The Korean Journal of Applied Statistics (2015) DOI: http://dx.doi.org/10.5351/KJAS.2015.28.6.1047
28(6), 1047–1061

A Study on Sample Allocation for Stratified Sampling

Ingue Lee[a] · Mingue Park[a,1]

[a]Department of Statistics, Korea University

(Received July 9, 2015; Revised September 1, 2015; Accepted October 19, 2015)

Abstract

Stratified random sampling is a powerful sampling strategy to reduce variance of the estimators by incorporating useful auxiliary information to stratify the population. Sample allocation is the one of the important decisions in selecting a stratified random sample. There are two common methods, the proportional allocation and Neyman allocation if we could assume data collection cost for different observation units equal. Theoretically, Neyman allocation considering the size and standard deviation of each stratum, is known to be more effective than proportional allocation which incorporates only stratum size information. However, if the information on the standard deviation is inaccurate, the performance of Neyman allocation is in doubt. It has been pointed out that Neyman allocation is not suitable for multi-purpose sample survey that requires the estimation of several characteristics. In addition to sampling error, non-response error is another factor to evaluate sampling strategy that affects the statistical precision of the estimator. We propose new sample allocation methods using the available information about stratum response rates at the designing stage to improve stratified random sampling. The proposed methods are efficient when response rates differ considerably among strata. In particular, the method using population sizes and response rates improves the Neyman allocation in multi-purpose sample survey.

Keywords: stratified sampling, sample allocation, Neyman allocation, non-response

1. 서론

층화표본추출(stratified sampling)은 모집단을 구성하는 층에 대한 정보를 표본설계에 반영함으로써 추정량의 분산을 낮추기 위한 표본추출 방법이다. 따라서 전체 표본크기에 대한 각 층별 배분방안의 선택이 표본설계의 핵심 요소라 할 수 있다. 특히 조사비용 또는 조사기간의 제약 등으로 표본 크기의 결정이 제한적일 때에는 적절한 표본배분 방법의 결정이 층화의 효과를 결정하는데 있어 매우 중요하다.

현재 층화표본추출에서 가장 널리 사용되는 표본배분 방법은 비례배분법(proportional allocation)과 네이만배분법(Neyman alloction)으로 전체 표본크기 n이 주어졌을 때 각 배분방법에 의한 층별 표본크기는 식 (1.1)과 같다. 이는 층별 추정량의 분산에 영향을 미치는 층별 모집단크기(N_h) 또는 모표준편차(S_h)에 대한 정보를 표본 배분비율에 반영하여 추정량의 정도(程度)를 개선하기 위한 것이다. 이론적으로는 층크기만을 반영하는 비례배분법보다 각 층의 표준편차를 함께 고려하는 네이만배분법이 모집

This research was supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education, Science and Technology (2013R1A1A2006363).
[1]Corresponding author: Department of Statistics, Korea University, 145 Anam-ro, Seongbuk-Gu, Seoul 02841, Korea. E-mail: mkpark2@korea.ac.kr

단에 대한 추가적인 정보를 반영함으로써 추정량의 분산을 낮추는데 더 효과적임이 알려져 있다 (Lohr, 2009).

$$n_h = \begin{cases} n \cdot \dfrac{N_h S_h}{\sum_{h=1}^{H} N_h S_h}, & \text{Neyman allocation,} \\ n \cdot \dfrac{N_h}{\sum_{h=1}^{H} N_h}, & \text{Proportional allocation.} \end{cases} \tag{1.1}$$

그러나 네이만배분법을 실제에 적용하기에는 몇가지 제약이 따른다. 통상적으로 각 층의 모표준편차는 예비조사(pilot survey)를 통해 추정하거나 과거 조사로부터 축적된 경험적 정보를 이용하는데, 이와 관련하여 Evans (1951)는 네이만배분의 기준이 되는 모표준편차의 추정오차가 일정 수준을 벗어날 경우 오히려 비례배분법보다 추정량의 분산이 더 커질 수 있음을 지적한 바 있다. 아울러 네이만배분법은 설계변수의 정보에만 의존하기 때문에 다른 관심 변수에 대한 층화분산을 최소화시키지 못하는 결과를 초래할 수 있다 (Kokan, 1963). 이러한 이유로 여러가지 관심변수를 동시에 조사하는 다목적 조사(multi-purpose survey)에서는 네이만배분법이 적합하지 않다는 주장이 제기되기도 한다.

한편 조사단계에서 발생하는 무응답(non-response)으로 인한 편향(bias)은 응답률 보정 방법을 통해 제거되는데, 이러한 추가적인 보정은 추정량의 전체 분산에 영향을 미친다. 그러나 기존의 전통적인 표본배분 방법들은 층별 무응답과 관련한 정보를 표본설계에 반영하지 않기 때문에 층별 응답율(response rate)에 차이가 크게 나타날 경우 층화표본에 의한 효과가 저하될 수 있다.

이에 본 연구에서는 층화표본추출에서 층간 응답률의 차이가 추정량의 분산에 미치는 영향을 살펴보고, 층별 응답률 정보를 표본설계에 반영하는 표본배분 방법을 제안한다. 본 논문의 구성은 먼저 2장에서 층간 응답률의 차이가 층화표본의 추정량에 미치는 영향을 살펴보고, 층별 응답률 정보를 활용한 새로운 표본배분 방법을 제안한다. 이어서 3장에서는 추정량의 분산 측면에서 새로운 표본배분 방법과 기존 방법과의 효율성을 비교한다. 4장에서는 몇가지 응답률 시나리오에 따라 각 표본배분법의 효과가 어떻게 나타나는지를 모의실험을 통해 환인한 후, 5장에서 연구의 결과를 요약한다.

2. 응답률을 반영한 표본배분

2.1. 층간 응답률 차이가 추정오차에 미치는 영향

먼저 H개의 층으로 구성된 모집단에서 모집단 총합($t = \sum_h \sum_k y_{hk}$)을 추정하기 위한 층화단순임의추출(stratified simple random sampling; STSI)을 고려하자. 층 $h\,(= 1, 2, \ldots, H)$의 모집단 크기와 표본수를 각각 N_h와 n_h라 하면, 층 h에 속한 개체들의 표본포함확률(inclusion probability)은 $\pi_h\,(= n_h/N_h)$이므로, 모집단 총합에 대한 호르비츠-톰슨 추정량 (Horvitz-Thompson, 1952)은 다음과 같다.

$$\hat{t}_{\mathrm{HT}} = \sum_{h=1}^{H} \hat{t}_h = \sum_{h=1}^{H} \frac{1}{\pi_h} \sum_{k=1}^{n_h} y_{hk}. \tag{2.1}$$

층 h에 속한 모집단 개체 중에서 조사에 응답할 집합(U_{r_h})의 개체 수가 N_{r_h}라 하면, 층 h의 응답률의 모수 p_h는 N_{r_h}/N_h이고, 이때 응답률 보정 후 층화추정량 $\hat{t}_{ST}$는 식 (2.2)와 같다. 여기서 $\{S\}$와 $\{S_r\}$은 각각 표본으로 추출된 집합과 표본 중에서 조사에 응답한 집합을 의미하며, I_{hk}와 R_{hk}는 층 h의 k번째 개체에 대한 표본포함여부와 응답여부를 나타내는 지시변수(indicator)이다.

$$\hat{t}_{\mathrm{ST}} = \sum_{h=1}^{H} \hat{t}_h = \sum_{h=1}^{H} \frac{1}{\pi_h} \frac{1}{p_h} \sum_{k \in \{S_r\}} y_{hk}$$

$$= \sum_{h=1}^{H} \frac{1}{\pi_h} \frac{1}{p_h} \sum_{k=1}^{N_h} y_{hk} I_{hk} R_{hk} \tag{2.2}$$

$$I_{hk} = \begin{cases} 1, & \text{if } k \in \{S\}\ , \\ 0, & \text{if } k \notin \{S\}\ , \end{cases} \qquad R_{hk} = \begin{cases} 1, & \text{if } k \in \{S_r\}\ , \\ 0, & \text{if } k \notin \{S_r\}\ . \end{cases}$$

