The Korean Journal of Applied Statistics
2023, Vol. 36, No. 4, 337–347

DOI: https://doi.org/10.5351/KJAS.2023.36.4.337
Print ISSN1225-066X / Online ISSN 2383-5818

Measuring stratification effects for multistage sampling

Taehoon Kim[a], KeeJae Lee[b], Inho Park[1,c]

[a]Busan Public Health Policy Institute;
[b]Department of Statistics & Data Science, Korea National Open University;
[c]Department of Statistics & Data Science, Pukyong National University

Abstract

Sampling designs often use stratified sampling, where elements or clusters of the study population are divided into strata and an independent sample is chosen from each stratum. The stratification strategy consists of stratification and sample allocation, which are important issues that are repeatedly considered in survey sampling. Although a stratified multistage sample design is often used in practice, the literature tends to discuss simple sampling in terms of stratum effects or stratum efficiency. This study examines an existing stratum efficiency measure for two-stage sampling and further proposes additional stratum efficiency measures using the design effect model. The proposed measures are used to evaluate the stratification strategy of the sample design for high school students of the 4^{th} Korean National Environmental Health Survey (KoNEHS).

Keywords: stratification effects, multistage sampling, sample allocation, design effect models, Korean National Environmental Health Survey (KoNEHS)

1. 서론

조사연구를 위한 표본설계에서는 흔히 개체(element) 혹은 개체 묶음인 집락(cluster)을 몇 개의 층(strata)으로 나누고 각 층으로부터 독립적으로 표본을 추출한다. 층화추출(stratified sampling)을 채택하는 주된 이유(예, Lohr, 2010, p. 86, p. 310)는 다음과 같이 정리할 수 있다. 첫째, 층별로 개체 혹은 집락을 추출하여 특정 집단이 과도하게 포함되는 불균형적 표본을 방지할 수 있다. 둘째, 전체는 물론 영역별 분석이 가능하도록 모집단을 층으로 구분하고 층별로 적절한 표본할당을 통해 추정량의 정도수준을 정할 수 있다. 셋째, 조사운영을 층화와 연계하여 고려하여 행정적 편리성을 높일 수 있다. 또한, 집락을 일차추출단위(primary sampling unit; PSU)로 사용하는 지역표집의 경우에는 분산추정을 고려한 층구성을 고려하기도 한다 (Valliant 등, 2019, p. 258).

층화표본설계(stratified sample design)는 표본추출과 자료분석의 측면에서 다음의 네 단계로 구성된다 (Heeringa 등, 2010, p. 32). 첫째, N개의 개체(집락)로 이루어진 모집단을 $N_h(h = 1, \ldots, H)$개 개체(집락)로 이루어진 H개의 층으로 구분한다. 둘째, 전체 표본개체(집락)수 n을 층별로 할당하고 층별로 서로 독립적으로 n_h개의 표본개체(집락)를 확률추출한다. 셋째, 층별로 표본평균을 산출하고 층상대크기 $W_h = N_h/N$를 이용한 가중합으로 전체 모집단 평균의 표본추정치를 산출한다. 개체표집의 경우에 각 층별로 표본평균

This paper is based upon the first author's master thesis from the Pukyong National University, Busan, Korea.
Inho Park's work was supported by a Research Grant of Pukyong National University (2021).
[1] Corresponding author : Department of Statistics & Data Science, Pukyong National University, 45 Yongso-ro, Nam-gu, Busan 48513, Korea. E-mail: ipark@pknu.ac.kr

Published 31 August 2023 / journal homepage: http://kjas.or.kr
© 2023 The Korean Statistical Society. All rights reserved.

$\bar{y}_h = \sum_{i=1}^{n_h} y_{hi}/n_h$을 구한 후, 다음과 같이 모평균 추정량을 산출할 수 있다.

$$\bar{y}_{str} = \sum_{h=1}^{H} W_h \bar{y}_h.$$

표본론 교재 (예, Heeringa 등, 2010, p.33; Lohr, 2010, p.86; Letonen과 Pahkiton, 2004, p.63)에서는 종종 개체표집의 층효율성은 층내 동질성(within-stratum homogeneity)이 크고 층간 평균 이질성(between-stratum heterogeneity)이 클수록 좋음을 소개한다. 특히 비례할당(proportional allocation), 즉 $n_h = n \cdot W_h$을 적용하고 유한모집단 수정계수를 무시할 수 있다면, 평균추정량 $\bar{y}_{str}$의 분산 $V_{prop}(\bar{y}_{str})$은 단순확률추출(simple random sampling; SRS)의 표본평균 $\bar{y}_{srs}$의 분산 $V_{srs}(\bar{y}_{srs})$와 다음과 같이 관계를 가짐을 보일 수 있다.

$$V_{srs}(\bar{y}_{srs}) = V_{prop}(\bar{y}_{str}) + \frac{1}{n}\sum_{h=1}^{H} W_h (\bar{y}_{Uh} - \bar{y}_U)^2, \tag{1.1}$$

여기서 $V_{srs}(\bar{y}_{srs}) = S^2_{yU}/n$, $S^2_{yU} = \sum_{h=1}^{H}\sum_{i=1}^{N_h}(y_{hi} - \bar{y}_U)^2/(N-1)$, y_{hi}는 층 h 내 i 번째 개체의 조사값, $V_{prop}(\bar{y}_{str}) = \sum_{h=1}^{H} W_h S^2_{yUh}/n$, $S^2_{yUh} = \sum_{i=1}^{N_h}(y_{hi} - \bar{Y}_{Uh})^2/(N_h - 1)$, $\bar{y}_{hU}$와 $\bar{y}_U$는 각각 층 h와 모집단의 평균을 나타낸다. 식 (1.1)을 이용하면 비례할당의 층화추출에 따른 표본평균의 설계효과(design effect)는 다음과 같이 정의할 수 있다 (예, Lohr, 2010, p.210; Letonen과 Pakiton, 2004, p.63).

