The Korean Communication in Statistics
Vol. 13 No. 2, 2006, pp. 327-341

Allocation in Multi-way Stratification by Linear Programing

Pyong Namkung[1], Jae Hyuk Choi[2]

Abstract

Winkler (1990, 2001), Sitter and Skinner (1994), Wilson and Sitter (2002) present a method which applies linear programing to designing surveys with multi-way stratification, primarily in situation where the desired sample size is less than or only slightly larger than the total number of stratification cells.

A comparison is made with existing methods both by illustrating the sampling schemes generated for specific examples, by evaluating sample mean, variance estimation, and mean squared errors, and by simulating sample mean for all methods. The computations required can, however, increase rapidly as the number of cells in the multi-way classification increase. In this article their approach is applied to multi-way stratification using real data.

Keywords : Linear programing; Iterative Proportional Fitting(IPF), Generalized Iterative Fitting Procedure(GIFP).

1. 서론

모집단을 두 개 이상의 변수를 사용해서 층화하는 경우를 다차원 층화라 하며, 다차원 층화의 경우 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 비례배정을 통해 각 층에 표본을 배정한다. 하지만 표본크기가 층화된 층의 개수보다 크기 않을 경우 기대되는 층의 크기가 매우 작을 뿐더러 정수로 마무리되어야 하는 비율적 할당의 특성에 큰 방해가 된다.

다차원 층화에서의 표본배정은 Winkler (1990, 2001)가 제안한 반복 비례 적합 방법(Iterative Proportional Fitting : IPF)과 일반화 반복 적합 방법(Generalized Iterative Fitting Procedure : GIFP) 그리고 Sitter와 Skiner (1994), Wilson과 Sitter(2002)가 제안한 SS방법이 있는데 이 방법들은 표본크기가 층화된 층의 개수보다 크지 않는 경우에 한에서 사용되는 표본 배정 방법으로 모두 선형계획법(Linear Programing)을 이용

1) Professor, Department of Statistics, Sungkyunkwan University, Seoul, 110-745, Korea. Correspondence : namkung@skku.ac.kr
2) Graduate Student, Department of Statistics, Sungkyunkwan University, Seoul, 110-745, Korea.

한다.

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

2. 선형계획법을 이용한 표본배정

2.1 선형계획법

선형계획법이란 목적함수와 제약조건을 통해 방정식의 해를 구하는 방법이다. 이 방법을 통해 표본배정을 결정할 수 있는데 목적함수에서 필요한 계수는 각 가능한 표본배정 결과로 인해 손실되는 정보량이고 미지수가 표본배정확률 $p(x)$이 되어 해를 구함으로써 표본배정을 결정하는 방법이다. 표본배정에 사용되는 목적함수 및 제약식은 다음과 같다.

$$\text{① 목적함수 : } \min\{c_1 p(x_1) + c_2 p(x_2) + \cdots + c_n p(x_n)\} \tag{2.1}$$

$$\text{② 제약조건 : } \begin{cases} \sum_{x=1}^{n} p(x_i) = 1 \\ \sum_{x=1}^{n} m_i p(x_i) = n_i \\ p(x_i) \geqq 0,\ i = 1,2,...,n \end{cases} \tag{2.2}$$

여기서 $p(x_i)$는 i번째 표본배정 결과가 선택될 확률이므로 총합은 1이고 c_i는 표본배정에서 주변합이 반올림되면서 생기는 손실값이다. m_i는 모든 표본배정 결과들이고 n_i는 비례배정 또는 다른 방법으로 적합하여 얻어진 결과이다.

위의 목적함수와 제약 식을 통해 표본배정 확률이 결정되면 그 값이 가장 큰 표본배정 결과가 최종 표본배정이 된다. 이와 같은 방법으로 표본배정을 결정하는 것이 다차원 층화에서 선형계획법을 이용한 표본배정방법이다.

2.2 반복 비례 적합 방법과 일반화 반복 비례 적합 방법

2.2.1 칸 도수 추정

반복비례적합은 주변합을 통해 셀 크기를 추정한다. 먼저 x_{ij}^s를 s번째 반복의 표준화 칸 값이라고 하면 $x_{ij}^0 = x_{ij}$은 비례배정을 통한 셀 크기인 관측값이 된다. 그리고 x_{i+}^*와 x_{+j}^*는 최소충분통계량이므로 충분주변합(sufficient marginal sum)이며, 다음 식을 통해 반복적합하게 된다.

$$x_{ij}^s = \left(\frac{x_{ij}^{s-1}}{x_{i+}^{s-1}}\right) x_{i+}^*,\ \ x_{ij}^{s+1} = \left(\frac{x_{ij}^s}{x_{+j}^s}\right) x_{+j}^*,\ \ \mid x_{ij}^{2k} - x_{ij}^{2k-2} \mid < \delta$$

여기서 k는 반복주기이다. 위 두 식을 반복의 첫 번째 주기로 하여 허용오차 δ에 이

를 때까지 반복하게 된다. 이러한 주변합에 의한 반복계산법은 정확도 δ에 이를 때까지 정확한 값에 수렴하게 되며, 이 수렴된 값을 통해 선형계획법을 이용하여 표본배정을 실시하게 된다.

한편, 일반화 반복 비례 적합 방법은 분류적 적합으로 Dykstra'의 일반화 반복 비율 적합)에 의해 실현된다. 그 적합 과정은 다음과 같다.(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.

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

여기서 p_{ij} 는 ij 셀의 확률이고 s_{ij} 은 ij 셀의 크기이다.

