



저작자표시 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.
- 이차적 저작물을 작성할 수 있습니다.
- 이 저작물을 영리 목적으로 이용할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.

- 귀하는, 이 저작물의 재이용이나 배포의 경우, 이 저작물에 적용된 이용허락조건을 명확하게 나타내어야 합니다.
- 저작권자로부터 별도의 허가를 받으면 이러한 조건들은 적용되지 않습니다.

저작권법에 따른 이용자의 권리는 위의 내용에 의하여 영향을 받지 않습니다.

이것은 [이용허락규약\(Legal Code\)](#)을 이해하기 쉽게 요약한 것입니다.

[Disclaimer](#) 

이 학 석 사 학 위 논 문

다단 층화추출 표본설계의 층 효율성 연구



2023년 2월

부 경 대 학 교 대 학 원

통 계 학 과

김 태 훈

이 학 석 사 학 위 논 문

다단 층화추출 표본설계의 층 효율성 연구

지도교수 박 인 호

이 논문을 석사 학위논문으로 제출함.



2023년 2월

부 경 대 학 교 대 학 원

통 계 학 과

김 태 훈

김태훈의 이학석사 학위논문을 인준함.

2023년 2월 17일



위 원 장 이학박사 장대흥 (인)

위 원 이학박사 박인호 (인)

위 원 이학박사 엄태웅 (인)

목 차

표 목차	ii
그림 목차	iii
논문 요약	iv
1. 서론	1
2. 층화단순추출의 표본설계의 층 효율성	3
2.1 층화추출 개요	3
2.2 층별 표본할당	4
2.3 층 경계 결정	6
2.4 층 효율성 평가	7
3. 다단추출의 층 효율성	11
3.1 층화다단확률추출 개요	11
3.2 일차추출단위 층화	12
3.3 설계효과 모형	13
3.4 표본할당	15
3.5 층 효율성 평가	17
4. 사례연구	20
4.1 국민 환경 보건 기초조사 개요	20
4.2 기하평균 추정량	23
4.3 결과 요약	24
5. 결론 및 향후과제	31
참고문헌	32
부록 A.1 프로그램	34
부록 A.2 Abstract	50

표 목 차

표 1. 증화단순확률추출 표본설계의 분산분석표	8
표 2. 국민 환경 보건 기초조사 고등학생 수 현황	21
표 3. 국민 환경 보건 기초조사 표본 고등학생 수 현황(비례배분)	22
표 4. 국민 환경 보건 기초조사 표본 고등학생 수 현황(제곱근 비례배분)	22
표 5. 국민 환경 보건 기초조사 중·고등학생 대상조사 결과 (혈중 납)	24
표 6. 국민 환경 보건 기초조사 중 효율 측도	27



그 립 목 차

그림 1. 층화 효과의 적극성 측도	25
그림 2. 층화 전략별 신규 층 효율성 측도	27
그림 3. 층화 전략 Sqrt의 층별 SE 상대 편차	29
그림 4. 층화 전략 Global의 층별 SE 상대 편차	29
그림 5. 층화 전략 Local의 층별 SE 상대 편차	30



다단 층화추출 표본설계의 층 효율성 연구

김 태 훈

부경대학교 대학원 통계학과

요 약

설계효과(design effect)란 같은 크기의 단순 확률표본 대비 현 표본설계 하의 추정량 분산비로 정의되며, 단순 확률표본과 비교했을 때 현 표본설계의 효율성을 의미한다. 따라서, 층화단순확률 표본설계의 설계효과는 단순 확률표본 대비 층화에 따른 효율성을 의미하여, 층 효율성으로 해석할 수 있다. 하지만, 층화다단확률 표본설계의 설계효과는 층화는 물론 다단추출 등의 복수의 설계요소를 복합적으로 반영하기 때문에 층 효율성의 측도로는 부적절하다고 할 수 있다. 본 연구에서는 층화다단확률 표본설계의 주요 설계요소로 층화, 집락, 불균등 가중치의 3가지를 고려한 설계효과 모형을 사용하되 표본할당을 중점적으로 기존 층화 전략이 적절한지를 다른 대안적 층화 전략과 비교할 수 있는 측도를 제시하였다. 신규 측도의 적합성을 평가하기 위해 국립환경 보건 연구원이 주관하는 제 4기 국민 환경 보건 기초조사의 고등학생 대상 조사자료를 사용하여 층 효율성 평가 측도를 계산하였다. 제안된 층 효율성 평가 측도는 조사수행 후 조사자료에 대한 다양한 조사 변수에 대해 사후적 평가를 실시할 수 있고 이를 바탕으로 표본 재설계를 위한 개선된 층화 전략을 제시할 수 있도록 고안되었다.

주요용어(Keywords) : 층화다단추출(stratified multistage sampling), 층화(stratification), 표본할당(sample allocation), 설계효과 모형(design effect model), 층 효과(stratification effect), 집락 효과(cluster effect), 가중치 효과(weighting effect)

1. 서론

표본설계(sampling design)는 연구 대상이 되는 모집단에 대하여 표본 추출틀을 설정하고 적절한 추출 방법을 통해 조사 방법을 선택하는 일련의 과정을 말한다. 모집단 특성인 모수를 추정할 때, 표본추출이 얼마나 효율적인가에 대해서 사후적으로 평가한다면, 이후 이를 반영한 표본설계 개선을 가능하게 할 수 있다. 따라서, 기존 연구에서는 표본설계의 적절성 및 효율성에 대한 다양한 측도들이 제시되었다.

층화 표본설계(stratified sampling)에서는 먼저 모집단을 서로 겹치지 않는 부모집단(sub-population)인 표본 층으로 구분하고 층별로 적절한 표본 규모를 결정한 후, 층별로 독립적인 확률표본을 추출한다. 이때, 층화 변수에 따라 조사특성의 층 내 및 층 간 동질성이 결정된다. 효과적인 층화 전략이란 층화 변수 선택에 따른 층 경계점 결정 및 표본할당을 통합적으로 일컫는다.

설계효과(design effect)는 표본설계의 상대적 효율성을 수치적으로 나타낸다. 다시 말해, 동일 크기의 단순 확률표본과 비교하여 표본설계 하의 추정량 분산비를 통해 현 표본설계가 단순 확률표본과 비교했을 때 얼마나 효과적인가를 나타낸다.

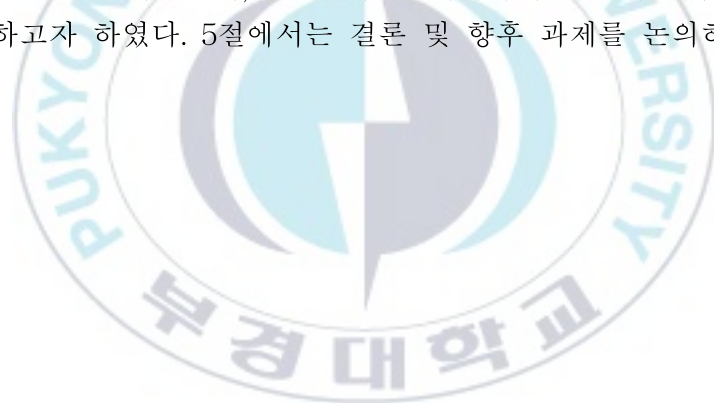
층 효율성은 층화 전략의 선택이 적절하였는지를 수치적으로 평가할 수 있을 것이다. 층화단순확률 표본설계에서는 층 효율성 특성치 변동성의 분산분해를 이용하면 단순 확률추출의 분산과 층화 단순추출의 분산비로 표현할 수 있다(Lohr, 2019). 이는 층화 단순추출의 설계효과와 동일하다. 하지만, 복합표본설계(complex sample design)의 설계효과는 층화 이외에 다른 설계요소를 포함하고 있기 때문에 층 효율성 측도로 사용하기에 부적절하다. Park(2015)와 Chen과 rust(2017)는 층화다단확률 표본설계의 설계효과는 층화 효과만이 아닌, 집락 효과, 가중치 효과와 같은 설계요소의 함수형태의 모형식(design effect formula)으로 근사적으로 표현될 수 있음을 보였다. 따라서, 층 효율성은 설계효과를 사용하여 평가하기에는 집락 효과와 같은 다른 설계요소의 영향력이 혼동되어 있다고 할 수 있다.

Krenzke et al. (2017)은 층화다단확률 추출에서 층 효율성을 계산하기 위해 프로그래밍적 접근으로 층화 옵션을 사용한 분산 추정값과 층화 옵션을

사용하지 않은 분산 추정값의 비교를 통해 층 효율성을 계산하였다. 하지만, 이러한 접근은 경험적 측면이 강조되고 분석적인 측면은 다소 결여되어 있다고 할 수 있다.

본 연구에서는 층화다단확률 표본설계의 설계효과 모형식을 통해 층별 분산을 계산하고, 층화 전략 간 비교를 위해 층화 효과 요소 중 표본 배분을 통한 효과를 층 효율성 측도로 제시하였다. 이를 실제 사례에 대입하여 분석하는 것으로 층화 전략의 개선 및 보완점을 제시하였다.

본 논문의 구성은 다음과 같다. 2절에서는 층화단순확률 표본설계에 대하여 층화 표본 배분 방법 및 층 경계점에 대해 알아보고, 층 효율성 평가 측도로 사용되는 설계효과 및 분산분석을 살펴보았다. 3절에서는 층화다단확률 표본설계의 설계효과에 대하여 설명하고, 층화 효과 요소인 층별 표본 배분을 통해 층별 분산을 활용하여 층 효율의 측도를 제시하였다. 4절에서는 신규 측도를 활용하여 제 4기 국민 환경 보건 기초조사의 층 효율 측도 계산 및 평가를 하였으며, 모의실험을 계획하여 측도의 적합성 및 안정성을 평가하고자 하였다. 5절에서는 결론 및 향후 과제를 논의하였다.



2. 층화단순추출 표본설계의 층 효율성

단순확률추출(simple random sampling, srs)은 가장 단순하지만 기본적인 확률추출이며, 크기 n 개의 모든 가능한 단순확률표본은 동일한 추출확률을 갖는다. 단순확률추출은 모집단 내 모든 개체에 대해 관심 특성과 관계없이 동일한 확률로 선택되지만 관심 특성의 특정한 값들이 집중적으로 포함될 가능성도 존재한다. 표본 추출에 앞서 모집단 개체들에 대해 관심 특성에 따른 동질적인 집단들로 나눈 후, 개별 집단들로부터 단순확률표본을 추출한다면 관심 특성에 대해 균형을 갖춘 표본을 추출할 수 있다. 본 장에서는 모집단을 특정 특성에 대해 층으로 분리한 후 개별 층으로부터 단순 확률추출을 통해 표본을 선택하는 층화 단순추출의 표본설계에서의 층 효율성에 대해서 논의한다.

2.1 층화추출 개요

층화추출(stratified sampling)은 모집단을 서로 겹치지 않는 복수의 층(strata)으로 나누고 각 층으로부터 독립적으로 단순확률 표본을 추출한다. 주요 관심 변수와 높은 상관관계를 갖는 특성을 층화 변수로 정의한다면 층간 이질성은 높게 하되, 층 내 이질성은 낮게 하여 표본 추정의 효율성을 높이는 효과적인 표본설계가 가능하다. 예를 들어, 임상시험과 같은 사람을 대상으로 한 조사는 성별 및 체중, 키에 대한 변수가 주요인이기 때문에 이러한 특성등을 층화 변수로 지정하여, 층내 동질성은 높이되 층간 이질성을 제고하여 층화 변수 기준의 균형을 갖춘 표본추출을 가능하게 할 수 있다. 단순확률추출에 비해 표본 추정량의 분산이 작게 된다. 이와 관련한 내용은 2.4절에서 추가로 논의하고자 한다.

