The Korean Journal of Applied Statistics
2021, Vol. 34, No. 5, 671–683

DOI: https://doi.org/10.5351/KJAS.2021.34.5.671
Print ISSN1225-066X / Online ISSN 2383-5818

Expansion of power allocation using response rate per stratum

Hyeonah Park[1,a]

[a]Division of Data Science, Yonsei University

Abstract

Power allocation is a technique that evenly allocates samples for each stratum, although the overall efficiency of the allocation is less than that of optimal allocation, and it is often used as a square root proportional allocation in real survey. Also, considering the non-response that occurs in real survey, a larger sample size is used than that in the theoretical formula. In this study, in determining the sample size for each stratum, we study the new methods of allocating by adding information on the response rate per each stratum to power allocation method. The proposed allocation methods are compare with proportional, optimal, and square root proportional allocation in simulation. In addition, the comparison with the proportional and optimal allocation to which the response rate was added is examined through simulation. As a result, we examine the advantages and disadvantages of the allocation methods.

Keywords: power allocation, response rate, stratum, sample size

1. 서론

여러 단계에 걸쳐 표본추출이 이루어지는 복합표본설계에서는 첫 번째 단계에서 층화추출이 이루어진다. 층을 나누어서 표본이 추출된다는 것은 모수에 관한 추정의 효율성을 담보할 수 있고 층별 추정의 목적을 수행하기 위한 것으로 생각할 수 있다. 그래서 표본추출을 위해 결정되는 표본크기 공식에는 표본의 배정에 관한 공식 제공도 중요한 결정 중의 하나가 된다.

표본배정의 방법은 다양하게 존재한다. 배정의 목적이 모집단에 대한 대표성을 가지는 것이라면 비례배정을 생각한다. 층별 모집단의 비율에 맞추어서 배정을 하기 때문에 모집단의 축소판이라 할 수 있다. 그러나 모수 추정의 효율성 측면에서는 비례배정이 좋지 않다. 그럼에도 불구하고 크기 비례 배정방법은 쉽게 받아들여지는 개념을 가지고 있기 때문에 사회 및 경제조사에서 많이 사용될 수 있다. 또한 표본설계를 통해 추출된 표본은 다변량 관점에서의 조사가 이루어지기 때문에 설계단계에서 다변량 관점에서 유익한 배정을 고려하게 되는 데 그것을 부합하기도 한다. 물론 이것이 다변량 관점에서 효율이 가장 좋은 배정이라는 것을 의미하는 것은 아니다.

이론적으로 설계나 추정에서 중요한 변수에 대해서 효율성을 강조해야 한다면 비례보다는 네이만 배정이나 최적배정을 생각해야 한다. 그러나 이것은 층별 표준편차가 추정되어야 하기 때문에 추정에서 오는 효율의 감소가 올 수도 있다. 추정에서 오는 효율의 감소를 수치적으로 계산한 연구로 Evans (1951)와 Park 등 (2007)이 있다. 또한 이것은 한 변수에 대해서만 효율성을 고려하는 것이기 때문에 다변량 관점인 다른 변수들에

[1] Division of Data Science, Yonsei University, 1, Yeonsedae-gil, Heungeop-myeon, Wonju-si, Gangwon-do 26493, Korea.
E-mail: hapk@daum.net

Published 31 October 2021 / journal homepage: http://kjas.or.kr

대해서는 최적의 배정이라 할 수 없다. 하지만 실제적인 표본설계에서는 다변량 관점에서 표본설계를 하기보다는 중요변수를 선정하고 그것에 대한 효율을 증대시키는 방향으로 배정의 방법이 적용되는 것을 많이 사용하는 편이다.

전체 추정량의 오차가 가장 작은 최적배정 또는 네이만 배정은 모든 층에서의 각각의 오차가 관리되는 것은 아니다. 그래서 전체 추정량의 오차를 관리하는 것이 아니라 층별 추정량의 오차를 관리하는 입장에서는 제곱근 비례배정, 멱배정(power allocation)을 생각할 수 있다. 제곱근 비례배정은 실제조사에서도 비례배정의 대안으로 많이 사용되는데 그 이유는 층별 크기의 효과를 줄이는 배정을 고려하기 위해서이다. Bankier (1988)은 크기측도 또는 층의 중요도와 추정량의 상대표준오차를 사용한 손실함수(loss function)을 최소화하는 배정방법으로 멱배정을 연구하였다. 그리고 층별 상대표준오차를 관리하면서 전체 상대표준오차를 줄일 수 있는 배정에 관한 연구로 Park과 Shin (2020)이 있다.

이와 같이 표본설계의 여러 목적에 따라 배정의 방법이 다양하게 제안되고 있다. 그런데 지금까지의 배정은 무응답의 발생이 고려되지 않은 상황 하에 생각할 수 있는 방법들이다. 실제로 조사 과정에서 무응답의 발생은 거의 필수적으로 생기기 때문에 이전 조사에서 응답률을 반영하여 표본의 크기를 결정하기도 한다. 예를 들어 지난 조사에서 응답률이 50%이었다면 현시점에서 실제로 추출하는 표본의 크기보다 두 배의 표본을 생각하는 것과 같은 이치라 할 수 있다. 이와 같은 결정을 표본의 배정 과정에 적용하여 연구한 결과가 Lee와 Park (2015)가 있고 계속조사에서 전체 표본크기에 응답률 보정을 고려한 연구로 Park과 Na (2018)가 있다.

본 연구에서는 층별 응답확률은 일정하다는 가정하에 응답률을 반영한 추정량의 상대표준오차, 크기측도 또는 층의 중요도를 고려한 손실함수와 비용함수를 최소화하는 배정방법을 연구한다. 다시 말하면 Bankier (1988)의 연구에 응답률을 고려한 배정방법이라 할 수 있다. 표본배정은 표본크기가 주어져 있다고 가정하고 층별 표본을 나누는 방법이다. 역으로 제안된 배정방법이 주어져 있을 때 표본크기를 구하는 공식도 같이 연구한다. 이와 같은 이론적 연구를 뒷받침하기 위해 가상의 층이 있는 모집단을 생성하여 각 층별 응답률에 따라 표본을 배정하고 그 배정된 크기에 따라 추정량의 상대표준오차를 계산하는 것을 기존의 여러 배정방법들과 이 연구에서 제안된 배정방법들을 비교하는 모의실험을 실시함으로써 모집단의 특성에 따라 배정방법들의 장단점을 살펴본다.

