



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

# 총화다단추출설계의 효율성 평가



부 경 대 학 교 대 학 원

통 계 학 과

박 성 진

이 학 석 사 학 위 논 문

# 총화다단추출설계의 효율성 평가

지도교수 박 인 호

이 논문을 석사 학위논문으로 제출함.

2018년 2월

부 경 대 학 교 대 학 원

통 계 학 과

박 성 진

박성진의 이학석사 학위논문을 인준함.

2018년 2월 23일



위원장	이학박사	장대흥 (인)
위원	이학박사	박인호 (인)
위원	이학박사	노맹석 (인)

# 목 차

표 차례	ii
그림 차례	v
제 1장 서론 .....	1
제 2장 복잡표본설계 .....	3
2.1 층화이단표본설계 .....	3
2.2 집락효과 .....	6
2.3 층화효과 .....	9
제 3장 가구조사 사례적용 및 평가 .....	12
3.1 복잡가구표본설계 사례 .....	12
3.1.1 농어업인복지실태조사 .....	12
3.1.2 경제활동인구조사 .....	13
3.1.2 우리나라 가구표본의 설계를 위한 추출단위 선택 .....	13
3.2 집락 효율성 평가 .....	14
3.2.1 이단추출 .....	14
3.3 층화 효율성 평가 .....	22
3.4 종합적 효율성 평가 .....	53
제 4장 결론 .....	55
참고문헌 .....	56
부록 A.1 R코드 .....	58
A.2 SAS코드 .....	59

# 표 차례

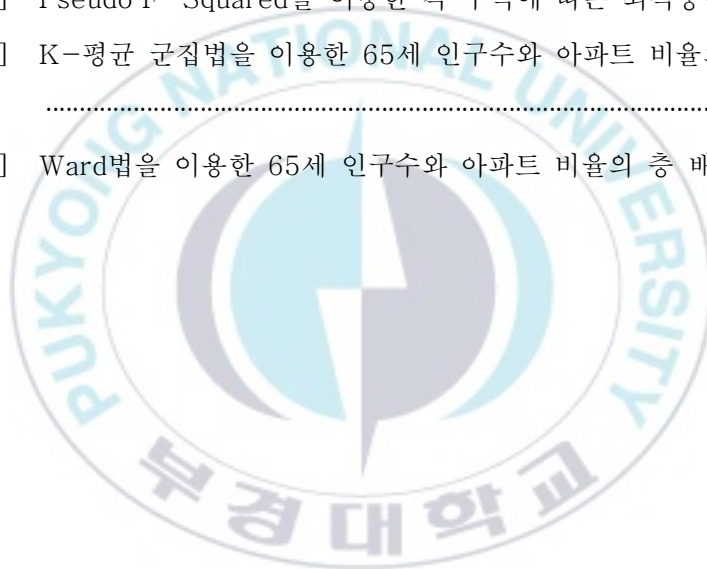
[표 3.1] 분석변수의 모비율과 모총합 .....	15
[표 3.2] 집계구 정보의 기술통계량 .....	16
[표 3.3] 집계구정보 표본할당에 따른 상대분산과 설계효과 .....	17
[표 3.4] 단순임의추출/단순임의추출의 이단표본설계에서의 분산분해요소와 동질성계수 .....	19
[표 3.5] 비례확률추출/단순임의추출의 이단표본설계에서의 분산분해요소와 동질성계수 .....	20
[표 3.6] 권역별 표본집계구 설계방식 .....	22
[표 3.7] 1인가구수 변수를 권역별로 층을 구분한 비례확률추출/단순임의추출 의 이단표본설계에서의 분산분해요소와 동질성계수 .....	24
[표 3.8] 자가가구수 변수를 권역별로 층을 구분한 비례확률추출/단순임의추출 의 이단표본설계에서의 분산분해요소와 동질성계수 .....	24
[표 3.9] 대졸학력인구수 변수를 권역별로 층을 구분한 비례확률추출/단순임의 추출의 이단표본설계에서의 분산분해요소와 동질성계수 .....	25
[표 3.10] 65세 이상 인구수 변수를 권역별로 층을 구분한 비례확률추출/단순 임의추출의 이단표본설계에서의 분산분해요소와 동질성계수 .....	25
[표 3.11] 10대인구수 변수를 권역별로 층을 구분한 비례확률추출/단순임의추 출의 이단표본설계에서의 분산분해요소와 동질성계수 .....	26
[표 3.12] 권역당 동읍면별 표본집계구 설계방식 .....	27
[표 3.13] 1인가구수 변수를 권역별 동 읍면으로 층을 구분한 비례확률추출/ 단순임의추출의 이단표본설계에서의 분산분해요소와 동질성계수 ..	28
[표 3.14] 자가가구수 변수를 권역별 동 읍면으로 층을 구분한 비례확률추출/ 단순임의추출의 이단표본설계에서의 분산분해요소와 동질성계수 ..	29