층화단순추출의 표본설계는 크기 N 의 모집단을 먼저 H 개의 서로 겹치지 않는 크기 N_h 의 층으로 나눈 후 층별로 서로 독립적인 크기 n_h 의 표본을 확률추출한다. 유한모집단 내 $h(=1, 2, \dots, H)$ 번째 층의 $j(=1, 2, \dots, N_h)$ 번째 개체인 조사특성 y_{hj} 를 가질 때, 표본 평균 및 표본분산은 다음과 같이 정의한다.

$$\bar{y}_{str} = \frac{1}{N} \sum_{h=1}^H N_h \bar{y}_h = \sum_{h=1}^H W_h \bar{y}_h \quad (2.1.1)$$

$$V(\bar{y}_{str}) = \sum_{h=1}^H W_h^2 V(\bar{y}_h) = \sum_{h=1}^H (1-f_h) W_h^2 \frac{S_h^2}{n_h} \quad (2.1.2)$$

여기서 $W_h = N_h/N$ 는 모집단의 층 규모, $f_h = n_h/N_h$ 는 층별 유한모집단 수정계수, $S_h^2 = (N_h - 1)^{-1} \sum_{j=1}^{N_h} (y_{hj} - \bar{Y}_h)^2$ 는 층별 분산, $N = \sum_{h=1}^H N_h$ 와 $n = \sum_{h=1}^H n_h$ 는 각각 모집단과 표본의 총 크기이다.

또한, 층 표본 평균 $\bar{y}_h = \sum_{j=1}^{n_h} y_{hj}/n_h$ 의 분산은 $V(\bar{y}_h) = S_h^2 (1-f_h)/n_h$ 이다. 층 분산의 불편추정량은 다음과 같다.

$$v(\bar{y}_h) = \frac{1-f_h}{n_h} \frac{1}{n_h-1} \sum_{j=1}^{n_h} (y_{hj} - \bar{y}_h)^2$$

2.2 층별 표본할당

층화단순 확률추출의 표본설계 하에서는 층화 변수를 통해 층을 결정한 후, 조사예산 및 추정량 분산을 고려하여 전체 표본 수를 적절한 기준에 따라 층별로 할당한다.

층별 표본배분은 연구 목적에 맞게 층별 비용과 분산을 고려하게 된다. 일반적으로 일정 수준의 분산을 갖출 수 있도록 하되 비용을 최소화할 수 있도록 고려하거나, 일정 비용하에서 추정량의 분산을 최소화할 수 있는 표본할당을 고려한다. 분산과 비용 함수식을 이용하면 최적 배분법을 정의할 수 있다. 먼저, 층화단순추출의 표본설계의 표본평균의 분산식 (2.1.2)와 총 조사비용 C 를 다음과 같이 층별 분산과 비용의 선형함수임을 가정한다.

$$V(\hat{t}_{str}) = V_0 + \sum_{h=1}^H \frac{V_h}{n_h}$$

$$C = C_0 + \sum_{h=1}^H C_h \times n_h$$

V_h 와 C_h 는 층화단순확률 표본설계 시 층별 분산 및 비용이며, V_0 와 C_0 는 이전 조사연구를 통해 알려지거나, 연구에서 표본할당과 관계없이 상수로 정의된다. 이때 $V_h = (1-f_h) \frac{S_h^2}{n_h}$ 이다. 해당 함수식들을 통해 층 할당 시 층 크기 n_h 를 계산하여, 층 크기를 조절하는 것으로 층별 분산 및 층별 비용의 적정값을 설정할 수 있다. 여기서 분산과 비용을 모두 고려한 할당을 최적 할당(optimal allocation)이라 한다(Lohr, 2010. p. 89).

$$n_h = \left(\frac{\frac{N_h S_h}{\sqrt{c_h}}}{\sum_{l=1}^H \frac{N_l S_l}{\sqrt{c_l}}} \right) n$$

표본할당은 다양하게 제시되고 있으며 대표적인 방법으로는 다음과 같이 네이만 배분법, 비례 배분법, 균등 배분법이 있다.

- 네이만 배분법(neyman allocation)

$$n_h = \frac{N_h S_h}{\sum_{l=1}^H N_l S_l} \times n$$

- 비례 배분법(proportion allocation)

$$n_h = \frac{N_h}{\sum_{l=1}^H N_l} \times n$$

- 균등 배분법(equal allocation)

$$n_h = \frac{1}{H} \times n$$

층화는 개체들의 조사특성이 층 내는 동질적이고, 층 간은 이질적일수록 효율적이다. 따라서, 층별 분산을 고려하는 네이만 배분법은 층별 분산을 고려하지 않는 비례 배분법보다 효율적이라고 할 수 있다. 그러나, 만약 분석 변수가 다양하다면, 층화 변수의 분산을 고려하는 네이만 배분은 층화 변수가 아닌 다른 변수에서 층화가 비효율적일 가능성이 존재한다. 따라서, 적절한 층화 변수의 설정이 필요하다.

2.3 층 경계 결정

층화 변수는 표본설계 시 조사연구의 목적에 의해 결정된다. 예를 들어, 청소년들의 연령별 흡연 실태조사와 같이 특정 집단의 조사연구를 목적으로 하는 경우, 층화 변수로써 연령대를 고려할 수 있다. 기존 연구 결과가 존재한다면 지역이나 성적을 층화 변수로써 고려할 수도 있다.

층 경계점 결정 방법은 집단 간의 이질성이 클수록 또한, 집단 내 분산이 작을수록 층화는 더욱 효과적이라는 관점에서 고려될 수 있다. 전통적인 방법으로 누적 도수 제공근 법(Dalenius & Hodges, 1959)이 있는데, 이는 임의의 구간별 관측치 도수 f_i ($i=1, 2, \dots, L$)에 대해 구간 내 누적 빈도수의 제공근이 같도록 임의의 층 경계를 설정한다. 해당 방법은 분산을 축소하는 등의 효과가 존재하나 도수가 0인 구간의 존재성을 고려하는 등 선행 과제가 존재한다.

최적 층 경계 결정 알고리즘(Lavalley-Hidirolou, 1988)은 일정한 목표 상대오차값 $CV=cv_0$ 를 갖도록 하는 경계점 $b_1, b_2, \dots, b_{L-1}$ 을 특정 할당법 하에서 찾는다.

두 경우는 비용의 차이 및 표본 크기 차이가 존재할 수 있으나 결과 단계에서 정밀도, 즉, 층 분산에 영향을 미치지 않는다. 층화 표본설계는 이전 조사연구의 결과를 반영하거나, 목적에 의해 층화를 할 때의 복잡성이 결

과 분석 단계에서 층화를 한 이후의 정밀도에 영향을 미치지 않기 때문이다.

2.4 층 효율성 평가

층화단순확률추출의 표본설계에서 층 효율성은 중요한 논점 중 하나이며 층화 전략이 효율적이었는지 기준 및 측도를 통해 판단할 수 있다. 일반적으로 표본설계 간의 효율성 비교는 분산추정량을 바탕으로 하며, 층화 표본설계에서는 다음의 두 가지 방법을 통해 층 효율성을 평가한다.

1) 분산분석

효율성을 계산하기 위해 층화단순확률추출과 단순확률추출의 두 가지 표본설계의 분산추정량 비를 층 효율성 측도로 사용할 수 있다. 계산의 편의성을 위해 두 표본설계에서 모두 복원추출과 비례할당으로 동일 크기를 갖는다고 가정할 때, 각각의 표본설계 하에서의 평균추정량의 분산추정량은 다음과 같이 주어진다.

$$V(\bar{y}_{strwr}) = \frac{1}{n} \sum_{h=1}^H W_h S_{yh}^{*2} \quad (2.4.1)$$

$$V(\bar{y}_{srswr}) = \frac{1}{n} S_y^{*2} \quad (2.4.2)$$

여기서 $S_{yh}^{*2} = \sum_{k=1}^{N_h} (y_{hk} - \bar{Y}_h)^2 / N_h$ 이고 $S_y^{*2} = \sum_{k=1}^N (y_k - \bar{Y})^2 / N$ 이다.

이때, 단순확률 표본설계의 편차 제곱 합은 층화단순확률 표본설계의 층별 편차 제곱 합과 층화 평균 편차의 제곱 합으로, 다음과 같이 분해할 수 있다.

$$\sum_{k=1}^N (y_k - \bar{Y})^2 = \sum_{h=1}^H \sum_{i=1}^{N_h} (y_{hi} - \bar{Y}_h)^2 + \sum_{h=1}^H N_h (\bar{Y}_h - \bar{Y})^2 \quad (2.4.3)$$

식 (2.4.5)을 분산추정량 (2.4.1)과 (2.4.2)에 적용하면 다음의 관계식을 유도할 수 있다.

$$V(\bar{y}_{srsur}) = V(\bar{y}_{strur}) + \frac{1}{n} \sum_{h=1}^H W_h (\bar{Y}_h - \bar{Y})^2 \quad (2.4.4)$$

단순확률 표본설계의 분산추정량은 층화단순확률 표본설계의 분산추정량과 층별 편차의 제곱 합으로 나타낼 수 있다. 일반적으로 서로 다른 층 ($h \neq h'$)에 대해서 $\bar{Y}_h \neq \bar{Y}_{h'}$ 이므로, 평균추정량에 대해 층화단순확률 표본설계의 분산이 단순확률 표본설계의 분산보다 작은 값, 즉 $V(\bar{y}_{srsur}) \geq V(\bar{y}_{strur})$ 이 됨을 보일 수 있다. 이를 분산분석표에서 다음과 같이 표현할 수 있다.

표 1. 층화단순확률추출 표본설계의 분산분석표

모형	자유도(df)	제곱합(sum of square)
층간 분산	H-1	$SSB = \sum_{h=1}^H N_h (\bar{Y}_h - \bar{Y})^2$
층 내 분산	N-H	$SSW = \sum_{h=1}^H \sum_{i=1}^{N_h} (y_{hi} - \bar{Y}_h)^2 = (N_h - 1) S_{yh}^{*2}$
총 분산	N-1	$SST = \sum_{k=1}^N (y_k - \bar{Y})^2 = (N-1) S_y^{*2}$

이때, 식 (2.4.4)을 활용하여 층 내 분산비 RSW_y , 층간 분산비 RSB_y 를 통해 식 (2.4.5)를 유도할 수 있다. 층별 변동을 제거한 층 효과 비 RSW_y 는 층 효율 측도로 고려할 수 있다. (Lohr, 2010. 86p)

$$1 = \sum_{h=1}^H W_h \left(\frac{S_{yh}^{*2}}{S_y^{*2}} \right) + \sum_{h=1}^H W_h \left(\frac{(\bar{Y}_h - \bar{Y})^2}{S_y^{*2}} \right) \quad (2.4.5)$$

$$= RSW_y + RSB_y$$

2) 설계효과

설계효과는 복합 표본설계(complex sampling)의 추정량 분산이 동일한 크기의 단순 확률 표본설계의 추정량 분산과 비교한 상대적 크기인 비로 다음과 같이 정의된다.

$$Def_f = \frac{V_p(\bar{y}_p)}{V_{srs}(\bar{y})} \quad (2.4.6)$$

따라서 설계효과는 단순확률추출에 의한 표본설계에서 층화 요소를 추가적으로 사용함에 따른 층 효율성으로 해석할 수 있다.

복원추출과 비례배분을 가정하면, 식 (2.4.4)와 (2.4.5)에 의해 설계효과는 다음과 같이 유도할 수 있다. (Lehtonen과 Pakiton, 2004)

$$Def_f(\bar{y}_{stsr}) \doteq \frac{\sum_{h=1}^H W_h S_{yh}^2}{\sum_{h=1}^H W_h [S_{yh}^2 + (\bar{Y}_h - \bar{Y})^2]} \quad (2.4.7)$$

이를 2.4.3과 유사한 방식으로 설계효과식 2.4.3을 다음과 같은 식 2.4.4로 유도할 수 있다.

$$Def_f(\bar{y}_{stsr}) \doteq \frac{SSW}{S_y^2} \quad (2.4.8)$$

따라서, 분산분해와 동일하게 층 효과 비 RSW_y 를 층 효율의 척도로서 사용할 수 있다.

Kalton et al. (2005)은 유사한 논리를 사용하여, 표본설계 단계에서 복원추출을 통한 표본설계 및 층 비례할당을 가정할 때, 층화 표본설계의 분산을 다음과 같이 정의하였다.

$$V(\bar{y}_{st}) = \sum_h \frac{W_h^2 S_h^2}{n}$$

여기서 $S_h^2 = \sum_i (Y_{hi} - \bar{Y}_h)^2 / (N_h - 1)$ 이다.