2. 응답률을 반영한 멱배정

H개의 층과 각 층에서의 모집단의 크기는 N_h이고 그 층에서 단순임의추출(simple random sampling)되는 표본의 크기를 n_h가 되도록 하는 층화임의추출(stratified random sampling)을 생각한다.

각 층의 표본의 크기를 결정하기 위해서 추정량의 분산 및 비용을 최소화할 수 있거나 적당한 손실함수(loss function)를 고려하여 최소화할 수 있다. 추정량의 분산을 최소화하도록 표본배정을 실시한다는 것은 전체적인 측면에서의 효율 조정이라 볼 수 있고 손실함수를 최소화하도록 배정을 한다는 것은 층별 효율을 중요시하는 것이라 할 수 있다.

배정된 표본은 조사과정에서 무응답이 발생하게 되며 그것의 보정방법으로 조사가 끝난 후 모수 추정의 과정에서 추정의 가중치를 보정함으로써 적용된다. 그러나 실제로 표본의 수가 줄어든다는 것은 효율의 감소를 가져오는 것이 되므로 표본크기 결정단계에서 무응답에 관한 정보가 포함되도록 한다 (Lee와 Park, 2015; Park과 Na, 2018).

먼저, 각 층에서의 표본의 크기를 결정하기 위해 층별 모평균 $\mu_h = \sum_{i=1}^{N_h} y_{hi}/N_h$의 추정량으로

$$\bar{y}_{hR} = \left(\sum_{i=1}^{N_h} N_h n_h^{-1} p_h^{-1} t_{hi} r_{hi} \right)^{-1} \left(\sum_{i=1}^{N_h} N_h n_h^{-1} p_h^{-1} t_{hi} r_{hi} y_{hi} \right) \tag{2.1}$$

이며 이것은 근사 비편향성을 만족한다. 여기서 y_{hi}는 관심변수를 나타내고 $N_h n_h^{-1}$는 설계가중치에 해당하

며 p_h는 h층에서의 응답률로 개체별 응답확률을 고려하는 것이 아니라 층별 응답확률이 일정하다는 가정을 한다. 또한 t_{hi}는 h층에서 i번째 개체가 표본에 포함되는 여부를 나타내는 지시변수(indicator variable)로써 표본에 포함되면 1을 가지고 표본에 포함되지 않으면 0을 가진다. 그리고 r_{hi}는 h층에서 i번째 개체의 응답여부를 나타내는 변수로써 표본으로 추출된 개체가 응답을 하면 1을 가지고 응답하지 않으면 0을 가진다.

다음으로, 층별 추정량의 상대표준오차(relative standard error)를 조절하기 위해 고려되는 손실함수로

$$F = \sum_{h=1}^{H} \left(Z_h^q CV(\bar{y}_{hR}) \right)^2 \tag{2.2}$$

을 고려하는 데 이것은 Bankier (1988)가 제안한 형태와 비슷하다. 여기서 Z_h는 h층에서의 크기(size) 또는 중요도(importance)를 나타내고 추정량의 표준오차를 추정량의 평균으로 나눈 $CV(\bar{y}_{hR}) = \sqrt{V(\bar{y}_{hR})}/E(\bar{y}_{hR})$ 로 층별 추정량의 상대표준오차를 나타내며 상수인 q의 범위는 $0 \le q \le 1$이다. 또한 층별 추정량의 분산을 조절하기 위해 고려되는 손실함수로 식 (2.2)의 추정량의 상대표준오차 $CV(\bar{y}_{hR})$를 추정량의 표준오차 $\sqrt{V(\bar{y}_{hR})}$으로 교체하는 것을 생각하는 데 이것은 Yates (1960)가 제안한 손실함수를 변형한 형태이다.
그리고 배정을 위해서 적절한 비용함수를 생각하는데,

$$C = c_0 + \sum_{h=1}^{H} n_h c_h \tag{2.3}$$

로써 고정비용 c_0와 h층에서 개체별 비용(단위당 비용)인 c_h 등이 사용되어 총비용인 C가 결정된다.

전체 표본의 크기 $n = \sum_{h=1}^{H} n_h$이 주어져 있고 식 (2.3)이 고정되어 있을 때 식 (2.2)를 최소화하는 배정 또는 식 (2.2)가 고정되어 있을 때 식 (2.3)을 최소화하는 배정으로,

$$n_h = n \frac{Z_h^q CV_h / \sqrt{p_h c_h}}{\sum_{h=1}^{H} Z_h^q CV_h / \sqrt{p_h c_h}} \tag{2.4}$$

이며 여기서 h층에서의 모표준편차는 $S_h = \sqrt{\sum_{i=1}^{N_h} (y_{hi} - \mu_h)^2/(N_h - 1)}$이고 $CV_h = S_h/\mu_h$는 h층에서의 모집단의 변동계수(coefficient of variation)이다.

Theorem 1. *층별 응답률이 일정상수보다 클 경우 식* (2.1)*은*

$$\bar{y}_{hR} = \mu_h + N_h^{-1} \left(\sum_{i=1}^{N_h} N_h n_h^{-1} p_h^{-1} t_{hi} r_{hi} y_{hi} - \mu_h \sum_{i=1}^{N_h} N_h n_h^{-1} p_h^{-1} t_{hi} r_{hi} \right) + o_p\left(n_h^{-\frac{1}{2}} \right) \tag{2.5}$$

으로 전개할 수 있고 식 (2.2)*과 식* (2.3)*의 곱을 최소화하는 h층의 표본크기 n_h는,*

$$n_h \propto \frac{Z_h^q CV_h}{\sqrt{p_h c_h}} \tag{2.6}$$

이다.

Proof: $\bar{y}_{hR}$의 테일러 전개를 사용하여 2차항을 무시한다면

$$\bar{y}_{hR} = \mu_h + N_h^{-1} \left(\sum_{i=1}^{N_h} N_h n_h^{-1} p_h^{-1} t_{hi} r_{hi} y_{hi} - N_h \mu_h \right) - N_h^{-1} \mu_h \left(\sum_{i=1}^{N_h} N_h n_h^{-1} p_h^{-1} t_{hi} r_{hi} - N_h \right) + o_p\left(n_h^{-\frac{1}{2}} \right)$$

이고 이것을 정리하면 식 (2.5)가 유도된다.