층 h에 속한 개체들의 모평균과 모표준편차를 각각 μ_h와 S_h라 하고 각 개체의 응답확률이 층내에서 동일하다면($p_{hk} = p_h$), 식 (2.2)의 추정량 $\hat{t}_{ST}$의 분산은 $x_{hk} = y_{hk} - \mu_h$의 정의를 통해 다음과 같이 유도된다.

$$\begin{aligned} V\left(\hat{t}_{ST}\right) &= \sum_{h=1}^{H} \left[V_{\mathrm{I}} \left(E_R \left(\hat{t}_h | I_{hk} \right) \right) + E_I \left(V_R \left(\hat{t}_h | I_{hk} \right) \right) \right] \\ &= \sum_{h=1}^{H} N_h^2 \left(\frac{N_h - n_h}{N_h} \right) \frac{S_h^2}{n_h} + \sum_{h=1}^{H} \frac{N_h}{n_h} \left(\frac{1}{p_h} - 1 \right) \sum_{k=1}^{N_h} x_{hk}^2 \\ &= \underbrace{\sum_{h=1}^{H} N_h^2 \left(\frac{1}{n_h} - \frac{1}{N_h} \right) S_h^2}_{\text{variation within stratum}} + \underbrace{\sum_{h=1}^{H} \frac{N_h}{n_h} \left(\frac{1}{p_h} - 1 \right) (N_h - 1) S_h^2}_{\text{added variation by non-response}} . \end{aligned} \tag{2.3}$$

식 (2.3)은 추정량의 전체 분산이 두가지 변동으로 구성되어 있음을 나타낸다. 여기서 V_I와 E_I는 각각 표본설계(sampling design)에 의한 분산과 기댓값을 의미하고, E_R과 V_R은 표본의 응답구조(response mechanism)에 의한 기대값과 분산을 의미한다. 따라서 식 (2.3)의 첫 번째 항은 각 층의 표본오차(sampling error)에 의해 야기되는 변동(variation)이고, 두 번째 항은 무응답오차(non-response error)에 의한 변동의 기댓값이다. 이는 층화추출에서 추정량의 분산을 최소화할 목적으로 표본을 배분한다면 각 층의 모표준편차와 함께 층별 응답률을 함께 고려할 필요가 있음을 의미한다.

2.2. 층별 응답률을 반영한 표본배분 방법

전체 모집단에 대한 관심변수, θ의 추정을 위하여 층화추출법이 사용되었다면, 주어진 제약조건 하에서 최적 표본배분(optimum sample allocation)은 식 (2.4)와 같이 추정량의 일반적인 형태의 분산을 최소화하는 문제로 귀결된다 (Sandal 등, 1992).

$$V\left(\hat{\theta}_{ST}\right) = \sum_{h=1}^{H} \frac{A_h}{n_h} + B, \tag{2.4}$$

여기서 관심모수를 모집단 총합($\theta = t$)으로, 식 (2.3)의 두 번째 항에서 $N_h - 1 \approx N_h$로 대체하면, 추정량의 분산은 식 (2.5)와 같이 근사시킬 수 있다.

$$V\left(\hat{t}_{ST}\right) \approx \sum_{h=1}^{H} \frac{N_h^2 S_h^2}{p_h} \frac{1}{n_h} - \sum_{h=1}^{H} N_h S_h^2. \tag{2.5}$$

이는 $A_h = N_h^2 S_h^2 / p_h$ 그리고 $B = -\sum_{h=1}^{H} N_h S_h^2$으로 정의할 때 식 (2.4)의 형태와 동일하다. 따라서 전체 표본크기 n $(= \sum_{h=1}^{H} n_h)$이 주어졌을 때 이를 제약식으로 추정량 $\hat{t}_{\mathrm{ST}}$의 분산을 최소화 하는 최적 표본 배분은 다음과 같다.

$$n_h = n \cdot \frac{N_h S_h / \sqrt{p_h}}{\sum_{h=1}^{H} N_h S_h / \sqrt{p_h}}. \tag{2.6}$$

식 (2.6)에 의한 표본배분법은 층의 모집단 크기 N_h와 모표준편차 S_h가 클수록 그리고 응답률 p_h가 낮은 층일수록 표본을 더 많이 배분하게 된다. 본 논문에서는 앞으로 식 (2.6)으로 정의한 표본배분 방법을 NSR배분법(NSR allocation)으로 명명한다. 각 층의 모표준편차 S_h에 대한 가용한 정보가 없거나 이를 무시할 수 있는 경우 NSR배분법은 식 (2.7)과 같이 층크기와 응답률만을 반영한 배분방법으로 간소화 할 수 있다. 앞으로 이 방법을 NR배분법(NR allocation)으로 명명한다. 만약 각 층의 응답률이 모두 동일하다면 NSR배분법은 식 (1.1)의 네이만배분법과 동일하고, NR배분법은 비례배분법과 같음을 알 수 있다.

$$n_h = n \cdot \frac{N_h/\sqrt{p_h}}{\sum_{h=1}^{H} N_h/\sqrt{p_h}}. \tag{2.7}$$

한편, 층화표본이 층별 모수를 추정(domain estimation)하기 위해서도 사용된다면 층별 표본배분은 각 층의 추정오차가 균형을 이루도록 배분하는 것이 바람직할 것이다. 이를 위해서 층화의 효율성이 설계변수의 선택에 의존하지 않도록 층별 추정량의 변동계수(coefficient of variation)를 목적함수로 하여 표본배분식을 도출하였다.

$\bar{y}_h$를 층 h의 모평균(μ_h)에 대한 불편추정량이라 하자. 추정량 $\bar{y}_h$의 변동계수는 식 (2.8)과 같이 층크기, 응답률 그리고 층의 모변동계수(population coefficient of variation)의 함수로 표현된다.

$$\mathrm{CV}(\bar{y}_h) = \frac{\sqrt{V(\bar{y}_h)}}{E(\bar{y}_h)} = \sqrt{\left(\frac{1}{p_h n_h} - \frac{1}{N_h}\right)}\ \frac{S_h}{\mu_h} = \sqrt{\left(\frac{1}{p_h n_h} - \frac{1}{N_h}\right)}\ \mathrm{CV}_h. \tag{2.8}$$

이로부터 식 (2.9)와 같이 전체 표본수 n이 주어졌을 때, 각 층별 추정량의 변동계수의 가중합을 최소화하는 최적 표본배분을 유도할 수 있다. 여기서 w_h는 층의 상대적 중요도를 나타내는 가중치($\sum w_h = 1$)를 나타낸다.

$$\underset{n_h}{\operatorname{argmin}} \sum_{h=1}^{H} w_h \left[\mathrm{CV}(\bar{y}_h)\right]^2, \quad \text{subject to } \sum n_h = n, \quad n_h = n \cdot \frac{\mathrm{CV}_h\sqrt{w_h/p_h}}{\sum_{h=1}^{H} \mathrm{CV}_h\sqrt{w_h/p_h}}. \tag{2.9}$$

3. 표본 배분방법에 따른 효율성 비교

본 절에서는 모집단 총합을 추정하기 위한 층화단순임의추출을 예제로 전통적인 표본배분방법과 응답률을 반영한 배분방법을 각각 적용할 때 추정량의 분산을 도출하고 각 배분방법들의 상대효율을 비교한다. 각 배분방법에 따른 비교의 편의(便宜)를 위해 먼저 모집단 전체의 분산(S^2), 각 층내 분산(S_h^2)과 층별 모평균의 층간 분산(S_μ^2)을 다음과 같이 정의하자. 여기서 μ는 각 층의 모평균을 층크기로 가중평균한 모집단 전체 평균을 나타낸다.

$$\begin{cases} S^2 = \dfrac{1}{N-1}\displaystyle\sum_{h=1}^{H}\sum_{k=1}^{N_h}(y_{hk} - \mu)^2, \\ S_h^2 = \dfrac{1}{N_h - 1}\displaystyle\sum_{k=1}^{N_h}(y_{hk} - \mu_h)^2, \\ S_\mu^2 = \displaystyle\sum_{h=1}^{H}\frac{N_h}{N}(\mu_h - \mu)^2. \end{cases} \tag{3.1}$$

