



저작자표시-비영리-변경금지 2.0 대한민국

이용자는 아래의 조건을 따르는 경우에 한하여 자유롭게

- 이 저작물을 복제, 배포, 전송, 전시, 공연 및 방송할 수 있습니다.

다음과 같은 조건을 따라야 합니다:



저작자표시. 귀하는 원저작자를 표시하여야 합니다.



비영리. 귀하는 이 저작물을 영리 목적으로 이용할 수 없습니다.



변경금지. 귀하는 이 저작물을 개작, 변형 또는 가공할 수 없습니다.

- 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

복잡표본설계를 위한 설계효과모형
및 설계요소 평가



2019년 2월

부 경 대 학 교 대 학 원

통 계 학 과

황 현 길

이 학 석 사 학 위 논 문

복잡표본설계를 위한 설계효과모형
및 설계요소 평가

지도교수 박 인 호

이 논문을 석사 학위논문으로 제출함.

2019년 2월

부 경 대 학 교 대 학 원

통 계 학 과

황 현 길

황현길의 이학석사 학위논문을 인준함.

2019년 2월 22일



위원장	이학박사	하 일 도 (인)
위원	이학박사	노 맹 석 (인)
위원	이학박사	박 인 호 (인)



목 차

표 차례	ii
그림 차례	iii
제 1장 서론	1
1.1 설계효과 정의	1
1.2 설계효과 용례	4
1.3 연구동기 및 목적	5
제 2장 설계효과 모형의 기존 연구	7
2.1 Kish 모형	7
2.2 Park / Chen and Rust 모형	9
제 3장 모의실험	10
3.1 유한모집단 생성	10
3.1.1 모집단 구성 : 층화	11
3.1.2 모집단 구성 : 집락	12
3.1.3 모집단 구성 : 불균등가중치	13
3.1.4 유한모집단 구성 및 특성	14
3.2 표본설계, 설계효과 및 설계요소 추정	16
3.3 모의실험결과	18
제 4장 결론	30
참고문헌	31
부록 A.1 R코드	33

표 차례

[표 3.1.3] 추출틀 정보(M_{hi})와 실제 정보(M_{hi}^*)	13
[표 3.1.4.1] 유한모집단 구성 및 특성	15
[표 3.1.4.2] 설계효과 및 층별 설계요소	15
[표 3.3.1] 설계효과($deff_M, deff_D$) 분포	21
[표 3.3.2] 층화설계요소(δ_{sh}) 분포	22
[표 3.3.3] 가중치설계요소($deff_{wh}$) 분포	23
[표 3.3.4] 집락설계요소($deff_{ch, Kish}$) 분포	23
[표 3.3.5] 급 내 상관계수($\rho_{h, Kish}$) 분포	24
[표 3.3.6] 집락설계요소($deff_{ch, AOV}$) 분포	24
[표 3.3.7] 급 내 상관계수($\rho_{h, AOV}$) 분포	25

그림 차례

[그림 3.3.1]	설계효과 산점도	19
[그림 3.3.2]	층 별 설계효과에 대한 상자그림	20
[그림 3.3.3]	층화, 가중치, 집락(ICC)효과요소의 상대편향	26
[그림 3.3.4]	층화, 가중치, 집락(ICC)효과요소의 RRMSE	27



Evaluation of design effect models for complex surveys

Hyeon Gil Hwang

Department of Statistics, The Graduate School,
Pukyong National University

Abstract

The design effect is defined by a ratio of the variance of the estimates of interest under a complex sample design to the variance of a hypothetical simple random sample of the same sample size. It is often used to evaluate the effectiveness of design components of the exiting surveys or to plan new sample designs. Complex sample design usually contains complex design features such as stratification, clustering, and unequal weighting. Kish(1987) presented design effect model in the form of multiplication for each model of the clustering and the unequal weighting effects based on intuition. Gabler *et al.*(1999) gave justification to Kish's design effect model based on the one-way random effect model. We study the design effect model by Park(2015), Chen and Rust(2017) that extends of Gabler *et al.*(1999)'s model to include stratification effects as well as Kish's model. A limited simulations are carried out to assess the design effect model and the

design component estimates under stratified multistage sample design.

keywords : complex sample design, design effect, stratification, simulation.



제 1장 서론

1.1 설계효과 정의

조사연구(survey research)를 위한 자료 수집은 집락화(clustering), 불균등 가중치(unequal weighting), 층화(stratification)등과 같은 복잡한 설계요소를 포함한 복잡표본설계(complex sample design)를 통해 이루어진다. 예를 들어, 농촌진흥청이 주관하는 농어업인 복지실태조사의 표본설계는 층화삼단추출을 활용하고 있는데, 동지역은 서울, 광역, 기타로 읍면지역은 경기, 강원, 충북, 충남, 전북, 전남, 경북, 경남, 제주로 층 구분이 되어있다. 첫 단계에서 층별로 동읍면을 크기비례로 추출하며, 두 번째 단계에서는 조사구를 크기비례로 추출하고, 마지막으로 가구를 단순확률로 추출한다.

설계효과는 Cornfield(1951)에 의해 처음 소개되었고 Kish(1965, 1987, 1994)를 통해 널리 알려졌다. 설계효과는 복잡표본설계에 대한 분산과 동일한 표본크기를 갖는 가상의 단순임의추출에서의 분산 비로 정의된다. 예로 모평균 추정량에 대한 설계효과는 다음과 같이 정의가 된다.

$$deff_p(\bar{y}) = \frac{v_p(\bar{y})}{v_{rs}(\bar{y})} \quad (1.1)$$

여기서 v_p 는 복잡표본설계 $p(s)$ 하의 분산,

v_{sts} 는 단순임의추출 하의 분산,

$$\bar{y} = \sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} y_{hij} / \sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} \text{는 모평균 추정량이다.}$$

예시 1 (층화효과) 층화단순임의추출 하에서 표본평균은 다음과 같이 정의된다.

$$\bar{y}_{st} = \sum_h \frac{N_h}{N} \sum \frac{y_{hi}}{n_h} = \sum_h A_h \bar{y}_h$$

여기서 n_h 는 h 층에서 N_h 에서 추출된 표본의 크기이고 $N = \sum N_h$ 는 모집단 크기, $A_h = N_h/N$ 이다. 층화 평균의 분산은 다음과 같다.

$$V(\bar{y}_{st}) = \sum_h \frac{A_h^2 S_h^2}{n_h} (1 - f_h) \quad (1.1.1)$$

여기서 $S_h^2 = \sum_i (Y_{hi} - \bar{y}_h)^2 / (N_h - 1)$ 는 h 층 내에서 모집단 분산이다. 모집단 전체에 대한 특성치 분산이 S^2 이라면 동일한 규모의 표본크기를 갖는 가상적인 단순(복원)확률추출 하에서 단순평균의 분산은 S^2/n 이므로, 층화표본평균의 설계효과는 다음과 같이 정의된다.

$$deff(\bar{y}_{st}) = \frac{V(\bar{y}_{st})}{S^2/n} = \sum_{h=1}^H \frac{A_h^2}{a_h} \frac{S_h^2}{S^2} (1 - f_h)$$

예시 2 (집락효과) 국가 전체 또는 지역에 걸쳐 널리 퍼져있는 표본을 추출하는 것은 비경제적이므로 자료 수집 비용을 줄이기 위해 주로 지역기반의 집락을 사용한다. 일반적으로 1차 추출단위(primary sampling unit, PSU)를 추출하고 선택된 PSU 내에서 다음 단위를 추출한다. 실제로 PSU의 크기는 항상 다르며 크기비례로 추출하게 되며, 집락평균은 다음과 정의된다.

$$\bar{y}_{cl} = \frac{1}{mn} \sum_i \sum_j y_{ij}$$

여기서 n 은 단순확률추출로 뽑힌 집락 수, m 은 집락표본크기, y_{ij} 는 i 번째 집락의 j 번째 개체이다. 만약 동일한 집락크기와 유한모집단 수정계수를 고려하지 않는다면 집락 평균에 대한 분산은 다음과 같다 (Lohr, 2010).

$$v(\bar{y}_{cl}) = \frac{1}{nm} MSB$$

여기서 $MSB = m \sum (\bar{y}_i - \bar{y}_{cl})^2 / (n-1)$ 은 집락 간 평균제곱이다. 집락평균에 대한 설계효과는 정의 (1.1)에 의해 다음과 같이 표현된다.

$$deff(\bar{y}_{cl}) = \frac{MSB}{S^2} = 1 + (M-1)ICC.$$

여기서, ICC 는 집락 내 상관계수이고, M 은 집락 크기이다.