이때, 표본설계 단계에서 모든 가중치가 동일하다는 가정하에, 단순 확률 표본설계의 평균은 $\bar{y}_{prop} = \sum \sum y_{hi} / n$ 으로 계산할 수 있고, 분산은 다음의 식 (2.4.9)와 같이 유도된다.

$$V(\bar{y}_{prop}) = \sum_h \frac{W_h S_h^2}{n} \quad (2.4.9)$$

따라서, 이들의 분산비인 설계효과는 식 (2.4.10)으로 유도할 수 있다.

$$Def_{st} = \frac{V(\bar{y}_{st})}{V(\bar{y}_{prop})} = S_{yw}^2 / S_y^2 \quad (2.4.10)$$

여기서 $S_{yw}^2 = \sum_{h=1}^H W_h S_{yh}^2$ 이다.

3. 다단추출의 총 효율성

층화단순확률 표본설계는 단순확률표본보다 효율적인 것이 증명되었으나, 층 내에서 치우친 표본이 추출될 경우 모집단에 대한 반영이 제대로 되지 않기 때문에, 비효율적인 표본의 가능성이 존재한다. 이는 조사연구 시 비용상의 문제로 인해, 많은 수행을 시도할 수 없어, 치우친 표본에 대한 해결방안을 고려할 필요성이 있다. 층화다단확률 표본설계는 주로 일차 추출단위(PSU; Primary Sampling Unit)의 층화를 고려하여 층별 치우침을 방지함과 동시에, 추가적인 정보를 활용하여 층화단순확률 표본설계와 비교했을 때 이점을 갖는다.

3.1 층화다단확률추출 개요

가구 조사와 같은 조사연구에서 다단추출이 사용되는 주된 이유는 면접원의 여행경비 및 시간적 비용을 축소하며, 목표 모집단의 대상 목록을 일일이 파악하지 않아도 표본추출에 문제가 없기 때문이다. 일례로, 국민 환경과학원의 국민 환경기초조사에서는 2단 추출에 의한 표본대상을 선정하며, 일차 추출단위로 통계청의 조사구를 사용하고, 추출된 조사구 내 일부 가구를 이차 추출단위로 최종 추출한다 (국립환경과학원, 2021).

층화다단확률 표본설계 시 면접원은 전국을 방문할 필요가 없고, 선택된 조사구만을 방문한다. 추출된 조사구들에 대한 가구명단을 확보하여 추정치를 계산하는 것으로 모집단을 추정할 수 있다. 농촌진흥청의 농어업인 등에 관한 실태조사는 3단 추출로 표본대상을 선정하여, 읍면동을 일차 추출단위로, 조사구와 가구를 각각 이차 추출단위와 최종 추출단위로 결정한다 (농촌진흥청, 2021). 이는 인구밀도가 상대적으로 희박한 읍면지역을 주된 조사대상으로 하는 목적을 주어진 비용하에 달성하기 위해 다단추출을 사용한다.

다단추출 표본설계에 따른 최종개체의 집락화는 비집락 추출로, 최종개체에 대한 직접추출인 단순임의추출과 비교할 때 같은 규모의 표본 크기라면 집락 내 존재하는 개체 간 상관관계로 인해 추정량의 변동성이 증가할 가능성이 있다(Korn & Graubard, 1999, 16).

이로 인해, 집락 내 모든 개체를 조사하기보다는 일부만을 조사하는 것이 더욱 효율적일 수 있다. 일례로, 국내에서 수행되는 가구 조사는 조사구를 사용하여 10개 가구만을 추출하여 변동성을 최소로 하고자 하였다. 집락사용에 따른 비효율성과 일부 개체만을 사용함에 따른 비효율의 감소에 대한 설명은 3.2에서 추가적으로 논의한다.

층화다단확률 표본설계 $p(s)$ 의 평균 추정량은 층 $h(=1, 2, \dots, H)$ 의 집락 $i(=1, 2, \dots, N_h)$ 의 개체 $j(=1, 2, \dots, M_{hi})$ 에 대하여, 관측치 y_{hij} 와 표본 가중치 w_{hij} 가 주어졌을 때, 평균 추정량은 다음과 같이 계산한다.

$$\hat{y}_p = \frac{\sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} \times y_{hij}}{\sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij}}$$

이때, 표본설계의 분산추정량으로 다음의 선형화 근사 추정량을 사용한다.

$$\hat{v}(\hat{y}_p) = \sum_{h=1}^H \frac{n_h(1-f_h)}{n_h-1} \sum_{i=1}^{n_h} (e_{hi} - \bar{e}_{h..})^2$$

여기서, $e_{hi} = \sum_{j=1}^{m_{hi}} w_{hij} (y_{hij} - \hat{Y}) / w_{hi..}$, $\bar{e}_{h..} = \frac{1}{n_h} \sum_{i=1}^{n_h} e_{hi}$ 이다.

3.2 일차추출단위 층화

층화 표본설계에서 표본추출 시 치우친 표본을 가진 층이 있다면, 해당 표본 층은 층 내의 특정 집단만을 대표할 수 있다. 이는 결과 분석 시 모집단의 층에 대한 오해를 유발한다. 이를 방지하고자 일차 추출단위의 층화를 고려한다.

일차 추출단위의 층화는 층 내에서 치우친 표본분포를 방지할 수 있어, 효과적인 표본을 추출할 수 있다. 따라서, 조사연구 시 일차 추출단위의 층

화를 고려하는 것은 층화단순확률 표본설계에 비해 효율적인 표본설계 방법이다.

일차 추출단위의 층화는 관리상, 행정상의 장점 또한 존재한다. 학교조사에서 학생들에 대한 설문 조사를 수행할 때, 해당 지역의 교육감에게 사전 허락을 받아야 할 경우가 존재한다. 이때, 일차 추출단위인 학교를 권역별로 층화를 고려함으로써 행정상의 절차를 줄일 수 있다. 또한, 특정 영역들에 대한 추정이 필요할 때, 일차 추출단위를 표본 층으로 구성하는 것을 고려할 수 있다.

일차 추출단위의 층화는 층별 PSU의 설정에 대한 논의가 필수적이다. 이는 층화 단순 표본설계에서 표본할당 시 층별 비용 및 분산을 고려하는 것을 확장했을 때, PSU에 대한 논의는 층별 비용과 분산을 고려하는 것과 동일하다. 예를 들어, 조사연구 시 예산이 적다면, PSU의 수를 줄이거나 PSU 내 개체 수를 줄여 층별 표본의 크기를 줄이는 것을 고려할 수 있다. 일차 추출단위의 층화에 대한 표본할당은 3.5에서 추가적으로 논의한다.

3.3 설계효과 모형

설계효과는 모집단의 구조와 설계요소들이 추정량의 분산에 주는 개별적 영향을 명백히 나타내는 함수적 형태로 표현될 때 유용성이 극대화될 수 있다. (Rust & Broene, 2010) Park(2015)와 Chen&rust(2017)는 다단추출에 대한 설계효과 모형식을 다음과 같이 제시하여 층별 설계요소들의 영향력을 분류하여 계산한다.

$$deff_{\xi}(\bar{y}_p) = \sum_{h=1}^H \delta_{ysh} \delta_{ych} \delta_{wh} \quad (3.3.1)$$

$$\delta_{ysh} = \frac{\hat{A}_h^2 \sigma_{yh}^2}{a_h \sigma_y^2} \quad (3.3.2)$$

$$\delta_{ych} = 1 + (\bar{m}_h - 1) \rho_{yh} \quad (3.3.3)$$

$$\delta_{wh} = 1 + cv_{wh}^2 \quad (3.3.4)$$

- [식 3.3.2] $\hat{A}_h = \hat{M}_h / \hat{M}$, $a_h = n_h / n$ 는 층 h 의 모집단과 표본의 상대적 크기로, $\hat{M}_h = \sum_i \sum_j w_{hij}$, $\hat{M} = \sum_h \hat{M}_h$ 는 모집단 간 할당 크기 추정량, $n_h = \sum_i \sum_j 1$, $n = \sum_h n_h$ 는 표본 간 할당 크기를 각각 나타낸다. 이를 통해 표본할당 및 층별 영향력을 계산한다.
- [식 3.3.3] ρ_{yh} 는 층 내 동질성 척도이며 $\bar{m}_h$ 는 층별 집락 당 평균 표본 크기, n_h / m_h 를 의미하여, 이를 통해 층별 집락 효과를 계산한다.
- [식 3.3.4] $cv_{wh} = \frac{sd(w_{hij})}{mean(w_{hij})}$ 은 표본 가중치의 상대분산을 의미한다. 이를 통해 층별 가중치 효과를 계산한다.

다단집락추출에 의한 표본설계 시 집락 내 동질성(ρ_{yh})와 평균 집락 크기(m_h^*)가 클수록 집락 효과가 크고, 표본 가중치 cv_{wh}^2 의 변동이 클수록 가중치 효과가 크게 나타남을 알 수 있다(Kish, 1965). Park(2015)는 층화다단추출에 의한 표본설계 시 설계효과 모형을 통해 집락 효과와 가중치 효과에 더불어 층별 표본할당 및 크기, 분산비를 통해 층화 효과를 형성하고, 이에 대해 논의하고자 하였다. 단, 위 모형은 표본 가중치와 관측치 간의 상관관계가 없으면, 즉 $cor(w, y)$ 값이 0에 가까울수록 정당성을 가질 수 있다.

만약, 유한모집단 수정계수를 무시할 수 있을 정도로 모집단이 클 경우, 또는 층별 표본의 수가 매우 작을 경우, 다음 설계효과 식 3.3.5를 만족한다.

$$deff_{\xi}(\bar{y}_p) = \frac{\sum_{h=1}^H \frac{N_h^2 S_{yh}^2}{n_h}}{\frac{N^2 S_y^2}{n}} \quad (3.3.5)$$

층화 단순추출에서 비례배분을 가정한다면 식 (2.4.3) - (2.4.6)의 동일한

과정을 거쳐 식 3.3.6을 유도할 수 있다.

$$def_{\xi}(\bar{y}_p) = S_{yw}^2 / S_y^2 \quad (3.3.6)$$

3.4 표본할당

층화다단확률 표본설계의 설계효과 모형식 (3.4.1)을 이용하여 다음과 같은 분산식을 유도할 수 있다.

$$\begin{aligned} v(\hat{\bar{y}}_p) &\doteq \frac{S_y^2}{n} \sum_{h=1}^H \frac{W_h^2}{a_h} \frac{S_{yh}^2}{S_y^2} L_{wh} [1 + r_{yh} (\bar{m}_h - 1)] \\ &\propto \sum_{h=1}^H N_h^2 \frac{S_{yh}^2}{n_h} L_{wh} [1 + r_{yh} (\bar{m}_h - 1)] \\ &= \sum_{h=1}^H N_h^2 S_{yh}^2 L_{wh} \left[\frac{1}{n_h} (1 - r_{yh}) + \frac{1}{\bar{m}_h} r_{yh} \right] \end{aligned} \quad (3.4.1)$$

1) Global, 전역 할당

3.3절에서 평균 추정량의 분산은 층 효과, 가중치 효과, 집락효과에 대한 함수적 영향을 받는다는 것을 밝혔다. 또한, 층별로 개체 수 n_h 와 집락 수 m_h 가 클수록 각각 집락 동질성 r_{yh} 과 비동질성 $(1 - r_{yh})$ 을 상쇄한다. Park & Hwang(2019)는 식 3.4.1과 비용 식 $n = \sum_{h=1}^H n_h$ 에 대해 Cauchy-Schwartz 부등식을 적용(Cochran, 1977), 다음과 같은 최적 할당식을 유도할 수 있음을 보인다.

$$\begin{aligned} n_h &\propto N_h S_{yh} \sqrt{\delta_{wh} \delta_{ych}} \\ &= N_h S_{yh} \sqrt{L_{wh}} \sqrt{(1 - r_{yh}) + \bar{m}_h r_{yh}} \end{aligned} \quad (3.4.2)$$

이때 2.2절에서 제시한 층화추출에서 표본할당 방식 중 하나인 네이만 할당은 식 3.4.2의 특수형태로, 표본의 크기와 표준편차의 비례식을 통해 층

크기를 계산한다.