모집단의 총변동은 층내변동과 층간변동의 합으로 표시할 수 있으므로 식 (3.1)의 분산식에서 유한모집단(finite population)에 의한 자유도 손실을 무시($N-1 \cong N$, $N_h-1 \cong N_h$)하면, 모집단 전체 분산은 다음과 같이 층간 모평균의 분산과 층내분산의 가중합으로 표현할 수 있다.

$$S^2 \approx S_\mu^2 + \sum_{h=1}^{H} \frac{N_h}{N} S_h^2. \tag{3.2}$$

3.1. 각 배분방법에 의한 추정량의 분산

전체 표본크기 n이 주어졌을 때, 각 배분방법에 의한 추정량의 분산은 식 (2.5)에 각 배분방법에 의한 층별 표본크기를 대입하여 유도할 수 있다. 분산식의 자세한 전개 과정은 부록을 참고하고, 본문에서는 각 분산식의 전개 결과에 대해 설명한다. 먼저 비례배분법에 의한 추정량의 분산 $V_{\mathrm{P}}(\hat{t})$은 다음과 같다.

$$V_P\left(\hat{t}\right) = \frac{1}{n}\sum_{h=1}^{H} \frac{N_h^2 S_h^2}{p_h}\frac{N}{N_h} - \sum_{h=1}^{H} N_h S_h^2 \tag{3.3}$$

$q_h(=1/\sqrt{p_h})$를 응답률 역수의 제곱근이라 하고, 이를 층크기와 층분산의 곱($N_h S_h^2$)으로 가중한 평균과 분산을 각각 $\tilde{q}$와 $S_{\tilde{q}}^2$라 하자. 그러면 비례배분법에 의한 분산은 식 (3.4)와 같이 표현할 수 있다.

$$V_P\left(\hat{t}\right) \approx \frac{N}{n}\left(N\tilde{q}^2 - n\right)\left[S^2 - S_\mu^2\right] + \frac{N}{n}\left(\sum_{h=1}^{H} N_h S_h^2\right) S_{\tilde{q}}^2, \tag{3.4}$$

$$\tilde{q} = \sum_{h=1}^{H} \frac{N_h S_h^2}{\left(\sum_{h=1}^{H} N_h S_h^2\right)} q_h, \quad S_{\tilde{q}}^2 = \sum_{h=1}^{H} \frac{N_h S_h^2}{\left(\sum_{h=1}^{H} N_h S_h^2\right)} (q_h - \tilde{q})^2.$$

식 (3.4)의 첫번째 항은 층화에 의한 효과로 전체분산에서 층간 평균의 분산을 줄이는 효과가 있음을 나타낸다. 두번째 항은 $S_{\tilde{q}}^2$에 비례하여 증가하는데, 이는 비례배분법이 각 층의 응답률 차이를 반영하지 않기 때문에 층간 응답률의 이질성(heterogeneity)이 커질수록 추정량의 분산이 늘어남을 의미한다.

한편, NR배분법에 의한 추정량의 분산 $V_{NR}(\hat{t})$은 식 (3.5)와 같다. 여기서 $\tilde{q}$는 식 (3.4)에서와 같이 층크기와 층분산의 곱으로 가중한 q_h의 평균이고, $\bar{q}$는 층크기로 가중한 q_h의 평균을 의미한다. NR배분법은 층간 응답률 차이를 표본배분에 반영하기 때문에 각 층의 응답률 차이에 의한 분산이 제거되었음을 확인할 수 있다. 또한 두 개의 q_h에 대한 가중평균 $\bar{q}$와 $\tilde{q}$가 유사하다면 즉, 각 층의 모표준편차 S_h의 차이를 무시할 수 있다면 NR배분법에 의한 추정량이 비례배분법보다 더 작음을 알 수 있다.

$$\begin{aligned} V_{NR}\left(\hat{t}\right) &= \frac{1}{n}\sum_{h=1}^{H} \frac{N_h^2 S_h^2}{p_h} \frac{\left(\sum_{h=1}^{H} N_h/\sqrt{p_h}\right)}{N_h/\sqrt{p_h}} - \sum_{h=1}^{H} N_h S_h^2 \\ &\approx \frac{N}{n}\left(N\bar{q}\,\tilde{q} - n\right)\left[S^2 - S_\mu^2\right], \\ \bar{q} &= \sum_{h=1}^{H} \frac{N_h}{N} q_h\,, \quad \tilde{q} = \sum_{h=1}^{H} \frac{N_h S_h^2}{\left(\sum_{h=1}^{H} N_h S_h^2\right)} q_h. \end{aligned} \tag{3.5}$$

식 (3.6)은 네이만배분법에 의한 추정량의 분산 $V_N(\hat{t})$이다. 여기서 $\check{q}$와 $S_{\check{q}}^2$은 각각 q_h를 층크기와 층표준편차의 곱으로 가중한 평균과 분산을 의미하며, S_S^2은 층표준편차 S_h를 층크기로 가중한 분산으로 층간 표준편차의 이질성을 의미한다. 식 (3.6)의 첫 번째 항은 네이만배분법이 층화의 효과와 함께 표준편

차(S_h)에 대한 층간 차이를 표본배분에 반영함으로써 추정량의 분산을 줄이는 효과를 나타낸다. 그러나 네이만배분법 역시 층별 응답률을 감안하지 않기 때문에 층간 응답률 차이에 의한 추정량의 분산은 여전히 남아 있음을 확인할 수 있다. 물론 층표준편차와 층응답률이 모든 층에서 동일하다면, 각각의 분산 S_S^2와 $S_{\bar{q}}^2$는 0이 되므로 네이만배분법과 비례배분법에 의한 추정량의 분산과 같아진다.

$$
\begin{aligned}
V_N\left(\hat{t}\right) &= \frac{1}{n}\sum_{h=1}^{H}\frac{N_h^2 S_h^2}{p_h}\frac{\left(\sum_{h=1}^{H}N_h S_h\right)}{N_h S_h} - \sum_{h=1}^{H}N_h S_h^2 \\
&\approx \frac{N}{n}\left(N\check{q}^2 - n\right)\left[S^2 - S_\mu^2 - \frac{N\check{q}^2}{(N\check{q}^2-n)}S_S^2\right] + \frac{1}{n}\left(\sum_{h=1}^{H}N_h S_h\right)^2 S_{\bar{q}}^2, \qquad (3.6)\\
\check{q} &= \sum_{h=1}^{H}\frac{N_h S_h}{\left(\sum_{h=1}^{H}N_h S_h\right)}q_h, \\
S_{\bar{q}}^2 &= \sum_{h=1}^{H}\frac{N_h S_h}{\left(\sum_{h=1}^{H}N_h S_h\right)}\left(q_h-\check{q}\right)^2, \quad S_S^2 = \sum_{h=1}^{H}\frac{N_h}{N}\left(S_h - \sum\frac{N_h}{N}S_h\right)^2.
\end{aligned}
$$

마지막으로 NSR배분법에 의한 추정량의 분산 $V_{NSR}(\hat{t})$은 식 (3.7)과 같이 전개할 수 있다. NSR배분법은 각 층의 크기는 물론 표준편차와 응답률을 모두 반영하여 표본을 배분하기 때문에 그로 인한 추정량의 분산을 효과적으로 줄이고 있음을 이론적으로 확인할 수 있다.

$$
\begin{aligned}
V_{NSR}\left(\hat{t}\right) &= \frac{1}{n}\sum_{h=1}^{H}\frac{N_h^2 S_h^2}{p_h}\frac{\left(\sum_{h=1}^{H}N_h S_h/\sqrt{p_h}\right)}{N_h S_h/\sqrt{p_h}} - \sum_{h=1}^{H}N_h S_h^2 \\
&\approx \frac{N}{n}\left(N\check{q}^2 - n\right)\left[S^2 - S_\mu^2 - \frac{N\check{q}^2}{(N\check{q}^2-n)}S_S^2\right]. \qquad (3.7)
\end{aligned}
$$