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

2.2.2 표본배정과 추정량

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

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

여기서 I_j 는 주변값 m_j $(j=1,...,s)$ 에 대한 I의 부분집합이다. 위의 조건하에 양의 정수로 주변값 m_j $(j=1,...,s)$ 을 가지고, 양의 정수값 $p_k(k=1,...,t)$ 가 존재하는 배

열 $M_{ik}(i\in I,\ k=1,...,t)$ 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 만들어 낼 수 있는 구조를 산출한다. 이런 배열을 찾았다면 확률적으로 비례크기 p_1를 가지는 배열 $M_{i1}(i\in I)$를 선택하고 나서 모든 $i\in I$인 칸 i에서 크기 M_{i1}인 표본을 선택한다. 이와 같은 방법으로 적합시킨 결과를 가지고 선형계획법을 이용하여 표본을 배정한다. M_0는 IPF 방법 및 GIFP 방법에 의한 배열과 같고 계속해서 $M_k(k=1,...,t)$을 계산하면 선형계획법 과정에 의해 양의 정수로만 이루어진 배열을 찾을 수 있다.

이 경우에 표본평균과 분산추정량은 포함확률을 이용하는 호르비츠와 톰슨(Horvitz-Thompson) 추정량을 이용하여 계산한다(Winkler, 2001). 즉, 표본평균과 분산추정량은 다음과 같다.

$$\bar{y} = \frac{1}{N}\sum_{i}^{I}\left(\frac{N_i}{M_0}\right)\sum_{j}^{n_i} y_{ij} \tag{2.3}$$

$$\widehat{Var}(\bar{y}) = \frac{1}{N^2}\sum_{i}^{I}\left(\frac{n_i}{M_0}-1\right)\left(\sum_{j}^{n_i} y_{ij}\right)^2 + \frac{1}{N^2}\sum_{i\neq i'}^{I}\left(\frac{n_i}{M_0}-1\right)\left(\frac{n_{i'}}{M_0}-1\right)\left(\sum_{j}^{n_i} y_{ij}\right)\left(\sum_{j}^{n_{i'}} y_{i'j}\right) \tag{2.4}$$

여기서 N은 모집단 크기, i는 각 셀, n_i은 각 셀의 표본크기, M_0은 IPF방법 및 GIFP 방법으로 적합시킨 셀 크기, y_{ij}는 i번째 셀에서 선택된 표본값이다.

따라서 모집단 총합 추정량은 $\hat{\tau} = \sum_{i}^{I}(N_i/M_0)\sum_{j}^{n_i} y_{ij}$ 이 된다.

2.3 Sitter와 Skinner 방법 : SS방법

2.3.1 칸 도수 추정

2차원 층화를 보면 N개의 모집단은 $R\times C$의 분할표로 분류되어 있다. 2단 추출과정을 고려해 볼 때 첫 번째로 표본크기 n_{ij}를 특별한 랜덤 과정을 따르는 셀이라 하고, s를 $R\times C$ 배열$(n_{ij}, i=1,...,R,\ j=1,...,C)$로 정의하여 가능한 배열 S집합에서 각 s의 확률 $p(s)$을 배정한다. s에 대한 n_{ij}의 독립을 강조하기 위해 $n_{ij}(s)$라고 하자. 두 번째로 $n_{ij}(s)$값들의 단순임의표본은 셀 ij으로부터 추출되며, 모든 배열의 집합 S_n이 되는 S로 제한한다. 여기서 $P_{ij}=N_{ij}/N$는 각 셀의 비율이며, 식(2.5)와 식(2.6)이 제약조건이 된다.

$$\sum_{i=1}^{R}\sum_{j=1}^{C} n_{ij}(s) = n \tag{2.5}$$

$$\sum_{s\in S_n} n_{ij}(s)p(s) = nP_{ij} \quad \text{for } i=1,...,R,\ j=1,...,C \tag{2.6}$$

따라서 다음과 같은 방법으로 표본 s가 정수로 적합되면서 손실되는 정보를 최소화

하는 표본설계 $p(s)$을 선택한다.

$$\underset{p\in P}{\text{mimimize}} \sum_{s\in S_n} w(s)p(s), \quad 0 \leqq p(s) \leqq 1 \text{ for all } s\in S_n \tag{2.7}$$

$w(s)$는 선택된 표본 s에 대한 손실함수이고, P는 S_n에 대한 가능한 표본설계의 일정한 제약조건에 종속되는 문제점이 있다. 위의 제약조건은 $\sum_{s\in S_n} p(s)=1$를 의미하며, 목적함수와 일정한 제약조건 모두 $p(s)$에 대해 선형관계이다.

반면 이러한 문제점은 미지의 $p(s), s\in S_n$에 대한 선형계획법을 통해 직접 해결될 수 있다. 그러나 이 해결방법은 S_n에서 원소의 수가 매우 크거나 미지의 수가 매우 클 때 선형계획법에 의한 결과가 도출되기 어려운 계산상의 문제점이 있다. 그러므로 S_n의 부분집합에 관심을 두고 제한하는 것은 바람직한 방법이다.

또 하나의 제약조건은 $n_{ij}(s)$가 $I_{ij}=[nP_{ij}]$와 같거나 가장 큰 정수가 nP_{ij}나 $I_{ij}+1$보다 적은 배열 s대해서만 오직 고려한다는 것이다. $\tilde{n}_{ij}(s)=n_{ij}(s)-I_{ij}$ $r_{ij}=nP_{ij}-I_{ij}$라고 하면 제약조건은 다음과 같다.

$$\underset{p\in P}{\text{mimimize}} \sum_{s\in \tilde{S}_{\tilde{n}}} w(s)p(s) \tag{2.8}$$

$$\sum_{s\in \tilde{S}_{\tilde{n}}} \tilde{n}_{ij}(s)p(s)=r_{ij} \tag{2.9}$$

$$\sum_{s\in \tilde{S}_{\tilde{n}}} p(s)=1, \quad 0 \leqq p(s) \leqq 1 \quad \text{for all } s\in \tilde{S}_{\tilde{n}} \tag{2.10}$$

여기서 $\tilde{S}_{\tilde{n}}$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 $\tilde{S}_{\tilde{n}}$의 원소의 수는 이제 ${}_{RC}C_n$이다. 이 숫자는 여전히 매우 클 수 있다. 그러나 축소된 이 배열에서 손실함수 $w(s)$의 알맞은 선택으로 보다 쉽게 나타낼 수 있다. 이러한 접근법을 통해 손실함수 $w(s)$을 선택하는 방법으로 선택된 표본 s는 다음과 같은 주변합 제한이 요구된다.