$$\text{deff}_{prop}(\bar{y}_{str}) = \frac{V_{prop}(\bar{y}_{str})}{V_{srs}(\bar{y}_{srs})} = \frac{\sum_{h=1}^{H} W_h S^2_{yUh}}{\sum_{h=1}^{H} W_h \left[S^2_{yUh} + (\bar{y}_{Uh} - \bar{y}_U)^2\right]} = \frac{S^2_{y,within}}{S^2_{y,total}}, \tag{1.2}$$

여기서 $S^2_{y,within} = \sum_{h=1}^{H} W_h S^2_{yUh}$이고 $S^2_{y,total} = \sum_{h=1}^{H} W_h[S^2_{yUh} + (\bar{y}_{Uh} - \bar{y}_U)^2]$이다. 식 (1.2)는 비례할당을 적용한 층화추출을 고려할 때 층평균 $\bar{y}_{Uh}$간의 차이를 크게 하는 층을 구성한다면 단순추출에 비해 평균추정량 $\bar{y}_{str}$의 정도수준을 높일 수 있음을 나타낸다. 다시말해, 비례할당을 적용한 층화추출은 단순추출이 주는 총변동 중 층 평균 간의 제곱변동을 차감하는 효과를 얻을 수 있다. 만약, 네이만할당(Neyman allocation), 즉 $n_h \propto N_h S_{yUh}$을 적용한다면, 층표준편차 S_{yUh} 간의 차이가 클수록 표본추정량의 정도수준이 높아짐을 보일 수 있다 (예, Lohr, 2010, p.111).

층화 전략(stratification strategy)은 크게 층구분과 표본할당으로 구성되는데 이는 조사연구에서 반복적으로 고려되는 중요한 주제이다. 층효과(stratification effects) 혹은 층효율성(efficacy of stratification)을 평가하는 측도는 앞서 논의한 바와 같이 단순추출에 대해서 주로 다루어지고 있다. 본 연구는 이단추출(two-stage sampling)에 대한 층효율성 평가에 관하여 살펴보았다. 2장에서는 이단추출의 층효율성에 대한 기존 측도를 먼저 소개하고, 추가로 설계효과모형(design effect model)을 활용한 층효율성 측도를 제안하였다. 3장에서는 2장에서 제안한 측도를 활용하여 제 4기 국민환경기초조사의 고등학생 대상 표본설계에 적용하여 층효율성을 평가하였다. 4장에서는 결론 및 향후연구에 대해 간단히 기술하였다.

2. 이단추출 표본설계의 층효율성

2.1. 기존연구

Krenzke와 Kali (2016, p. 43)는 다단추출의 층효율성을 평가하기 위해 다음의 측도를 고려하였다.

$$\Delta_{KK}(\bar{y}) = \frac{\text{층화를 반영한 분산추정값}}{\text{층화를 반영하지 않은 분산추정값}}. \tag{2.1}$$

이때 식(2.1)의 분모인 "층화를 반영하지 않은 분산추정값"이란 복합표본설계에서 층구분이 없음을 가정한 분산추정값으로 복합조사자료(complex survey data) 분석을 위한 전문 통계분석 패키지에서 층 옵션을 생략하여 산출한 분산값 의미한다(예로, SAS의 survey 프로시져에서 strata 옵션을 생략한다).

Krenzke와 Kali의 접근을 좀 더 분석적으로 표현하기 위해 층화이단추출에 대한 기본적인 정의를 먼저 소개한다. 모집단은 N개의 집락을 포함하며 N_h개의 집락으로 이루어진 H개 층으로 구분된다. 층별로 표본집락 $n_h(\leq N_h)$개를 추출하고 추출된 집락 내 M_{hi} 개체 중 표본개체 $m_{hi}(\leq M_{hi})$개를 추출한다. 표본개체(hik)에 대해 조사특성값 y_{hik}와 표본가중치 w_{hik}가 주어진다고 가정하자. 여기서 $h = 1,\ldots,H$, $i = 1,\ldots,n_{hi}$이고 $k = 1,\ldots,m_{hi}$이다. 층화이단추출을 $p(s)$으로 표기하면 평균추정량 $\bar{y} = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{k=1}^{m_{hi}} w_{hik}y_{hik}/M$의 분산추정량은 다음과 같이 주어진다.

$$v_{p(s)}(\bar{y}) = \frac{1}{M^2}\sum_{h=1}^{H}\frac{n_h}{n_h-1}\sum_{i=1}^{n_h}\left(\hat{t}_{hi\cdot} - \hat{\bar{t}}_{h\cdot\cdot}\right)^2, \tag{2.2}$$

여기서 $\hat{t}_{hi\cdot} = \sum_{k=1}^{m_{hi}} w_{hik}y_{hik}$이고 $\hat{\bar{t}}_{h\cdot\cdot} = n_h^{-1}\sum_{i=1}^{n_h}\hat{t}_{hi\cdot}$이다 (예, SAS, 2010, p. 7433). 만약 모든 층에 대해 추출률이 무시할 수 있을만큼 작고, $n_h/(n_h-1) \approx 1$이며 $n/(n-1) \approx 1$을 만족한다면, 식 (2.1)의 분모인 층화을 반영하지 않은 분산추정량은 다음과 같이 표현할 수 있다.

$$\begin{aligned} v_{nst}(\bar{y}) &\approx \frac{1}{M^2}\sum_{h=1}^{H}\sum_{i=1}^{n_h}\left(\hat{t}_{hi\cdot} - \hat{\bar{t}}_{\cdots}\right)^2 \\ &= \frac{1}{M^2}\sum_{h=1}^{H}\left\{\sum_{i=1}^{n_h}\left(\hat{t}_{hi\cdot} - \hat{\bar{t}}_{h\cdot\cdot}\right)^2 + n_h\left(\hat{\bar{t}}_{h\cdot\cdot} - \hat{\bar{t}}_{\cdots}\right)^2\right\} \\ &\approx v_{p(s)}(\bar{y}) + \frac{1}{M^2}\sum_{h=1}^{H} n_h\left(\hat{\bar{t}}_{h\cdot\cdot} - \hat{\bar{t}}_{\cdots}\right)^2, \end{aligned} \tag{2.3}$$