다음으로 식 (2.5)를 가지고 $V(\bar{y}_{hR})$를 계산해 보면

$$\begin{aligned} V(\bar{y}_{hR}) &= V\left(E_R(\bar{y}_{hR})\right) + E\left(V_R(\bar{y}_{hR})\right), \\ &= V\left(n_h^{-1}\sum_{i=1}^{N_h} y_{hi}t_{hi}\right) + E\left(n_h^{-2}\left(p_h^{-1}-1\right)\sum_{i=1}^{N_h} t_{hi}^2(y_{hi}-\mu_h)^2\right), \\ &= \frac{S_h^2}{n_h}\left(1-\frac{n_h}{N_h}\right) + \frac{1}{n_h}\left(p_h^{-1}-1\right)\frac{\sum_{i=1}^{N_h}(y_{hi}-\mu_h)^2}{N_h} + o\left(n_h^{-1}\right), \end{aligned}$$

이다. 여기서 E_R과 V_R은 표본이 주어져 있을 때 응답여부에 관련된 기댓값과 분산이며 E와 V는 표본설계에 대한 기댓값과 분산이다. 만약 $N_h \approx N_h - 1$이 성립된다면

$$V(\bar{y}_{hR}) \simeq S_h^2\left(\frac{1}{n_h p_h} - \frac{1}{N_h}\right) \tag{2.7}$$

가 나오며 이것은 Lee와 Park (2015)의 식 (2.8)에도 언급되어 있다.
배정공식을 유도하기 위해 식 (2.2)에 식 (2.7)를 넣으면

$$F = \sum_{h=1}^{H}\left(Z_h^q CV(\bar{y}_{hR})\right)^2 = \sum_{h=1}^{H} Z_h^{2q}\left(\frac{S_h}{\mu_h}\right)^2\left(\frac{1}{n_h p_h} - \frac{1}{N_h}\right)$$

이고 식 (2.3)과 더불어 Cauchy-Schwarz 부등식을 이용하면

$$\left(F + \sum_{h=1}^{H}\frac{Z_h^{2q}CV_h^2}{N_h}\right)(C-c_0) = \sum_{h=1}^{H}\frac{Z_h^{2q}CV_h^2}{n_h p_h}\sum_{h=1}^{H} n_h c_h \geq \left(\sum_{h=1}^{H} Z_h^q CV_h\sqrt{\frac{c_h}{p_h}}\right)^2,$$

이며 손실함수와 비용함수의 곱의 최솟값을 유도할 수 있다.
그리고 이 최솟값이 나오게 되는 조건은 임의의 상수 k에 대해서

$$\frac{Z_h^q CV_h/\sqrt{n_h p_h}}{\sqrt{n_h c_h}} = k$$

일 경우이다. 따라서 최종적으로 식 (2.6)이 유도되며 $n = \sum_{h=1}^{H} n_h$의 조건하에 식 (2.4)의 배정공식이 나오게 된다. □

식 (2.4)의 배정공식은 추정량의 분산을 최소화하는 목적을 가진 것이 아니라 층별 추정의 효율을 조절하려는 목적을 지니고 있다. 식 (2.4)를 여러 조건에 대해 분리해 보면 다음과 같은 배정공식을 유도할 수 있다.

그리고 식 (2.2)의 추정량의 상대표준오차 $CV(\bar{y}_{hR})$를 추정량의 표준오차인 $\sqrt{V(\bar{y}_{hR})}$로 교체하는 것을 생각하면 비용과 손실함수를 고려한 최적배정은,

$$n_h = n\frac{Z_h^q S_h/\sqrt{p_h c_h}}{\sum_{h=1}^{H} Z_h^q S_h/\sqrt{p_h c_h}} \tag{2.8}$$

가 된다.

$Z_h = W_h = N_h/\sum_{h=1}^{H} N_h$로 크기를 나타내고 단위당 비용 $c_h = c$로 일정하면 식 (2.4)는

$$n_h = n\frac{N_h^q CV_h/\sqrt{p_h}}{\sum_{h=1}^{H} N_h^q CV_h/\sqrt{p_h}} \tag{2.9}$$

가 되며 Bankier (1988)의 멱배정(power allocation)에 응답률이 반영된 배정공식이 된다. 이와 더불어 식 (2.8)은 $Z_h = W_h = N_h / \sum_{h=1}^{H} N_h$로 크기를 나타내고 단위당 비용 $c_h = c$로 일정하면

$$n_h = n \frac{N_h^q S_h / \sqrt{p_h}}{\sum_{h=1}^{H} N_h^q S_h / \sqrt{p_h}}, \tag{2.10}$$

가 되고 q의 범위는 $0 \le q \le 1$이다. 또한, $Z_h = W_h = N_h / \sum_{h=1}^{H} N_h$로 크기를 나타내고 단위당 비용과 모변동계수가 $c_h = c, CV_h = CV$로 일정하면

$$n_h = n \frac{N_h^q / \sqrt{p_h}}{\sum_{h=1}^{H} N_h^q / \sqrt{p_h}}, \tag{2.11}$$

이며 이것은 $q = 1/2$인 경우 제곱근비례배정공식에 응답률이 반비례하게 반영된 공식이 된다.

표본배정은 표본크기 n이 주어져 있을 때 결정되는 데 실제로 표본크기가 사전에 결정되어야 한다. 표본배정방법이 주어져 있을 때 두 가지 방향으로 표본크기 공식이 정해질 수 있다.

첫째, 총비용이 주어져 있을 경우 즉 $C = C'$인 경우 식 (2.4)를 사용한 표본크기 공식은

$$n = (C' - c_0) \frac{\sum_{h=1}^{H} \left(Z_h^q CV_h / \sqrt{p_h c_h} \right)}{\sum_{h=1}^{H} \left(Z_h^q CV_h \sqrt{c_h} / \sqrt{p_h} \right)},$$

이며 둘째, 손실함수가 일정한 값 $F = F'$라면 식 (2.4)를 사용한 표본크기 공식은

$$n = \frac{\left(\sum_{h=1}^{H} Z_h^q CV_h / \sqrt{p_h c_h} \right) \sum_{h=1}^{H} \left(Z_h^q CV_h \sqrt{c_h} / \sqrt{p_h} \right)}{F' + \sum_{h=1}^{H} Z_h^{2q} CV_h^2 / N_h}$$

이다.