$$| n_{i\cdot}(s)-nP_{i\cdot} | < 1 \quad i=1,\ldots,R \tag{2.11}$$

$$| n_{\cdot j}(s)-nP_{\cdot j} | < 1 \quad j=1,\ldots,C \tag{2.12}$$

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

이런 제한조건은 S_n 집합으로부터의 선택된 표본 s가 식(2.8), (2.9), (2.10)의 제약조건을 만족하지 못하면 그 표본을 배제시키거나 $w(s)$을 유한하게 적용함으로써 접근방법을 조절할 수 있다. 이런 전통적인 접근은 최적화를 해결할 방법이 존재하지 않는다는 문제점이 있으나 다음과 같은 손실함수를 사용한다면 문제점이 해결될 수 있다.

$$w(s)=\sum_{i=1}^{R}(n_{i\cdot}(s)-nP_{i\cdot})^2+\sum_{j=1}^{C}(n_{\cdot j}(s)-nP_{\cdot j})^2 \tag{2.13}$$

최적 해결 방법은 충분히 큰 S_n 집합 사이에 항상 존재할 것이다. 실제로 유한한 집합 S_n을 위의 조건식에 따르게 하거나 부분집합인 표본으로 제한하는 것은 컴퓨터 계산상이나 표본들의 집합을 확장하는데 매우 유리하게 작용된다.

2.3.2 표본배정 및 추정량(SS방법)

선형계획법을 이용하여 앞서 언급한 문제를 해결하는데 필요한 손실함수는 다음과 같다.

$$w(s)=\sum_{i=1}^{R}(n_{i\cdot}(s)-nP_{i\cdot})^2+\sum_{j=1}^{C}(n_{\cdot j}(s)-nP_{\cdot j})^2 \tag{2.14}$$

I_{ij}값은 $n_{ij}=I_{ij}+\tilde{n}_{ij}(\tilde{s})$ 으로 계산된다. 그것은 $p(s)>0$인 각 s가 정확한 실험설계의 주변값과 대응되는 주변값 $n_{i\cdot}(s)$과 $n_{\cdot j}(s)$를 가지는 방법으로 완성된다. 이와 같은 방법으로 가능한 표본 S_n은 각 셀에 $[nP_{ij}]$나 $[nP_{ij}]+1$의 값을 가지는 표본크기 n_{ij}을 갖게 된다. $[nP_{ij}]$는 nP_{ij}와 같거나 보다 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 1이고 $0\leq r_{ij}\leq 1$이다. 따라서 선형계획법에 의해 $\tilde{n}_{ij}$을 구할 수 있고 각 셀의 표본크기는 $[nP_{ij}]+\tilde{n}_{ij}$가 된다. 그러므로 표본크기는 일반적인 손실 없이 $n_{ij}=0,1$나 $0\leq r_{ij}=nP_{ij}<1$로 계산한다.

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

$$w(s)=\gamma_1\sum_{i=1}^{R_1}(n_{i\cdots}(s)-nP_{i\cdots})^2+\cdots+\gamma_L\sum_{k=1}^{R_L}(n_{\cdots k}(s)-nP_{\cdots k})^2 \tag{2.15}$$

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

SS방법의 분산추정량을 구하는 방법은 앞선 방법과 마찬가지로 포함확률을 이용하는 호르비츠와 톰슨 추정량을 이용한다(Sitter와 Skinner, 1994, Wilson과 Sitter, 2002). 분산추정량을 구하기 위해 우선 각 표본 셀의 포함확률을 계산하여야 한다. 만약 표본크기가 2 이상인 경우는 각각의 표본 단위에 대해 포함확률을 따로 계산해야 한다. 따라서 표본크기가 1 이상인 표본 셀의 포함확률은 다음과 같이 계산한다. A_c은 주변포함확률이고 B_c은 결합포함확률이다.

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

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

여기서 c는 셀$(c=1,\cdots,ij)$, N_c는 모집단 셀 크기, N은 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. 위의 식으로 계산된 포함확률을 이용하여 다음과 같이 분산추정량 식을 유도할 수 있다.

따라서 SS 방법에서의 평균과 분산추정량은 다음과 같이 계산된다.

$$\bar{y} = \frac{1}{N} \sum_{c}^{ij} N_c \bar{y}_c \tag{2.16}$$

$$\widehat{Var}(\bar{y}) = \frac{1}{2n^2} \sum_{c}^{ij} \sum_{k}^{n_c} \sum_{k' \neq k}^{n_c} \left(\frac{n^2}{N^2} - A_c \right) (y_{ck} - y_{ck'})^2 + \frac{1}{2n^2} \sum_{c}^{ij} \sum_{c' \neq c}^{ij} \sum_{kk'}^{n_c} \left(\frac{n^2}{N^2} - B_{cc'} \right) (y_{ck} - y_{c'k'})^2 \tag{2.17}$$

3. 사례연구

3.1 2차원 층화

2003년 코스닥 상장 일반기업 495개 회사의 자료를 통해 2차원 층화를 실시하였다. 먼저 첫 번째 층화변수는 일반자본금이고 두 번째 층화변수는 발행주식수이며, 추정하고자 하는 종속변수는 총자본금이다. 일반자본금과 발행주식수에 대해 크기로 소, 중소, 중, 중대, 대로 층화 분류한 모집단의 분포와 각각의 셀을 총 표본크기 20으로 비례배정한 결과는 <표 1>과 같다. 표본크기 20으로 비율 배정한 결과 주변합의 값이 모두 정수가 아니기 때문에 반올림을 통해 표본크기를 정수로 조정하였다.