여기서 $n_{\cdot} = \sum_{h=1}^{H} n_h$는 총 표본집락수이다. 따라서 식 (2.1)는 분산식 (2.2)와 (2.3)을 적용하여 다음과 같이 표현할 수 있다.

$$\begin{aligned} \Delta_{KK}(\bar{y}) &= \frac{v_{p(s)}(\bar{y})}{v_{nst}(\bar{y})} \\ &\approx \frac{v_{p(s)}(\bar{y})}{v_{p(s)}(\bar{y}) + (1/M^2)\sum_{h=1}^{H} n_h\left(\hat{\bar{t}}_{h\cdot\cdot} - \hat{\bar{t}}_{\cdots}\right)^2}. \end{aligned} \tag{2.4}$$

식 (2.4)은 식 (1.2)와 유사하게 정의됨을 알 수 있다. Krenzke와 Kali (2016, p.43)는 $\Delta_{KK}(\bar{y})$을 영향률(impact rate)이라 정의하였다.

2.2. 설계효과모형을 활용한 층효율성 측도

표본추정에 대해 설계요소가 갖는 효율성은 흔히 단순추출과 비교한 분산의 상대적 크기로 설계효과로 평가한다. 예로, 식 (1.2)은 비례할당을 적용한 층화단순추출의 표본설계가 평균추정량에 대한 갖는 설계효과인데, 층평균간의 차이가 클수록 단순추출과 비교하여 작은 분산, 즉 작은 설계효과를 가짐을 보여준다. 설계효과가 표본설계의 주요 설계요소별 영향력 함수로 표현될 수 있다면 기존의 표본설계를 평가하고 재설계를 위한 개선안을 도출하는데 매우 유용할 수 있다 (예, Rust와 Broene, 2010).

Park (2015)과 Chen과 Rust (2017)는 복합표본설계에서 주로 고려하는 세 가지 설계요소인 층화, 집락추

출, 불균등가중치를 포함한 설계효과모형식을 다음과 같이 제안하였다.

$$\begin{aligned}\text{deff}_{PCR}(\bar{y}) &= \sum_{h=1}^{H}\left[\frac{W_h^2}{(m_{h\cdot}/m_{\cdot\cdot})}\frac{\sigma_{yh}^2}{\sigma_y^2}\right]\left[1+\rho_{yh}(m_h^*-1)\right]\left[1+cv_{wh}^2\right]\\ &= \sum_{h=1}^{H} d_{ysh}d_{ych}d_{wh}\\ &= \sum_{h=1}^{H} d_{ysh}\,\text{deff}(\bar{y}_h),\end{aligned} \tag{2.5}$$

여기서 $d_{ych} = 1+\rho_{yh}(m_h^*-1)$, $d_{wh} = 1+cv_{wh}^2$는 각각 표본층 h의 집락효과(effect due to cluster)와 가중치효과(effect due to weighting)이고 $d_{ysh} = W_h^2(\sigma_{yh}^2/\sigma^2)/(m_{h\cdot}/m_{\cdot\cdot})$는 층효과 혹은 층기여(effect due to stratification)을 나타내며, $\text{deff}(\bar{y}_h) = d_{ych}d_{wh}$는 층표본평균의 설계효과이다. 또한 $W_h = M_h/M$는 (개체수 기준)상대크기이고, $M_h = \sum_{i=1}^{N_h} M_{hi}$와 $M = \sum_{h=1} M_h$이며, $m_{h\cdot}/m_{\cdot\cdot}$는 (개체수 기준) 층 표본할당비(allocation rate)이며, σ_y^2 와 σ_{yh}^2는 각각 특성치 y의 (단위)모분산과 층분산이다. ρ_{yh}는 층 집락내 개체간 상관관계, 동질성계수(intracluster correlation; ICC)이며 $m_h^* = [\sum_{i=1}^{n_h}(\sum_{k=1}^{m_{hi}} w_{hik})^2]/\sum_{i=1}^{n_h}\sum_{k=1}^{m_{hi}} w_{hik}^2$으로 간단히 평균집락표본크기로 해석할 수 있고, cv_{wh}는 층내 개체가중치의 상대표준오차를 나타낸다.

만약 층구분이 없다면, 설계효과모형식 (2.5)은 다음과 같이 층별 차별성은 없는 집락효과와 가중치효과만으로 정리할 수 있다 (Kim, 2023; Valliant 등, 2019).

$$\text{deff}_{KG}(\bar{y}) = \left[1+\rho_{y\cdot}^*(m_{\cdot}^*-1)\right]\left[1+cv_{w\cdot}^2\right] = d_{yc}^*\,d_w^*, \tag{2.6}$$

여기서 $\rho_{y\cdot}^*$, $m_{\cdot}^*$, $cv_{w\cdot}$는 각각 층구분이 없는 표본설계의 동질성계수, 집락당 평균 표본수, 표본가중치 상대표준편차, $d_{yc}^* = 1+\rho_{y\cdot}^*(m_{\cdot}-1)$이 $d_w^* = 1+cv_{w\cdot}^2$이다. 식 (2.6)은 Kish (1987)가 직관적으로 제안하여 널리 쓰이고 있고 이후 Gabler 등 (1999)가 일원램덤효과모형(one-way random effect model)을 통해 입증한 설계효과모형식이다.