3.2. 분산 절약 효과

식 (3.8)은 NSR배분법이 비례배분법과 비교하여 추정량의 분산을 얼마나 줄일 수 있는지를 나타낸다. NSR배분법 하에서 추정량의 분산이 비례배분법 하에서 동일한 추정량의 분산보다 항상 작고, 그 차이는 층별 표준편차와 응답률의 함수인 $S_h/\sqrt{p_h}$의 분산에 비례함을 알 수 있다. 즉, NSR배분법의 분산절약 효과는 $S_h/\sqrt{p_h}$의 층간 이질성(heterogeneity)에 비례하여 커진다.

$$
\begin{aligned}
V_P\left(\hat{t}\right) - V_{NSR}\left(\hat{t}\right) &= \frac{1}{n}\sum_{h=1}^{H}\left[\frac{N_h^2 S_h^2}{p_h}\frac{N}{N_h} - \frac{N_h^2 S_h^2}{p_h}\frac{\left(\sum_{h=1}^{H}N_h S_h/\sqrt{p_h}\right)}{N_h S_h/\sqrt{p_h}}\right] \\
&= \frac{N^2}{n}\left[\sum_{h=1}^{H}\frac{N_h}{N}\left(\frac{S_h}{\sqrt{p_h}}\right)^2 - \left(\sum_{h=1}^{H}\frac{N_h}{N}\frac{S_h}{\sqrt{p_h}}\right)^2\right] \qquad (3.8)\\
&= \frac{N^2}{n}\mathrm{V}\left(\frac{S_h}{\sqrt{p_h}}\right) \geq 0.
\end{aligned}
$$

다음으로 식 (3.9)를 통해 NSR배분법과 네이만배분법을 비교해 보면 역시 NSR배분법에 의한 추정량의 분산이 네이만배분법에 의한 추정량의 분산보다 항상 작다는 것을 이론적으로 증명할 수 있다. 아울러 층간 응답률의 편차가 커질수록 네이만배분법을 사용할 때보다 NSR배분법을 사용할 때 분산 절약효

과가 커짐을 확인할 수 있다.

$$V_N\left(\hat{t}\right) - V_{NSR}\left(\hat{t}\right) = \frac{1}{n}\sum_{h=1}^{H}\left[\frac{N_h^2 S_h^2}{p_h}\frac{\left(\sum_{h=1}^{H} N_h S_h\right)}{N_h S_h} - \frac{N_h^2 S_h^2}{p_h}\frac{\left(\sum_{h=1}^{H} N_h S_h/\sqrt{p_h}\right)}{N_h S_h/\sqrt{p_h}}\right] \tag{3.9}$$
$$= \frac{1}{n}\left(\sum_{h=1}^{H} N_h S_h\right)^2 \mathrm{Var}\left(\frac{1}{\sqrt{p_h}}\right) \geq 0.$$

비례배분법과 네이만배분법, NR배분법 사이의 효율은 층의 표준편차와 응답률 간의 관계에 따라 달라진다. 먼저 비례배분법과 네이만배분법의 상대효율은 전술한 바와 같이 층간 응답률에 차이가 없는 경우 네이만배분법이 더 좋은 추정량을 제공하지만, 식 (3.10)에서 보는 바와 같이 표준편차가 작은 층에서 응답률이 상대적으로 낮게 나타나면 오히려 비례배분법이 더 좋은 추정량을 제공할 수 있다.

$$\begin{aligned} V_P\left(\hat{t}\right) - V_N\left(\hat{t}\right) &= \frac{1}{n}\sum_{h=1}^{H}\left[\frac{N_h^2 S_h^2}{p_h}\frac{N}{N_h} - \frac{N_h^2 S_h^2}{p_h}\frac{\left(\sum_{h=1}^{H} N_h S_h\right)}{N_h S_h}\right] \\ &= \frac{N}{n}\sum_{h=1}^{H}\frac{N_h S_h}{p_h}\left[S_h - \sum_{h=1}^{H}\frac{N_h}{N}S_h\right] \\ &= \frac{N}{n}\sum_{h=1}^{H}\frac{N_h S_h}{p_h}\left[S_h - \bar{S}\right] \\ &= \frac{N}{n}\left[\sum_{h\in\{S_h>\bar{S}\}}\frac{N_h S_h}{p_h}\left(S_h - \bar{S}\right) - \sum_{h\in\{S_h<\bar{S}\}}\frac{N_h S_h}{p_h}\left(\bar{S} - S_h\right)\right]. \end{aligned} \tag{3.10}$$

NR배분법의 경우도 마찬가지로 층별 응답률을 반영함으로써 추정량의 분산을 줄일 수 있지만, 응답률이 낮은 층에서 표준편차가 충분히 작아서 응답률을 고려하는 배분 효과가 상쇄되어 버리면 비례배분법이 더 효율적일 수 있다.

$$\begin{aligned} V_P\left(\hat{t}\right) - V_{NR}\left(\hat{t}\right) &= \frac{1}{n}\sum_{h=1}^{H}\left[\frac{N_h^2 S_h^2}{p_h}\frac{N}{N_h} - \frac{N_h^2 S_h^2}{p_h}\frac{\left(\sum_{h=1}^{H} N_h/\sqrt{p_h}\right)}{N_h/\sqrt{p_h}}\right] \\ &= \frac{N}{n}\sum_{h=1}^{H}\frac{N_h S_h^2}{\sqrt{p_h}}\left[\frac{1}{\sqrt{p_h}} - \sum_{h=1}^{H}\frac{N_h}{N}\frac{1}{\sqrt{p_h}}\right] \\ &= \frac{N}{n}\sum_{h=1}^{H}\frac{N_h S_h^2}{\sqrt{p_h}}\left[q_h - \bar{q}\right] \\ &= \frac{N}{n}\left[\sum_{h\in\{q_h>\bar{q}\}}\frac{N_h S_h^2}{\sqrt{p_h}}(q_h - \bar{q}) - \sum_{h\in\{q_h<\bar{q}\}}\frac{N_h S_h^2}{\sqrt{p_h}}(\bar{q} - q_h)\right]. \end{aligned} \tag{3.11}$$

끝으로 네이만배분법과 NR배분법의 상대효율은 식 (3.12)에서 보는 바와 같이 표준편차 S_h와 응답률 역수의 제곱근 q_h의 변동계수(coefficient of variation)의 상대적 크기에 따라 달라진다. q_h의 변동계수가 S_h의 변동계수를 상회할 경우 즉, 층별 응답률이 층별 표준편차보다 더 이질적일 때는 NR배분법이 네이만배분법보다 더 효과적이고, 반대의 경우는 네이만배분법이 더 효과적임을 알 수 있다. 따라서,

NR배분법과 네이만배분법의 상대효율은 응답률과 표준편차 중에서 어느 쪽이 추정량의 분산에 더 큰 영향을 미치는가에 따라 달라진다.

$$\begin{aligned} V_{N}(\hat{t}) - V_{NR}(\hat{t}) &= \frac{1}{n}\sum_{h=1}^{H}\left[\frac{N_h^2 S_h^2}{p_h}\frac{\left(\sum_{h=1}^{H} N_h S_h\right)}{N_h S_h} - \frac{N_h^2 S_h^2}{p_h}\frac{\left(\sum_{h=1}^{H} N_h/\sqrt{p_h}\right)}{N_h/\sqrt{p_h}}\right] \\ &= \frac{N}{n}\bar{S}\sum_{h=1}^{H}\frac{N_h S_h}{p_h} - \frac{N}{n}\bar{q}\sum_{h=1}^{H}\frac{N_h S_h^2}{\sqrt{p_h}} \\ &= \frac{N^2}{n}\bar{S}^2\left[S_{\check{q}}^2 - \check{q}^2\right] - \frac{N^2}{n}\bar{q}^2\left[S_{\tilde{S}}^2 - \tilde{S}^2\right] \\ &\approx \frac{N^2}{n}(\bar{S}\,\bar{q})^2\left[\frac{S_{\check{q}}^2}{\bar{q}^2} - \frac{S_{\tilde{S}}^2}{\bar{S}^2}\right] \qquad \left(\text{단, } \bar{q}\approx\check{q},\ \bar{S}\approx\tilde{S}\right) \\ &\approx \frac{N^2}{n}(\bar{S}\,\bar{q})^2\left[\mathrm{CV}_q^2 - \mathrm{CV}_S^2\right]. \end{aligned} \tag{3.12}$$