<표 1> 모집단 분포 및 비례배정-(): 반응비율 / []:표본크기

자기주식 / 자본금	소	중소	중	중대	대	계	소	중소	중	중대	대	계
소	44 (0.089)	39 (0.079)	18 (0.036)	14 (0.028)	26 (0.053)	141 (0.285)	1.78	1.58	0.72	0.57	1.05	5.70 [6]
중소	19 (0.038)	18 (0.036)	14 (0.028)	16 (0.032)	18 (0.036)	85 (0.172)	0.77	0.72	0.57	0.65	0.73	3.44 [3]
중	11 (0.022)	39 (0.079)	21 (0.042)	18 (0.036)	4 (0.008)	93 (0.188)	0.44	1.58	0.85	0.72	0.16	3.75 [4]
중대	10 (0.020)	32 (0.065)	27 (0.055)	27 (0.055)	19 (0.038)	115 (0.232)	0.40	1.29	1.09	1.09	0.77	4.64 [5]
대	3 (0.006)	3 (0.006)	6 (0.010)	12 (0.024)	38 (0.077)	62 (0.123)	0.12	0.12	0.20	0.49	1.54	2.47 [2]
계	87 (0.176)	131 (0.265)	85 (0.172)	87 (0.176)	105 (0.212)	495	3.51 [4]	5.29 [5]	3.43 [3]	3.52 [4]	4.25 [4]	20

3.2 선형계획법을 이용한 표본배정

3.2.1 셀 크기 추정

2차원 층화의 모집단 분포에 의해 비례배정을 실시한 각 셀 값으로 각각의 방법에 의해 셀 크기를 추정하면 먼저 IPF방법과 GIFP방법에 의해 추정된 셀 값은 다음의

<표 2>와 같다. 비례배정과 비교해 보면 전체적으로 균일한 분포로 조정되었다 높은 비례값은 낮아졌고 낮은 비례값은 조금 높아졌다. 층화1 요인과 층화2 요인이 서로 종속되어 있으면 2차원 층화하여 표본을 추출하는 의미가 희석되기 때문에 교차적비가 0이 되게 하는 방법으로 셀 크기를 추정하게 된다. 따라서 IPF 방법은 주변합이 로그선형인 경우에 자료구조가 손상되지 않는 교차적비 0으로 적합되게 된다. GIFP 방법은 주변합에 의한 방법이 아닌 셀 간의 변량을 최소화하는 방법으로 IPF 방법을 보완하는 방법이다. IPF 방법이 균일한 분포로 가는 방법이므로 비율이 작은 칸에 대해 모집단의 비율보다 IPF 방법에 의한 비율이 너무 커지는 것을 방지한다. 이 방법은 비율의 각 셀이 서로 값의 결정에 영향을 준다. 그 결과를 보면 주변합이 약간 조정되었다. 이 자료는 주변합이 선형관계가 아니기 때문에 셀 크기가 조정되었지만 주변합이 선형관계라면 IPF 방법과 동일한 결과를 얻게 된다.

<표 2> IPF방법과 GIFP방법에 의한 셀 크기 추정

층화2 / 층화1	IPF 방법에 의한 셀 크기						GIFP 방법에 의한 셀 크기					
	I	II	III	IV	V	계	I	II	III	IV	V	계
A	1.00	1.51	0.98	1.00	1.21	5.70	1.90	1.62	0.71	0.52	0.84	5.59
B	0.60	0.91	0.59	0.61	0.73	3.44	0.86	0.77	0.58	0.62	0.61	3.43
C	0.66	0.99	0.64	0.66	0.80	3.75	0.46	1.61	0.82	0.65	0.13	3.67
D	0.81	1.23	0.80	0.82	0.99	4.64	0.45	1.39	1.12	1.04	0.65	4.65
E	0.43	0.65	0.42	0.43	0.52	2.47	0.16	0.15	0.25	0.56	1.54	2.66
계	3.51	5.29	3.43	3.52	4.25	20	3.83	5.55	3.48	3.38	3.77	20

<표 3> SS방법에 의한 $r_{ij}(\tilde{n}=12)$

층화2 / 층화1	SS 방법의 셀 크기						비례 배정					
	I	II	III	IV	V	계	I	II	III	IV	V	계
A	0.78	0.58	0.72	0.57	0.05	2.70	1.78	1.58	0.72	0.57	1.05	5.70
B	0.77	0.72	0.57	0.65	0.73	3.44	0.77	0.72	0.57	0.65	0.73	3.44
C	0.44	0.58	0.85	0.72	0.16	2.75	0.44	1.58	0.85	0.72	0.16	3.75
D	0.40	0.29	0.09	0.09	0.77	1.64	0.40	1.29	1.09	1.09	0.77	4.64
E	0.12	0.12	0.20	0.49	0.54	1.47	0.12	0.12	0.20	0.49	1.54	2.47
계	2.51	2.29	2.43	3.52	2.25	12	3.51	5.29	3.43	3.52	4.25	20

SS 방법에 의한 셀 크기 추정 방법은 비례배정의 비례값을 그대로 이용한다. 그것은 SS 방법이 계산상의 신속성에 중점을 둔 방법이기 때문이다. 이 자료의 층은 모두 25개이고 표본크기는 20이다. 따라서 가능한 표본배정의 경우의 수는 ${}_{25}C_{20}$인데 반해 1보다 큰 셀의 비례값에서 1을 빼서 모든 셀의 크기가 모두 1보다 작게 하는 SS 방법의 표본배정의 모든 경우에 수는 ${}_{25}C_{12}$가 되어 계산상의 많은 이점이 있다.

3.2.2 각 셀의 표본 배정

셀 크기를 추정한 후 선형계획법의 방법에 의해 표본크기를 적합시키기 위해 각각

의 주변합의 손실을 최소화하는 선에서 주변합을 정수로 적합하는 선형계획법(Linear Programing)을 이용한다.

IPF 방법 및 GIFP 방법을 통해 얻어진 셀 크기로 이루어진 배열을 M_0로 놓고 주변합의 손실을 최소화하면서 각각의 선택확률을 가지는 배열 $M_i\ (i=1,\cdots,k)$을 계산한다.