예시 3(랜덤가중효과) 불균등한 가중치를 사용하면 그 변동성으로 인해 그렇지 않은 경우와 비교할 때 추정량의 분산을 증가시키는 것으로 알려져 있다. Kish(1992)는 이러한 상황을 표현하기 위해 불균등한 가중치가 층별로 다를

수 있음을 가정하여 층화평균에 대한 설계효과로 표현하는 방법을 제시하였다. 이러한 설계효과는 분석단계에서 불균등 층화의 효과를 산출하는데 매우 유용하다. 층 별 평균과 분산이 같다고 가정하면, 층화평균의 설계효과는 다음과 같이 표현된다.

$$def(\bar{y}_{st}) = \frac{n \sum_h \sum_i w_{hi}^2}{(\sum_h \sum_i w_{hi})^2} = 1 + cv^2(w_{hi})$$

여기서 $cv(w_{hi})$ 는 가중치의 변동계수, $cv^2(w_{hi}) = \sum \sum (w_{hi} - \bar{w})^2 / n \bar{w}^2$, $\bar{w} = \sum \sum w_{hi} / n$ 이다. 불균등 가중치의 상황을 나타내는 형태는 다음 식 (1.1.3)과 같다.

$$def(\bar{y}_{st}) = \frac{n \sum_j w_j^2}{(\sum_j w_j)^2} = 1 + cv^2(w_j) \quad (1.1.3)$$

1.2 설계효과 용례

설계효과와 관련하여 또 하나의 유용한 개념은 표본의 크기를 결정과 관련 있다. 일반적으로 표본설계는 조사비용, 시간 등의 현실적인 이유로 인해 이를 보완하기 위해 이루어진다. 표본 크기 결정은 복잡표본설계 하에서 같은 정도를 갖는 표본의 크기를 구하기 위해서는 단순임의추출 하에서 구한 표본의 크기에 설계효과(design effect)를 곱해 표본설계에 맞는 표본크기를 결정한다. 이러한 개념을 유효표본크기(effective sample size)라 하는데 다음과 같이 정의된다.

$$n_{eff} = \frac{n}{def_f_p(\bar{y})} \quad (1.2)$$

여기서 n_{eff} 는 유효표본크기

n 은 복잡표본설계 하에서 표본의 크기

$def_f_p(\bar{y})$ 는 모평균 추정량에 대한 설계효과

예를 들어, 캐나다 TIMSS는 수학·과학 성취도 추이변화 국제비교 연구로 캐나다 학생의 수학 점수의 평균에 대한 설계효과 값은 12.99, 표본크기는 16,002로 나타난다(Gonzalez and Foy, 1997). 이는 단순확률추출과 비교하여 복잡표본설계에 의해 평균추정의 분산이 약 13배가량 큼을 나타내며 단순확률 하의 자료분석을 기준으로 보면 유효표본크기 식 (1.2)는 1,232개, 즉 실제 표본수를 표본 설계효과를 나눈 값밖에 되지 않는 것을 의미한다.

1.3 연구동기 및 목적

앞에서 언급한 설계효과와 유효표본크기는 복잡한 표본설계요소(집락화, 불균등 가중치, 층화 등)의 함수형태로 나타내어진다면 그 유용성은 더욱 더 커지게 된다. 예로, Rust and Broene(2010)은 다음과 같이 언급하였다.

“설계효과의 개념은 할 수만 있다면 모집단의 구조와 설계요소들이 갖는 개별적 영향을 명백히 나타낼 수 있는 함수적 형태로 표현될 때 그

유용성은 극대화 된다. 이렇게 될 때 향후 적용 시 설계효율을 증진시킬 수 있는 지침으로 사용될 수 있기 때문이다.”

집락추출을 위한 설계효과모형은 주로 경험적 제안(Kish, 1965, 1994)이나 모수 모형 식에 근거한 근사 추정식(Gabler *et al.*, 1999, 2006)만이 제시되었고 층화다단추출을 위한 설계효과모형으로 Park(2015)과 Chen and Rust(2017)에 의해 제시 되었다. 본 연구에서는 몇 가지 설계효과모형의 적절성과 안정성에 대해 모의실험을 통해 평가하고자 한다. 2장에서는 Kish(1994)의 설계효과 모형식과 Park(2015)과 Chen and Rust(2017)가 제시한 설계효과모형식에 대하여 정리한다. 3장에서는 층화다단추출을 위한 집락효과, 가중치효과, 층화효과가 반영된 유한모집단을 구축하고 몬테카를로 모의실험을 통한 반복적 표본추출 및 추정치 산출을 통해 설계효과 모형식을 평가한다. 4장에서는 모의실험에 따른 기본적 논의와 향후 연구과제에 대하여 언급하고자 한다.

제 2장 설계효과모형의 기존 연구

본 장에서는 Kish와 Park 등이 제시한 설계효과모형의 적절성과 안정성에 대해 모의실험을 통해 평가하기에 앞서 기존 연구를 논의한다. 2.1절에서는 집락과 불균등가중치를 반영한 Kish(1965, 1987, 1994)와 Gabler 등(1999)의 모형을 살펴보고 2.2절에서는 층화, 집락, 불균등가중치를 모두 반영한 Park(2015)과 Chen and Rust(2017)의 모형을 살펴본다.

2.1 Kish 모형

다단추출(이단추출)에 의한 표본설계 $p(s)$ 를 통해 집락 i 의 조사단위 j 에 대한 y_{ij} 와 그에 대한 가중치 w_{ij} 가 주어진다. 이때 관심특성 y_{ij} 와 표본가중치 w_{ij} 는 상관관계가 없다고 가정하자. 즉, $cor(w_{ij}, y_{ij}) = 0$ 이다. Kish(1987, 1994)는 랜덤가중치효과와 집락효과를 고려한 설계효과 모형식을 다음과 같이 제안하였다.

$$deff_K(\bar{y}) = deff_{w, Kish} deff_{c, Kish} \quad (2.1.1)$$

여기서 $deff_{c, Kish} = 1 + (\bar{m} - 1)\rho_K$ 는 집락효과이고 $deff_{w, Kish} = 1 + cv_w^2$ 은 랜덤가중효과이고 $\bar{m}$ 은 평균집락표본크기이고 cv_w^2 은 표본가중치의 상대분산이고 ρ_K 는 급 내 상관계수(Intraclass correlation, ICC)이다. 집락표본크기와 급 내 상관계수가 클 경우 설계효과식이 증가함을 명확히 나타내고 있다. 급 내 상관계수와 유사한 의미로 동질성 계수(Valliant

et al., 2013)가 있으며 다음과 같이 정의된다.

$$\delta = \frac{S_B^2}{S_B^2 + S_W^2} \quad (2.1.2)$$

여기서 S_B^2 는 집락 간 분산을 의미하고 S_W^2 는 집락 내 분산을 의미한다. 집락 간 분산이 집락 내 분산보다 클 경우 동질성 계수는 큰 값을 가지게 되며 반대로 집락 간 분산이 집락 내 분산보다 작을 경우 작은 값을 가진다. 집락은 동질적인 개체로 집락 내부는 이질적으로 개체구성이 이루어지면 동질성계수가 작은 값을 가져 추정에 있어서 좋은 효과를 나타낸다.

Kish 식 (2.1.1)은 경험적 바탕으로 직관에 의해 제시되었다. Gabler *et al.* (1999)는 일원확률효과모형에 근거하여 Kish의 설계효과모형식 (2.1.1)에 대하여 정당화하였다. 또한 다음과 같은 상한값을 가짐을 보였다.

$$deff_G(\bar{y}) = [1 + (m^* - 1)\rho][1 + cv_w^2] \quad (2.1.3)$$

여기서 $m^* = \sum_{i=1}^n (\sum_{j=1}^{m_i} w_{ij})^2 / \sum_{i=1}^n \sum_{j=1}^{m_i} w_{ij}$ 는 일종의 평균집락크기로 모든 집락표본이 동일할 때 식 (2.1.1)의 $\bar{m}$ 와 같다.