Remark 1. 응답률, 단위당 비용 $p_h = p,\ c_h = c$로 일정하면 식 (2.4)는 Bankier (1988)의 식 (2.2)와 같다. 모변동계수, 응답률, 단위당 비용이 $CV_h = CV,\ p_h = p,\ c_h = c$로 일정하며 $Z_h = W_h = N_h / \sum_{h=1}^{H} N_h$로 크기를 나타내면 식 (2.4)는

$$n_h = n \frac{N_h^q}{\sum_{h=1}^{H} N_h^q},$$

로 단순화 되며 여기서 $q = 1$인 경우가 비례배정이며 $q = 1/2$인 경우가 제곱근비례배정이 된다.

Remark 2. 응답률 $p_h = p$로 일정하며 $q = 1$이고 $Z_h = W_h = N_h / \sum_{h=1}^{H} N_h$로 크기를 나타내면 식 (2.8)은 최적배정을 나타낸다. 그리고 단위당 비용이 $c_h = c$로 일정하며 $q = 1$이고 $Z_h = W_h = N_h / \sum_{h=1}^{H} N_h$로 크기를 나타내면 식 (2.8)은 Lee와 Park (2015)의 식 (2.6)을 나타내며 여기에 더하여 모표준편차가 $S_h = S$로 일정하면 Lee와 Park (2015)의 식 (2.7)을 나타낸다. 단위당 비용이 $c_h = c$로 일정하며 $q = 1/2$이고 $Z_h = W_h = N_h / \sum_{h=1}^{H} N_h$로 크기를 나타내면 식 (2.4)는 Lee와 Park (2015)의 식 (2.9)를 나타낸다.

3. 모의실험

가상의 층의 개수는 8개로 하며 각 층별 정규분포를 가정하고 1,838,000개의 유한모집단을 발생하여 모의실험을 위한 표본의 크기는 5,000개와 500개를 고려한다. Table 1의 층별 정규분포에서 추출한 가상 모집단 구조를 살펴보면 stratum 6을 큰 층으로 stratum 7, 8을 작은 층으로 생각하며 층화의 타당성을 위해 층별 모평균을 다르게 고려한다. 또한 층별 모표준편차가 차이가 나도록 하되 층이 큰 곳과 층이 작은 곳에서 산포가 다른 층에 비해 큰 것을 가정하고 평균적 응답률은 0.56이며 층별 응답률의 차이가 있도록 생각한다.

Table 1: Values of population 1

Value	Stratum							
	1	2	3	4	5	6	7	8
Size	200,000	280,000	115,000	112,000	190,000	820,000	81,000	40,000
Mmean	450.1	490.0	499.9	520.0	530.2	620.0	740.3	839.0
Standard deviation	64.0	55.1	94.9	98.7	85.9	104.9	149.9	156.0
Response rate	0.30	0.50	0.55	0.60	0.65	0.40	0.70	0.75

배정을 위해서 층별 모집단의 크기 또는 모표준편차 및 모변동계수의 정보 또는 층별 응답률의 정보를 사용한다. 본 논문에서는 모의실험에서 비교하는 배정의 방법으로 12가지 방법을 비교하는 데 첫째, 비례배정을 사용한 것으로 “A1”의 기호를 사용하고 둘째, 네이만배정을 사용한 것으로 “A2”의 기호를 사용하고 제곱근 비례배정을 사용한 것을 “A3”로 하며 세제곱근 비례배정을 사용한 것을 “A4”로 한다. 그리고 식 (2.11)에서 $q = 1$인 경우를 “A5”의 기호를 사용하고 식 (2.10)에서 $q = 1$인 경우를 “A6”를 이용한다. 또한 본 논문에서 제안된 배정방법에 대한 기호를 다음과 같이 정리한다.

- A7 : 식 (2.11)의 배정공식에서 $q = 1/2$인 경우
- A8 : 식 (2.11)의 배정공식에서 $q = 1/3$인 경우
- A9 : 식 (2.10)의 배정공식에서 $q = 1/2$인 경우
- A10 : 식 (2.10)의 배정공식에서 $q = 1/3$인 경우
- A11 : 식 (2.9)의 배정공식에서 $q = 1/2$인 경우
- A12 : 식 (2.9)의 배정공식에서 $q = 1/3$인 경우

단, 여기서 $A5, A6, A11$의 기호를 사용하는 배정공식은 Lee와 Park (2015)에서 제안된 것이다. 추정량의 형태는 층별로는 식 (2.1)이며 전체 추정량은

$$\bar{y}_R = \left(\sum_{h=1}^{H} \sum_{i=1}^{N_h} N_h n_h^{-1} p_h^{-1} t_{hi} r_{hi} \right)^{-1} \left(\sum_{h=1}^{H} \sum_{i=1}^{N_h} N_h n_h^{-1} p_h^{-1} t_{hi} r_{hi} y_{hi} \right)$$

을 사용한다. 5,000개와 500개의 표본을 12개의 배정공식에 의해 배분된 표본크기를 가지고 층별 가상의 모집단에서 표본을 추출하여 추정량의 값을 계산한다. 단 표본배정공식에 의해 나온 표본크기는 반올림하여 정수화한다.

이와 같은 추정량의 값을 5,000개 발생하여 다음과 같은 공식에 의해 배정방법들의 효율을 계산한다.

$$\text{MSE} = \frac{\sum_{i=1}^{5000} (\text{estimator}_i - \text{Population mean})^2}{5000}. \tag{3.1}$$

단, 이 여기서 Table 1의 전체 모평균은 568.9이며 Table 4의 전체 모평균은 535.7이다.

Table 2와 Table 5는 배정을 위한 총표본크기는 5,000이며 괄호는 500의 표본크기를 나타내는 데 배정공식들에 의해 층별 표본크기를 살펴보면 5,000 표본크기의 거의 10%에 해당하는 것이 500의 표본크기이다. 또한, Table 3과 Table 6은 배정방법에 따른 MSE를 계산한 것인데 총 표본크기는 5,000이며 괄호에 있는 것이 총 표본크기 500에 대한 MSE를 나타낸다.