[IPF 방법]

$$M_0=\begin{vmatrix} 1.00 & 1.51 & 0.98 & 1.00 & 1.21 \\ 0.60 & 0.91 & 0.59 & 0.61 & 0.73 \\ 0.66 & 0.99 & 0.64 & 0.66 & 0.80 \\ 0.81 & 1.23 & 0.80 & 0.82 & 0.99 \\ 0.43 & 0.65 & 0.42 & 0.43 & 0.52 \end{vmatrix}$$

[GIFP 방법]

$$M_0=\begin{vmatrix} 1.90 & 1.62 & 0.71 & 0.52 & 0.84 \\ 0.86 & 0.77 & 0.58 & 0.62 & 0.61 \\ 0.46 & 1.61 & 0.82 & 0.65 & 0.13 \\ 0.45 & 1.39 & 1.12 & 1.04 & 0.65 \\ 0.16 & 0.15 & 0.25 & 0.56 & 1.54 \end{vmatrix}$$

위의 M_0을 통해 선형계획법을 실시한 결과 IPF방법은 반복 13, GIFP 방법은 반복 9로 배열 M_i들을 얻었다. 여기서 각각의 배열 M_i들은 모두 정수로 이루어진 배열이다. 각각의 M_i들은 선택확률 $p(M_i)$를 가진다. 배열 M_i들과 선택확률 $p(M_i)$은 다음의 조건을 만족하게 된다.

$$\sum_{i=1}^{13} p(M_i)=1\,(0<p(M_i)<1),\ \sum_{i=1}^{13} p(M_i)M_i=M_0,\ E(M_i)=M_0 \tag{3.1}$$

$$\sum_{i=1}^{9} p(M_i)=1\,(0<p(M_i)<1),\ \sum_{i=1}^{9} p(M_i)M_i=M_0,\ E(M_i)=M_0 \tag{3.2}$$

식(3.1)은 IPF방법의 배열의 조건의 결과이고 식(3.2)는 GIFP방법의 배열의 조건의 결과이다. 그 결과 얻어진 배열 중 선택확률이 가장 높은 배열은 다음과 같다.

[IPF 방법]

$$M_9=\begin{vmatrix} 1 & 2 & 1 & 1 & 1 \\ 0 & 1 & 1 & 1 & 0 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 1 & 1 \\ 1 & 0 & 0 & 0 & 1 \end{vmatrix}$$

$p(M_9)=0.6$

[GIFP 방법]

$$M_7=\begin{vmatrix} 2 & 1 & 1 & 1 & 1 \\ 1 & 1 & 0 & 0 & 1 \\ 0 & 2 & 0 & 1 & 1 \\ 1 & 1 & 2 & 1 & 0 \\ 0 & 0 & 0 & 1 & 1 \end{vmatrix}$$

$p(M_7)=0.46$

따라서 IPF방법 및 GIFP방법의 결과 얻어진 표본크기는 다음의 <표 4>와 같다. 이것은 선형계획법에 의해 반복 계산된 표본배정 결과이다. IPF방법에 의해 추정된 셀 크기를 이용하여 선형계획법을 통해 얻어진 표본크기는 13번 반복 계산되어 이 표본 배열의 선택확률은 0.6이다. 표본크기가 0인 셀이 6개이며 표본크기가 2인 셀은 1개이고 나머지 셀은 모두 표본크기가 1이다.

<표 4> IPF방법 및 GIFP방법에 의한 표본배정

층화2 / 층화1	IPF 방법						GIFP 방법					
	I	II	III	IV	V	계	I	II	III	IV	V	계
A	1	2	1	1	1	6	2	1	1	1	1	6
B	0	1	1	1	0	3	1	1	0	0	1	3
C	1	1	0	1	1	4	0	2	0	1	1	4
D	1	1	1	1	1	5	1	1	2	1	0	5
E	1	0	0	0	1	2	0	0	0	1	1	2
계	4	5	3	4	4	20	4	5	3	4	4	20
	반복 : $k=13$, $p(s)=0.6$						반복 : $k=9$, $p(s)=0.46$					

GIFP방법에 의해 추정된 셀 크기를 이용하여 선형계획법을 통해 얻어진 표본크기는 총 9번 반복 계산되었고 이 표본배열의 선택확률은 0.46이다. 표본크기가 0인 셀은 8개 셀이고 표본크기가 2인 셀은 3개이며 나머지 모든 셀은 표본크기가 1이다.

SS방법에 의해 결정된 셀 크기 $r_{ij}(\tilde{n}=12)$을 이용하여 선형계획법을 실시하면 다음과 같은 비례배정을 통해 $[nP_{ij}]$로 이루어진 배열이 얻어지고 그 배열에서 $r_{ij}=nP_{ij}-I_{ij}$로 이루어진 배열을 계산해 냄으로써 표본이 배정된다.

$$[nP_{ij}]=\begin{vmatrix}1.78 & 1.58 & 0.72 & 0.57 & 1.05\\ 0.77 & 0.72 & 0.57 & 0.65 & 0.73\\ 0.44 & 1.58 & 0.85 & 0.72 & 0.16\\ 0.40 & 1.29 & 1.09 & 1.09 & 0.77\\ 0.12 & 0.12 & 0.20 & 0.49 & 1.54\end{vmatrix} \quad [I_{ij}]=\begin{vmatrix}1 & 1 & 0 & 0 & 1\\ 0 & 0 & 0 & 0 & 0\\ 0 & 1 & 0 & 0 & 0\\ 0 & 1 & 1 & 1 & 0\\ 0 & 0 & 0 & 0 & 1\end{vmatrix} \quad [r_{ij}]=\begin{vmatrix}0.78 & 0.58 & 0.72 & 0.57 & 0.05\\ 0.77 & 0.72 & 0.57 & 0.65 & 0.73\\ 0.44 & 0.58 & 0.85 & 0.72 & 0.16\\ 0.40 & 0.29 & 0.09 & 0.09 & 0.77\\ 0.12 & 0.12 & 0.20 & 0.49 & 0.54\end{vmatrix}$$