2.2 Park / Chen and Rust 모형

다단설계에 대하여 Kish(1987, 1994)에 의해서 설계효과에 대한 식이 제안되었지만 층화효과에 대해서는 고려되지 않았다. 하지만 표본조사에서 일반적으로 층화추출법은 가장 많이 이용되는 추출법이다. h 층 집락 i 의 조사단위 j 에 대한 y_{hij} 와 그에 대한 가중치 w_{hij} 가 주어진다고 가정하자. Park(2015)과 Chen and Rust(2017)은 층의 효과를 고려한 층화 다단추출에 대한 설계효과모형식(이하 PCR 모형식)을 다음과 같이 제안하였다.

$$deff_{PCR}(\bar{y}) = \sum_{h=1}^H \delta_{sh} deff_{ch} deff_{wh} \quad (2.2.1)$$

여기서 $\delta_{sh} = (A_h^2/a_h)/(\sigma_{yh}^2/\sigma_y^2)$ 는 층화효과(설계요소)모형이고 $A_h = M_h/M$ 는 층별 상대크기, $a_h = m_h/m$ 는 층별 표본상대크기, σ_{yh}^2 는 h 층의 분산, σ_y^2 는 전체 분산을 의미한다. $deff_{wh} = 1 + cv_h^2(w)$ 는 가중효과(설계요소)모형이고 $deff_{ch} = 1 + (m_h^* - 1)\rho_h$ 는 집락효과(설계요소)모형이고 m_h^* 는 층별 평균집락표본크기이며 층별 집락크기가 동일하다면 Kish모형식에서 $\bar{m}$ 과 같아진다. 또한 집락과 층화에 대한 효과(설계요소)모형은 조사단위 특성(y)와 연관이 되어있지만 랜덤가중으로 인한 가중(설계요소)모형은 연관되어 있지 않다. 즉, $cor(y_{hik}, w_{hik}) = 0$ 이다. 여기서 $h = 1, \dots, H$, $i = 1, \dots, n_h$, $j = 1, \dots, m_{hi}$ 이다.

제 3장 모의실험

본 장에서는 설계효과모형식의 적절성 및 안정성을 평가하기 위해 층화이단추출 표본설계를 적용할 가상의 유한모집단을 생성하고 몬테카를로실험을 통해 반복적 표본추출 및 추정치 산출하고자 한다. 3.1절에서는 랜덤효과모형을 이용한 유한모집단의 생성에 대해 논의한다. 3.2절에서는 표본설계는 물론 설계효과 및 설계요소 추정에 대해 기술한다. 3.3절에서는 모의실험을 통한 설계효과모형식에 대한 적절성 및 안정성에 대한 평가결과를 요약한다.

3.1 유한모집단 생성

2장에서 논의한 설계효과모형의 적절성 및 안정성을 평가하기 위해 Feng(2006)의 혼합랜덤효과모형을 참고하여 복잡설계요소인 층화효과와 집락효과를 갖춘 유한모집단을 생성하였다. 먼저, h 번째 층의 집락수 $N_h \sim iid N(100, 5^2)$ 생성하였다. 그 결과 N_h 는 92부터 109 ($N_h = 98, 92, 100, 101, 109, 99$)까지 나타난다. 다음으로, h 번째 층 내 i 번째 집락의 개체 수 $M_{hi} \sim iid gamma(\alpha = 2, \beta = 300)$ 생성하였다. 이렇게 생성된 유한모집단의 크기는 6개의 층으로 각 층별로 N_h 개의 집락으로 이루어져 있고 총 376,884개의 조사단위가 생성되었다. 조사단위 특성을 부여하기 위해서 다음의 혼합랜덤효과모형을 가정하였다.

$$y_{hij} = A_{hi} + \epsilon_{hij} \quad (3.1)$$

여기서, $h = 1, \dots, H$, $i = 1, \dots, N_h$ 에 대하여 집락특성은 $A_{hi} \sim iid N(\mu_{Ah}, \sigma_{Ah}^2)$ 으로 평균값 μ_{Ah} (각 층의 평균을 의미)는 $\mu_{Ah} \sim iid N(235, 5^2)$ 으로부터 생성하여 다음의 값을 산출하였다. ($\mu_{Ah} = 225.24, 227.28, 238.65, 233.61, 226.72, 233.74$). 그리고 $\sigma_{Ah}^2 = \rho_h \nu_h^2$ 이다. 여기서 ρ_h 는 급 내 상관관계수(IntraCluster Correlation, ICC)로 층별로 다른 상황을 고려하기 위해서 본 연구에서는 층별로 $\rho_h = 0.01, 0.02, 0.05, 0.10, 0.20, 0.50$ 인 상황을 고려하였다. ν_h^2 는 $\nu_h \sim iid N(20, 2^2)$ 에서 발생한 고정된 값이다($\nu_h = 16.10, 16.91, 21.46, 19.44, 16.69, 19.50$). $h = 1, \dots, H$, $i = 1, \dots, N_h$, $j = 1, \dots, M_{hi}$ 에 대하여 개체특성 $\epsilon_{hij} \sim iid N(0, \sigma_{chi}^2)$ 을 생성하였다. 여기서 σ_{chi}^2 는 $E(\sigma_{chi}^2) = (1 - \rho_h) \nu_h^2$, $V(\sigma_{chi}^2) = \nu_h^2 / 10$ 인 *gamma*분포에서 생성하였다. 그리고 A_{hi} 와 ϵ_{hij} 는 독립이다.

3.1.1 모집단 구성 : 층화

층화는 모집단에 대한 대표성을 높이기 위해서 사용된다. 이를 위해 모집단을 집단 내는 동질적이고 집단 간은 이질적으로 나누는 것이 효과적이다. 앞서 언급했듯이 층화설계요소는 층 분산비 $\sigma_{yh}^2 / \sigma_y^2$, 층별 상대크기 A_h 그리고 층별 표본할당 a_h 로 이루어져 있다. 3.1절 유한모집단에서 조사단위 특성에 대한 분산은 다음과 같다.

$$V(y_{hij}) = V(A_{hi}) + V(\epsilon_{hij}) = \nu_h^2 \quad (3.1.1)$$

여기서 유한 모집단에서 ν_h^2 는 고정된 값으로 초모집단의 총 분산을 의미한다.

3.1.2 모집단 구성 : 집락

집락은 층화와는 다르게 모집단을 이질적인 요소를 포함하는 여러 개의 집락으로 구분하며 집단 내는 이질적이고 집단 간은 동질적인 특성을 가지도록 하는 것이 특징이다. 급 내 상관계수는 집락 내 분산과 집락 간 분산의 상대적 크기로 정의 할 수 있다. 급 내 상관계수 ρ_h 는 모형 (3.1)에 의해 다음 식 (3.1.2)으로 정의될 수 있다(Lohr, 2010).

$$\rho_h = \frac{\sigma_{Ah}^2}{\sigma_{Ah}^2 + \sigma_{\epsilon hi}^2} \quad (3.1.2)$$

즉, 층별 조사특성 y 의 분산 대비 집락간 분산비를 나타낸다. 주의할 것은 식 (3.1.2)에서의 ρ_h 는 비음(nonnegative)의 값만을 갖게 된다.

3.1.3 모집단 구성 : 불균등가중치

랜덤가중(haphazard weight)의 발생원인으로 무응답, 추출틀의 정보 오류 등이 있다. 그 중 추출틀의 정보오류를 반영하여 표본틀 집락규모 오류를 고려하였다. 예를 들어, 3.1절 표본추출틀 내의 집락 (hi)내 총 개체수 M_{hi} 의 정보는 실사과정을 통해 오류가 있었음을 확인할 수 있다. 이러한 고려는 일반적으로 표본추출틀의 정보는 노후화로 인해 최신정보와 차이가 발생할 수 있기 때문이다. 이러한 상황을 반영하기 위해 조건부 감마분포를 이용하였다. 먼저, 표본추출틀 내 집락의 보고된 규모 M_{hi} 와 실제 규모 M_{hi}^* 를 각각 x 와 y 로 표기한다면, 후자는 평균 $\mu(x) = a + bx + cx^2$ 이고 분산 $\sigma^2(x) = d^2 x^{2g}$ 를 갖는 감마분포를 따른다는 가정 하에 생성할 수 있다(Lee *et al.*, 1994). 논의를 단순화 하기 위해 조건부 평균과 분산은 선형모형을 가정하였다($a=0$, $b=1$, $c=0$, $d=2$, $g=0.5$). 추출틀에 주어진 정보와 실제값 간의 상관관계는 0.9이상을 유지하였고 추출틀 정보와 실제 정보의 평균과 표준편차는 표 3.1.3과 같다.

표 3.1.3 추출틀 정보(M_{hi})와 실제 정보(M_{hi}^*)

	평균	표준편차
M_{hi}	631.3	430.4
M_{hi}^*	631.3	430.7

h 층 집락 i 의 조사단위 j 에 대한 불균등 가중치는 식 (3.1.3)과 같이 정의가 된다.

$$w_{hij} = \frac{M_h M_{hi}^*}{n_h M_{hi} m_o} \quad (3.1.3)$$

3.1.4 유한모집단 구성 및 특성

층화, 집락, 불균등가중치 효과가 모두 반영된 층화다단추출을 위한 유한 모집단은 표 3.1.4.1 와 같이 생성되었다. 총 집락의 수는 597개이며 생성된 총 조사단위 수는 376,884개이며 층별 표본할당은 동일하게 주었다. 표 3.1.4.2는 유한모집단에서 층별 설계효과, 층화효과, 집락효과, 가중치효과를 각각 나타내고 있다.