3.3. 표본 절약 효과

앞 절에서는 전체 표본크기 n이 동일할 때 층별 응답률을 반영하는 표본배분 방법이 추정의 측면에서 더 효율적임을 확인하였다. 여기서 전체 표본크기 n은 층화표본의 표본설계 단계에서 표본배분을 위해 사전에 결정된 조사 표본수(size of sample)를 의미한다. 그러나 추정에 사용되는 표본수는 조사 과정에서 발생하는 무응답을 제외한 실제 응답한 표본수(size of respondents)이다. 층 h에 대한 응답 표본수의 기대값 $E(r_h)$는 층내 개체의 응답확률이 p_h로 동일하다고 전제하면 $n_h p_h$와 같다. 식 (3.13)은 NR배분법과 비례배분법을 사용할 때 전체 응답 표본수에 대한 기댓값을 비교한 것으로 NR배분법 하에서의 응답 표본수의 기댓값이 더 작다는 것을 알 수 있다. 층별 응답률을 고려하는 배분 방법은 상대적으로 응답률이 낮은 층에 표본수를 더 배분하기 때문에 전체 응답 표본수에 대한 기대값이 비례배분법이나 네이만배분법을 사용할 때보다 더 작게 나타난다. 이는 비록 추정량의 추정오차가 비슷한 수준이라 하더라도 상대적으로 적은 수의 응답 표본을 사용하는 것이 조사기간이나 비용 등을 절약할 수 있다는 점에서 응답률을 고려하는 배분방법이 조사의 효율성 측면에서도 더 유리(有利)함을 의미한다.

$$\begin{aligned} E_{NR}(r) &= \sum_{h=1}^{H} n\cdot\frac{N_h/\sqrt{p_h}}{\sum_{h=1}^{H} N_h/\sqrt{p_h}}\cdot p_h \\ &\le \sum_{h=1}^{H} n\cdot\frac{N_h}{\sum_{h=1}^{H} N_h}\cdot\sqrt{p_h} \qquad [p_h \le 1] \\ &\le \sum_{h=1}^{H} n\cdot\frac{N_h}{N}\cdot p_h = E_P[r]. \end{aligned} \tag{3.13}$$

4. 모의실험을 통한 효율성 비교

3장에서 층별 표준편차와 응답률을 고려하여 표본을 배분하는 NSR배분법이 기존 비례배분법이나 네이만배분법에 비해 추정량의 분산을 효과적으로 줄일 수 있음을 이론적으로 설명하였다. 본 장에서는 층별 표준편차와 응답률의 관계에 따라 나머지 배분법들 간의 상대효율이 어떻게 나타나는지를 모의실험을 통해 확인해보기로 한다. 아울러 복수의 관심변수를 조사하는 다목적 조사에서 각 표본배분방법들의 효과를 비교한다.

Table 4.1. Variation of survey variables

	Y1	Y2	Y3	Y4	Y5	Y6	Y7
Total variation(S^2)	2.273	2.353	2.360	2.460	2.534	2.703	2.717
Between variation(S_μ^2)	1.989	2.002	1.958	1.983	1.984	2.070	2.013
Heterogeneity(S_S^2)	0.001	0.004	0.007	0.014	0.020	0.024	0.040

Table 4.2. Heterogeneity of response rates

	R0	R1	R2	R3	R4	R5	R6	R7
Heterogeneity(S_p^2)	0.000	0.005	0.010	0.018	0.029	0.041	0.056	0.073

4.1. 모의실험 설계

모의실험을 위한 모집단을 10개의 층(H)으로 구성하고, 모집단 총합($t = 100{,}000$)은 같으나 층간 표준편차(S_h)간 이질성이 서로 다른 7개의 관심변수(survey variable)를 생성하였다. Table 4.1은 7개 변수의 전체변동(S^2), 층간변동(S_μ^2), 층간 표준편차의 이질성(S_S^2)을 요약한 것이다. 여기서 층간 표준편차의 이질성 정도는 층별 표준편차(S_h)를 층크기(N_h)로 가중한 가중분산을 사용하였다. Y1이 가장 작은 경우미고 Y7로 갈수록 층간 표준편차의 이질성이 커진다.

다음으로 관심변수와는 독립적으로 전체 응답률($p = 0.55$)은 같지만 층간 응답률의 이질성이 커지도록 하여 총 8개의 응답률 시나리오를 생성하였다. 마찬가지로 층별 응답률(p_h)을 층크기(N_h)로 가중한 가중분산(S_p^2)을 이용하여 층간 이질성의 정도를 측정하였으며, 각 시나리오별 응답률의 이질성은 Table 4.2와 같다.

모의실험은 7개의 관심변수와 8개의 응답률 시나리오의 모든 조합에 대해 4가지 표본배분 방법을 비교하였다. 모의실험을 위한 표본크기(n)는 전체 모집단크기(N)의 10%인 1,000개로 동일하게 하였으며, 표본추출은 층화 임의표본 추출을 이용하였다. 매 경우마다 각각 5,000번의 독립 표본을 반복 추출한 후 평균제곱오차(Mean Square Error; MSE)를 계산하여 표본배분 방법의 상대 효율을 비교하였다.

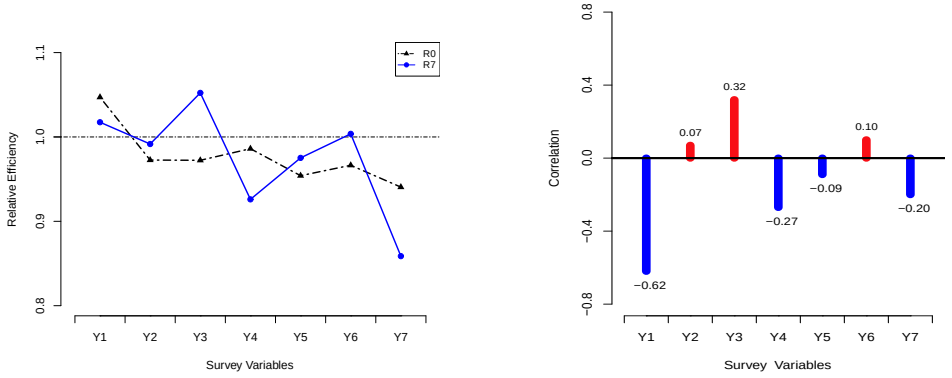
4.2. 모의실험 결과

먼저 층별 응답률 편차가 비례배분법과 네이만배분법의 상대효율에 미치는 영향을 확인해 보자. Figure 4.1은 층간 응답률의 차이가 없는 시나리오(R0)와 층간 응답률 편차가 가장 큰 경우(R7)의 비례배분법 대비 네이만배분법의 상대효율을 나타내고 있다. 여기서 상대효율($= \mathrm{MSE}_N/\mathrm{MSE}_P$)은 네이만배분법에 의한 평균제곱오차를 비례배분법의 평균제곱오차로 나누어 계산하였으므로, 상대효율 값이 1보다 작을 때 네이만배분법이 비례배분법보다 더 효과적임을 의미한다.

Figure 4.1의 좌측 그래프를 보면 층간 응답률에 차이가 없는 R0에서는 층별 표준편차가 이질적일수록 네이만배분법의 상대효율이 더 좋아짐을 알수 있다. 반면 층별로 응답률이 서로 다른 R7에서는 R0에서와 달리 비례배분법에 의한 추정오차가 더 작은 경우가 발생한다. 이를 좌측의 층간 표준편차와 응답률간 상관계수와 함께 확인해 보면, 응답률과 표준편차 사이에 정(正)의 관계에 있는 Y3과 Y6에서 네이만배분법의 효율이 오히려 떨어지는 것을 확인할 수 있다.