선형계획법을 이용하여 위의 배열 $[r_{ij}]$을 7번 반복하면 배열 $[\tilde{n}_{ij}]$을 얻을 수 있고 배열 $[\tilde{n}_{ij}]$을 통해 $[n_{ij}]=[\tilde{n}_{ij}]+[I_{ij}]$의 배열을 얻을 수 있다. 배열 $[\tilde{n}_{ij}]$의 원소는 0 또는 1의 값만을 가지며 원소의 합인 총 표본크기는 $\tilde{n}=n-\sum_{i=1}^{5}\sum_{j=1}^{5}I_{ij}$으로 계산된다. 따라서 원하는 총 표본크기는 20이지만 $[I_{ij}]$의 배열의 총 원소의 합이 8이므로 SS방법에 의해 조정된 표본크기는 $\tilde{n}=12$이다. 선형계획법에 의해 얻어진 배열은 $\tilde{n}_{ij}(s_7)$이고 $p(s_7)=0.48$이다.

$$[\tilde{n}_{ij}]=\begin{vmatrix}1 & 0 & 1 & 1 & 0\\ 1 & 1 & 0 & 0 & 1\\ 0 & 1 & 1 & 1 & 0\\ 1 & 0 & 0 & 0 & 1\\ 0 & 0 & 0 & 1 & 0\end{vmatrix} \quad [n_{ij}]=\begin{vmatrix}2 & 1 & 1 & 1 & 1\\ 1 & 1 & 0 & 0 & 1\\ 0 & 2 & 1 & 1 & 0\\ 1 & 1 & 1 & 1 & 1\\ 0 & 0 & 0 & 1 & 1\end{vmatrix}$$

여기서 이 배열은 다음의 조건을 만족한다.

$$\sum_{k=1}^{7}\tilde{n}_{ij}(s_k)p(s_k)=r_{ij} \tag{3.3}$$

$$\sum_{k=1}^{7}p(s_k)=1,\ 0\leq p(s_k)\leq 1 \quad k=1,\cdots,7 \tag{3.4}$$

$$E\left[\tilde{n}_{ij}(s_k)\right]=r_{ij} \tag{3.5}$$

<표 5> SS방법에 의한 표본배정

층화2 / 층화1	SS 방법 $\tilde{n}_{ij}(\tilde{n}=14)$						SS 방법 $n_{ij}(n=20)$					
	I	II	III	IV	V	계	I	II	III	IV	V	계
A	1	0	1	1	0	3	2	1	1	1	1	6
B	1	1	0	0	1	3	1	1	0	0	1	3
C	0	1	1	1	0	3	0	2	1	1	0	4
D	1	0	0	0	1	2	1	1	1	1	1	5
E	0	0	0	1	0	1	0	0	0	1	1	2
계	3	2	2	3	2	12	4	5	3	4	4	20
	반복 : $k=7$, $p(s)=0.48$											

얻어진 $\tilde{n}_{ij}(s_7)$에 대해 $[I_{ij}]$을 더하면 셀 (1,1), (1,2), (1,5), (3,2), (4,2), (4,3), (4,4), (5,5)의 원소가 1씩 늘어나 배열 $[n_{ij}]$을 얻을 수 있고 이것이 표본크기의 배정 결과 <표 5>이다. 총 표본크기는 7번 반복 계산되었고 이 표본 배열의 선택확률은 0.48이다. 표본크기가 0인 셀이 7개이며 표본크기가 2인 셀은 2개이고 나머지 셀은 모두 표본크기가 1이다.

3.3 분산추정량 및 MSE 비교

2차원으로 층화를 했을 경우와 1차원으로 층화했을 때의 추정에 대한 비교를 위해 먼저 각 방법에 대해 최종 표본배정 결과를 이용하여 표본을 1회 추출했다. 여기서 각 층에서의 표본은 단순임의추출방법으로 추출하였다. 층1 요인(일반자본금)에 따른 층화추출 추정량을 계산하고 두 번째로 층2 요인(발행주식수)에 따른 층화추출 추정량을 계산하고 마지막으로 2차원 층화에 의한 추정량 계산법으로 추정한 추정량들을 서로 비교한 결과 <표 6>과 같다.

<표 6> 각 방법의 표본추출에 따른 추정량 비교 단위 : 백만원

	IPF방법			GIFP방법			SS 방법		
	층1	층2	2차원	층1	층2	2차원	층1	층2	2차원
평균	202	221	237	138	142	145	177	187	184
분산추정량	3,018	1,166	10	646	625	9	1,501	1,708	8
MSE	3,054	1,796	1,708	3,981	3,468	2,634	1,872	1,779	158

각 방법에 의해 셀의 표본크기를 결정하여 추출한 표본의 추정량을 비교해 본 결과 각 층별로 1차원 층화한 표본보다 각 방법으로 층화추출한 표본의 평균의 분산추정량이 더 작았다. 평균제곱오차(MSE) 또한 가장 작아 1차원 층화보다는 각 방법으로 2차원 층화한 추정량이 훨씬 효율적인 추정량임을 알 수 있다.

각 방법을 비교해 보면 평균은 모평균(196)을 중심으로 IPF방법은 크고 GIFP방법은 작았으며 SS방법은 작지만 모평균에 근접한 값을 얻었다. 평균이 모평균과 거리가 상당히 먼 것은 모집단의 분포가 0의 값 쪽으로 매우 몰려있는 분포이기 때문이다. 따라서 표본이 편의를 가질 확률이 매우 높다. 분산추정량 및 MSE를 비교해 보면 분

산추정량은 SS방법 < GIFP방법 < IPF방법 순이었고 MSE 추정량 역시 SS방법 < IPF방법 < GIFP방법 순이었다. 따라서 세 방법 중 SS 방법이 가장 효율적인 방법이고 GIFP방법은 IPF방법에 비해 분산추정량은 더 작지만 MSE는 더 크므로 더 효율적이다 말할 수는 없다.