Table 2를 먼저 살펴보면 비례배정인 A1은 가장 큰 층인 stratum 6에 가장 많이 배정되며 크기가 작은 층인 stratum 7, 8에 가장 적게 배정되었다. 최적배정인 A2는 stratum 6, 7, 8의 모표준편차가 크기 때문에 비례보다 더 많이 배정되었으며 stratum 1은 모표준편차가 작기 때문에 비례보다 작은 수가 배정되었다. 제곱근

Table 2: Sample size for allocation methods (population 1)

Allocation	Stratum							
	1	2	3	4	5	6	7	8
A1	544	762	313	305	517	2231	220	109
	(54)	(76)	(31)	(30)	(52)	(223)	(22)	(11)
A2	375	451	318	324	478	2515	355	184
	(38)	(45)	(32)	(32)	(48)	(252)	(35)	(18)
A3	643	761	488	481	627	1302	409	288
	(64)	(76)	(49)	(48)	(63)	(130)	(41)	(29)
A4	651	728	541	536	640	1042	482	381
	(65)	(73)	(54)	(54)	(64)	(104)	(48)	(38)
A5	667	724	283	264	431	2369	177	84
	(67)	(72)	(28)	(26)	(43)	(237)	(18)	(8)
A6	464	433	291	284	402	2696	287	144
	(46)	(43)	(29)	(28)	(40)	(270)	(29)	(14)
A7	817	749	457	432	541	1432	340	231
	(82)	(75)	(46)	(43)	(54)	(143)	(34)	(23)
A8	837	726	514	488	559	1161	406	310
	(84)	(73)	(51)	(49)	(56)	(116)	(41)	(31)
A9	565	445	467	461	501	1620	550	391
	(57)	(45)	(46)	(46)	(50)	(162)	(55)	(39)
A10	571	426	518	514	512	1296	647	517
	(57)	(43)	(52)	(52)	(51)	(130)	(65)	(52)
A11	717	519	534	506	540	1493	424	267
	(72)	(52)	(53)	(51)	(54)	(149)	(42)	(27)
A12	729	499	595	567	554	1200	502	354
	(73)	(50)	(59)	(57)	(55)	(120)	(50)	(35)

배례배정인 A3와 세제곱근 비례배정인 A4의 표본배정은 골고루 배정의 효과가 있기에 stratum 6의 표본이 반으로 줄어들고 다른 층들의 표본수가 늘어나는 효과가 발생하였다. A5에서부터 A12까지의 배정방법들은 층별 응답률이 반영되었으며 $A5, A7, A8$ 배정방법은 층의 크기와 응답률이 반영된 배정으로 stratum 1이 응답률이 작은 것이 반영되어 A1에서부터 A4까지의 배정방법보다 표본크기가 크며 $A7, A8$의 배정방법이 A5 배정방법보다 골고루 표본이 배분되어 stratum 1에서 큰 표본수를 가졌다. 추가적으로 $A6, A9, A10, A11, A12$의 표본수는 층의 산포가 반영되므로 stratum 1에서 크기와 응답률이 반영된 $A7, A8$의 배정방법에 따른 표본수보다 작게 나타난다. $A7, A8$의 배정방법은 층의 크기가 반영되지만 제곱근으로 또는 세제곱근으로 반영되어 층의 표본크기가 골고루 배분된 효과가 발생하며 층별 응답률이 반영되어 응답률이 적은 곳에 더 많이 배정되는 효과가 추가적으로 적용됨으로 층별 고르게 배정하는 효과가 있는 $A3, A4$의 배정방법에 비해 정보의 적용이 더 많이 된 것이다. 또한 $A9, A10, A11, A12$의 배정방법은 층별 크기의 제곱근 및 세제곱근과 응답률 뿐만 아니라 층별 산포의 정보도 반영되어 표본이 배정된 것을 알 수 있으며 그것의 효과가 전체 층에서 반영되었으며 특히 stratum 1과 stratum 2를 보면 층의 산포가 다른 층에 비해 작아서 $A7, A8$ 배정방법에 비해 표본의 크기가 작아짐을 확인할 수 있었다.

이와 같은 표본배정의 결과를 가지고 5,000번의 모의 실험을 통해 계산된 MSE를 Table 3을 통해 살펴보면 먼저 전체에 대한 MSE를 보면 비례배정인 A1의 효과가 네이만배정인 A2의 것보다 약간 좋음을 표본크기 5,000에서 확인할 수 있으나 표본크기 500에서는 역으로 발생됨을 알 수 있었다. 이중 응답률을 반영한 네이만 배정인 A6가 표본크기 5,000과 500에서 가장 작은 MSE를 가진다. 하지만 층안으로 들어가면 MSE가 큰 곳이

Table 3: MSE according to allocation methods (population 1)

Allocation	Stratum								Total
	1	2	3	4	5	6	7	8	
A1	26.34	8.15	52.98	53.52	21.82	12.22	141.53	292.16	5.45
	(263.42)	(82.04)	(513.90)	(568.53)	(221.73)	(122.39)	(1549.11)	(3095.31)	(58.24)
A2	36.64	13.81	52.35	50.91	23.44	10.92	91.76	174.08	5.76
	(389.05)	(138.63)	(500.88)	(528.51)	(240.30)	(116.33)	(924.67)	(1884.42)	(57.85)
A3	21.23	7.78	33.30	34.00	17.46	20.97	77.20	113.73	6.83
	(228.78)	(81.79)	(339.03)	(349.42)	(183.84)	(213.41)	(823.09)	(1145.68)	(68.50)
A4	20.64	8.49	29.80	30.48	17.53	26.79	64.44	84.09	8.12
	(217.57)	(81.33)	(306.89)	(310.69)	(180.37)	(265.07)	(655.28)	(867.09)	(79.20)
A5	20.87	8.70	57.08	61.62	25.96	11.69	181.11	399.20	5.52
	(222.06)	(86.36)	(600.24)	(632.17)	(274.07)	(118.15)	(1882.64)	(4307.70)	(55.72)
A6	28.52	14.00	54.36	57.75	28.00	10.29	111.24	225.32	5.33
	(311.31)	(148.65)	(560.64)	(609.21)	(291.12)	(104.00)	(1103.25)	(2499.27)	(54.82)
A7	16.97	8.22	34.55	38.97	21.44	18.50	93.89	140.43	6.10
	(177.09)	(85.41)	(354.74)	(389.32)	(212.55)	(195.43)	(961.00)	(1424.69)	(64.77)
A8	16.61	8.32	31.45	33.48	20.23	24.03	80.04	106.70	7.26
	(170.52)	(85.22)	(323.58)	(336.95)	(204.95)	(247.08)	(788.69)	(1114.98)	(74.36)
A9	24.17	13.30	34.15	36.48	22.63	16.85	55.84	87.41	6.31
	(259.00)	(138.79)	(362.20)	(358.39)	(229.78)	(175.37)	(592.96)	(860.44)	(65.08)
A10	24.37	14.45	32.59	32.21	21.63	21.09	47.98	61.02	7.25
	(253.70)	(146.79)	(300.12)	(315.93)	(229.81)	(212.80)	(476.36)	(630.47)	(71.10)
A11	18.76	11.63	30.56	32.11	21.07	18.25	74.39	123.50	6.15
	(200.15)	(120.38)	(306.69)	(321.29)	(214.50)	(186.09)	(758.50)	(1236.75)	(62.13)
A12	18.38	12.02	27.90	28.41	20.45	22.41	62.49	94.85	7.32
	(199.15)	(126.61)	(278.42)	(288.81)	(215.65)	(233.39)	(664.90)	(985.01)	(72.09)