이때, 분산추정 식 $v(\bar{y}_p)$ 의 하한값은 식 3.4.2를 사용하여, 식 3.4.3을 유도할 수 있다.

$$\begin{aligned}
 v(\bar{y}_p) &= \frac{S_y^2}{n} \sum_{h=1}^H \frac{N_h^2}{N^2} \frac{n}{n_h} \frac{S_{yh}^2}{S_y^2} \delta_{ych} \delta_{wh} \\
 &= \sum_{h=1}^H \frac{N_h^2}{N^2} \frac{S_{yh}^2}{n_h} \delta_{ych} \delta_{wh} \\
 &\geq \sum_{h=1}^H \frac{N_h^2}{N^2} \frac{S_{yh}^2}{n \frac{N_h S_{yh} \sqrt{\delta_{ych} \delta_{wh}}}{B}} \delta_{ych} \delta_{wh} \\
 &= \frac{B^2}{n N^2} \\
 &:= v_{\min}
 \end{aligned} \tag{3.4.3}$$

여기서 $B = \sum_h N_h S_{yh} \sqrt{\delta_{ych} \delta_{wh}}$

2) Local, 국소 할당

층별 목표 CV를 설정하는 것으로 목표 CV를 달성하기 위한 층별 표본 수¹⁾를 계산할 수 있다. 층별 표본 평균의 상대분산, 즉 상대 표준오차의 제곱은 층화다단확률 설계효과 모형식인 다음과 같이 근사할 수 있다.

$$cv^2(\bar{y}_h) = \frac{1}{\bar{y}_h^2} \frac{S_{yh}^2}{n_h} \delta_{ych} \delta_{wh}$$

층별 목표 상대 표준오차 $cv(\bar{y}_h) = \alpha_h \kappa_h$ 라고 할 때, κ_h 는 표본 크기 조정 전 상대 표준오차의 목표값이며, α_h 는 층별 조정계수이다. 이는 목표 상대

1) $v(\bar{y}_p) = \sum_h W_h^2 v(\bar{y}_h)$ 이므로 ‘ n_h 동일’의 가정하에 식 (3.4.1)으로부터 유도

표준오차의 달성을 위한 표본 크기는 일반적으로 기존 표본 크기보다 작다. 따라서, 비교를 위하여 α_h 를 통해 기존 표본 크기와 동일한 크기를 계산한다. 모든 층 h 에 대하여 같은 목표 상대 표준오차를 가정한다면, $\alpha_h \kappa_h \equiv \alpha \kappa$ 이며 이를 통해 식 3.4.4를 유도할 수 있다. 다음 식을 만족한다.

$$n = \sum_{h=1}^H n_h^{str} = \frac{1}{(\alpha \kappa)^2} \sum_{h=1}^H \frac{S_{yh}^2}{y_h^2} \delta_{ych} \delta_{wh} \quad (3.4.4)$$

3.5 층 효율성 평가

3.5.1 기존 측도

1) 기존 측도 - 분석적 접근

층 효율 평가를 위해 2.4절의 분산분석과 동일하게 비례할당 복원추출 및 동일 확률의 집락추출을 가정한다. 동일한 조건의 단순 확률표본과 분산추정량의 비를 계산하여 층 효율의 측도로써 고려한다.

$$\begin{aligned} \sigma_y^2 &= \frac{1}{N} \sum_{k=1}^K (y_k - \mu_y)^2 = \sum_h A_h \sigma_{yh}^2 + \sum_h A_h (\mu_{yh} - \mu_y)^2 \\ &\left(\because \sigma_{yh}^2 = \sum_{i=1}^{n_h} \sum_{k=1}^{m_{hi}} w_{hik} (y_{hik} - \bar{y}_h)^2 / w_{\dots} \right) \\ V(\bar{y}_{str}) &= V(\bar{y}_p) + \sum_h A_h (\mu_{yh} - \mu_y)^2 \end{aligned}$$

이때 $A_h = N_h/N$ 는 층별 크기 비율이다. 식 2.4.3의 일치 통계량으로써 층내 상대분산비를 고려하고, 이를 층 효율 측도로써 고려한다.

$$\delta_{ystrI} = \widehat{RSW}_y \quad (3.5.1)$$

그러나, 해당 측도는 층 효과만이 아닌 가중치 및 집락 효과를 동시에 계산한 것으로, 모형의 설계효과로써 의미는 있으나 층 효율의 측도로 사용하는 것은 한계가 있다.

2) 기존 측도 - 경험적 접근

Krenzke et al.(2017)의 Westat 보고서에 따르면, 층화 요소의 유무에 따른 층화 효과를 수치화하기 위해, SAS등 조사자료 분석 시 옵션 선택에 따른 분산추정량의 비를 통해 층 효율을 정의하였다. SAS의 surveymeans 명령문은 층화다단확률 표본설계의 분석 시 사용하는 명령문으로, 옵션 중 strata, cluster를 통해 층, 집락을 정의하고 실행한다. 이때, strata 옵션을 추가 및 삭제하는 것으로 표본 설계모형의 분산추정량을 획득할 수 있으며, 이들의 분산추정량 비를 통해 층 효율 측도를 계산한다.

그러나, 층 규모가 작은 표본은 표본 크기가 작아 분산추정량의 불안정성이 높다. 따라서, 표본설계의 전체 분산추정량을 계산하여 측도로써 활용하였다.

$$\delta_{ystrII} = \frac{v(\bar{y}_p : strata\ option)}{v(\bar{y}_p : no\ strata\ option)} \quad (3.5.2)$$

2.4절에서 특정 조건(비례할당 및 복원추출)을 만족할 때 층화 요소를 사용하는 것은 일반적으로 분산을 줄이는 효과가 있다는 것을 증명하였다. 그러나, 해당 측도는 프로그래밍을 바탕으로 제시된 측도이기 때문에, 분석적으로 사용하는 것은 어려울 수 있다.

3.5.2 신규 측도

설계효과 모형식(3.3.1)은 이단추출설계에 대한 층화에 따른 평균 추정량 분산에 대하여 층별 영향력을 다음과 같은 식 3.5.3으로 표현하고 있다. 이를 활용한 측도는 매우 유용할 것으로 판단²⁾한다. (Rust & Broene, 2010)

$$\delta_{ysh} = \delta(A_h, a_h, S_{yh}) = \frac{W_h^2 S_{yh}^2}{a_h S_y^2} \quad (3.5.3)$$

식 3.5.3의 층화 효과 요소 중 표본할당비 a_h 에 중점을 두어, 현재의 표본할당 방법과 제안된 표본할당 방법 간 비교를 활용하여 층 효율성 측도를 제시한다. 전체 분산추정량 비를 사용하여 모형의 층 효율성을 판단하고, 층별 분산추정량의 상대 편차 평균을 사용하여 평균적인 층별 효율성을 판단하고자 한다. 다음의 식 3.5.4 - 식 3.5.6은 층화 표본할당 전략 str 에 대해 본 연구에서 제안하는 층 효율성 측도이다.

$$\delta_{ystrIII} = d_y(str, v) = \frac{v^{str}(\bar{y})}{v(\bar{y})} \quad (3.5.4)$$

$$\delta_{ystrIV} = d_y(str, n_h) = \frac{1}{H} \sum_{h=1}^H \frac{|n_h - n_h^{str}|}{n_h^{str}} \quad (3.5.5)$$

$$\delta_{ystrV} = d_y(str, se) = \frac{1}{H} \sum_{h=1}^H \frac{|se(\bar{y}_h) - se^{str}(\bar{y}_h)|}{se^{str}(\bar{y}_h)} \quad (3.5.6)$$

식 3.5.4는 층 분산을 비교한 것으로, 층 할당을 이용한 분산 대비 현 표본설계의 분산 증감비를 통해 모형의 층 효율성을 계산할 수 있을 것으로 기대되며, 식 3.5.5와 식 3.5.6은 층별 표본 크기와 층별 표준오차의 상대 편차 평균을 구한 것으로, 현 표본설계의 층별 효율성을 계산할 수 있을 것으로 기대된다.

또한, 식 3.5.3의 층화 요소의 변동량을 구하는 것으로 층별 층화 효과의 적극성을 구할 수 있다. 이는 층화 효과가 총 설계효과에 얼마나 적극적으로 개입하는지 확인을 위한 측도이다. 이를 층 효율성 측도로써 활용할 수 있을 것이다.

$$\delta_{ystrVI} = \frac{sd(\delta_{ysh}'s)}{ave(\delta_{ysh}'s)} \quad (3.5.7)$$

2) 단, 층 크기 N_h 와 표본 수 n_h 의 영향력이 작다고 가정 (fpc 영향무)

4. 사례연구

4.1. 국민 환경 보건 기초조사 개요

국민 환경 보건 기초조사(이하, 기초조사)는 환경 유해물질 노출 수준에 대해 원인 및 경로의 현황 파악을 목적으로 국가통계자료를 산출하여, 환경 보건정책 수립을 위한 기초자료 제공을 목적으로 조사한다. (국립환경과학원, 2021) 자료 수집 시 설문 조사 및 인체 유래물 시료수집을 통한 임상 검사를 통해 환경성 질환 발생과 연관성이 있는 납, 수은 등 중금속 및 유기 화합물 33종에 대한 체내 잔류량 및 노출 수준을 분석한다.

표본설계는 조사대상에 따라 표본추출 방법을 다르게 적용한다. 성인을 대상으로 한 기초조사는 인구주택총조사 시 구성한 조사구에 기초한 가구 조사를 통해 표본설계를 진행한다. 반면 영·유아, 어린이·청소년을 대상으로 한 조사는 대상조사 수행의 어려움으로 인해 보육·교육기관 현행자료를 통해 성인과 다른 표본설계를 진행한다.

본 연구에서는 기초조사의 고등학생을 대상으로 층화 효과를 분석하고, 이에 따른 기존 층화 전략의 층 효율 계산 및 층화 전략 개선 방법에 대해 논의하였다.

표 2. 국민 환경 보건 기초조사 고등학생 수 현황

항목		고등학생		고등학교	
		일반계	특성화계	일반계	특성화계
서울		185,864	38,452	250	70
경기, 인천, 강원권	시부	387,823	60,928	528	112
	군부	14,318	1,822	68	13
대전, 충청권	시부	110,545	22,771	171	46
	군부	14,788	3,970	46	20
광주, 전라권	시부	117,756	23,633	206	49
	군부	18,654	6,292	81	39
대구, 부산, 울산, 영남권	시부	245,846	49,190	429	109
	군부	29,424	5,236	99	31
계		1,125,018	212,294	1,878	489

고등학생을 대상으로 한 기초조사는 낮은 귀가 및 적격 가구의 비율이 낮아 성인과 같이 조사구를 일차 추출단위로써 사용하기에 어려움이 있다. 따라서, 고등학교를 일차 추출단위로 사용하는 표본조사 방식을 적용하였다. 표본설계를 위한 층화 변수는 지역을 사용하여, 권역 및 시 군 구분을 일차적인 층으로 선정했으며, 학교 구분(일반고, 특성화고)을 세부 층화 변수로 설정하였다. 전체 예산과 분석 비용을 고려하여, 최종적으로 고등학생 표본의 총 크기는 440으로 결정하였다. 이를 학급수에 비례하는 확률비례 계통추출법에 따라 표본을 추출했다.

표본 배분은 층화 변수에 따른 제곱근비례 방식을 선택하여 권역 간의 차이가 큰 국내의 사회적 상황을 고려해 지나치게 작거나 큰 표본의 생성을 방지하여 안정성을 높이고자 하였다.