표 3.1.4.1 유한모집단 구성 및 특성

층	N_h	M_{hi}	a_h	$\bar{y}_h$	S_h^2	σ_{Ah}^2	σ_{ehi}^2	ρ_h
1	97	59,933	0.167	225.3	258.5	2.2	256.2	0.008
2	91	57,655	0.167	226.9	288.9	6.4	283.2	0.022
3	100	63,210	0.167	239.3	473.8	19.1	458.5	0.040
4	101	63,991	0.167	233.2	405.1	31.7	373.0	0.078
5	109	69,960	0.167	227.4	331.8	53.4	277.8	0.161
6	99	62,135	0.167	234.3	540.3	179.5	373.9	0.324
전체	597	376,884	1.000	231.1	408.6			

표 3.1.4.2 설계효과 및 층별 설계요소

층	$Deff_h$	δ_{sh}^U	$deff_{ch}^U$	$deff_{wh}^U$
1	1.081	0.096	1.075	1.006
2	1.208	0.099	1.200	1.007
3	1.367	0.196	1.360	1.006
4	1.716	0.171	1.705	1.007
5	2.467	0.168	2.451	1.007
6	3.945	0.216	3.920	1.006
전체	2.051			

3.2 표본설계, 설계효과 및 설계요소 추정

모의실험에서는 앞서 생성된 유한모집단으로부터 층화이단크기비례확률추출을 실시하였다. 총 6개 층으로 이루어진 유한모집단으로부터 층화이단크기비례확률추출로 총 $R=1,000$ 개의 확률표본을 반복적으로 추출하였다. 먼저, 확률비례추출법(probability-proportional-to-size, pps)으로 층별로 일차추출단위(primary sampling unit, psu)를 50개씩($n_h = 50$) 추출하였다. 그 다음, 단순임의추출법(simple random sampling, srs)으로 이차추출단위(secondary sampling unit, SSU)를 10개씩($m_{hi} = 10$) 추출하였다. 이렇게 추출된 표본규모는 총 $m = 3000$ 개 개체를 갖는다.

설계효과는 식 (1.1)과 같이 정의가 되었으며, 이때 분자의 $v_p(\bar{y}_p)$ 는 테일러선형화 분산추정식(Lohr, 2010)을 사용하여 다음과 같이 계산하였다.

$$v_p(\bar{y}_p) = \sum_{h=1}^H \frac{n_h}{n_h - 1} \sum_{i=1}^{n_h} (e_{hi+} - \bar{e}_{h++})^2 \quad (3.2.1)$$

여기서 $e_{hi+} = (\sum_{j=1}^H w_{hij}(y_{hij} - \bar{y}_p))/w_{+++}$, $\bar{e}_{h++} = (\sum_{i=1}^{n_h} e_{hi+})/n_h$ 이고 $\hat{v}_{srs}(\bar{y}_{srs})$

는 단순임의추출 분산추정식을 사용하였으며 $\hat{\sigma}_y^2/m$ 으로 계산되고 $\hat{\sigma}_y^2$ 은 다음과 같은 일차추정량의 형태로 주어진다.

$$\hat{\sigma}_y^2 = \frac{\sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} (y_{hij} - \bar{y}_p)^2}{\sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij}} \quad (3.2.2)$$

Kish 모형식 (2.1.2)의 추정량으로 $deff_{w,Kish} = 1 + cv_w^2$ 가중치 효과모형과 $deff_{c,Kish} = deff_p(\bar{y}) / deff_{w,Kish}$ 집락효과모형 추정량을 얻을 수 있으며 급내 상관계수 ρ_K 는 식 (3.2.3)과 같이 추정할 수 있다(Park and Lee, 2004).

$$\hat{\rho}_{Kish} = \frac{deff_{c,Kish} - 1}{m - 1} \quad (3.2.3)$$

PCR 모형식(2.2.1)은 Kish 모형식에서 층의 효과가 고려된 모형으로 층 내 가중치 효과모형($deff_{wh}$)와 집락효과모형($deff_{ch,Kish}$)은 Kish모형식의 추정방법과 유사하게 추정량을 얻을 수 있으나 층 별 설계효과를 계산하여야 한다. 층 별 설계효과는 식 (3.2.1)과 식 (3.2.2)를 통해 얻을 수 있다. ρ 는 집락 내 조사특성 y 들이 얼마나 유사하는지를 나타내어 주는 하나의 동질성 척도이다. ρ 값이 크다는 것은 집락 내 요소들이 집락 간 요소들에 비해 더 유사하다는 것을 의미한다. 본 연구에서는 ρ 의 다른 추정량으로 식 (3.2.4)와 같이 ANOVA 추정량도 고려하였다(Ganninger, 2010).

$$\hat{\rho}_{h, AOV} = \frac{MSB - MSW}{MSB + (K_h - 1)MSW} \quad (3.2.4)$$

여기서 $MSB = SSB/(n_h - 1)$ 는 집락 간 평균제공합이고 $SSB = \sum_{i=1}^{n_h} m_{hi}(\bar{y}_{hi} - \bar{y}_h)^2$ 는 집락 간 제공합이며, $MSW = SSW/(m_h - n_h)$ 는 집락 내 평균제공합이고, $SSW = 1/(n_h - 1)(m_h - \sum_{i=1}^{n_h} \frac{m_{hi}^2}{m_h})$ 는 집락 내 제공합이고 $K_h = \frac{1}{n_h - 1}(m_h - \sum_{i=1}^{n_h} \frac{m_{hi}^2}{m_h})$ 이다. 층화효과모형의 추정량은 다음 식 (3.2.5)과 같이 계산되었다.

$$\hat{\delta}_{sh} = \frac{(\hat{M}_h/\hat{M})^2}{m_h/m} \frac{\hat{\sigma}_{yh}^2}{\hat{\sigma}_y^2} = \frac{\hat{A}_h^2}{a_h} \frac{\hat{\sigma}_{yh}^2}{\hat{\sigma}_y^2} \quad (3.2.5)$$

여기서 $\hat{M}_h = \sum_{i=1}^{n_h} \sum_{j=1}^m w_{hij}$, $\hat{M} = \sum_{h=1}^H \hat{M}_h$, $\hat{\sigma}_{yh}^2 = \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} (y_{hij} - \bar{y}_{h..})^2 / \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij}$,
 $\hat{\sigma}_y^2 = \sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij} (y_{hij} - \bar{y}_{...})^2 / \sum_{h=1}^H \sum_{i=1}^{n_h} \sum_{j=1}^{m_{hi}} w_{hij}$ 이다.

3.3 모의실험결과

표 3.3.1은 설계기반의 설계효과($deff_D$), 모형기반의 설계효과($deff_M$) 층별 설계효과($deff_{D,h}$, $h = 1, \dots, 6$)에 대하여 평균(mean), 표준오차(standard error), 최소(min), Q_1 , Q_2 , Q_3 , 최대(max)를 나타내고 있다. $deff_D$ 의 평균은

2.008, $deff_M$ 의 평균은 2.022로 나타나고 있다. 그림 3.3.1은 $deff_D$ 와 $deff_M$ 의 산점도를 나타낸 것이다.

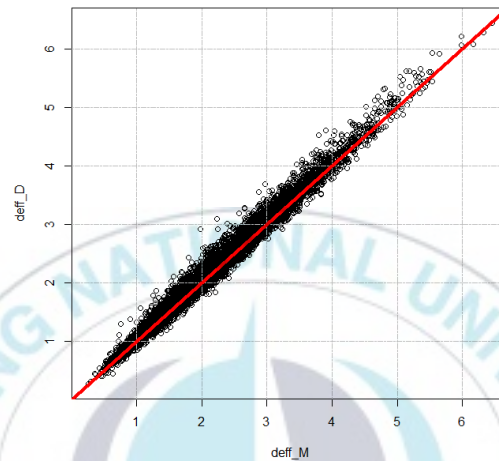


그림 3.3.1 설계효과 산점도
($deff_M = deff_M$, $deff_D = deff_D$)