Table 4: Values of population 2

Value	Stratum							
	1	2	3	4	5	6	7	8
Size	300,000	380,000	215,000	212,000	290,000	320,000	81,000	40,000
Mean	450.0	490.0	500.0	520.1	530.0	620.0	740.1	839.5
Standard deviation	64.0	55.0	95.2	99.2	85.8	104.9	149.7	155.6
Response rate	0.30	0.50	0.55	0.60	0.65	0.40	0.70	0.75

나타나는 데 A9와 A10의 배정은 층의 크기가 작은 stratum 7, 8에서 작은 MSE가 나타남을 확인할 수 있다.

전체적 MSE를 살펴볼 때 A10의 배정보다는 A9의 배정이 더 낮게 나타나며 이것은 $A1, A2, A5, A6, A7, A11$ 배정방법 보다는 높다고 할 수 있으나 다른 배정방법들 보다는 낮기에 A9 배정의 전체 MSE가 잘 조절된다고 볼 수 있다. 또한 층의 응답률이 제일 작은 stratum 1을 살펴보면 $A7, A8$의 배정에서의 MSE가 가장 작게 나타나며 그 다음으로 A11과 A12의 배정에서 가장 작게 나타난다. 이것을 살펴보면 멱배정 기법에 응답률을 사용한 공식이 층안에서 대체적으로 MSE를 잘 조절하고 있음을 확인할 수 있었다.

또 다른 모의실험을 위해 Table 4는 모평균,모표준편차,응답률은 Table 1과 비슷한 것을 생각하고 stratum의 크기에 대해 stratum 6의 크기를 다른 큰 층들과 비슷하게 하고 stratum 7과 stratum 8은 작은 층을 고려한다.

Table 5의 배정공식에 따른 표본크기를 차례대로 살펴보면 비례배정인 A1은 층의 크기에 따라 표본이

Table 5: Sample size for allocation methods (population 2)

Allocation	Stratum							
	1	2	3	4	5	6	7	8
A1	816	1034	585	577	789	871	220	109
	(82)	(103)	(58)	(58)	(79)	(87)	(22)	(11)
A2	606	660	646	664	786	1060	383	197
	(61)	(66)	(65)	(66)	(79)	(106)	(38)	(20)
A3	745	838	631	626	732	769	387	272
	(74)	(84)	(63)	(63)	(73)	(77)	(39)	(27)
A4	711	769	636	633	703	726	459	363
	(71)	(77)	(64)	(63)	(70)	(73)	(46)	(36)
A5	1031	1011	545	515	677	952	182	87
	(103)	(101)	(55)	(51)	(68)	(95)	(18)	(9)
A6	779	657	613	603	686	1180	322	160
	(78)	(66)	(61)	(60)	(69)	(118)	(32)	(16)
A7	957	834	598	569	639	856	326	221
	(96)	(83)	(60)	(57)	(64)	(86)	(33)	(22)
A8	921	771	608	580	618	815	389	297
	(92)	(77)	(61)	(58)	(62)	(81)	(39)	(30)
A9	683	512	635	629	612	1001	544	384
	(68)	(51)	(64)	(63)	(61)	(100)	(54)	(38)
A10	640	461	630	625	577	929	634	503
	(64)	(46)	(63)	(63)	(58)	(93)	(63)	(50)
A11	843	580	706	672	641	897	408	254
	(84)	(58)	(71)	(67)	(64)	(90)	(41)	(25)
A12	802	531	710	678	614	845	483	338
	(80)	(53)	(71)	(68)	(61)	(84)	(48)	(34)

배정되므로 stratum 2가 가장 많고 stratum 8이 가장 적게 배정되었다. 네이만배정인 A2는 stratum $1,2$의 표준편차가 작아서 표본크기가 비례보다 줄어들고 다른 층들은 대체적으로 표본크기가 비례보다 증가한다. 제곱근 및 세제곱근 비례배정인 $A3, A4$은 비례와 비교해 보면 상대적으로 층의 크기가 큰 stratum $1,2,5,6$은 표본크기가 감소하고 상대적으로 크기가 작은 층들은 표본크기가 증가한다. 응답률이 반영된 비례배정인 A5를 비례배정과 비교하면 응답률이 작은 곳의 표본크기가 증가하는 데 그 층이 stratum $1,6$이며 stratum $7,8$은 층의 크기가 작고 응답률이 상대적으로 높아서 표본크기가 줄어들었다. 응답률이 반영된 네이만배정인 A6은 A5 배정과 비교해 보면 표준편차가 작은 곳인 stratum$1,2$에서 표본크기가 감소한다. $A3, A4$ 배정방법과 $A7, A8$ 배정방법의 차이는 배정공식에 응답률의 포함 여부로써 응답률이 상대적으로 작은 층인 stratum $1,6$에서는 A3 배정방법보다 A7 배정방법의 표본크기가 더 크고 상대적으로 응답률이 큰 층인 straum $7,8$에서는 반대의 현상이 보인다. 이것은 A4의 배정과 A8의 배정의 비교에도 같은 현상이 발생한다. $A9, A10$의 배정방법은 층의 산포가 작은 곳인 straum $1,2$에서는 $A7, A8$배정방법에 따른 표본크기보다 줄어드는 경향이 있고 층의 산포가 큰 곳인 stratum $6,7,8$에서는 $A7, A8$의 배정방법과 비교해 보면 층의 크기가 커지는 현상이 발생한다. 이것은 $A7, A8$의 배정공식에 층별 모표준편차가 반영된 현상이라 할 수 있다. $A11, A12$의 배정방법은 $A7, A8$의 배정공식에 층별 상대표준오차가 반영되었으므로 층별 상대표준오차가 작은 곳인 stratum $1,2$에서는 표본크기가 $A7, A8$ 배정방법에 따른 표본크기보다 줄어드는 현상이 발생한다.