설계효과는 표본추정량에 대해 표본설계가 주는 단순추출과 비교한 분산증감분으로 정의되므로, Krenzke와 Kali의 층효율성 측도 (2.4)는 설계효과모형식 (2.5)과 (2.6)을 적용하여 다음과 같이 표현할 수 있다.

$$\Delta_{KK}(\bar{y}) = \frac{v_{p(s)}(\bar{y})}{v_{nst}(\bar{y})} \approx \frac{\text{deff}_{PCR}(\bar{y})}{\text{deff}_{KG}(\bar{y})} = \sum_{h=1}^{H} d_{ysh}\,r_{ych}\,r_{wh} = \sum_{h=1}^{H} d_{ysh}\,\text{reff}(\bar{y}_h), \tag{2.7}$$

여기서 $r_{ych} = d_{ych}/d_{yc}^*$, $r_{wh} = d_{wh}/d_w^*$이고 $\text{reff}(\bar{y}_h) = r_{ych}\cdot r_{wh}$이다. 식 (2.7)의 층효율성은 층화로 인한 차별적인 층별 상대적 집락효과 r_{ych}와 상대적 가중치 효과 r_{wh}에 대해 층기여도 d_{ysh}가 얼마나 효율적으로 조정하여 더해지는지로 결정됨을 보여준다. 다시말해, 집락효과 혹은 가중효과가 큰 층은 층기여도를 낮추고 그렇지 않은 층은 층기여도를 높인다면 층효율성을 높을 수 있다.

주어진 층구분 하에서 표본할당을 조정하면 보다 층효율성을 높일 수 있다. 예로, Park와 Hwang (2019)은 설계효과모형 (2.5)에 기반한 최적할당식을 다음과 같이 유도하였다. 먼저, 설계효과모형식 (2.5)를 이용하여 분산을 다음과 같이 추정 후,

$$v_{PCR}(\bar{y}) = \frac{s_y^{*2}}{m_{\cdot\cdot}}\text{deff}_{PCR}(\bar{y}) \tag{2.8}$$

표본수 조건식 $m_{\cdot\cdot} = \sum_{h=1}^{H} m_{h\cdot}$을 만족시키기 위한 최적할당식을 Cauchy-Schwarz 부등식으로 도출하면 다음과 같다.

$$m_{h\cdot}^{opt} \propto M_h\sigma_{yh}\sqrt{d_{ych}\,d_{wh}}, \tag{2.9}$$

여기서 $s_y^{*2} = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{k=1}^{m_{hi}} w_{hik}(y_{hik}-\bar{y})^2/\hat{M}$, $\bar{y} = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{k=1}^{m_{hi}} w_{hik}y_{hik}/\hat{M}$이고 $\hat{M} = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{k=1}^{m_{hi}} w_{hik}$이다. 따라서, 최적할당식 (2.9)을 설계효과모형에 적용하여 분산추정식을 정한다면, 표본할당을 수정하여 얻을 수 있는 평균추정량의 분산을 축소시킬 수 있는 최대비율이라는 관점에서 또 다른 층효율 측도를 다음과 같이 정의할 수 있다.

$$\Delta_{opt}(\bar{y}) = \frac{v_{opt}(\bar{y})}{v_{p(s)}(\bar{y})}. \tag{2.10}$$

2.3. 층구분 효율성 측도

조사특성 y의 모분산이 층구분에 따라 층내 개체간 변동과 층간 개체변동으로 어떻게 나뉘어지는지를 분산분해를 통해 살펴볼 수 있다. 우선, 모분산은 층구분에 따라 다음과 같이 분해할 수 있다.

$$\sigma_y^2 = \sigma_{yW}^2 + \sigma_{yB}^2,$$

여기서 $\sigma_y^2 = M^{-1}\sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{k=1}^{M_{hi}}(y_{hik}-\bar{y}_U)^2$, $\sigma_{yW}^2 = \sum_{h=1}^{H} W_h\sigma_{yh}^2$, $\sigma_{yB}^2 = \sum_{h=1}^{H} W_h(\bar{y}_{Uh}-\bar{y}_U)^2$, $W_h = M_h/M$, $M_h = \sum_{i=1}^{N_h} M_{hi}$, $\sigma_{yh}^2 = M_h^{-1}\sum_{i=1}^{N_h}\sum_{k=1}^{M_{hi}}(y_{hik}-\bar{y}_{Uh})^2$, $\bar{y}_U = M^{-1}\sum_{h=1}^{H}\sum_{i=1}^{N_h}\sum_{k=1}^{M_{hi}} y_{hik}$ 이고 $\bar{y}_{Uh} = M_h^{-1}\sum_{i=1}^{N_h}\sum_{k=1}^{M_{hi}} y_{hik}$ 이다. 따라서, 층구분에 따라 층평균간 차이를 분리하여 층내 개체변동량의 상대적 크기인 다음의 측도로 층효율성을 평가할 수 있을 것이다.

$$\Delta_{yW} = \frac{\hat{\sigma}_{yW}^2}{\hat{\sigma}_{yW}^2 + \hat{\sigma}_{yB}^2}, \tag{2.11}$$

여기서 $\hat{\sigma}_{yW}^2 = \sum_{h=1}^{H}\hat{W}_h\hat{\sigma}_{yh}^2$, $\hat{\sigma}_{yh}^2 = \sum_{i=1}^{n_h}\sum_{k=1}^{m_{hi}} w_{hik}(y_{hik}-\bar{y}_h)^2/\sum_{i=1}^{n_h}\sum_{k=1}^{m_{hi}} w_{hik}$이고 $\hat{\sigma}_{yB}^2 = \sum_{h=1}^{H}\hat{W}_h(\bar{y}_h-\bar{y})^2$이다. 층효율성 측도 (2.11)은 개체층화에 대한 비례할당을 적용한 층효율성인 설계효과 (1.2)와 유사함을 알 수 있다.

3. 자료분석

3.1. 제4기 국민환경기초조사 고등학교 조사자료

국민환경보건 기초조사(이하, 기초조사)는 환경 유해물질 노출 수준에 대한 원인 및 경로의 현황 파악을 목적으로 국가통계자료를 산출하여, 환경보건정책 수립을 위한 기초자료 제공을 목적으로 한다 (National Institute of Environmental Research, 2021). 자료수집 시 설문조사 및 인체 유래물 시료수집에 의한 임상검사를 통해 환경성 질환 발생과 연관성이 있는 납, 수은 등 중금속 및 유기화합물 33종에 대한 체내 잔류량 및 노출 수준을 분석한다.