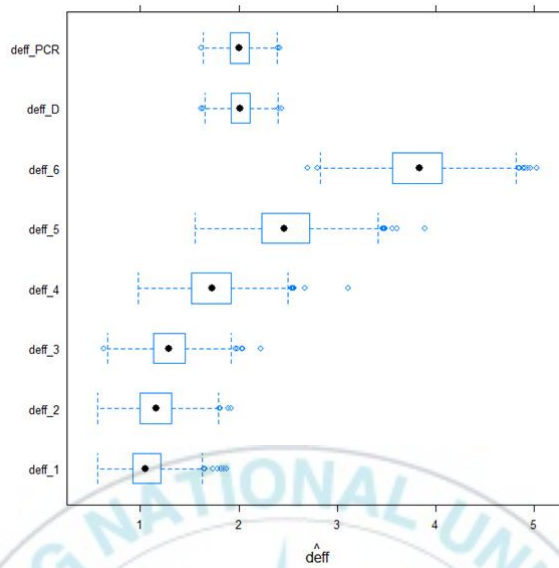


그림 3.3.2 층 별 설계효과에 대한 상자그림

표 3.3.1 설계효과($deff_{PCR}$, $deff_D$) 분포

		평균	표준오차	최소	Q_1	Q_2	Q_3	최대
	$deff_D$	2.022	0.144	1.621	1.925	2.014	2.115	2.435
	$deff_{PCR}$	2.008	0.143	1.615	1.910	2.000	2.102	2.412
	η_y	1.007	0.005	0.990	1.004	1.007	1.010	1.025
층	$deff_{D,1}$	1.069	0.212	0.565	0.922	1.054	1.205	1.872
	$deff_{D,2}$	1.166	0.227	0.564	0.997	1.159	1.318	1.923
	$deff_{D,3}$	1.295	0.235	0.632	1.137	1.288	1.450	2.219
	$deff_{D,4}$	1.736	0.307	0.983	1.517	1.730	1.920	3.115
	$deff_{D,5}$	2.484	0.355	1.553	2.236	2.466	2.718	3.885
	$deff_{D,6}$	3.821	0.388	2.329	3.557	3.833	4.063	5.026

※ 여기서 $\eta_y = deff_D / deff_M$

추출된 표본으로부터 구해지는 편향(bias), 상대편향(relative bias), RMSE(Root Mean Square Error), RRMSE(Relative Root Mean Square Error)을 안정성 평가 통계량으로 사용하였다.

$$bias(\hat{\theta}, \theta) = \hat{\theta} - \theta$$

$$\text{상대편향} = bias(\hat{\theta}, \theta) / \theta$$

$$RMSE(\hat{\theta}) = \sqrt{\sum_r^R (\hat{\theta}_r - \theta)^2 / R + bias^2}$$

$$RRMSE(\hat{\theta}) = RMSE(\hat{\theta}) / \hat{\theta}$$

여기서, $\hat{\theta} = \sum_r^R \hat{\theta}_r / R$, $\hat{\theta}_r$ = r번째 추정값, θ = 유한모집단의 모수이다.

표 3.3.2 층화설계요소(δ_{sh}) 분포

층	평균	표준 오차	편향	상대 편향	<i>rmse</i>	<i>rrmse</i>	Q_1	Q_2	Q_3
1	0.095	0.006	-0.001	-0.006	0.006	0.064	0.091	0.095	0.100
2	0.099	0.006	-0.001	-0.006	0.006	0.065	0.095	0.099	0.103
3	0.196	0.011	0.000	0.002	0.011	0.058	0.189	0.196	0.204
4	0.171	0.011	-0.001	-0.003	0.011	0.063	0.164	0.171	0.178
5	0.167	0.011	-0.001	-0.006	0.011	0.064	0.160	0.166	0.174
6	0.215	0.013	-0.001	-0.003	0.013	0.061	0.206	0.215	0.223

표 3.3.3 가중치설계요소(def_{wh}) 분포

		평균	표준 오차	편향	상대 편향	$rmse$	$rrmse$	Q_1	Q_2	Q_3
	$def_{w,K}$	1.012	0.001					1.011	1.012	1.013
층	1	1.006	0.001	0.000	0.000	0.001	0.001	1.005	1.006	1.006
	2	1.008	0.001	0.001	0.001	0.002	0.002	1.007	1.008	1.009
	3	1.006	0.001	0.001	0.001	0.002	0.002	1.005	1.006	1.007
	4	1.008	0.001	0.001	0.001	0.002	0.002	1.007	1.007	1.008
	5	1.008	0.003	0.001	0.001	0.003	0.003	1.006	1.007	1.010
	6	1.008	0.001	0.001	0.001	0.002	0.002	1.007	1.008	1.009

표 3.3.4 집락설계요소($def_{ch,Kish}$) 분포

		평균	표준 오차	편향	상대 편향	$rmse$	$rrmse$	Q_1	Q_2	Q_3
	$def_{c,Kish}$	1.997	0.142					1.902	1.991	2.090
층	1	1.063	0.211	-0.012	-0.011	0.212	0.199	0.917	1.047	1.198
	2	1.157	0.225	-0.043	-0.036	0.229	0.198	0.990	1.151	1.307
	3	1.286	0.232	-0.073	-0.054	0.243	0.190	1.130	1.280	1.441
	4	1.723	0.304	0.019	0.011	0.304	0.177	1.508	1.715	1.904
	5	2.464	0.353	0.013	0.005	0.353	0.143	2.220	2.445	2.697
	6	3.791	0.388	-0.129	-0.033	0.409	0.108	3.527	3.803	4.034

표 3.3.5 급 내 상관계수($\rho_{h,Kish}$) 분포

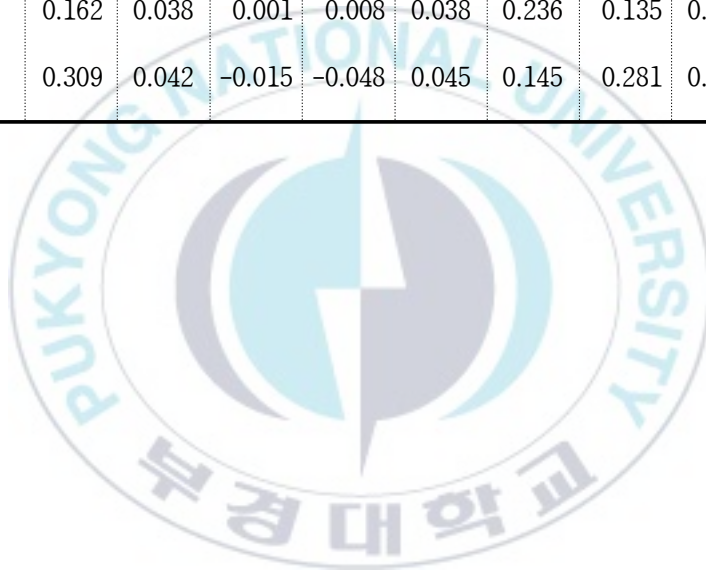
		평균	표준 오차	편향	상대 편향	$rmse$	$rrmse$	Q_1	Q_2	Q_3
	ρ_{Kish}	0.111	0.016					0.100	0.110	0.121
층	1	0.007	0.023	-0.001	-0.162	0.024	3.367	-0.009	0.005	0.022
	2	0.017	0.025	-0.005	-0.215	0.025	1.460	-0.001	0.017	0.034
	3	0.032	0.026	-0.008	-0.203	0.027	0.849	0.014	0.031	0.049
	4	0.080	0.034	0.002	0.026	0.034	0.421	0.056	0.080	0.100
	5	0.163	0.039	0.001	0.009	0.039	0.241	0.136	0.161	0.189
	6	0.310	0.043	-0.014	-0.044	0.045	0.146	0.281	0.311	0.337

표 3.3.6 집락설계요소($deff_{ch,AOV}$) 분포

		평균	표준 오차	편향	상대 편향	$rmse$	$rrmse$	Q_1	Q_2	Q_3
층	1	1.065	0.210	-0.010	-0.009	0.210	0.197	0.923	1.048	1.214
	2	1.162	0.223	-0.038	-0.032	0.226	0.195	0.996	1.156	1.312
	3	1.295	0.227	-0.064	-0.047	0.236	0.182	1.138	1.292	1.453
	4	1.693	0.291	-0.012	-0.007	0.291	0.172	1.481	1.687	1.866
	5	2.462	0.344	0.011	0.005	0.344	0.140	2.217	2.442	2.692
	6	3.781	0.380	-0.139	-0.035	0.404	0.107	3.526	3.775	4.041

표 3.3.7 급 내 상관계수($\rho_{h,AOV}$) 분포

		평균	표준 오차	편향	상대 편향	$rmse$	$rrmse$	Q_1	Q_2	Q_3
층	1	0.007	0.023	-0.001	-0.130	0.023	3.212	-0.009	0.005	0.024
	2	0.018	0.025	-0.004	-0.191	0.025	1.399	-0.000	0.017	0.035
	3	0.033	0.025	-0.007	-0.179	0.026	0.800	0.015	0.032	0.050
	4	0.077	0.032	-0.001	-0.017	0.032	0.420	0.053	0.076	0.096
	5	0.162	0.038	0.001	0.008	0.038	0.236	0.135	0.160	0.188
	6	0.309	0.042	-0.015	-0.048	0.045	0.145	0.281	0.308	0.338