3.4 모의실험

앞 절에서 표본크기 20으로 하여 각각의 방법에 의한 표본 셀 크기로 표본을 추출하여 통계량들을 계산하였다. 표본배정 방법에 따라 표본평균의 분포의 형태가 어떤지를 알아보기 위해 표본크기 20으로 표본을 총 10,000번 반복 추출하였다.

<표 7> 모의실험 결과(r= 10,000) 단위 : 백만원

	IPF	GIFP	SS	반올림	SRS	Population
$E(\bar{y})$	222	181	190	182	196	μ=196
$Var(\bar{y})$	5,461	2,385	3,016	2,680	4,361	σ^2/n= 4,613

모의실험을 10,000번 반복한 결과 정규분포보다 분산이 더 작은 분포임을 알 수 있다. 즉, 첨도가 3보다 더 큰 분포로 나타났다. SRS를 제외한 모든 방법에 대해 편의가 존재하였는데 모집단의 자료가 매우 심하게 한쪽으로 치우쳐진 자료이기 때문이다. 따라서 이 경우 중심에서 매우 떨어진 집단들을 더 많이 추출하여야 편의를 줄일 수 있지만 이 논문에서는 각 방법의 단순 비교를 위해 일반적 층화추출방법으로 추출하였다. 따라서 일반적 층화추출방법으로 각 배정방법들의 편의가 어떻게 존재하는지도 함께 비교해 보았다.

각 방법별로 비교해 보면 IPF방법은 앞서 연구된 대로 평균이 과대 추정되고 평균의 분산 역시 큰 값을 갖게 된다. 즉, 표본평균의 분포의 산포가 크다. 그것은 IPF방법의 셀 크기 조정은 균일한 분포로 만드는 것이므로 비율이 낮은 셀에서 표본이 뽑힐 확률이 높기 때문이다. 두 번째로 GIFP방법의 결과는 IPF방법의 문제점을 보완되어 평균이 약간 과소 추정되었지만 표본평균의 분산은 작았다.

SS 방법을 보면 평균은 모평균에 근사하는 값을 가졌지만 표본평균의 분산은 작지 않았다. 그러나 일반적인 표본평균의 분산인 σ^2/n보다는 작은 값을 가졌다. 즉, 표본평균의 산포가 작다는 뜻이다. 선형계획법이 아닌 반올림 방법으로 추출한 표본평균은 분산은 작지만 평균은 과소 추정되었다. 반올림 방법은 GIFP방법과 거의 비슷한 결과를 얻었다. 여기서 반올림 방법은 각 셀의 값을 수리적으로 0.5보다 크면 올림, 0.5보다 작으면 내림하는 방법이 아니라 각 열과 행을 비교하여 주변합의 손실을 최소화하면서 가장 큰 값을 올림하는 방법을 말한다. 즉, 0.6과 0.7의 경우 주변합은 1.3이고 주변합을 반올림하면 1.0이다. 0.6, 0.7 모두를 올림할 수 없으므로 손실이 더 적은 0.7을 올림하고 0.6은 내림한다. 반올림 방법을 SS방법과 비교해보면 SS방법은 편의가 작고 반올림 방법은 분산이 작았다. SRS방법의 표본평균은 모평균과 일치하지만 표본평균의 산포가 다차원 층화 방법에 비해 훨씬 컸다.

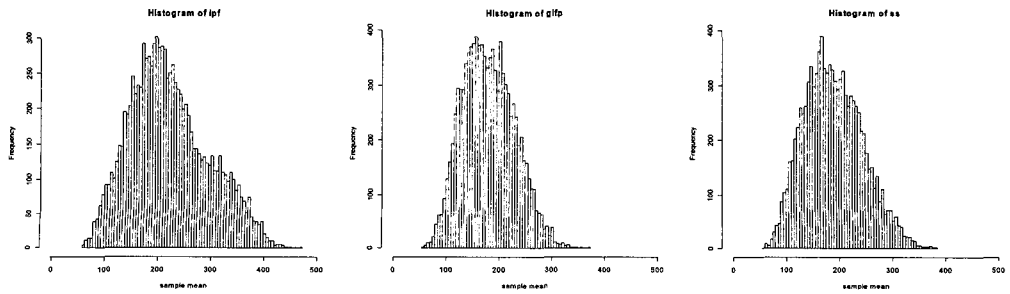
다음으로 각 방법의 표본평균들의 분포를 살펴보기 위해 모의실험 결과의 히스토그램을 그려보았다. IPF방법은 분포의 중심이 모평균보다 약간 오른쪽에 치우쳐 있다. 다른 방법과 달리 분포의 꼬리가 오른쪽으로 많이 치우쳐 있는데(왜도>0) 그 정도가 다른 방법들에 비해 더 컸다. 분포의 형태도 정규분포 같은 종모양이 아닌 오른쪽에 높은 빈도가 나타나서 평균이 과대 추정되는 모습을 보여준다. IPF방법은 각 셀을 균일분포로 근사시키기 때문에 추출확률이 낮은 셀의 추출확률을 더 커지게 하는 효과가 있다. 따라서 이상값에 영향을 받기 쉽다.