표 3. 국민 환경 보건 기초조사 표본 고등학생 수 현황(비례배분)

항목		고등학생		고등학교	
		일반계	특성화계	일반계	특성화계
서울		61	13	5	1
경기, 인천, 강원권	시부	127	21	6	2
	군부	5	1	1	0
대전, 충청권	시부	36	8	4	1
	군부	5	1	1	0
광주, 전라권	시부	38	8	4	1
	군부	6	2	1	1
대구, 부산, 울산, 영남권	시부	80	17	6	1
	군부	10	2	2	0
계		367	73	30	7

표 4. 국민 환경 보건 기초조사 표본 고등학생 수 현황(제곱근 비례배분)

항목		고등학생		고등학교	
		일반계	특성화계	일반계	특성화계
서울		57	12	5	1
경기, 인천, 강원권	시부	82	15	6	2
	군부	16	3	1	0
대전, 충청권	시부	44	9	4	1
	군부	16	4	1	0
광주, 전라권	시부	45	9	4	1
	군부	18	5	1	1
대구, 부산, 울산, 영남권	시부	66	13	6	1
	군부	23	4	2	0
계		367	73	30	7

4.2. 기하평균 추정량

기초조사에서 수집하는 자료의 특성상 노출 시 지속적인 축적으로 인해 체내에 다량으로 발견될 수 있으며, 노출되지 않으면 체내에 극소량으로 존재하기 때문에, 극단적인 분포를 갖는다. 따라서, 기하평균을 통해 분포의 안정성을 높이려고 하였다. 기초조사의 모평균에 대해 기하평균을 사용한 추정량은 다음과 같은 식을 따른다.

$$\bar{z}_p = \frac{\sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} z_{hij}}{\sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij}} = \frac{\sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} z_{hij}}{w_{\dots}} \quad (4.2.1)$$

여기서, w_{hij} 는 각 응답자에 부여된 가중치이고, $z_{hij} = \ln(y_{hij})$ 이다.

H 는 층의 수, n_h 는 층 h 에서의 1차 표본추출단위인 표본 조사구의 수, m_{hi} 는 층 h 내 i 번째 표본 조사구의 응답자 수, $w_{\dots} = \sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij}$ 은 전체 응답자에 대한 가중치의 합계이다. 이를 통해 특성별 모평균 추정량 $\hat{\theta}_p = \exp(\bar{z}_p)$ 를 계산하고, 분산추정량으로 식 3.3.2의 선형화 추정량을 사용하였다.

$$var(\hat{\theta}_p) = \hat{\theta}_p^2 var(\bar{z}_p) \quad (4.2.2)$$

$$var(\bar{z}_p) = \sum_{h=1}^L \frac{n_h(1-f_h)}{n_h-1} \sum_{i=1}^{n_h} (e_{hi.} - \bar{e}_{h..})^2 \quad (4.2.3)$$

표 5. 국민 환경 보건 기초조사 중·고등학생 대상조사 결과 (혈중 납)

구분		응답자	혈중 납(기하평균)		
			추정값	표준오차	상대 표준오차 (%)
전체		455	0.831	0.0201	2.42
성별	남자	211	0.928	0.0232	2.50
	여자	244	0.734	0.0261	3.56
지역1	서울	74	0.877	0.0446	5.09
	경기, 인천, 강원권	110	0.817	0.0420	5.14
	대전, 충청권	75	0.862	0.0441	5.12
	광주, 전라, 제주권	82	0.875	0.0267	3.05
	대구, 부산, 울산, 영남권	114	0.787	0.0384	4.88
지역2	시부	306	0.795	0.0250	3.14
	군부	75	1.022	0.0098	0.96

4.3. 결과 요약

1) 층화 효과

본 연구는 일차 추출단위로 학교를 지정하고, 층화 변수로 권역을 지정하여 표본설계의 분산추정량을 계산한다. 이때, 분산추정량 계산 시 층별로 단일 집락이 존재한다면 분산이 0으로 수렴하는 매우 작은 값을 가진다. 본 연구에서 실행하고자 하는 측도는 층별 분산을 사용하기 때문에, 이는 분석에 있어 매우 불안정하다. 따라서, 표 4에서 확인할 수 있는 단일 집락을 갖는 경기·인천·강원권 군부 층과 대전·충청권 군부 층을 제외하고 분석을 진행한다.

설계효과는 현 표본설계와 동일한 크기의 단순확률표본과 분산비를 통해

계산한다. 층화다단확률 표본설계의 설계효과 모형은 층화, 집락, 가중치 효과로 구성되어 있다. 본 연구에서는 설계효과 모형 중 층화 효과를 활용하여 층 효율성에 대한 측도를 제시하고자 한다. 층화 전략 중, 층 할당 a_h 를 전략별 층별 분산을 계산하고, 이를 통해 층 효율성을 비교한다. 먼저, 기초조사의 표본설계에 있어 층화의 적극성은 δ_{VI} 를 통해 계산한다.

항목	$\delta_{ystr VI}$
BPb	1.2070
BHg	0.9193
Uhg	0.8954
Ucd	1.0644
OHP	1.0887
NAP	0.9460

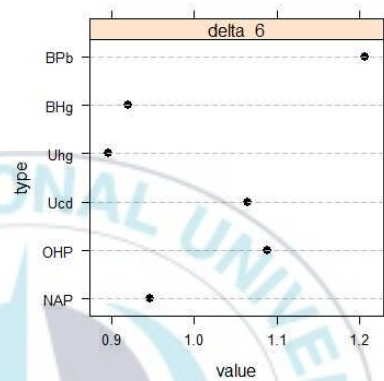


그림 1. 층화 효과의 적극성 측도

그림 1에서 BPb(혈중 납)이 표본설계의 층화 효과가 가장 활발하게 영향을 미치고 있는 것을 확인할 수 있고, Uhg(요중 수은)이 층화 효과의 영향력이 가장 적음을 확인할 수 있다. 이는 층화 전략의 작용이 크면 클수록 층 효율성이 크다고 해석할 수 있다.

2) 층 효율성 측도

층화 전략 str 은 3.4의 Global, Local 층화 전략과 기초조사에서 사용한 층화 전략인 제곱근 비례할당을 사용한다. 층화 전략별 층 효율 측도를 계산하여, 어떤 층화 전략이 효율적인지, 기존 층화 전략을 어떻게 개선해야 할지 방향성을 제안하고자 한다. 먼저, 기존 층 효율성 측도를 계산하여 항목별 층 효율을 계산한다.

표 2에서 $\delta_{ystr I}$ (이하, 측도 1)과 $\delta_{ystr II}$ (이하, 측도 2)는 기존 층 효율성 측

도이다. 측도 1은 단순 표본설계에서 사용한 분산분해를 층화다단확률 표본설계로 확대하여 RSW_y 를 층 효율 측도로 사용한다. 측도 2는 Krenzke(2017)의 층 효율성 측도로 제시한 strata 옵션 유무의 분산비를 사용한다.

측도 1은 모든 항목에서 층화가 효율적임을 나타내고 있다. 그러나, 3.5절에서 언급하듯, 해당 결과는 층화단순확률 표본에서의 측도를 층화다단확률 표본으로 확대하였기 때문에, 층 효율성을 판단하기 어렵다.

측도 2는 BPb(혈중 납) 항목만 층화가 효율적이고 다른 항목은 효율적이지 않음을 나타낸다. 이는 층화 옵션의 유무를 비율로써 계산하였기 때문에 신빙성은 있으나, 직관적인 방법이기 때문에 분석에 있어 적합성에 대한 근거가 빈약하다.

따라서, 층화 전략별 측도를 제시하여 이를 통해 기초조사의 표본설계에 대한 층 효율을 판단하고자 한다. 표 6의 $\delta_{ystrIII}$, δ_{ystrIV} , δ_{ystrV} (이하, 측도 3, 측도 4, 측도 5)는 층화 전략별 신규 측도의 결과다.

측도 3은 제시한 층화 전략으로 계산된 표본설계의 분산추정량을 분산비를 통해 층 효율을 계산한다. 측도 4는 제시한 층화 전략의 층별 표본 n_h 의 상대 편차의 평균을 계산한다. 측도 5는 제시한 층화 전략을 통해 계산된 층별 SE의 상대 편차의 평균을 계산한다.

표 6. 국민 환경 보건 기초조사 총 효율 측도

항목	기존 측도		층화 전략 - Sqrt		
	$\delta_{ystr I}$	$\delta_{ystr II}$	$\delta_{ystr III}^{sqrt}$	$\delta_{ystr IV}^{sqrt}$	$\delta_{ystr V}^{sqrt}$
BPb	0.9596	0.9242	0.9576	0.1267	0.0751
BHg	0.9716	1.0244	1.0129	0.1267	0.0751
Uhg	0.9799	1.0436	1.0503	0.1267	0.0751
Ucd	0.9779	1.0565	0.9971	0.1267	0.0751
OHP	0.9813	1.0271	1.0413	0.1267	0.0751
NAP	0.9731	1.0033	1.0838	0.1267	0.0751

항목	층화 전략 - Global			층화 전략 - Local		
	$\delta_{ystr III}^{glo}$	$\delta_{ystr IV}^{glo}$	$\delta_{ystr V}^{glo}$	$\delta_{ystr III}^{loc}$	$\delta_{ystr IV}^{loc}$	$\delta_{ystr V}^{loc}$
BPb	0.8200	0.8588	0.2054	1.0260	0.9914	0.2567
BHg	0.8756	2.0732	0.2725	1.7385	2.8106	0.5969
Uhg	0.8658	2.0846	0.2722	1.2812	2.1846	0.3933
Ucd	0.8776	0.8833	0.2373	2.0253	1.3473	0.4872
OHP	0.8948	0.7717	0.1866	2.1752	2.8035	0.5733
NAP	0.9805	0.4910	0.1887	1.5158	0.6407	0.3318

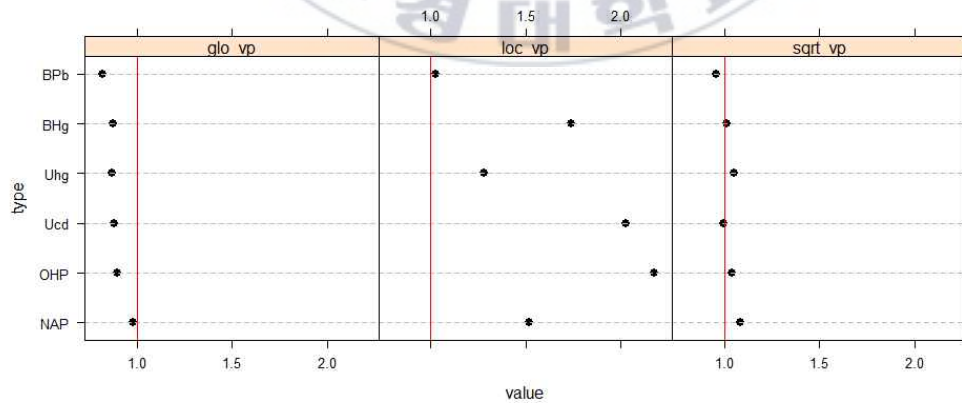


그림 2. 층화 전략별 신규 총 효율성 측도 δ_{III}

측도 3의 경우, 층화 전략 Sqrt는 기초조사의 확률비례 계통추출을 실시하지 않고, 층 규모의 제곱근비례만 실시하기 때문에 기초조사와 약간의 차이를 보인다. 층화 전략 Global의 경우 기초조사보다 약 10% 효율적임을 확인할 수 있으며, 이는 전략의 배경 상 분산의 최솟값을 갖는 할당 방법이기 때문임을 확인할 수 있다. 층화 전략 Local은 기초조사보다 비효율적임을 보이는데, 이는 전략상 층별로 동일한 영향력을 갖게끔 하기 위해 총 분산 면에서는 불이익을 갖는다.

측도 4와 측도 5의 경우, 층화 전략 Sqrt는 다른 전략에 비해 작은 값을 갖는다. 이는 전략과 기초조사의 전략이 거의 동일하기 때문에 표본 및 표준오차에서 큰 차이가 없기 때문이다. 층화 전략 Global은 층화 전략 Local에 비해 작은 값을 갖는다. 이는 기초조사의 전략이 층별 영향력 면에서 동일하지 않기 때문에 Local의 측도가 Global에 비해 크다는 것을 확인할 수 있다.

측도 3을 통해 층 효율성을 판단하고, 측도 4, 측도 5의 층별 편차를 통해 개선 사항을 파악하고자 한다. 다음은 측도 5의 전략별 층별 SE 상대 편차이다.