본 연구는 고등학생을 대상으로 실시한 기초조사에서 채택한 표본설계의 층화 전략을 평가하기 위해 2장에서 논의한 층효율성 측도를 적용하였다. 고등학생을 대상으로 한 기초조사는 학교와 학생을 차례로 표집하는 층화이단추출법을 채택하였다. 전국을 5개 권역(서울, 경기·인천·강원, 충청권, 호남권, 영남권)으로 나누고, 서울을 제외한 나머지 권역은 시부와 군부로 구분한 뒤 추가로 학교구분(일반고, 특성화고)에 따라 세분화하여 총 9개의 층을 구성하였다. 표본할당은 학생수의 제곱근에 비례하는 제곱근할당(square-root allocation)을 적용하였고, 층별로 남녀공학, 남자, 여자학교 순으로 정렬한 후 학교 소재지에 따라 2차 정렬하고 학급수에 비례하는 표본학교를 추출하였다.

기초조사에서 수집하는 자료의 특성상 노출 시 지속적인 축적으로 인해 체내에 다량으로 발견될 수 있으며, 노출되지 않으면 체내에 극소량으로 존재하기 때문에 조사특성은 주로 극단적인 분포를 갖는다. 따라서,

Table 1: Geometric means, standard errors, design effects & efficiency measures

Type	Statistics	Variables					
		BPb ($\mu g/dL$)	BHg ($\mu g/L$)	Uhg ($\mu g/L$)	Ucd ($\mu g/L$)	MEP ($\mu g/L$)	MMP ($\mu g/L$)
Geometric mean standard error	$\hat{\theta}$	0.8199	1.3754	0.3187	0.1564	6.7210	3.5711
	$v_{p(s)}^{1/2}(\hat{\theta})$	0.0222	0.0467	0.0159	0.0124	0.9586	0.3478
	$v_{nst}^{1/2}(\bar{z})$	0.0280	0.0338	0.0488	0.0776	0.1428	0.1031
	$v_{opt}^{1/2}(\bar{z})$	0.0202	0.0441	0.0148	0.0117	0.7861	0.3107
	$v_{srs}^{1/2}(\bar{z}_{srs})$	0.0330	0.0247	0.1566	0.5082	0.0212	0.0273
Design effect	$\text{deff}_{PCR}(\hat{\theta})$	2.1633	2.0098	2.3847	2.3509	2.0276	2.6558
	$\text{deff}_{KG}(\hat{\theta})$	2.1700	1.9840	2.4327	2.2883	2.0720	2.7570
	d_{zc}^{*}	1.8809	1.7196	2.1085	1.9834	1.7959	2.3896
	d_{zw}^{*}	1.1538	1.1538	1.1538	1.1538	1.1538	1.1538
Stratification	$\Delta_{KK}(\hat{\theta})$	0.9969	1.0130	0.9802	1.0273	0.9786	0.9633
Efficiency	$\Delta_{opt}(\hat{\theta})$	0.8334	0.8923	0.8605	0.8837	0.6726	0.7977
Measure	Δ_{yW}	0.9620	0.9710	0.9800	0.9772	0.9654	0.9510

자료분석의 안정성을 높이기 위해 모평균의 표본추정량으로 조사특성의 기하평균 θ을 다음과 같이 추정하였다.

$$\hat{\theta} = \exp(\bar{z}),$$

여기서 $z_{hik} = \ln(y_{hik})$는 조사특성 y의 로그(변환)값이고, 가중평균

$$\bar{z} = \frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{k=1}^{m_{hi}} w_{hik} z_{hik}}{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{k=1}^{m_{hi}} w_{hik}} = \frac{\sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{k=1}^{m_{hi}} w_{hik} z_{hik}}{w_{...}}$$

이며, w_{hik}는 각 응답자에 부여된 표본가중치이다. 표본가중치는 일반적으로 고려되는 요소인 추출확률, 무응답조정, 사후층화, 가중치 절사 등이 반영되어 산출되었다. H는 전체 층의 수, n_h는 층 h의 표본학교수, m_{hi}는 층 h내 i 번째 표본학교 응답자수, $w_{...} = \sum_{h=1}^{H}\sum_{i=1}^{n_h}\sum_{k=1}^{m_{hi}} w_{hik}$는 전체 표본응답자들의 가중치 합으로 총 학생수 추정치이다.

기하평균추정 $\hat{\theta}$의 (근사적)분산추정량은 다음과 같이 주어진다 (Wolter, 2007).

$$v_{p(s)}\left(\hat{\theta}\right) = \hat{\theta}^2 v_{p(s)}(\bar{z}), \tag{3.1}$$

여기서 $v_{p(s)}(\bar{z}) = \sum_{h=1}^{H} n_h(1-f_h)/(n_h-1)\sum_{i=1}^{n_h}(e_{hi\cdot} - \bar{e}_{h\cdot\cdot})^2$, $f_h = n_h/N_h$, $e_{hi\cdot} = \sum_{k=1}^{m_{hi}} w_{hik}(z_{hik} - \bar{z})/w_{...}$이고 $\bar{e}_{h\cdot\cdot} = \sum_{i=1}^{n_h} e_{hi\cdot}/n_h$이다.