GIFP방법에 의한 표본평균의 분포의 히스토그램을 그려본 결과 앞의 IPF방법에 비해 좀 더 종 모양에 가까운 분포의 형태를 띠었지만 모평균보다 왼쪽으로 약간 치우친 분포이고 특히 분포의 중심이 뭉툭한 형태로 정확한 평균을 추정하기 어려운 방법임을 보여주고 있다 그러나 다른 분포에 비해 표본평균의 분포의 퍼짐 정도가 작아 폭이 좁은 형태임을 알 수 있다.(첨도>3)

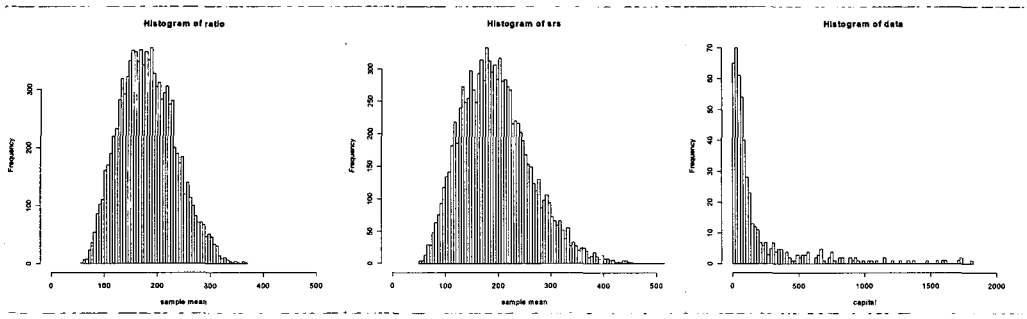
SS방법의 표본평균들의 분포는 정규분포와 흡사한 종모양의 분포 형태를 띠었지만 모평균에 비해 평균이 좀 더 왼쪽으로 치우쳐진 분포의 형태를 보였다. 정규분포의 형태와 많이 비슷하지만 GIFP방법과 마찬가지로 첨도가 3보다 큰 좀 더 뾰족한 형태이고 왜도가 0보다 큰 오른쪽으로 꼬리가 많이 치우쳐진 분포이다. SS방법은 GIFP방법과 IPF방법과 달리 표본평균의 평균은 모평균에 거의 수렴하는 것으로 나타났다. 그러나 최빈값은 모평균과 차이가 있었다.

반올림 방법에 의해 표본을 추출하여 나온 평균들의 분포를 보면 전체적으로 분포가 왼쪽으로 치우친 것을 알 수 있다. 그 형태도 중심이 뭉툭한 형태를 띠었다. 즉, 표본평균들이 모평균보다 작아 불편추정량은 만족할 수 없지만 분산은 작았다. 이것은 반올림의 경우 셀 크기가 작은 층의 경우 추출될 확률이 너무 낮아서 항상 추출되지 않기 때문에 그 층에서 설명하는 종속변수의 내용을 포함하고 있지 않기 때문에 분산은 작지만 모평균과는 차이가 있었다.

단순임의추출법에 의해 표본크기 $n=20$으로 추출한 표본들의 평균의 분포를 보면 다른 방법들에 비해 왜도가 훨씬 컸다. 즉, 오른쪽 꼬리가 매우 길게 나타났다. 표본평균들의 중심은 모평균과 같았지만 분산이 매우 컸다. 표본의 수를 늘려 가면 모평균에 근접하는 표본평균을 얻을 수 있지만 하나의 표본 수에서는 그 가능성이 매우 희박하다 할 수 있다. 따라서 이 자료에서는 SRS방법으로 효율적인 추정량을 얻기 힘들다.



<그림 1> IPF방법, GIFP방법, SS방법의 표본평균 분포



<그림 2> 반올림방법과 SRS의 표본평균 및 모집단 분포

4. 결론

셀 크기보다 표본크기가 더 적은 경우에 한정하여 주변합이 로그선형 모형인 경우 효과적인 IPF방법, 셀 간의 변량을 최소화시키는 GIFP방법, 셀 크기를 비율크기로 하여 선형계획법 실행횟수를 줄여주는 SS방법으로 표본 배정한 결과를 살펴보았다.

각 방법에 의해 각 셀에 표본배정한 결과 0, 1, 2의 표본크기가 배정되었고 배정결과로 코스닥 상장 일반기업의 자료를 표본추출하여 통계량들을 비교해 보고 모의실험을 실시한 결과, SS 방법이 가장 모평균에 가까웠고 분산추정량은 SS방법 < GIFP방법 < IPF방법 순이며 MSE 추정량은 SS방법 < IPF방법 < GIFP방법 순이었다.

SS방법은 계산 속도가 빠르고 효율적인 최소 분산추정량을 얻게 해 주며 MSE의 값이 작아 불편추정량을 구할 수 있는 가장 최적의 방법으로 나타났다. GIFP방법은 분산은 작으나 편의가 존재하여 MSE 값이 크고 IPF방법은 분산은 크지만 MSE 추정량이 작았다. 이 자료의 모집단은 주변합이 선형결합의 형태가 아니기 때문에 IPF방법이 가장 비효율적으로 나타났다. 만약 주변합이 선형결합의 형태라면 GIFP방법과 IPF방법은 동일한 셀 크기와 표본배정의 결과를 갖게 된다.

모의실험을 실시한 결과, 세 방법 모두 표본의 수를 늘려감에 따라 표본평균의 분포가 정규분포는 아니지만 종모양의 형태에 근접하였다. 특히, SS방법은 정규분포의 형태와 매우 비슷한 분포를 나타냈다. 하지만 모집단의 형태가 왼쪽으로 많이 치우쳐진 자료이기 때문에 왼쪽으로 조금 치우친 형태로 나타났다. SS방법은 주변합이 손실을 최소화하는 방법인 반올림 방법과 비교해 보았을 때는 편의는 더 작으나 분산은 더 큰 것으로 나타났다. 따라서 표본배정방법을 결정하기 위해서는 자료의 특성 및 목적에 따라 가장 효과적이고 최적화되는 방법을 선택하는 것이 매우 중요하다.

참고문헌

[1] Sitter, R.R. and Skinner, C.J. (1994). Multi-way Stratification by Linear Programing. *Survey Methodology*. Vol. 20, 65-73.

[2] Dykstra, R.L. (1985). An Iterative Procedure for Obtaining I-Projections onto the Intersection of Convex Sets. *The Annals of Probability*. Vol. 13, 975-984.

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

[4] Winkler, W.E. (1990). On Dykstra's Iterative Fitting Procedure. *The Annals of Probability*. Vol. 18, 1410-1416.

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

[6] Wilson, L. and Sitter, R.R. (2002). Multi-way Stratification by Linear Program -ming Made Practical. *Survey Methodology*. Vol. 28, 199-207.

[Received October 2005, Accepted June 2006]