w(s) = \sum_{i=1}^{R} (n_{i\cdot}(s) - nP_{i\cdot})^2 + \sum_{j=1}^{C} (n_{\cdot j}(s) - nP_{\cdot j})^2$$

최적 해결은 항상 충분히 큰 $S_n$ 집합 사이에 존재할 것이다. 실제로 유한한 집합 $S_n$를 위의 조건식에 따르게 하거나 혹은 이들의 부분집합인 표본으로 제한하는 것은 컴퓨터 계산상과 필요에 의해 집합을 확장하는데 유리함으로 작용할 것이다. 해결 방법을 찾을 때까지 위의 제한식에서 1에서 2로 바꿈으로써 얻어진다.

**4.2 표본배정 및 추정량(SS방법)**

선형계획법을 통한 문제 해결을 위한 손실함수는 다음과 같다.

$$w(s) = \sum_{i=1}^{R} (n_{i\cdot}(s) - nP_{i\cdot})^2 + \sum_{j=1}^{C} (n_{\cdot j}(s) - nP_{\cdot j})^2$$

$I_{ij}$ 값은 $n_{ij} = I_{ij} + \tilde{n}_{ij}(\tilde{s})$으로 계산된다. 그것은 $p(s) > 0$ 인 각 $s$가 정확한 실험설계의 주변값과 어울리는 주변값 $n_{i\cdot}(s)$과 $n_{\cdot j}(s)$를 가지는 해결방법으로 완성된다.

이와 같은 방법으로 가능한 표본 $S_n$은 각 셀에 $[nP_{ij}]$나 $[nP_{ij}]+1$의 값을 가지는 표본크기 $n_{ij}$을 갖게 된다. $[nP_{ij}]$은 $nP_{ij}$와 같거나 작은 정수값을 가진다.

$\tilde{n}_{ij} = n_{ij} - [nP_{ij}]$, $r_{ij} = nP_{ij} - [nP_{ij}]$로 표현되는 표본크기는 $E(\tilde{n}_{ij}) = r_{ij}$의 조건을 만족해야 한다. 여기서 $\tilde{n}_{ij} = 0$ or 1 이고 $0 \le r_{ij} \le 1$ 이다. 따라서 선형계획법에 의해 $\tilde{n}_{ij}$을 구할 수 있고 각 셀의 표본크기는 $[nP_{ij}] + \tilde{n}_{ij}$가 된다. 그러므로 표본크기는 일반적인 손실없이

$n_{ij}=0, 1$ 나 $0 \le r_{ij}=nP_{ij}<1$ 로 계산한다.

마지막으로 3차원 이상의 L차원의 층화추출에서는 같은 방법을 적용하는데 손실함수의 형태만 달라진다. 그것은 이전 정보에 의해 계산된 세 개 이상의 층화 요인들의 가중값이 포함된다.

$$w(s)=\gamma_1\sum_{i=1}^{R_1}(n_{i\cdots}(s)-nP_{i\cdots})^2+\cdots+\gamma_L\sum_{k=1}^{R_L}(n_{\cdots k}(s)-nP_{\cdots k})^2$$

여기서 $\gamma_1, \cdots, \gamma_L$ 은 이전 정보에 의한 L개의 층화 요인들의 평균의 층간 분산의 추정값으로 구성된다.

SS방법의 분산추정량을 구하는 방법은 앞선 방법과 마찬가지로 포함확률을 이용하는 호르비츠와 톰슨(Horvitz-Thompson) 추정량을 이용한다(Sitter, Skinner, 1994, 2002). 분산추정량을 구하기 위해 우선 각 표본 셀의 포함확률을 계산하여야 한다. 만약 표본크기가 2 이상을 경우는 각각의 표본 단위에 대해 포함확률을 따로 계산해야 한다. 따라서 표본크기가 1 이상인 표본 셀의 포함확률은 다음과 같이 계산한다. $A_c$은 주변포함확률이고 $B_c$은 결합포함확률이다.

$$\pi_c=\frac{n_c}{N_c},\ \pi_{cc'}=\frac{n_c(n_c-1)}{N_c(N_c-1)}$$

$$A_c=\frac{I_c(I_c+2r_c-1)}{N_c(N_c-1)},\ B_{cc'}=\frac{I_cI_{c'}+r_{c'}I_c+r_cI_{c'}+r_{cc'}}{N_cN_{c'}}$$

여기서 $c$는 셀( $c=1,\cdots,ij$), $N_c$는 모집단 셀 크기, $N$은 전체 모집단 크기, $I_c=n_c-\tilde{n}_c$, $r_c=nP_c-I_c$, $r_{cc'}=\tilde{n}_c\tilde{n}_{c'}$이다. 결국 위의 포함확률을 이용하여 다음과 같은 식으로 분산추정량을 유도할 수 있다. 따라서 SS 방법에서의 평균과 분산추정량은 다음과 같이 계산된다.

$$\bar{y}=\frac{1}{N}\sum_{c}^{ij}N_c\bar{y}_c$$

$$\widehat{Var}(\bar{y})=\frac{1}{2n^2}\sum_{c}^{ij}\sum_{k}^{n_c}\sum_{k\neq k}^{n_c}\left(\frac{n^2}{N^2}-A_c\right)(y_{ck}-y_{ck'})^2+\frac{1}{2n^2}\sum_{c}^{ij}\sum_{c\neq c}^{ij}\sum_{kk}^{n_c}\left(\frac{n^2}{N^2}-B_{cc'}\right)(y_{ck}-y_{c'k'})^2$$

## 참고문헌


[1] Randy R. Sitter and C.J. Skinner(1994). "Multi-way Stratification by Linear Programing", *Survey Methodology*.

[2] William E. Winkler(1987). "An Application of Multi-purpose Survey Sampling", *American Statistical Association, Proceeding of the Section on Survey Research Methods.*

[3] William E. Winkler (1990). "On Dykstra's Iterative Fitting Procedure", *The Annals of Probability.*

[4] William E. Winkler (2001). "Multi-way Survey Stratification and Sampling", *U.S. Bureau of the Census, Statistical Research Division Report.*

[5] William G. Cochran (1977). "Sampling Techniques-third edition", *John Wiley & Sons.*

[6] Wilson Lu and Randy R. Sitter (2002). "Multi-way Stratification by Linear Programing Made Practical", *Survey Methodology.*