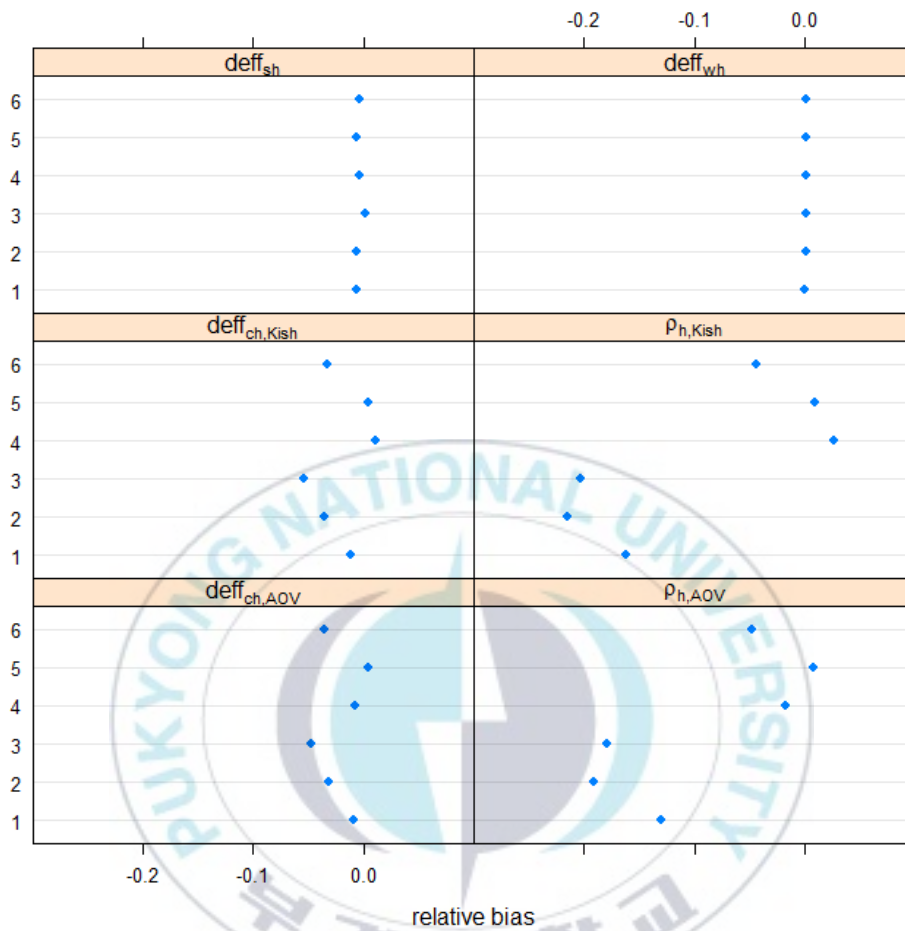


그림 3.3.3 층화, 가중치, 집락(ICC)효과요소의 상대편향

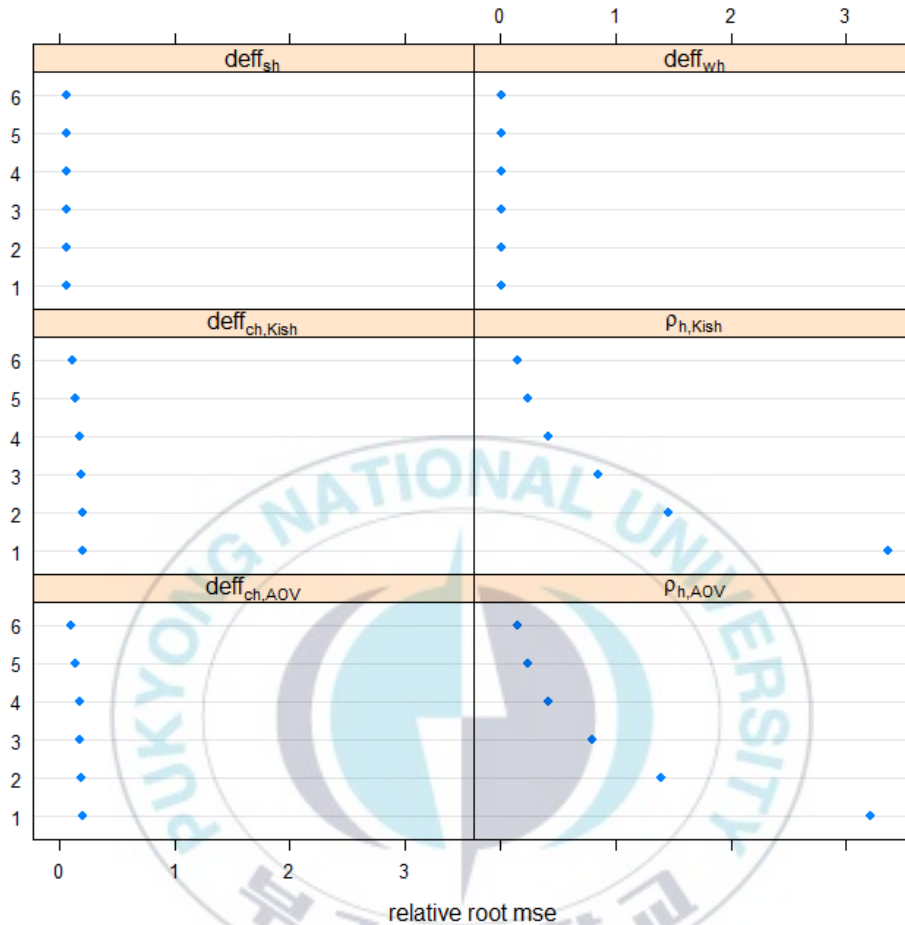


그림 3.3.4 층화, 가중치, 집락(ICC)효과요소의 RRMSE

표 3.3.1에서 설계효과($deff_D$)의 평균과 분포를 확인해보면 평균이 2.022이고 최소가 1.621 ~ 최대 2.435 까지 나타남을 확인할 수 있으며 PCR모형식의 설계효과는 평균이 2.008이고 최소가 1.615 ~ 최대 2.412 까지 나타나는 것을 확인할 수 있으며 PCR모형식이 설계효과를 잘 추정하고 있다고 볼 수 있다. 또한 $\eta_y = deff_D / deff_{PCR}$ 역시 평균이

1.007이고 표준오차가 0.005로 매우 작게 나타난다. 그리고 급 내 상관 계수가 커짐에 따라 설계효과도 역시 커지는 것을 확인할 수가 있다. 예를 들어, 6번째 층의 $\rho_6 = 0.324$ 로 $deff_{D,6} = 3.821$ 가장 크다. 그림 3.3.2는 전체 설계효과, PCR모형식, 층별 설계효과에 대하여 상자그림을 나타낸 것이다. 표 3.3.2 ~ 표 3.3.7는 설계요소(층화, 가중치, 집락(ICC))들에 대하여 평균, 표준오차, 편향, 상대편향, $rmse$, $rrmse$, Q_1 , Q_2 , Q_3 를 정리한 것이다. 상대편향이 양의 값을 가지면 과대추정을 의미하며 음의 값을 가지면 과소추정을 의미하고 상대편향의 절대 값이 작을수록 정확도가 높다고 할 수 있다. 층화설계요소(δ_{sh})는 상대편향은 $-0.7\% \sim 0.2\%$ 로 비교적 작은 값을 나타내고 $rrmse$ 도 $5.8\% \sim 6.5\%$ 로 작은 값을 나타나고 있다. 표 3.3.3에서 Kish의 가중치설계요소($deff_{w,K}$)는 평균이 1.012로 PCR모형식보다 다소 과대 추정하고 있으며 PCR 가중치설계요소($deff_{wh}$)의 상대편향과 $rrmse$ 는 $0.0 \sim 0.3\%$ 로 층화설계요소보다 작은 값을 보이고 있다. 표 3.3.4에서 Kish의 집락설계요소($deff_{c,Kish}$)는 평균이 1.997로 PCR모형식의 6개의 층의 평균적인 값을 보이고 있으며 PCR 집락설계요소($deff_{ch,Kish}$)의 상대편향은 $-5.4\% \sim 1.1\%$, $rrmse$ 는 $10.8\% \sim 19.9\%$ 로 나타나고 있다. 급 내 상관계수 표 3.3.5에서는 ρ_{Kish} 의 평균이 0.111로 6개의 층의 평균적인 값을 보이며 각 PCR모형($\rho_{h,Kish}$)의 6개의 층에 대한 상대 편향은 $-21.5 \sim 2.6\%$, $rrmse$ 는 $14.6\% \sim 336.7\%$ 로 나타나고 있다. 집락(ICC)설계요소는 급내 상관계수가 커짐에 따라 $rrmse$ 는 작아지는 현상을 확인할 수가 있다. 표 3.3.6 집락설계요소($deff_{ch,AOV}$)는 표 3.3.4 집락설계요소($deff_{ch,Kish}$)와

비교하였을 때 비슷한 경향을 나타내고 전체적으로 $deff_{ch,AOV}$ 의 상대편향과 $rrmse$ 가 비교적 낮게 나타난다. 표 3.3.7 급 내 상관계수($\rho_{h,AOV}$) 역시 표 3.3.5 급 내 상관계수($\rho_{h,Kish}$)와 비교하였을 때 비슷한 경향을 나타내고 전체적으로 $\rho_{h,AOV}$ 가 $\rho_{h,Kish}$ 보다 상대편향과 $rrmse$ 가 비교적 낮게 나타난다.