식 (3.1)에 의해 기하평균추정량 $\hat{\theta}$의 설계효과는 $\bar{z}$의 표본평균의 설계효과와 같고, 식 (2.4), (2.8), (2.10)의 층효율성 측도들은 각각 다음과 같이 표기될 수 있다.

$$\Delta_{KK}\left(\hat{\theta}\right) = \frac{v_{p(st)}(\bar{z})}{v_{nst}(\bar{z})},$$
$$\Delta_{opt}\left(\hat{\theta}\right) = \Delta_{opt}(\bar{z}),$$
$$\Delta_{yW}\left(\hat{\theta}\right) = \Delta_{zW}(\bar{z}).$$

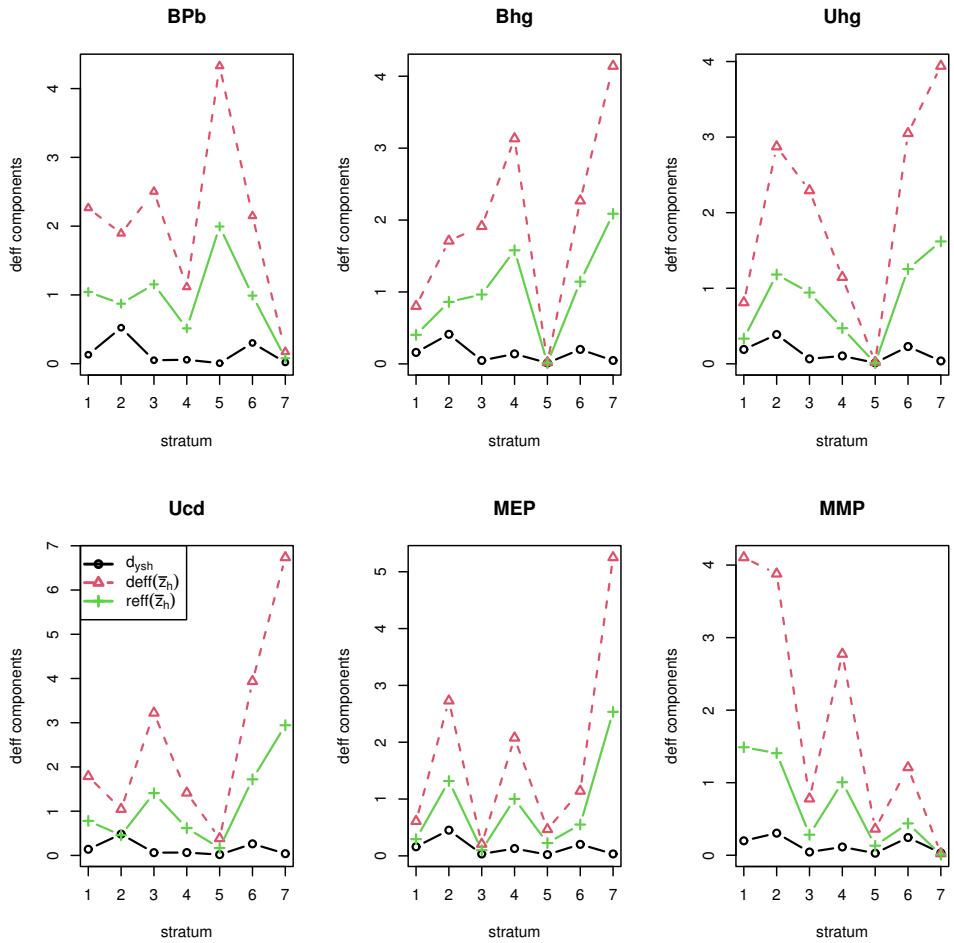

Figure 1: *Comparisons of stratification effects, within-stratum design effects and relative design effects.*

층효율성 분석을 위해 총 6가지 유해물질인 혈중 납(BPb), 혈중 수은(BHg), 요중 수은(UHg), 요중 카드뮴(UCd), 요중 모노 에틸 프탈레이트(MEP), 요중 모노 메틸 프탈레이트(MMP)의 농도 측정값을 고려하였다. 또한 분석의 편의를 위해 6가지 유해물질의 농도 측정값이 있는 경우만을 포함하였다.

3.2. 분석결과

Table 1은 주요 유해물질 농도에 대한 기하평균 추정값, 기존 및 가정적 표본설계에 따른 (예상)표준오차, 층화가 있는 다단추출과 층화가 없는 다단추출 가정에 따른 설계효과, 그리고 층효율성 측도를 비교하고 있다. 6개 유해물질 농도에 대한 기하평균 추정치의 표준오차 $v_{p(s)}^{1/2}(\hat{\theta})$는 0.0124−0.9586의 범위를 보이며 설계효과 $\text{deff}_{PCR}(\hat{\theta})$는 2.0276−2.6558의 범위를 갖는다. 층화없는 다단추출 가정하의 설계효과 $\text{deff}_{KG}(\hat{\theta})$는 1.9840−2.7570의 범위를 보이고 있는데, 층화를 포함할 때에 비해 혈중 수은(BHg)과 요중 카드뮴(UCd)에서 낮은 값을 보였고 혈중 납(BPb), 요중 수은(UHg), 요중 모노 에틸 프탈레이트(MEP), 요중 모노 메틸 프탈레이트(MMP)에