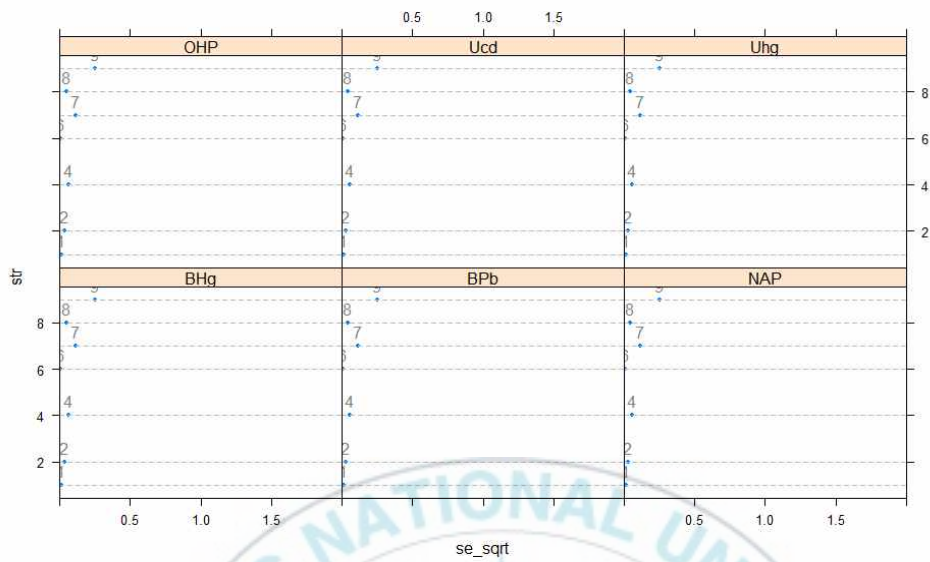


그림 3. 총화 전략 Sqrt의 층별 SE 상대 편차 $\frac{|se(\bar{y}_h) - se^{sqrt}(\bar{y}_h)|}{se^{sqrt}(\bar{y}_h)}$

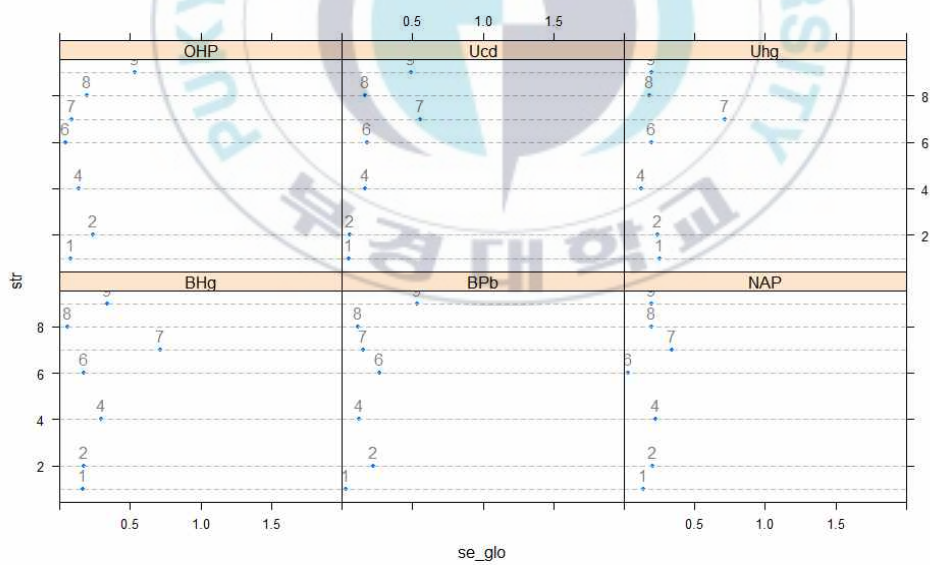


그림 4. 총화 전략 Global의 층별 SE 상대 편차 $\frac{|se(\bar{y}_h) - se^{glo}(\bar{y}_h)|}{se^{glo}(\bar{y}_h)}$

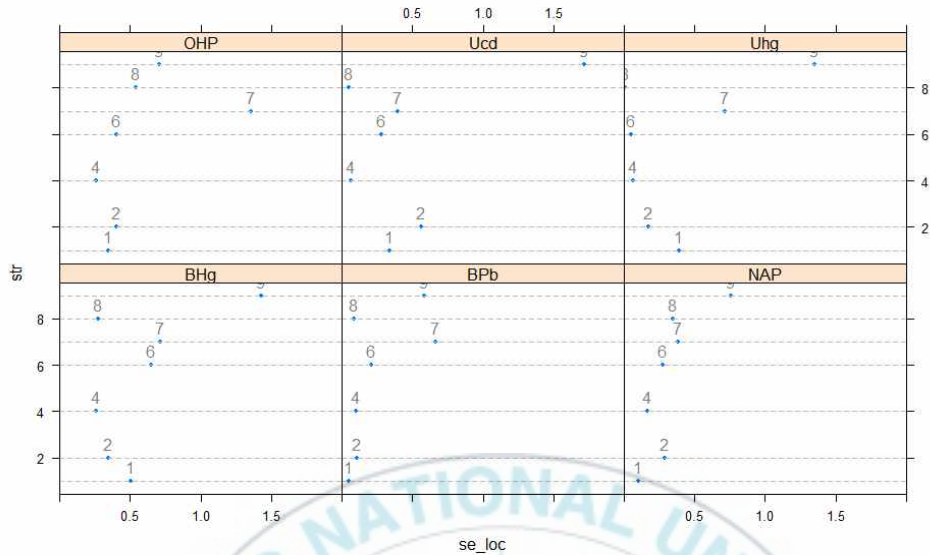


그림 5. 층화 전략 Local의 층별 SE 상대 편차 $\frac{|se(\bar{y}_h) - se^{loc}(\bar{y}_h)|}{se^{loc}(\bar{y}_h)}$

그림 3, 4, 5는 전략별 기초조사 대비 층별 상대 편차를 계산하여 그림으로 나타냈다. 그림 3의 층화 전략 Sqrt는 앞서서도 언급했듯, 층화 전략이 기초조사와 유사하기 때문에 층별 상대 편차에서 0에 가까운 값을 갖는다.

그림 4의 층화 전략 Global은 항목별로 7번 층과 9번 층의 편차가 큰 것으로 확인된다. 이때, 측도 3에서 층화전략 Global이 효율적인 전략임이 밝혀졌기 때문에, 기초조사의 7번 층과 9번 층의 층화전략을 수정하는 것으로 더 높은 층화전략을 가질 수 있을 것이다.

그림 5의 층화전략 Local은 측도 3에서 효율적인 전략이 아님이 밝혀졌다. 만약 층별로 서로 동일한 영향력을 갖게끔 층화 전략을 수정한다면 편차를 줄이는 것으로 효율적인 층화전략이 될 수 있을 것이다. 이때, 7번 층과 9번 층을 중점적으로 층화 전략을 수정하면 층 효율을 높일 수 있을 것이다.

5. 결론 및 향후 과제

층화 단순확률 표본추출의 층 효율성 측도로 설계효과를 사용할 수 있지만, 층화다단확률추출은 식 (3.4.1)의 설계효과모형을 통한 설계요소 분해 결과, 다른 설계요소의 효과로 인해, 설계효과를 층 효율 측도로써 사용할 수 없다는 것을 확인할 수 있었다.

기존 층 효율성 측도는 층화다단확률 표본설계에서 사용할 때, 다른 설계요소의 영향으로 적절치 않다는 것은 3.5.1절에서 확인할 수 있었다. 따라서, 신규 층 효율 측도로 설계효과모형을 고려하여 층화 전략 중 표본 배분 방법에 따른 분산추정량을 계산하여 상대 편차 평균을 활용해 측도를 제시했다.

이를 3.5.2절에서 소개한 몇 가지 층화 전략을 사용해 제 4기 국민 환경기초보건조사의 고등학교 자료에 적용하여 층 효율을 평가했다. 또한, 층별 상대 편차를 이용하여 표본 배분을 통한 개선 방법에 대해 추정할 수 있었다.

기초조사는 자료의 특성상 기하평균을 사용하여 층 효율에 대해 평가하였으나, 기하평균의 분산추정량이 선형화 근사 추정량이므로 해석 시 주의가 필요하다. 또한, 단일 집락 층의 존재로 인해 층별 분산이 불안정하다. 따라서, 해당 층을 제외하고 분석을 진행하였으므로, 이에 따른 보완이 필요할 것이다. 본 논문에서는 고등학생만을 대상으로 분석하여, 이후 성인과 유아에 대한 추가적인 분석이 필요하다.

향후 연구로 모의실험을 고려할 수 있다. 모의실험은 측도의 적합성 및 안정성을 평가할 수 있을 것이다. Feng(2006)이 고려한 혼합 랜덤 효과모형을 참고하여 유한모집단을 생성한 후, 층화다단 표본설계를 통한 층 효율성 측도값을 계산하여 안정성을 평가할 수 있을 것이다.

6. 참고문헌

1. 국립환경과학원(2019) 제 3기('15 ~ '17) 국민환경보건 기초조사
2. 농촌진흥청(2021) 2021 농어업인 등에 대한 복지 실태조사 보고서
3. Sharon L. Lohr (2019) *Sampling Design And Analysis*, CRC press
Taylor & Francis Group
4. Risto Lehtonen, Erkki Pahkinen (2004) *Practical Methods for Design and Analysis of Complex Surveys*. John Wiley
5. Kalton et al.(200) *Estimating components of design effects for use in sample design* Household Sample Surveys in Developing and Transition Countries. UN stats
6. Edward L. Korn, Barry I. Graubard (1999) *Analysis of Helth Surveys*. John Wiley
7. Richard Valliant, Jill A. Dever, Frauke Kreuter (2013) *Practical Tools for Designing and Weighting Survey Samples*. Springer
8. 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, 2174-2181.
9. 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.
10. Park (2015). *Understanding Complex Design Features via Design Effect Models*. The Korean Journal of Applied Statistics, 28(6), 1217 - 1225
11. Kish, L. (1965). *Survey Sampling*, John Wiley & Sons, New York.
12. Park, Hwang (2019) *Sample size determination using design effect formula for repeated surveys*. The Korean Journal of Applied Statistics, 32, 643-652
13. Cochran, W. G. (1977). *Sampling techniques*, 3rd ed., John Wiley & Sons, New York.

14. Tom Krenzke, Jennifer Kali (2017) *Review of the FoodAPS 2012 Sample Design*. Westat An Employee-Owned Research Corporation® 1600 Research Boulevard



A.1. 프로그램

```
# ===== 필수 패키지 ===== #
# install.packages('survey')
library(survey)

# install.packages('dplyr')
library(dplyr)

# install.packages('sampling')
library(sampling)

# install.packages('sas7bdat')
library(sas7bdat)

# install.packages('lattice')
library(lattice)

# ===== data ===== #
# konehs4 파일의 집락 = j_id, 층화 = st_var, 가중치 = wgt

data_mh = read.sas7bdat('C:/Users/User/Desktop/연구과제/제4기 원시자료
/konehs4_mh.sas7bdat')

# ===== data 정리 ===== #
data_high1 = data_mh %>%
  filter(grade > 3) %>%
  filter(!is.na(BPb)) %>%
  filter(!is.na(BHg)) %>%
  filter(!is.na(Ucd)) %>%
  filter(!is.na(Uhg)) %>%
  filter(!is.na(OHP)) %>%
  filter(!is.na(NAP))
```

```

#===== function =====#
int_fun = function(var, int, data, wt, str, psu, domain){

# data : 자료
# var : 분석 변수 ( BPb, BHg, ... )
# int : 관심 변수 ( gender, st_var, grade, ... )
# wt : 설정된 가중치
# psu : 일차 추출단위 PSU
# domain : 층화 효과 추가 분석 변수
# ( 0 = 추가 분석 실시 X, 1 = 추가 분석 실시 )

# 단일 집락 층 확인 후 해당 층 데이터 제거
cls_ind = apply(ifelse(table(psu, str)>0, 1, 0), 2, sum)

data$var1 = var
data$int1 = int
data$psu1 = psu
data$str1 = str
data$wt1 = wt

data1 = data %>%
  arrange(str1) %>%
  filter(!str1 %in% which(cls_ind == 1))

var = data1$var1
int = data1$int1
psu = data1$psu1
str = data1$str1
wt = data1$wt1

if(domain == 1){
  int = int
} else if(domain == 0){
  int = str
}