제 4장 결론

본 논문에서는 복잡표본설계를 위한 설계효과모형과 설계요소들의 정도수준을 평가하였다. 복잡표본설계는 주로 층화, 집락, 불균등확률 등과 같은 설계요소로 구성되어진다. Kish와 PCR모형식은 복잡표본설계의 설계요소를 반영한 함수식을 제공한다. 여기서 Kish모형식은 층화설계요소를 미반영하는 반면 PCR모형식은 층화, 집락, 불균등가중 설계요소를 모두 반영하였다. 층화다단설계에서는 층마다 동일하지 않은 급 내 상관계수(ICC)을 가지고 있다. 이러한 층화다단설계에서 층을 고려하지 않는 Kish 모형식을 사용하고 Kish의 급 내 상관계수를 적용하게 된다면 매우 잘못된 결과를 가져올 수 가 있다. 예를 들어, 표 3.3.5에서 ρ_{Kish} 는 0.111이지만 전체 설계효과는 2.022이다. 하지만 첫 번째 층의 급 내 상관계수는 0.007이며 설계효과는 1.063이고 여섯 번째 층의 급 내 상관계수는 0.310이며 설계효과는 3.791이다. 따라서 향후 첫 번째 층의 표본크기를 늘리거나, 여섯 번째 층의 표본크기를 줄일 수가 있다.

PCR모형식에서 집락설계요소는 가중치, 층화설계요소에 비해 다소 불안정 모습을 보이기도 하지만 층화에 따른 층 규모, 표본할당, 조사변수 분산 구조와 층별 집락효과와 가중치효과로 함수관계를 명백히 표현하고 있는 것을 확인할 수 있었다. 향후 대안적 ρ 적용을 통한 연구와 PCR모형식을 활용한 표본할당 연구를 논의할 수 있다.

참고문헌

- [1] 박인호 (2014). 설계효과모형 적용에 관한 연구. 한국데이터정보과학회지, 25권 3호, 523-531.
- [2] Inho Park (2015). Understanding Complex Design Features via Design Effect Models. *The Korean Journal of Applied Statistics*, 28권 6호, 1217-1225.
- [3] Inho Park (2015). Assessing Complex Sample Designs via Design Effect Decompositions. *Quantitative Bio-Science*, Vol. 34(2), 51-61.
- [4] Sixia Chen and Keith Rust (2017). An Extension of Kish's Formula for Design Effects to Two- and Three-Stage Designs With Stratification. *Journal of Survey Statistics and Methodology*, Vol. 5, 111-130.
- [5] Cindy Feng (2006). On Confidence Intervals for Proportions with Focus on the U.S. National Health and Nutrition Examination Surveys. Simon Fraser University dissertation.
- [6] Ganninger, Matthias (2010). Design effects : model-based versus design-based approach. GESIS-Leibniz Institute for the So

cial Science.

- [7] Hyunshik Lee, Eric Rancourt, and Carl E. Sarndal (1994). Experiments with Variance Estimation from Survey Data with Imputed Values. *Journal of Official Statistics*, Vol. 10, No3, 231–243.
- [8] Lohr, S. L. (2010). *Sampling : design and analysis 2nd edition.*, Brooks/Cole, Boston.
- [9] Valliant, R., Dever J. A. and Kreuter, F.(2013). Practical Tools for Designing and Weighting Survey Samples, Springer New York Heidelberg Dordrecht London.
- [10] Eugenio J. Gonzalez, Pierro Foy. (1997). Estimation of Sampling Variability, Design Effects, and Effective Sample Sizes.
- [11] Graham Kalton, J.Michael Brick, Thanh Lê. (2005). Estimating components of design effect for use in sample design.
- [12] Kish, L. (1987). Weighting in Deft2. *The Survey Statistician*, June, 1987.
- [13] Cornfield, J. (1951). Modern methods in the sampling of human populations, *American Journal of Public Health*,, 41, 654–661.

부 록

A.1 R코드

```
library(sampling)

# h층에 PSU(집락) 수 ( $N_h$ )
Nh=round(rnorm(6,100,5))

# h층에 i 번째 집락규모 ( $M_{hi}$ )
for (i in 1:length(Nh)){
  result <-round(rgamma(Nh[i],2,scale=300))
  assign(paste("H",i,sep=""), subset(result,result>0))
}

#  $M$ 
sum(H1,H2,H3,H4,H5,H6)

# 표본틀 정보오류 반영 ( $M^*$ )
b=1
d=2
g=0.5
alpha=(b*Mhi_F) ^ 2/(d ^ 2*Mhi_F ^ (2*g))
beta=d ^ 2*Mhi_F ^ (2*g)/(b*Mhi_F)

Mhi_A=vector()
for (i in 1:length(Mhi_F)){
  Mhi_A[i]=round(rgamma(1,shape = alpha[i],scale = beta[i]))
}

#  $V(y_{hij}) = \nu_h^2$ 
nuh=round(rnorm(6,20,2),2)
```

```

#  $y_{hij} = A_{hi} + \epsilon_{hij}$ 
#  $A_{hi} \sim iid N(\mu_{Ah}, \sigma_{Ah}^2)$ 
#  $\mu_{Ah} \sim iid N(235, 5^2)$ 
muAh=round(rnorm(6,235,5),2)

#  $\rho_h$  급 내 상관계수(Intraclass correlation)
rhoh=c(0.01, 0.02, 0.05, 0.10, 0.20, 0.50)

#  $\sigma_{Ah}^2 = \rho_h \nu_h^2$ 
sigmaAh=sqrt(rhoh)*nuh

for (h in 1:length(Nh)){
  for (i in 1:Nh_new[h]){
    result <-rnorm(1,muAh[h],sigmaAh[h])
    assign(paste("A","H",h,"I",i,sep=""), subset(result,result>0))
  }

  A_hi=vector()
  for (h in 1:length(Nh)){
    for (i in 1:Nh_new[h]){
      A_hi=c(A_hi,eval(parse(text=paste("A","H",h,"I",i,sep=""))))
    }

    #  $\epsilon_{hij} \sim iid N(0, \sigma_{ehi}^2)$ ,
    #  $\sigma_{ehi}^2 \sim iid Gamma$  s.t.  $E(\sigma_{ehi}^2) = (1 - \rho_h) \nu_h^2$   $V(\sigma_{ehi}^2) = \nu_h^2 / 10$ 
    for (h in 1:length(Nh)){
      for (i in 1:Nh_new[h]){
        for (j in 1:H[[h]][i]){
          sigmaehi2<-rgamma(1,10*(1-rhoh) ^ 2*nuh[h] ^ 2,scale=1/(10*(1
            -rhoh)))
          result <-rnorm(1,0,sqrt(sigmaehi2))
          assign(paste("e","H",h,"I",i,"J",j,sep=""), subset(result,is.numeric(res
            ult))==TRUE))
        }
      }
    }
  }

```

```

#  $y_{hij} = A_{hi} + \epsilon_{hij}$ 
for (h in 1:length(Nh)){
  for (i in 1:Nh_new[h]){
    for (j in 1:H[[h]][i]){
      assign(paste("x","H",h,"I",i,"J",j,sep=""),
        eval(parse(text=paste("A","H",h,"I",i,sep="")))+eval(parse(text=paste("e","H",h,"I",i,"J",j,sep=""))))
    }
  }
}

```

```

#  $y_{hij}$ 
x_hij=vector()
for (h in 1:length(Nh)){
  for (i in 1:Nh_new[h]){
    for (j in 1:H[[h]][i]){
      x_hij=c(x_hij,eval(parse(text=paste("x","H",h,"I",i,"J",j,sep=""))))
    }
  }
}

```

```

# strata number
strata=vector()
for(i in 1:length(Nh)){
  strata=c(strata,rep(i,sum(H[[i]])))
}

```

```

# PSU number
PSU=vector()
for(i in 1:sum(Nh_new)){
  PSU=c(PSU,rep(i,pop_new$Mhi_A[i]))
}