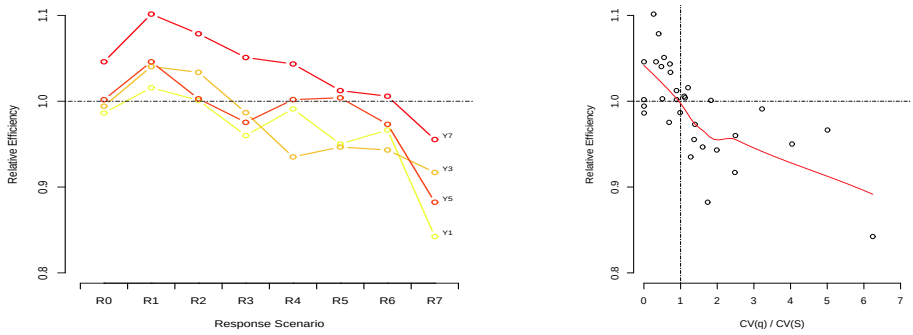
Figure 4.2는 NR배분법과 네이만배분법을 비교한 것이다. 앞서 확인한 바와 같이 응답률의 변동계수가 표준편차의 변동계수보다 더 큰 경우 NR배분법의 상대효율이 더 좋고, 반대로 표준편차의 변동계수가 더 큰 경우는 네이만배분법이 더 효과적임을 보여준다.

그러나 층별 추정(domain estimation)의 측면에서 두 방법을 비교해 보면, 네이만배분법에 비해 NR배



(a) Relative efficiency of Neyman (b) Correlation between S_h and p_h in R7

Figure 4.1. Weakened efficiency of Neyman when high differences among stratum response rates.



(a) Relative efficiency (b) Ratio of C.V.

Figure 4.2. Efficiency of NR allocation compared with Neyman.

분법이 층별로 보다 균형적인 추정량을 제공하고 있음을 확인할 수 있다. Figure 4.3은 동일한 관심변수(Y4)에 대해 두 방법의 층별 추정오차를 비교한 것이다. 여기서 층별 추정오차의 범위(range)는 추정오차가 가장 큰 층과 작은 층 사이의 차이로 이 값이 클수록 불균형이 크다고 이해할 수 있다. 네이만배분법의 경우 층별 응답률을 고려하지 않기 때문에 상대적으로 응답률이 낮은 층에 대한 추정의 정도(程度)가 낮아지는 문제가 발생한다. 특히, 이러한 문제는 층간 응답률의 격차가 커질수록 심화된다. 이에 반해 NR배분법은 상대적으로 응답률이 낮은 층에 표본을 더 배분하기 때문에 이러한 문제를 보다 완화할 수 있다. 이는 모집단 전체 추정량의 오차가 비슷한 수준이더라도 균형적인 층별 추정을 위해서는 층별 응답률을 반영한 표본배분 방법이 더 유용함을 의미한다.

마지막으로 다목적조사에서 표본배분 방법의 효과를 확인하기 위해 하나의 표본 설계변수를 가지고 표본을 배분하여 추출한 표본으로부터 4개의 관심변수를 동시에 추정할 때 추정량의 정도를 비교해 보았다. Figure 4.4의 첫 번째 행에서 보는 바와 같이 각 관심변수의 층별 표준편차는 설계변수와의 유사성이 서로 다르다. 첫 번째 변수의 경우 층별 표준편차가 설계변수와 완전한 선형관계에 있는 경우이며, 두 번째 변수는 상당히 유사한 경우, 세 번째는 특별한 상관성이 없는 경우이고, 마지막 변수는 반대로 설계변수의 층별 표준편차와는 부(負)의 관계에 있는 경우이다. 먼저 모집단 총합에 대한 추정효과를

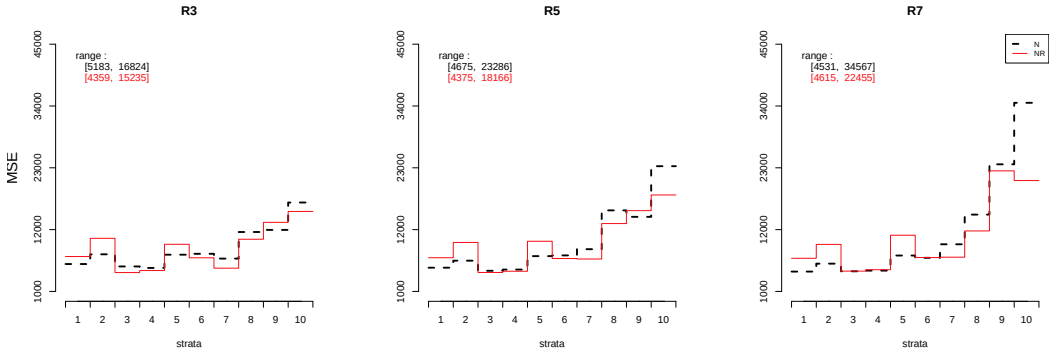


Figure 4.3. Mean Square Error of domain estimation for Y4 in NR and Neyman allocation.

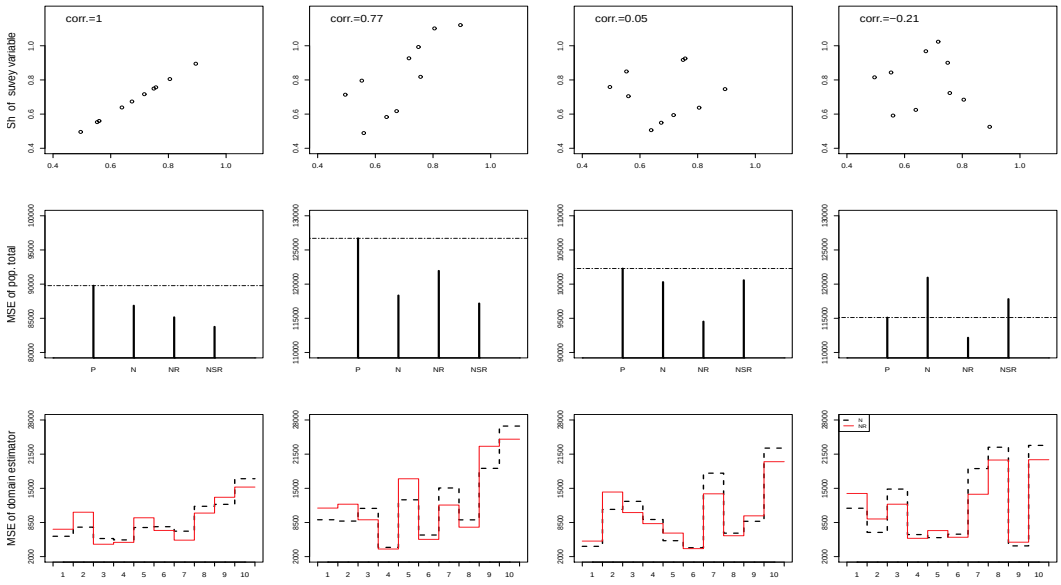


Figure 4.4. Efficiency in multi-purpose survey.

비교해 보면 표본배분에 사용된 설계변수와 층별 표준편차의 상대적 크기가 유사할 경우 층크기만을 고려한 비례배분법보다 층별 표준편차와 응답률 등 추가적인 정보를 활용한 나머지 배분방법들의 효율성이 더 높게 나타난다. 그러나 세 번째의 경우 처럼 표본설계에서의 표준편차에 대한 정보가 관심변수의 모집단 특성을 잘 반영하지 못할 때는 네이만배분법이나 NSR배분법의 효과는 상대적으로 낮아지는 것을 확인할 수 있다. 네 번째 경우와 같이 관심변수의 층별 표준편차가 표본 설계변수와는 반대의 양상을 가지고 있다면 오히려 표준편차에 대한 정보를 사용하지 않은 비례배분법과 NR배분법이 더 나은 결과를 나타낸다. 세 번째 행은 네이만배분법과 NR배분법의 층별 추정오차를 비교한 것인데, 4변수 모두 NR배분법이 더 균형적인 층별 추정량을 제공하고 있음을 보여준다. 이는 NR배분법이 다목적조사에서 네이만배분법의 제약점을 해소하면서도 비례배분법보다 추정량의 정도를 개선할 수 있는 훌륭한 대안이 될 수 있음 시사한다.