```

```

} else {
print('잘못된 domain 지정입니다.')
}

# 모형 설계
mh.dsgn = svydesign(ids = ~psu,
strata = ~str,
nest = TRUE,
data = data1,
weights = ~wt)

options(survey.lonely.psu = "adjust")

A = svyby(log(var), by = ~factor(int), design = mh.dsgn, FUN = svymean,
na.rm = TRUE)
B = svyby(var, by = ~factor(int), design = mh.dsgn, FUN = svymean, na.rm =
TRUE)
C = table(int[!is.na(var)])

options(survey.lonely.psu = "certainty")

D = svymean(~log(var), design = mh.dsgn, na.rm = TRUE)
E = svymean(~var, design = mh.dsgn, na.rm = TRUE)

nh = data.frame(C)[,2]
nh_total = sum(nh)

# survey 패키지를 사용한 모수
mu_LM = data.frame(A)[,2]
se_LM = data.frame(A)[,3]
sd_LM = se_LM * sqrt(nh - 1)

mu_LM_total = data.frame(D)[,1]
se_LM_total = data.frame(D)[,2]

```

```

sd_LM_total = se_LM_total * sqrt(nh_total - 1)

mu_AM = data.frame(B)[,2]
se_AM = data.frame(B)[,3]
sd_AM = se_AM * sqrt(nh - 1)

mu_AM_total = data.frame(E)[,1]
se_AM_total = data.frame(E)[,2]
sd_AM_total = se_AM_total * sqrt(nh_total - 1)

mu_GM = exp(mu_LM)
se_GM = mu_GM * se_LM
sd_GM = se_GM * sqrt(nh - 1)

mu_GM_total = exp(mu_LM_total)
se_GM_total = mu_GM_total * se_LM_total
sd_GM_total = se_GM_total * sqrt(nh_total - 1)

# 계산을 통해 층화단순확률 계산
z_hij = log(var)
mean_z = sum(wt * z_hij) / sum(wt)

mu_srs_LM = tapply(z_hij*wt, int, sum) / tapply(wt, int, sum); mu_srs_LM.rep
= rep(mu_srs_LM, nh)
sd_srs_LM = sqrt((tapply((wt*(z_hij - mu_srs_LM.rep)^2), int, sum) /
tapply(wt, int, sum)))
se_srs_LM = sd_srs_LM / sqrt(nh - 1)

mu_srs_LM_total = mean_z
sd_srs_LM_total = sqrt((sum(wt*(z_hij - mu_srs_LM_total)^2) / sum(wt)))
se_srs_LM_total = sd_srs_LM_total / sqrt(nh_total - 1)

mu_srs_AM = tapply(var*wt, int, sum) / tapply(wt, int, sum)
mu_srs_AM.rep = rep(mu_srs_AM, nh)

```

```

sd_srs_AM = sqrt((tapply((wt*(var - mu_srs_AM.rep)^2), int, sum) / tapply(wt,
int, sum)))
se_srs_AM = sd_srs_AM / sqrt(nh - 1)

mu_srs_AM_total = sum(wt * var) / sum(wt)
sd_srs_AM_total = sqrt((sum(wt*(var - mu_srs_AM_total)^2) / sum(wt)))
se_srs_AM_total = sd_srs_AM_total / sqrt(nh_total - 1)

mu_srs_GM = exp(mu_srs_LM)
se_srs_GM = mu_srs_GM * se_srs_LM
sd_srs_GM = se_srs_GM * sqrt(nh - 1)

mu_srs_GM_total = exp(mu_srs_LM_total)
se_srs_GM_total = mu_srs_GM_total * se_srs_LM_total
sd_srs_GM_total = se_srs_GM_total * sqrt(nh_total - 1)

# 층화 효과 설명 & 관심 변수에 대한 추가 설명
rse_GM = 100 * se_GM / mu_GM
rse_AM = 100 * se_AM / mu_AM

rse_GM_total = 100 * se_GM_total / mu_GM_total
rse_AM_total = 100 * se_AM_total / mu_AM_total

deff_GM = (sd_GM / sd_srs_GM)^2
deff_AM = (sd_AM / sd_srs_AM)^2
deff_LM = (sd_LM / sd_srs_LM)^2

deff_GM_total = (sd_GM_total / sd_srs_GM_total)^2
deff_AM_total = (sd_AM_total / sd_srs_AM_total)^2
deff_LM_total = (sd_LM_total / sd_srs_LM_total)^2

# 분산 분해를 통한 층화 효과 계산
Mh = tapply(wt, int, sum)
M = sum(Mh)

```


Ah = Mh / M

s2_z = sd_srs_LM_total^2

RSW_z = sum(Ah * sd_srs_LM^2, na.rm = TRUE)

RSB_z = sum(Ah * (mu_LM - mu_LM_total)^2, na.rm = TRUE)

RSR_z = RSW_z + RSB_z

P_RSB = 100 * RSB_z / s2_z

RSW = sum(Ah * sd_srs_LM^2, na.rm = TRUE) / sd_srs_LM_total^2

RSB = sum(Ah * (mu_LM - mu_LM_total)^2, na.rm = TRUE) /
sd_srs_LM_total^2

data_totalA = round(data.frame(nh, mu_GM, se_GM, rse_GM, deff_GM, mu_AM,
se_AM), 4)

data_totalB = round(data.frame(nh = nh_total, mu_GM = mu_GM_total, se_GM
= se_GM_total, rse_GM = rse_GM_total, deff_GM = deff_GM_total, mu_AM =
mu_AM_total, se_AM = se_AM_total), 4)

rownames(data_totalB) = "total"

기하평균 추정 결과

data_total1 = rbind(data_totalB, data_totalA)

층 효율 평가 결과

data_total2 = round(data.frame(nh_total, mu_GM = mu_GM_total, se_GM =
se_GM_total, se_LM = se_LM_total, SSR = RSR_z, RSW_z, RSB_z, P_RSB), 4)

가중치 효과

wt_mu = sum(wt) / nh_total

```

wt_sd = sqrt(sum((wt - wt_mu)^2) / (nh_total - 1))
wt_cv = wt_sd / wt_mu

wh_mu = tapply(wt, int, sum) / nh
wh_mu.rep = rep(wh_mu, nh)
wh_sd = sqrt(tapply((wt - wh_mu.rep)^2, int, sum) / (nh - 1))
wh_cv = wh_sd / wh_mu

deff_w = 1 + wt_cv^2
deff_wh = 1 + wh_cv^2

# 집락 효과
bar_m = nh_total / length(unique(psu))

rho_z = (1 / (bar_m - 1)) * (deff_GM_total/deff_w - 1)
deff_c = 1 + (bar_m - 1) * rho_z

# 설계효과 평가
data_total3 = round(data.frame(nh_total, mu_GM = mu_GM_total, se_GM =
se_GM_total, deff_GM = deff_GM_total, deff_w, deff_c, rho_z), 4)

# 설계효과 모형의 추가 분석
if(domain == 0){

# no strata option in svymeans
mh.dsgn_nostr = svydesign(ids = ~psu,
nest = TRUE,
data = data1,
weights = ~wt)

options(survey.lonely.psu = "adjust")
W = svyby(var, by = ~factor(psu), design = mh.dsgn_nostr, FUN = svymean,
na.rm = TRUE)

```

```
options(survey.lonely.psu = "certainty")
X = svymean(~log(var), design = mh.dsgn_nostr, na.rm = TRUE)
```

```
cls_name = data.frame(W)[,1]
cls_mu = data.frame(W)[,2]
cls_se = data.frame(W)[,3]
```

```
mu_LM_nostr = data.frame(X)[,1]
se_LM_nostr = data.frame(X)[,2]
sd_LM_nostr = se_LM_nostr * sqrt(nh_total - 1)
```

```
mu_GM_nostr = exp(mu_LM_nostr)
se_GM_nostr = mu_GM_nostr * se_LM_nostr
sd_GM_nostr = se_GM_nostr * sqrt(nh_total - 1)
```

```
# no_str 통계량
data_total4 = data.frame(조사구ID = cls_name, cls_mu, cls_se)
```

```
# 다단 층화 표본설계의 설계효과 분해
int_cluster = apply(as.matrix(ifelse(table(psu, int)>0, 1, 0)), 2, sum)
ah_str = nh / nh_total
Ah_str = Mh / M
bar_mh = nh / int_cluster
rho_yh = (1 / (bar_mh - 1)) * (deff_LM/deff_wh - 1)
```

```
delta_sh = (Ah_str^2 / ah_str) * (sd_srs_LM / sd_srs_LM_total)^2
delta_wh = 1 + wh_cv^2
delta_ch = 1 + (bar_mh - 1) * rho_yh
design_eff = delta_sh * delta_wh * delta_ch
```

```
cls_eff = delta_ch * delta_wh / delta_sh
cv_eff = sd(cls_eff) / mean(cls_eff)
delta_7 = sd(delta_ch) / mean(delta_ch)
```

```

# design effect
data_total5 = round(data.frame(str = sort(unique(int)), nh, deff_GM, design_eff,
delta_sh, delta_ch, delta_wh, ah_str, cls_eff), 4)

# multistage design effect with no-strata option
deff_ms = sum(Ah_str * delta_ch * delta_wh)
v_ms = (sd_srs_GM_total^2 * deff_ms) / nh_total
d_ms = deff_GM_total / deff_ms

# no-strata design effect
data_total6 = data.frame(deff_ms, v_ms, d_ms)

# new stratified effect measure
# comparison strata method
# global optimum
B = Mh * sd_srs_GM * sqrt(delta_ch * delta_wh)

nh_glo = sum(nh) * (B / sum(B))
adj_glo = ifelse(nh_glo > 2, nh_glo, 0)/sum(nh_glo[nh_glo > 2]) *
(sum(nh_glo)-sum(nh_glo < 2)*2)
nh_glo1 = round(ifelse(nh_glo < 2, 2, adj_glo))

vh_glo = sd_srs_GM^2 * delta_ch * delta_wh / nh_glo1

cvh_glo = sqrt(vh_glo / mu_GM^2)

v_glo = sum(Ah_str^2 * vh_glo)
cv_glo = v_glo / mu_GM_total^2

# local optimum
ak = sqrt(sum(sd_srs_GM^2 * delta_ch * delta_wh / mu_GM^2) / sum(nh))

nh_loc = (sd_srs_GM^2 * delta_ch * delta_wh) / (mu_GM^2 * ak^2)

```

```

adj_loc = ifelse(nh_loc > 2, nh_loc, 0)/sum(nh_loc[nh_loc > 2]) *
(sum(nh_loc)-sum(nh_loc < 2)*2)
nh_loc1 = round(ifelse(nh_loc < 2, 2, adj_loc))

vh_loc = sd_srs_GM^2 * delta_ch * delta_wh / nh_loc1
cvh_loc = sqrt(vh_loc / mu_GM^2)

v_loc = sum(Ah_str^2 * vh_loc)
cv_loc = sqrt(v_loc / mu_GM_total^2)

# sqrt proportion optimum
nh_sqrt = round(sum(nh) * sqrt(Mh) / sum(sqrt(Mh)))
adj_sqrt = ifelse(nh_sqrt > 2, nh_sqrt, 0)/sum(nh_sqrt[nh_sqrt > 2]) *
(sum(nh_sqrt)-sum(nh_sqrt < 2)*2)
nh_sqrt1 = round(ifelse(nh_sqrt < 2, 2, adj_sqrt))

vh_sqrt = sd_srs_GM^2 * delta_ch * delta_wh / nh_sqrt1
cvh_sqrt = sqrt(vh_sqrt / mu_GM^2)

v_sqrt = sum(Ah_str^2 * vh_sqrt)
cv_sqrt = sqrt(v_sqrt / mu_GM_total^2)

# 현 표본설계
v_h = se_GM^2
cv_h = sqrt(v_h / mu_GM^2)

v = se_GM_total^2
cv = sqrt(v / mu_GM_total^2)

# 측도 계산
delta_1 = RSW
delta_2 = (sd_LM_total / sd_LM_nostr)^2

# Global 측도