```

```

# weight (  $n_h=50$ ,  $m_0=10$  )
w=c(sum(fH1)*H1/(50*fH1*10),
sum(fH2)*H2/(50*fH2*10),
sum(fH3)*H3/(50*fH3*10),
sum(fH4)*H4/(50*fH4*10),
sum(fH5)*H5/(50*fH5*10),
sum(fH6)*H6/(50*fH6*10))

w_hij=vector()
for(i in 1:sum(Nh_new)){
w_hij=c(w_hij,rep(w[i],pop_new$Mhi_A[i]))
}

# 유한모집단 생성
Feng=cbind(strata,PSU,x_hij,w_hij)
Feng=data.frame(Feng)

# Design effect component
# PCR 모형식
A=vector()           #  $A$ 
a=vector()           #  $a$ 
sigma__yh2=vector()  #  $\sigma_{yh}^2$ 
sigma__y2=vector()   #  $\sigma_y^2$ 
delta_Sh=vector()    #  $\delta_{sh}$ 

CV_wh2=vector()      #  $cv_{wh}^2$ 
delta_Wh=vector()    #  $def_{wh}$ 

m_h_Star=vector()    #  $m_h^*$ 
rho_h=vector()       #  $\rho_h$ 
delta_Ch=vector()    #  $def_{ch}$ 

eta_y=vector()       #  $\eta_y$ 

```

```

bary__h=vector()      #  $\bar{y}_h$ 
barY=vector()         #  $\bar{y}$ 

# Kish모 형식
deff_w_k=vector()     #  $deff_{w,k}$ 
deff_c_k=vector()     #  $deff_{c,k}$ 
rho_k=vector()        #  $\rho_k$ 

# R=1,000번 반복 시행
for (i in 1:1000){

# 층, PSU별로 합
Feng_PSU=aggregate(Feng_pop$x_hij, list(strata=Feng_pop$strata,
PSU = Feng_pop$PSU), sum)

#  $n_h = 50$ 개의 집락 추출
psu_sample=vector()
for (i in 1:6) {
Feng_PSU1=Feng_PSU[Feng_PSU$strata==i,]
psu_s=sample(Feng_PSU1$PSU,replace=FALSE,prob=c(Feng_PSU1
$x),50)
psu_sample=c(psu_sample,psu_s)}

psu_sample1=data.frame()
for (i in 1:300){
bc=Feng_pop[Feng_pop$PSU==psu_sample[i],]
psu_sample1=rbind(psu_sample1,bc)}

obs_new=seq(1,dim(psu_sample1)[1])
psu_sample2=data.frame(psu_sample1,obs_new)

```

```

#  $m_0 = 10$ 개의 최종단위 추출
st.sample.index <- strata(data = psu_sample2,
stratanames = "PSU",          # 층화로 사용되는 변수명,
size = c(rep(10,300)) ,      # 300개의 그룹에서 각각 10개
method = "srswor")           # simple random sampling
                                # without replacement

Feng_sample=data.frame()
for (i in 1:3000){
bc=psu_sample2[psu_sample2$obs_new==st.sample.index $ID_unit[i],]
Feng_sample=rbind(Feng_sample,bc)}
#  $A_h = M_h/M, A_h^2$ 
M_h=aggregate(Feng_sample$w_hij, list(strata=Feng_sample$strata), sum)
A_h=M_h$x/sum(M_h$x)
A_h2=(A_h)^2

#  $a_h = 500/3000$ 
a_h=rep(500/3000,6)

#  $\bar{y}_h$ 
a2=data.frame(st=Feng_sample$strata,xw=Feng_sample$w_hij*Feng
_sample$x_hij)
a3=aggregate(a2$xw, list(strata=a2$st), sum)
bary_h=a3$x/M_h$x

#  $\sigma_{yh}^2$ 
a4=Feng_sample$w_hij*(Feng_sample$x_hij-rep(bary_h,each=45
0))^2
a5=data.frame(st=Feng_sample$strata,a4)
a6=aggregate(a5$a4, list(strata=a5$st), sum)
sigma_yh2=a6$x/M_h$x

#  $\bar{y}$ 
bary=sum(a2$xw)/sum(M_h$x)

```

```

#  $\sigma_y^2$ 
a7=Feng_sample$w_hij*(Feng_sample$x_hij-rep(bary,2700))^2
sigma_y2=sum(a7)/sum(M_h$x)

#  $\delta_{sh}$ 
delta_sh=(A_h2/a_h)*(sigma_yh2/sigma_y2)

#  $\bar{w}_h$ 
barw_h=M_h$x/450

#  $s_{wh}^2$ 
a8=data.frame(st=Feng_sample$strata,a9=(Feng_sample$w_hij-rep
(barw_h,each=450))^2)
a10=aggregate(a8$a9, list(strata=a8$st), sum)
s_wh2=a10$x/450

#  $cv_{wh}^2$ 
cv_wh2=s_wh2/(barw_h)^2

#  $def_{wh}$ 
delta_wh=1+cv_wh2

#  $m_h^*$ 
a11=aggregate(Feng_sample$w_hij,list(strata=Feng_sample$strata,p
su=Feng_sample$PSU),sum)
a12=data.frame(a11$strata,a11$x^2)
a13=aggregate(a12$a11.x.2,list(strata=a12$a11.strata),sum)
a14=data.frame(st=Feng_sample$strata,w2=Feng_sample$w_hij^2)
a15=aggregate(a14$w2,list(st=a14$st),sum)
m_h_star=a13$x/a15$x

```

```

#  $\bar{y}_{hi}$ 
a16=aggregate(Feng_sample$w_hij,list(psu=Feng_sample$PSU),sum)
a17=aggregate(Feng_sample$w_hij*Feng_sample$x_hij,list(st=Feng_sample$strata,psu=Feng_sample$PSU),sum)
bary_hi=a17$x/a16$x

#  $e_{hi}$ 
a18=data.frame(st=Feng_sample$st,psu=Feng_sample$PSU,t=Feng_sample$w_hij*(Feng_sample$x_hij-bary))
a19=aggregate(a18$t,list(st=a18$st,pus=a18$psu),sum)
e_hi=a19$x/sum(Feng_sample$w_hij)

#  $\bar{e}_h$ 
a21=data.frame(st=a19$st,e_hi)
a20=aggregate(a21$e_hi,list(st=a19$st),sum)
bare_h=a20$x/50

#  $V_p(\bar{y})$ 
a21=data.frame(st=a19$st,t=(e_hi-rep(bare_h,each=50))^2)
a22=aggregate(a21$t,list(st=a21$st),sum)
sigma_bary=sqrt(sum(a22$x*50/49))

#  $e_{hi} - \frac{\sum}{M_h} x$ 
a23=data.frame(st=Feng_sample$st,psu=Feng_sample$PSU,t=Feng_sample$w_hij*(Feng_sample$x_hij-rep(bary_h,each=500)))
a24=aggregate(a23$t,list(st=a23$st,pus=a23$psu),sum)
e_hi_h=a24$x/rep(M_h$x,each=50)

#  $\bar{e}_h - \frac{\sum}{M_h} x$ 
a25=data.frame(st=a24$st,e_hi_h)
a26=aggregate(a25$e_hi_h,list(st=a25$st),sum)
bare_h_h=a26$x/50

```

```

#  $V_p(\bar{y}_h)$ 
a27=data.frame(st=a24$st,t=(e_hi_h-rep(bare_h_h,each=50))^2)
a28=aggregate(a27$st,list(st=a27$st),sum)
sigma_bary_h=sqrt(a28$x*50/49)

#  $deff_h$ 
deff_h=sigma_bary_h^2/sigma_yh2*499

#  $deff$ 
deff=sigma_bary^2/sigma_y2*2999;deff

#  $deff_{ch}$ 
delta_ch=deff_h/delta_wh

#  $\hat{deff}$ 
hat_deff=sum(delta_sh*delta_ch*delta_wh);hat_deff

#  $\eta_y$ 
eta_yw=deff/hat_deff

#  $\rho_y$ 
rho_yh=(m_h_star-1)^-1*(delta_ch-1)

A=c(A,A_h)
a=c(a,a_h)
sigma__yh2=c(sigma__yh2,sigma_yh2)
sigma__y2=c(sigma__y2,sigma_y2)
delta_Sh=c(delta_Sh,delta_sh)
CV_wh2=c(CV_wh2,cv_wh2)
delta_Wh=c(delta_Wh,delta_wh)
m_h_Star=c(m_h_Star,m_h_star)
rho_h=c(rho_h,rho_yh)
delta_Ch=c(delta_Ch,delta_ch)

```

```

eta_y=c(eta_y,eta_yw)
bary__h=c(bary__h,bary_h)
barY=c(barY,bary)

#  $deff_{w,k}$ 
cv_w2=1+(sqrt(sum((Feng_sample$w_hij-mean(Feng_sample$w_hi
j))^2)/3000)/mean(Feng_sample$w_hij))^2

#  $deff_{c,k}$ 
deff_c=deff/cv_w2

#  $\rho_k$ 
rho_kish=(deff_c-1)/9

deff_w_k=c(deff_w_k,cv_w2)
deff_c_k=c(deff_c_k,deff_c)
rho_k=c(rho_k,rho_kish)
}

```