5. 결론

표본조사에서 표본수는 목표한 신뢰수준을 충족할 수 있도록 충분한 크기를 확보하는 것이 바람직하다. 하지만, 조사비용과 시간의 제약 등으로 인해 표본크기를 무한정 늘릴 수만은 없는 것이 현실이다. 많은 경우 제반 여건상 가용한 수준 이내에서 표본수가 결정되고 나면 주어진 표본크기를 최대한 활용하여 통계량의 정도(精度)를 최대화하는 것이 표본설계의 주목적이 되곤 한다. 층화표본추출에서도 마찬가지로 전체 표본수가 정해졌을 때 이를 층별로 어떻게 배분할 것인가가 곧 표본설계의 핵심이라 할 수 있는데, 표본배분 기준을 정함에 있어 층에 대한 정보를 더 많이 고려할수록 더 효과적인 표본배분이 가능하리라는 것은 자명한 일이다. 본 연구는 이 점에 착안하여 전통적인 표본배분 방법에서 고려되어 온 층크기 및 층분산과 함께 층별 응답률을 반영하는 표본배분 기준을 이론적으로 도출하였다.

본 연구에서 제안한 표본배분 방법은 각 층의 응답률보정 과정에서 늘어나는 추정량의 분산을 고려하여 표본을 배분하기 때문에 층간 응답률 편차에 따른 기존 배분 방법의 효율성 저하를 개선할 수 있다. 아울러 상대적으로 응답률이 낮은 층에 표본을 더 배분함으로써 기존 방법에 비해 층별 추정오차의 불균형을 개선하는 효과를 기대할 수 있다. 이는 지역별 또는 산업별 통계 등 층별 추정을 위한 표본조사에서 응답률을 반영한 표본배분 방법이 더 효과적일 수 있음을 시사한다.

아울러 응답률에 대한 정보는 정기적으로 반복되는 표본조사는 물론이고, 새로운 관심변수에 대한 신규조사의 경우에도 같은 추출틀을 사용한 선행 조사의 결과를 활용할 수 있다는 이점(利點)이 있다. 이처럼 응답률에 대한 정보는 표준편차에 비해 상대적으로 정보 수집이 용이하고, 조사시점 등 여타 환경적 요인에도 덜 민감하다는 점에서 표본배분의 기준으로써 중요한 정보가 될 수 있다. 특히 조사변수의 척도(scale)나 개수에 크게 의존하지 않기 때문에 여러가지 관심변수를 동시에 조사하는 다목적 조사의 표본설계에 더 유용하다.

본 연구는 응답률에 대한 정보를 표본설계에 활용하기 위한 실무적 고민에 대해 이론적 토대를 제공하고, 실제 통계생산 현업에서 바로 활용할 수 있는 개선된 표본배분 방법을 제안하였다는 점에서 의의가 있다.

부록 A: 표분배분 방법별 추정량의 분산

(A1) 비례배분법

$$
\begin{aligned}
V_P\left(\hat{t}\right) &= \frac{1}{n}\sum_{h=1}^{H}\frac{N_h^2 S_h^2}{p_h}\frac{N}{N_h} - \sum_{h=1}^{H} N_h S_h^2 \\
&= \frac{N}{n}\left(\sum_{h=1}^{H} N_h S_h^2\right)\sum_{h=1}^{H}\frac{N_h S_h^2}{\left(\sum_{h=1}^{H} N_h S_h^2\right)}\frac{1}{p_h} - \sum_{h=1}^{H} N_h S_h^2 \\
&= \frac{N}{n}\left(\sum_{h=1}^{H} N_h S_h^2\right) S_{\tilde{q}}^2 + \frac{N}{n}\tilde{q}^2\left(\sum_{h=1}^{H} N_h S_h^2\right) - \sum_{h=1}^{H} N_h S_h^2 \\
&= \frac{N}{n}\left(\sum_{h=1}^{H} N_h S_h^2\right) S_{\tilde{q}}^2 + \frac{N}{n}\left(N\tilde{q}^2 - n\right)\sum_{h=1}^{H}\frac{N_h}{N} S_h^2 \\
&\approx \frac{N}{n}\left(N\tilde{q}^2 - n\right)\left[S^2 - S_\mu^2\right] + \frac{N}{n}\left(\sum_{h=1}^{H} N_h S_h^2\right) S_{\tilde{q}}^2.
\end{aligned}
$$

(A2) 네이만배분법

$$
\begin{aligned}
V_{N}\left(\hat{t}\right) &= \frac{1}{n}\sum_{h=1}^{H}\frac{N_h^2 S_h^2}{p_h}\frac{\left(\sum_{h=1}^{H}N_h S_h\right)}{N_h S_h} - \sum_{h=1}^{H}N_h S_h^2 \\
&= \frac{1}{n}\left(\sum_{h=1}^{H}N_h S_h\right)^2 \sum_{h=1}^{H}\frac{N_h S_h}{\left(\sum_{h=1}^{H}N_h S_h\right)}\frac{1}{p_h} - \sum_{h=1}^{H}N_h S_h^2 \\
&= \frac{N^2}{n}\left(\sum_{h=1}^{H}\frac{N_h}{N}S_h\right)^2\left[S_{\check{q}}^2 + \check{q}^2\right] - \sum_{h=1}^{H}N_h S_h^2 \\
&= \frac{1}{n}\left(\sum_{h=1}^{H}N_h S_h\right)^2 S_{\check{q}}^2 + \frac{N^2}{n}\bar{S}^2\check{q}^2 - \sum_{h=1}^{H}N_h S_h^2 \\
&= \frac{1}{n}\left(\sum_{h=1}^{H}N_h S_h\right)^2 S_{\check{q}}^2 + \frac{N^2}{n}\check{q}^2\left[\sum_{h=1}^{H}\frac{N_h}{N}S_h^2 - S_S^2\right] - \sum_{h=1}^{H}N_h S_h^2 \\
&= \frac{1}{n}\left(\sum_{h=1}^{H}N_h S_h\right)^2 S_{\check{q}}^2 - \frac{N^2}{n}\check{q}^2 S_S^2 + \frac{N}{n}\left(N\check{q}^2 - n\right)\sum_{h=1}^{H}\frac{N_h}{N}S_h^2 \\
&\approx \frac{N}{n}\left(N\check{q}^2 - n\right)\left[S^2 - S_\mu^2 - \frac{N\check{q}^2}{\left(N\check{q}^2 - n\right)}S_S^2\right] + \frac{1}{n}\left(\sum_{h=1}^{H}N_h S_h\right)^2 S_{\check{q}}^2
\end{aligned}
$$

(A3) NR배분법

$$
\begin{aligned}
V_{NR}\left(\hat{t}\right) &= \frac{1}{n}\sum_{h=1}^{H}\frac{N_h^2 S_h^2}{p_h}\frac{\left(\sum_{h=1}^{H}N_h/\sqrt{p_h}\right)}{N_h/\sqrt{p_h}} - \sum_{h=1}^{H}N_h S_h^2 \\
&= \frac{1}{n}\left(\sum_{h=1}^{H}\frac{N_h}{\sqrt{p_h}}\right)\sum_{h=1}^{H}\frac{N_h S_h^2}{\sqrt{p_h}} - \sum_{h=1}^{H}N_h S_h^2 \\
&= \frac{N}{n}\left(\sum_{h=1}^{H}\frac{N_h}{N}\frac{1}{\sqrt{p_h}}\right)\left(\sum_{h=1}^{H}\frac{N_h S_h^2}{\left(\sum_{h=1}^{H}N_h S_h^2\right)}\frac{1}{\sqrt{p_h}}\right)\left(\sum_{h=1}^{H}N_h S_h^2\right) - \sum_{h=1}^{H}N_h S_h^2 \\
&= \frac{N}{n}\bar{q}\tilde{q}\sum_{h=1}^{H}N_h S_h^2 - \sum_{h=1}^{H}N_h S_h^2 \\
&= \frac{N}{n}\left(N\bar{q}\tilde{q} - n\right)\sum_{h=1}^{H}\frac{N_h}{N}S_h^2 \\
&\approx \frac{N}{n}\left(N\bar{q}\tilde{q} - n\right)\left[S^2 - S_\mu^2\right]
\end{aligned}
$$