Table 5의 표본크기 공식을 종합해서 살펴보면 stratum $1,2$를 살펴보면 비례배정인 A1의 표본크기보다

Table 6: MSE according to allocation methods (population 2)

Allocation	stratum								Total
	1	2	3	4	5	6	7	8	
A1	17.40	5.68	28.25	28.90	14.50	31.71	144.07	301.23	4.64
	(170.23)	(56.74)	(287.87)	(278.59)	(143.69)	(321.52)	(1466.01)	(2998.94)	(44.95)
A2	22.50	9.20	25.60	24.16	14.14	25.62	87.60	167.57	4.26
	(236.86)	(92.32)	(245.33)	(250.56)	(145.98)	(268.10)	(839.58)	(1689.46)	(43.88)
A3	18.05	7.06	26.15	26.50	15.62	37.39	81.73	118.58	4.60
	(189.57)	(73.18)	(262.11)	(264.75)	(158.40)	(351.45)	(803.17)	(1213.12)	(45.17)
A4	19.46	7.75	25.67	26.58	15.62	37.49	70.13	85.08	4.80
	(199.74)	(80.09)	(259.37)	(265.64)	(166.23)	(378.62)	(698.91)	(890.07)	(45.98)
A5	12.67	5.89	31.19	31.80	16.13	27.51	175.70	374.03	4.28
	(134.88)	(60.66)	(298.42)	(319.58)	(166.03)	(305.47)	(1868.19)	(3837.56)	(44.76)
A6	17.14	9.47	26.91	26.94	16.38	24.26	97.63	199.56	4.24
	(177.87)	(91.79)	(262.88)	(273.35)	(167.35)	(236.26)	(1007.70)	(2171.50)	(42.16)
A7	14.17	7.06	28.84	27.72	17.99	31.89	97.24	144.01	4.18
	(145.16)	(76.77)	(288.13)	(283.42)	(179.43)	(324.39)	(987.10)	(1488.33)	(43.35)
A8	14.69	8.02	26.97	28.84	18.17	34.72	81.49	106.40	4.39
	(149.73)	(78.66)	(275.13)	(288.25)	(183.80)	(338.57)	(813.78)	(1079.43)	(44.15)
A9	19.99	11.76	26.47	26.30	19.39	26.91	59.40	85.66	4.41
	(207.65)	(121.06)	(265.78)	(260.76)	(184.52)	(285.79)	(598.20)	(877.56)	(44.45)
A10	21.85	12.43	27.21	26.07	19.72	28.13	49.72	66.79	4.58
	(226.69)	(131.79)	(261.27)	(264.17)	(192.31)	(305.01)	(506.03)	(637.37)	(46.64)
A11	15.83	10.81	22.80	24.79	17.94	30.53	78.32	125.74	4.19
	(158.44)	(106.17)	(240.90)	(252.54)	(172.04)	(306.41)	(793.99)	(1303.91)	(42.21)
A12	16.98	11.24	22.70	24.17	18.56	31.68	66.44	95.16	4.41
	(176.96)	(115.39)	(235.84)	(237.95)	(194.78)	(329.89)	(681.16)	(959.51)	(44.99)

Table 7: Values of population 3

Value	Stratum							
	1	2	3	4	5	6	7	8
Size	300,000	380,000	215,000	212,000	290,000	320,000	81,000	40,000
Proportion	0.15	0.22	0.19	0.05	0.11	0.17	0.56	0.59
Response rate	0.30	0.50	0.55	0.60	0.65	0.40	0.70	0.75

네이만배정의 A2가 더 표본크기가 작은 것은 층의 표준편차가 다른 층에 비해 작기 때문이며 stratum 1의 응답률이 작기 때문에 응답률이 반영된 A5의 표본크기가 가장 크게 나타난다. stratum $7,8$은 다른 층에 비해 층의 크기가 작아서 비례배정인 A1의 표본크기가 작은 데 비해 층의 산포가 반영된 네이만 배정인 A2는 그 층의 모표준편차가 크기 때문에 표본의 크기가 커지며 제곱근 크기 및 세제곱근 크기가 반영된 배정인 $A3, A4$도 표본의 크기가 커진다. 이와 같은 성질은 응답률이 반영된 곳에서도 적용되어 층의 제곱근 크기 및 세 제곱근 크기 그리고 층의 표준편차등이 반영된 A7에서 A12까지의 배정방법에 따른 표본이 A1 배정과 비교하여 stratum $7,8$에서 크게 나타난다.

Table 6은 Table 5의 표본크기로 표본을 뽑아서 계산된 추정량의 값들의 산포 즉 MSE로 전체 층에서 가장 작은 것을 살펴보면 A7과 A11 배정방법이며 네이만배정의 영향에 있는 A2와 A6 보다 작다. 그리고 제일 큰 MSE를 가진 곳은 A4 배정방법이다. 이와 반대로, 층마다는 제일 작은 MSE를 가지는 곳이 다른 데 각 층에서