```

```
delta_31 = v_glo / v
```

```
delta_32 = sum(abs(nh - nh_glo1) / nh_glo1, na.rm = T) / length(nh)
```

```
delta_33 = sum(abs(sqrt(v_h) - sqrt(vh_glo)) / sqrt(vh_glo), na.rm = T) /  
length(nh)
```

```
delta_32_str = abs(nh - nh_glo1) / nh_glo1
```

```
delta_33_str = abs(sqrt(v_h) - sqrt(vh_glo)) / sqrt(vh_glo)
```

```
# local 측도
```

```
delta_41 = v_loc / v
```

```
delta_42 = sum(abs(nh - nh_loc1) / nh_loc1, na.rm = T) / length(nh)
```

```
delta_43 = sum(abs(sqrt(v_h) - sqrt(vh_loc)) / sqrt(vh_loc), na.rm = T) /  
length(nh)
```

```
delta_42_str = abs(nh - nh_loc1) / nh_loc1
```

```
delta_43_str = abs(sqrt(v_h) - sqrt(vh_loc)) / sqrt(vh_loc)
```

```
# sqrt 측도
```

```
delta_51 = v_sqrt / v
```

```
delta_52 = sum(abs(nh - nh_sqrt1) / nh_sqrt1, na.rm = T) / length(nh)
```

```
delta_53 = sum(abs(sqrt(v_h) - sqrt(vh_sqrt)) / sqrt(vh_sqrt), na.rm = T) /  
length(nh)
```

```
delta_52_str = abs(nh - nh_sqrt1) / nh_sqrt1
```

```
delta_53_str = abs(sqrt(v_h) - sqrt(vh_sqrt)) / sqrt(vh_sqrt)
```

```
# 층화 효과 적극성
```

```
delta_6 = sd(delta_sh) / mean(delta_sh)
```

```
# 결과 값 정리
```

```
data_total7 = data.frame(delta_1, delta_2)
```

```

delta_3 = c(delta_31, delta_32, delta_33)
delta_4 = c(delta_41, delta_42, delta_43)
delta_5 = c(delta_51, delta_52, delta_53)

delta_nh = data.frame(str = sort(unique(int)), nh_glo = delta_32_str, nh_loc =
delta_42_str, nh_sqrt = delta_52_str)
delta_seh = data.frame(str = sort(unique(int)), se_glo = delta_33_str, se_loc =
delta_43_str, se_sqrt = delta_53_str)

data_total8 = data.frame(rbind(delta_3, delta_4, delta_5))
colnames(data_total8) = c("vp", "str_nh", "str_se")
rownames(data_total8) = c("Global", "Local", "Sqrt")

data_total9 = data.frame(rbind(nh, nh_glo1, nh_loc1, nh_sqrt1));
colnames(data_total9) = unique(str)
data_total10 = round(data.frame(rbind(v_h, vh_glo, vh_loc, vh_sqrt)),5);
colnames(data_total10) = unique(str)

data_total11 = data.frame(delta_1, delta_2, delta_31, delta_32, delta_33, delta_41,
delta_42, delta_43, delta_51, delta_52, delta_53, delta_6, cv_eff, delta_7)
names(data_total11) = c("delta_1", "delta_2", "glo_vp", "glo_nh", "glo_se",
"loc_vp", "loc_nh", "loc_se", "sqrt_vp", "sqrt_nh", "sqrt_se", "delta_6", "cv_eff",
"delta_7")

return(list(int_stat = data_total1, int_str = data_total2, int_deff = data_total3,
nostr_stat = data_total4, dsgn_eff = data_total5, nostr_dsgn_eff = data_total6,
old_measure = data_total7, new_measure = data_total8, total_measure =
data_total11, nh_str = data_total9, vh_str = data_total10, delta_nh = delta_nh,
delta_seh = delta_seh, str_ah = str_ah))

} else {
return(list(int_stat = data_total1, int_str = data_total2, int_deff = data_total3))
}

```



```

}

# ===== 함수 실행 (high school data) ===== #
attach(data_high1, warn.conflicts = FALSE)
A = int_fun(var = BPb, int = gender, data = data_high1, wt = wgt, str =
st_var, psu = j_id, domain = 0)
B = int_fun(var = BHg, int = gender, data = data_high1, wt = wgt, str =
st_var, psu = j_id, domain = 0)
C = int_fun(var = Uhg, int = gender, data = data_high1, wt = wgt, str =
st_var, psu = j_id, domain = 0)
D = int_fun(var = Ucd, int = gender, data = data_high1, wt = wgt, str =
st_var, psu = j_id, domain = 0)
E = int_fun(var = OHP, int = st_var, data = data_high1, wt = wgt, str =
st_var, psu = j_id, domain = 0)
F = int_fun(var = NAP, int = st_var, data = data_high1, wt = wgt, str =
st_var, psu = j_id, domain = 0)
detach(data_high1)

# 1. 표 5. 항목별 측도 (delta_1 ~ delta_6)

total_measure.A = A$total_measure
total_measure.B = B$total_measure
total_measure.C = C$total_measure
total_measure.D = D$total_measure
total_measure.E = E$total_measure
total_measure.F = F$total_measure

total_measure.A = data.frame(value = t(total_measure.A), type = rep(1,14),
total_measure = colnames(total_measure.A))
total_measure.B = data.frame(value = t(total_measure.B), type = rep(2,14),
total_measure = colnames(total_measure.B))
total_measure.C = data.frame(value = t(total_measure.C), type = rep(3,14),
total_measure = colnames(total_measure.C))
total_measure.D = data.frame(value = t(total_measure.D), type = rep(4,14),

```

```

total_measure = colnames(total_measure.D))
total_measure.E = data.frame(value = t(total_measure.E), type = rep(5,14),
total_measure = colnames(total_measure.E))
total_measure.F = data.frame(value = t(total_measure.F), type = rep(6,14),
total_measure = colnames(total_measure.F))

total_measure = rbind(total_measure.A, total_measure.B, total_measure.C,
total_measure.D, total_measure.E, total_measure.F)

```

2. 항목별 층화 전략별 층별 SE 상대 편차

```

delta_seh.A = A$delta_seh
delta_seh.B = B$delta_seh
delta_seh.C = C$delta_seh
delta_seh.D = D$delta_seh
delta_seh.E = E$delta_seh
delta_seh.F = F$delta_seh

delta_seh.A = data.frame(delta_seh.A, type = rep("BPb", nrow(delta_seh.A)))
delta_seh.B = data.frame(delta_seh.B, type = rep("BHg", nrow(delta_seh.B)))
delta_seh.C = data.frame(delta_seh.C, type = rep("Uhg", nrow(delta_seh.C)))
delta_seh.D = data.frame(delta_seh.D, type = rep("Ucd", nrow(delta_seh.D)))
delta_seh.E = data.frame(delta_seh.E, type = rep("OHP", nrow(delta_seh.E)))
delta_seh.F = data.frame(delta_seh.F, type = rep("NAP", nrow(delta_seh.F)))

delta_seh = rbind(delta_seh.A, delta_seh.B, delta_seh.C, delta_seh.D, delta_seh.E,
delta_seh.F)

dsgn_eff.A = A$dsgn_eff
dsgn_eff.B = B$dsgn_eff
dsgn_eff.C = C$dsgn_eff
dsgn_eff.D = D$dsgn_eff
dsgn_eff.E = E$dsgn_eff
dsgn_eff.F = F$dsgn_eff

```

```
dsgn_eff.A = data.frame(dsgn_eff.A, type = rep("BPb", nrow(dsgn_eff.A)))
dsgn_eff.B = data.frame(dsgn_eff.B, type = rep("BHg", nrow(dsgn_eff.B)))
dsgn_eff.C = data.frame(dsgn_eff.C, type = rep("Uhg", nrow(dsgn_eff.C)))
dsgn_eff.D = data.frame(dsgn_eff.D, type = rep("Ucd", nrow(dsgn_eff.D)))
dsgn_eff.E = data.frame(dsgn_eff.E, type = rep("OHP", nrow(dsgn_eff.E)))
dsgn_eff.F = data.frame(dsgn_eff.F, type = rep("NAP", nrow(dsgn_eff.F)))
```

```
dsgn_eff = rbind(dsgn_eff.A, dsgn_eff.B, dsgn_eff.C, dsgn_eff.D, dsgn_eff.E,
dsgn_eff.F)
```

3. 그림 1. 층화 효과의 변동성

```
p1 = xyplot(desc(type) ~ value | total_measure, data =
total_measure[total_measure$total_measure %in% c("delta_6"),], scales = list(y =
list(at = -1 : -6, labels = c("BPb", "BHg", "Uhg", "Ucd", "OHP", "NAP"))),
layout = c(1,1), ylab = "type",
panel = function(x, y, ...){
panel.xyplot(x, y, pch = 20, cex = 1.5, col = "black" , ...)
panel.abline(h = c(-1:-6), col = "grey", lty = 2)
}
)
```

4. 그림 2. 층화 전략별 신규 층 효율성 측도 v_p

```
p2 = xyplot(desc(type) ~ value | total_measure, data =
total_measure[total_measure$total_measure %in% c("glo_vp", "loc_vp",
"sqrt_vp"),],
scales = list(y = list(at = -1:-6, labels = c("BPb", "BHg",
"Uhg", "Ucd", "OHP", "NAP"))), layout = c(3,1), ylab = "type",
panel = function(x, y, ...){
panel.xyplot(x, y, pch = 20, cex = 1.5, col = "black" ,xlim = c(0, 2), ...)
panel.abline(h = c(-1:-6), col = "grey", lty = 2)
panel.abline(v = 1.0, col = "red")
}
```

```

    }
)

```

5. 그림 3, 4, 5. 층화 전략별 층별 SE의 상대편차

```

p3 = xyplot(str ~ se_glo | type, data = delta_seh, layout = c(3,2), xlim = c(0,2),
  panel = function(x, y, ...){
    panel.xyplot(x, y, pch = 20, ...)
    panel.text(x, y, labels = sort(unique(delta_seh$str)), col = grey(0.5), cex = 1,
pos = 3)
    panel.abline(h = c(1,2,4,6,7,8,9), col = "grey", lty = 2)
  }
)

```

```

p4 = xyplot(str ~ se_loc | type, data = delta_seh, layout = c(3,2), xlim = c(0,2),
  panel = function(x, y, ...){
    panel.xyplot(x, y, pch = 20, ...)
    panel.text(x, y, labels = sort(unique(delta_seh$str)), col = grey(0.5), cex = 1,
pos = 3)
    panel.abline(h = c(1,2,4,6,7,8,9), col = "grey", lty = 2)
  }
)

```

```

p5 = xyplot(str ~ se_sqrt | type, data = delta_seh, layout = c(3,2), xlim =
c(0,2),
  panel = function(x, y, ...){
    panel.xyplot(x, y, pch = 20, ...)
    panel.text(x, y, labels = sort(unique(delta_seh$str)), col = grey(0.5), cex = 1,
pos = 3)
    panel.abline(h = c(1,2,4,6,7,8,9), col = "grey", lty = 2)
  }
)

```

Stratification Effects under Multistage Sampling

Tae Hun Kim

Department of Statistics, The Graduate School,
Pukyong National University

Abstract

Design effect is defined as the ratio of the variance estimate under the current sample design to that under a simple random sample of the same size, enabling to measure the efficiency of the current sample design as compared to a simple random sample. Therefore, the design effect can be used to evaluate the efficiency of stratification under unclusterd simple random sampling situation. However, the design effect can be inappropriate for measuring the efficiency of stratification under multistage stratified sample design due to the presence of the effects of clustering and weighting. In this paper, we propose a measure for evaluating the efficiency of stratification strategy utilizing the design effect model proposed by Park(2015) and Chen and Rust(2017). In order to evaluate the suitability of the new measure of the efficiency of stratification, we used the real survey data for high school students of the 4th National Health Basic Survey organized by the National Institute of Health. The stratification efficiency evaluation measure is designed to allow post-evaluation of various survey variables on survey data after survey and to present an improved stratification strategy for sample redesign.

Keywords : stratified multistage sampling, stratification, sample allocation, design effect model, stratification effect, cluster effect, weighting effect