(A4) NSR배분법

$$
\begin{aligned}
V_{NSR}\left(\hat{t}\right) &= \frac{1}{n}\sum_{h=1}^{H}\frac{N_h^2 S_h^2}{p_h}\frac{\left(\sum_{h=1}^{H}N_h S_h/\sqrt{p_h}\right)}{N_h S_h/\sqrt{p_h}} - \sum_{h=1}^{H}N_h S_h^2 \\
&= \frac{1}{n}\left(\sum_{h=1}^{H}\frac{N_h S_h}{\sqrt{p_h}}\right)^2 - \sum_{h=1}^{H}N_h S_h^2
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{n}\left(\sum_{h=1}^{H} N_h S_h\right)^2 \left(\sum_{h=1}^{H} \frac{N_h S_h}{\left(\sum_{h=1}^{H} N_h S_h\right)} \frac{1}{\sqrt{p_h}}\right)^2 - \sum_{h=1}^{H} N_h S_h^2 \\
&= \frac{N^2}{n}\left(\sum_{h=1}^{H} \frac{N_h}{N} S_h\right)^2 \check{q}^2 - \sum_{h=1}^{H} N_h S_h^2 \\
&= \frac{N^2}{n} \bar{S}^2 \check{q}^2 - \sum_{h=1}^{H} N_h S_h^2 \\
&= \frac{N^2}{n} \check{q}^2 \left[\sum_{h=1}^{H} \frac{N_h}{N} S_h^2 - S_S^2\right] - \sum_{h=1}^{H} N_h S_h^2 \\
&= \frac{N}{n}\left(N\check{q}^2 - n\right) \sum_{h=1}^{H} \frac{N_h}{N} S_h^2 - \frac{N^2}{n} \check{q}^2 S_S^2 \\
&\approx \frac{N}{n}\left(N\check{q}^2 - n\right)\left[S^2 - S_\mu^2 - \frac{N\check{q}^2}{(N\check{q}^2 - n)} S_S^2\right].
\end{aligned}
$$

References

Bankier, M. D. (1988). Determining sample sizes for subnational areas, *The American Statistician*, **42**, 174–177.

Choudhry, G. H., Rao, J. N. K. and Hidiroglou, M. A. (2012). On sample allocation for efficient domain estimation, *Survey Methodology*, **38**, 23–29.

Cornfield, J. (1944). On samples from finite populations, *Journal of the American Statistical Association*, **39**, 236–239.

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

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association*, **47**, 663–685.

Kokan, A. R. (1963). Optimum allocation in multivariate surveys, *Journal of the Royal Statistical Society, Series A (General)*, **126**, 557–565.

Lohr, S. L. (2009). *Sampling: Design and Analysis*, John Wiley & Cengage Learning.

Neyman, J. (1934). On the two different aspects of the representative method: The method of stratified sampling and the method of purposive selection, *Journal of Royal Statistical Society*, **97**, 558–625.

Sarndal, C. E., Swensson, B. and Wretman, J. (1992). *Model Assisted Survey Sampling*, Springer-Verlag, New York.

Sukhatme, B. V. and Tang, V. K. T. (1975). Allocation in stratified sampling subsequent to preliminary test of significance, *Journal of the American Statistical Association*, **70**, 175–179.

Sukhatme, P. V. (1935). Contribution to the theory of representative method, *Supplement to the Journal of the Royal Statistical society*, **2**, 253–268.

층화표본에서의 표본 배분에 대한 연구

이인규[a] · 박민규[a,1]

[a]고려대학교 통계학과

(2015년 7월 9일 접수, 2015년 9월 1일 수정, 2015년 10월 19일 채택)

요 약

층화표본추출(stratified sampling)은 모집단을 구성하는 층에 대한 정보를 표본설계에 반영함으로써 추정량의 분산을 낮추기 위한 표본추출 방법으로, 표본배분 방안의 선택이 층화표본의 효과를 결정하는데 매우 중요한 요소이다. 전통적인 표본배분 방법으로는 비례배분법(proportional allocation)과 네이만배분법(Neyman alloction)이 주로 사용되는데, 이는 층별 추정량의 분산에 영향을 미치는 요인들을 표본 배분에 반영함으로써 전체 추정량의 분산을 최적화하기 위한 것이다. 이론적으로는 층크기(size of strata)만을 반영하는 비례배분법보다 층별 표준편차(standard deviation)를 함께 고려하는 네이만배분법이 추정량의 분산을 낮추는데 더 효과적임이 알려져 있다. 그러나 층별 표준편차에 대한 사전 정보가 모집단을 잘 반영하지 못하면 네이만배분법의 효과를 기대할 수 없으며, 특히 복수의 관심변수를 조사하는 다목적조사(multi-purpose survey)에서는 각 관심변수들의 층별 표준편차가 서로 다른 양상을 나타내기 때문에 네이만배분법이 적합하지 않다는 주장이 제기되기도 한다. 한편 표본조사에서는 조사단계에서 발생하는 무응답으로 인한 추정량의 편향을 제거하기 위해 응답률 보정 방법이 사용되는데, 이 또한 추정량의 분산에 영향을 미치는 주요한 요인 중에 하나이다. 그러나 전통적인 표본배분 방법은 응답률(response rate)을 감안하지 않기 때문에 층별 응답율에 차이가 크게 나타날 경우 층화표본에 의한 효과가 저하될 수 있다. 이에 본 연구는 층화표본추출에서 층간 응답률의 차이가 추정량의 분산에 미치는 영향을 살펴보고, 층별 응답률 정보를 표본설계에 반영하는 새로운 표본배분 방법을 제안하였다. 모의실험을 통해 확인한 결과 네이만배분법은 당초 표본배분 시에 적용한 층별 표준편차의 구조가 각 층의 응답률 보정과정에서 증가하는 분산을 반영하지 못하기 때문에 층간 응답률의 편차가 커질수록 효율이 저하되는 것으로 나타났다. 반면 층 크기와 층별 응답률을 함께 반영한 배분방법은 비례배분법에 비해 효율이 개선되며, 층간 응답률의 편차가 클수록 그 효과는 커진다. 특히 층별 응답률의 변동계수(coefficient of variance)가 층별 표준편차의 변동계수를 상회하는 경우는 네이만배분법 보다도 효율적인 추정량을 제공함을 확인하였다. 아울러 응답률을 반영한 배분방법은 기존 배분방법에 비해 각 층별 추정량을 보다 안정적으로 추정할 수 있기 때문에 층별 추정을 목적으로 하는 층화표본조사에서는 여타 추정방법보다 더 효과적이다. 층별 응답률에 대한 정보는 관심변수가 다르더라도 추출틀이 유사한 기존 조사의 결과를 활용할 수 있다는 점에서 표준편차에 비해 비교적 정보 수집이 용이한 장점이 있고, 다목적조사에서도 관심변수의 척도(scale)나 개수와 관계없이 적용 가능하기 때문에 활용도가 높을 것으로 생각된다.

주요용어: 층화표본, 표본배분, 네이만배분법, 무응답

이 논문은 2013년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (2013R1A1A2006363).

[1]교신저자: (02841) 서울특별시 성북구 안암로 145, 고려대학교 통계학과. E-mail: mpark2@korea.ac.kr