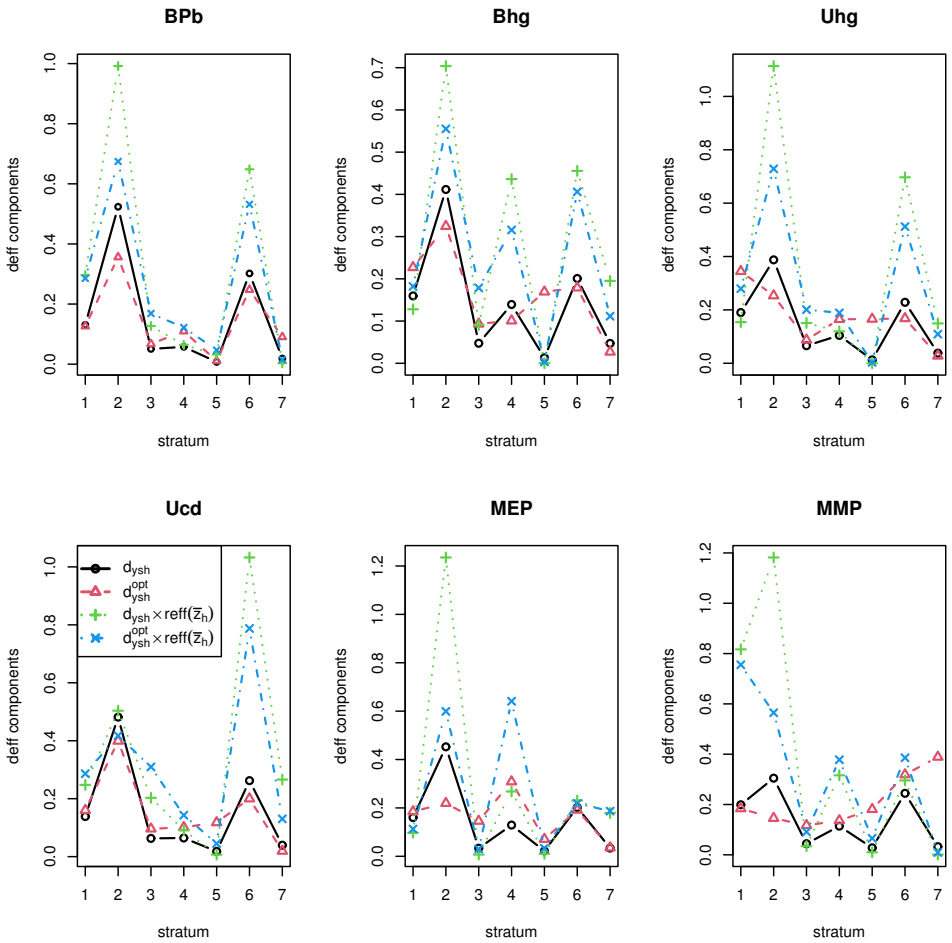

Figure 2: *Comparisons of current stratum effects, optimal stratum effects and their stratum contributions.*

대해 다소 높게 나타났고 게 나타났다. Kish-Gabler 설계효과모형식 (2.6)에 기초하여 층별 차별성을 무시하고 측정한 가중치효과 d_w^*는 1.1538이고 집락효과 d_{zc}^*는 1.7196–2.3896 범위의 값을 보여 설계효과의 대부분이 집락효과로 인한 결과임을 알 수 있다.

6개 유해물질 농도의 개체분산은 층구분에 따라 층평균 차이에 따른 변동량을 제거함에 따라 95.1–98.0% 정도의 층내 개체간 변동비 Δ_{yW}를 갖는다. 표본할당을 추가한 층화전략을 통해 혈중 납(BPb), 요중 수은(UHg), 요중 모노 에틸 프탈레이트(MEP), 요중 모노 메틸 프탈레이트(MMP)는 0.9969, 0.9802, 0.97, 0.9633의 좋은 층효율성 $\Delta_{KK}(\hat{\theta})$을 보인 반면, 혈중 수은(BHg), 요중 카드뮴(UCd)은 1.0130, 1.0273의 낮은 층효율성을 보였다.

기존 표본설계는 층별 제곱근할당을 고려하여 전체 대비 층별 표본추정량이 어느정도 정도수준을 확보할 수 있도록 총 학생수의 제곱근비례로 표본크기를 절충하였다. 만약, 층별 절충적 표본할당을 고려하지 않고 전국 수준의 기하평균 추정값의 정도수준이 최적화 할 수 있도록 한다면 층효율성 측도 $\Delta_{opt}(\hat{\theta})$에서 보듯 6개 유해물질 농도 모두에 대해서 67.3–89.2% 정도로 분산값을 낮출수 있음 알 수 있다.

Figure 1은 각각의 유해물질 농에 대해 층별로 층기여도 d_{ysh}, 층평균의 설계효과 deff($\bar{z}_h$), 그리고 식 (2.7)의 상대적 설계효과 reff($\bar{z}_h$)를 비교하고 있다. 유해물질에 따른 층별 표본평균의 (상대)설계효과가 매우 상이함을 확인할 수 있으며 층화 전략에 따른 해당 유해물질 농도에 대한 층별 영향력이 어떻게 결정되었는지를 알 수 있다. 예로, 혈중 납(BPb)의 경우에는 층 5의 집락효과가 매우 크지만 층기여도를 낮게하여 전체적인 설계효과를 낮출 수 있었음을 확인할 수 있다.

앞절에서 언급한대로 기초조사의 층화전략은 지역에 따른 층분리와 지역별 규모차이를 절충하는 제곱근비례 표본할당을 채택하고 있다. Figure 2은 개별 유해물질 농도에 대해 기존 표본할당은 물론 전국수준의 기하평균 추정량의 정도수준을 최적화하는 식 (2.9)의 표본할당을 채택할때 예상되는 층기여도를 비교하고 있다. 전반적으로 층 2, 4와 6의 기존 층기여도 d_{ysh}에 비해 최적할당에 따른 층기여도 d_{ysh}^{opt}가 낮아질 수 있다. Table 1의 층효율성 측도 $\Delta_{opt}(\hat{\theta})$에서 알 수 있드시 요중 모노 에틸 프탈레이트(MEP)는 현재의 분산값을 67.3%으로 낮출수도 있을 것으로 기대할 수 있다.

4. 결론

본 연구는 다단추출 표본설계에서 층화전략에 대해 Krenzke와 Kali가 제시한 층효율성 측도 $\Delta_{KK}(\hat{\theta})$에 대하여 살펴보고, 이에 Park (2015)과 Chen과 Rust (2017)이 제시한 설계효과모형(design effect model)을 적용하여 층효율성 측도를 복합표본설계의 주요한 3개의 설계요소인 층화, 집락추출, 가중치효과로 분해하여 살펴보았다. 또한, 전체수준의 평균추정량의 분산이 최소화될 수 있는 최적할당을 적용하여 기대할 수 있는 분산축소율로 해석되는 층효율성 측도 $\Delta_{opt}(\hat{\theta})$를 제안하였다. 마지막으로 층구분에 따른 층간 변동량 제거로 얻게 되는 층내 개체 변동량 비율인 층효율성 측도 Δ_{yW}을 제안하였다.