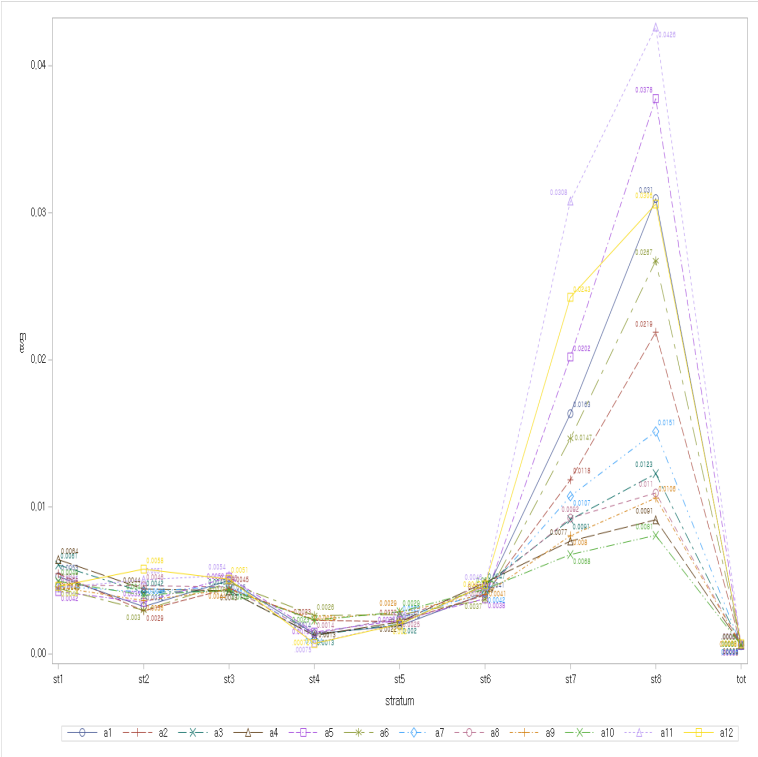

Figure 1: *Chart of MSE (population 3, samplesize=500).*

배정방법에 따라 표본크기가 많이 배정된 곳이 가장 작은 MSE를 가짐을 확인할 수 있었다. 예를 들어 층의 크기가 다른 층들에 비해 작은 stratum $7, 8$에서는 $A9, A10$ 배정방법을 사용한 추정량의 MSE가 다른 배정들에 비해 작게 나타나는데 그 이유는 다른 배정기법들과 비교해서 배정된 표본수가 많기 때문이다. 그 다음으로 층의 크기가 작은 stratum $3, 4$에서는 $A11, A12$의 MSE가 다른 배정기법들과 비교하여 작게 나타난다. 층의 크기가 상대적으로 큰 층인 stratum $1, 2, 5, 6$에서는 응답률을 사용한 비례배정인 A5가 대략적으로 MSE가 작게 나타난다.

자료를 구분하면 연속형과 범주형으로 구분할 수 있다. 지금까지 연속형 자료에 대하여 배정기법들을 비교하였으며 범주형자료에 대한 분석을 위해 층별 일정 모비율에 대한 베르누이 분포를 가정한다. 예를 들어 범죄율이나 특정 질병에 대한 유병율과 같이 낮은 비율을 추정하려고 할 때의 가상집단을 위해 Table 7을 생각한다. 전체 모비율이 18.3%가 되도록 층별 대체적으로 작은 모비율을 가진 모집단으로 Table 7을 고려하며 모집단의 크기와 응답률은 Table 4와 같은 것을 설정한다. 배정방법에 따른 MSE를 계산하기 위해 식 (3.1)의 population mean은 Table 7에서의 proportion과 전체 집단의 모비율 0.183을 사용한다. 표본의 크기가 500인 경우의 모의 실험 결과를 그래프로 나타낸 Figure 1에서 st1부터 st8은 8개의 층을 나타내고 tot는 전체를 나타내고 a1이 A1을 a2가 A2 등등을 나타낼 때 배정기법인 A1에서 A12까지 MSE를 살펴보면 층의 크기가 작은 층인 층7과 층8에서 A10 배정기법의 MSE가 가장 작게 나타나며 A9와 A10 배정기법의 MSE가 모든 층에서 대체적으로 작은 편에 속해있음을 확인할 수 있으며 전체에서의 MSE도 작은 편에 속해 있음을 확인할 수 있었다.

4. 결론

전체 층을 조절하기 위해서는 층의 크기 정보와 층별 모표준편차와 층별 응답률이 반영된 A6의 배정방법이 작은 MSE를 가짐을 확인할 수 있었다. 그러나 층 안으로 들어갔을 때는 최적의 개념과 응답률이 반영된 A6의 배정방법의 효율이 좋지 않은 층이 발생하였다. 그래서 층마다 가장 좋은 배정방법을 찾았을 때 일관성은 없었으나 한가지 층이 작은 곳에서는 A9와 A10의 배정방법의 MSE가 작음을 확인할 수 있었다.

본 연구는 전체적인 집단에서의 효율은 약간 떨어질 수 있으나 조사의 목적을 전체 층에서의 효율과 층별 효율을 적절하게 맞추고자 하는 것과 무응답 여부를 반영하고자 층별 응답률을 반영한 여러 멱배정기법을 제공함으로써 전체 효율과 층안에서의 효율을 증대시키는 기법을 찾아보고자 하였다. 전체적으로 일정부분 작아지는 배정기법을 찾지는 못했지만 제안되는 배정기법들이 층의 크기가 작은 곳에서는 효과가 있음을 확인할 수 있었다. 또한 연구의 내용을 살펴볼 때 층별 효과가 좋은 배정기법이 층마다 다른 것에 착안하여 향후 연구과제로는 모든 층에 대해 동일한 배정방법을 적용하기보다는 여러 배정을 결합하여 적용하는 방법을 생각해 보고자 한다.

References

Bankier MD (1988). Power allocations: Determining sample sizes for subnational areas, *The American Statistician*, **42**, 174–177.

Evans WD (1951). On stratification and optimum allocations, *Journal of the American Statistical Association*, **46**, 95–104.

Lee I and Park M (2015). A study on sample allocation for stratified sampling, *The Korean Journal of Applied Statistics*, **28**, 1047–1061.

Park H, Na S, and Jeon J (2007). Compromise allocation in univariate stratified sampling, *Communications in Statistics-Theory and Methods*, **36**, 265–271.

Park H and Na S (2018). Sample size using response rate on repeated surveys, *The Korean Journal of Applied Statistics*, **31**, 587–597.

Park J and Shin KI (2020). A Study on determination of stratum sample size in stratified sample design, *The Korean Official Statistics*, **25**, 32–53.

Yates F (1960). *Sampling methods for censuses and surveys*, London: Charles W. Griffin.

Received April 20, 2021; Revised June 11, 2021; Accepted July 5, 2021

층별 응답률을 사용한 멱배정 방법의 확장

박현아[1,a]

[a]연세대학교 데이터사이언스학부

요 약

멱배정의 전체 효율은 최적배정보다 감소하지만, 층별 표본을 골고루 배정하는 기법이며 조사 현장에서는 제곱근 비례배정으로 많이 사용된다. 그리고 실제조사에서는 무응답을 고려하여 이론적 공식에서 나오는 것보다 더 많은 표본크기를 추출한다. 본 연구에서는 층별 표본크기를 결정함에 있어서 멱배정 방법에 층별 응답률의 정보를 추가하여 배정하는 기법을 연구한다. 제안된 배정기법들을 비례,최적,제곱근 비례 배정과 비교하며 응답률이 추가된 비례, 최적배정과도 비교하는 것을 모의실험을 통해 살펴봄으로써 배정방법들의 장단점을 살펴본다.

주요용어: 멱배정, 응답률, 층, 표본크기

[1](26493) 강원도 원주시 연세대길 1, 연세대학교 데이터사이언스학부. E-mail: hapk@daum.net