제안된 층효율성 측도들은 고등학교 학생을 대상으로한 제4기 국민환경기초조사의 표본설계를 평가하는데 적용하였으며 조사특성별로 층화에 따른 설계요소간 효과를 분리한 효율성을 이해할 수 있음을 확인하였다. 제안된 층효율성 평가 측도는 다단추출 표본설계에 의해 수행된 조사결과를 이용한 사후적 평가를 가능하게 할 것으로 기대한다. 또한 Krenzke와 Kali (2016)은 그들이 제안한 측도에서 층정보를 무시하고 계산하는 다단추출 가정 하의 분산추정치 $v_{nst}(\bar{y})$이 층화추출로 인해 갖는 집락간 변동성(dispersion)을 적절히 반영하지 못해 분산값이 과대하게 계상할 수 있다고 논하고 있다.

향후 연구에서는 우선 앞서 논의한 점을 개선한 분산추정치 혹은 설계효과모형식을 살펴볼 수 있을 것이다. 또한 제안된 층효율성 측도에 대해 모의실험을 통해 적합성 및 안정성을 평가할 수 있을 것이다. 예로, 혼합 랜덤 효과모형 (예, Feng, 2006)을 통해 유한모집단을 생성한 후, 층화다단 표본설계에 따른 반복적 표집을 통해 층효율성의 평가를 고려할 수 있을 것이다. 더불어 Krenzke와 Kali (2016)이 실질적으로 고려한 일차추출단위(PSU) 추출의 명시적 층화(explicit stratification)가 아닌 크기비례추출(probability proportionate to size; PPS)로 적용된 내재적 층화(implicit stratification)의 경우에 적합한 연속적 차이(successive differences) 방식의 분산추정 (Wolter, 2007)에 대해서도 다룰 수 있을 것으로 판단한다.

References

Chen S and Rust K (2017). An extension of Kish's formula for design effects to two- and three-stage designs with stratification, *Journal of Survey Statistics and Methodology*, **5**, 111–130.

Feng X (2006). On confidence intervals for proportions with focus on the U.S. National Health and Nutribution Examination Surveys (Master's thesis), Simon Fraser University, Burnaby, BC, Canada.

Gabler S, Häder S, and Lahiri P (1999). A model based justification of Kish's formula for design effects for weighting and clustering, *Survey Methodology*, **25**, 105–106.

Heeringa SG, West BT, and Berglund PA (2010). *Applied Survey Data Analysis*, Chapman and Hall/CRC, New York.

Kim TH (2023). Stratification effects under multistage sampling (Master's thesis), Pukyong National University, Busan, Korea.

Kish L (1987). Weighting in $Deft^2$, *Survey Statistician*, **June**, 26–30.

Krenzke T and Kali J (2016). *Review of the FoodAPS 2012 Sample Design*. A technical report by Westat prepared for Economic Research Service, U.S. Department of Agriculture.

Lehtonen R and Pahkinen E (2004). *Practical Methods for Design and Analysis of Complex Surveys* (2nd ed), Chapman and Hall/CRC, New York.

Lohr SL (2010). *Sampling: Design and Analysis* (2nd ed), Brooks/Cole, Boston.

National Institute of Environmental Research (2021). *Sample documentation for cycle 4 of the Korean National Environmental Health Survey (KoNEHS)*.

Park I (2015). Understanding complex design features via design effect models, *The Korean Journal of Applied Statistics*, **28**, 1217–1225.

Park I and Hwang HG (2019). Sample size determination using design effect formula for repeated surveys, *The Korean Journal of Applied Statistics*, **32**, 1–10.

Rust K and Broene P (2010). Design effects for totals in multi-stage samples, In *Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods, American Statistical Association*, Vancouver, British Columbia, 2174–2181.

SAS Institute Inc. (2010). *SAS/STAT 9.3 User's Guide*, Cary, NC: SAS Institute Inc.

Valliant R, Dever JA, and Kreuter F (2019). *Practical Tools for Designing and Weighting Survey Samples* (2nd ed), Springer, New York.

Wolter KM (2007). *Introduction to Variance Estimation* (2nd ed), Springer, New York.

Received June 17, 2023; Revised June 18, 2023; Accepted June 19, 2023

다단추출 표본설계의 층효율성 연구

김태훈[a], 이기재[b], 박인호[1,c]

[a]부산광역시 공공보건의료지원단; [b]한국방송통신대학교 통계데이터사이언스학과;
[c]부경대학교 데이터정보과학부 통계데이터사이언스 전공

요 약

표본설계는 개체 혹은 집락을 층으로 나눈후 층별로 독립적으로 표본추출하는 층화추출을 종종 채택한다. 층화 전략은 크게 층구분과 표본할당으로 구성되는데 이는 조사연구에서 반복적으로 고려되는 중요한 주제이다. 조사연구에서는 층화다단추출 방식의 복합표본설계를 채택하고 있지만 층효과 혹은 층효율성과 관련하여서 표본론 교재들에서 주로 단순추출에 대해서 다루어지고 있다. 본 연구는 이단추출에 대한 기존 층효율성 측도를 살펴보며 설계효과모형을 적용한 추가적인 층효율성 측도들을 제안하였다. 제안된 측도들을 활용하여 제4기 국민환경기초조사의 고등학교 대상 표본설계의 층화전략에 대해 평가하였다.

주요용어: 층화효과, 다단추출, 표본할당, 설계효과모형, 국민환경기초조사

본 논문은 제1저자의 부경대학교 석사학위 논문에 기초하였고, 교신저자의 부경대학교 자율창의학술연구비 (2021년)에 의하여 연구되었음.

[1]교신저자 : (48513) 부산광역시 남구 용소로 45, 부경대학교 데이터정보과학부 통계데이터사이언스 전공. E-mail: ipark@pknu.ac.kr