[표 3.15] 대졸학력인구수 변수를 권역별 동 읍면으로 층을 구분한 비례확률 추출/단순임의추출의 이단표본설계에서의 분산분해요소와 동질성계 수 .....	30
[표 3.16] 65세 이상 인구수 변수를 권역별 동 읍면으로 층을 구분한 비례확 률추출/단순임의추출의 이단표본설계에서의 분산분해요소와 동질성 계수 .....	31
[표 3.17] 10대인구수 변수를 권역별 동 읍면으로 층을 구분한 비례확률추출/ 단순임의추출의 이단표본설계에서의 분산분해요소와 동질성계수 ..	32
[표 3.18] 시도당 동읍면별 표본집계구 설계방식 .....	33
[표 3.19] 1인가구수 변수를 시도별 동 읍면으로 층을 구분한 비례확률추출/ 단순임의추출의 이단표본설계에서의 분산분해요소와 동질성계수 ..	35
[표 3.20] 자가가구수 변수를 시도별 동 읍면으로 층을 구분한 비례확률추출/ 단순임의추출의 이단표본설계에서의 분산분해요소와 동질성계수 ..	36
[표 3.21] 대졸학력인구수 변수를 시도별 동 읍면으로 층을 구분한 비례확률 추출/단순임의추출의 이단표본설계에서의 분산분해요소와 동질성 계수 .....	37
[표 3.22] 65세 이상 인구수 변수를 시도별 동 읍면으로 층을 구분한 비례확 률추출/단순임의추출의 이단표본설계에서의 분산분해요소와 동질성 계수 .....	38
[표 3.23] 10대 인구수 변수를 시도별 동 읍면으로 층을 구분한 비례확률추출 /단순임의추출의 이단표본설계에서의 분산분해요소와 동질성 계수 .....	39
[표 3.24] 1인가구수 변수를 K-평균 군집법으로 층을 구분한 비례확률추출/ 단순임의추출의 이단표본설계에서의 분산분해요소와 동질성계수 ..	48

[표 3.25]	자가가구수 변수를 K-평균 군집법으로 층을 구분한 비례확률추출/ 단순임의추출의 이단표본설계에서의 분산분해요소와 동질성계수 ..	48
[표 3.26]	대졸학력인구수 변수를 K-평균 군집법으로 층을 구분한 비례확률 추출/단순임의추출의 이단표본설계에서의 분산분해요소와 동질성 계수 .....	49
[표 3.27]	65세 이상 인구수 변수를 K-평균 군집법으로 층을 구분한 비례확 률추출/단순임의추출의 이단표본설계에서의 분산분해요소와 동질성 계수 .....	49
[표 3.28]	10대 인구수 변수를 K-평균 군집법으로 층을 구분한 비례확률추출 /단순임의추출의 이단표본설계에서의 분산분해요소와 동질성 계수 .....	50
[표 3.29]	1인가구수 변수를 Ward법으로 층을 구분한 비례확률추출/단순임의 추출의 이단표본설계에서의 분산분해요소와 동질성계수 .....	50
[표 3.30]	자가가구수 변수를 Ward법으로 층을 구분한 비례확률추출/단순임 의추출의 이단표본설계에서의 분산분해요소와 동질성계수 .....	51
[표 3.31]	대졸학력인구수 변수를 Ward법으로 층을 구분한 비례확률추출/단 순임의추출의 이단표본설계에서의 분산분해요소와 동질성계수 .....	51
[표 3.32]	65세 이상 인구수 변수를 Ward법으로 층을 구분한 비례확률추출/ 단순임의추출의 이단표본설계에서의 분산분해요소와 동질성계수 ..	52
[표 3.33]	10대 인구수 변수를 Ward법으로 층을 구분한 비례확률추출/단순임 의추출의 이단표본설계에서의 분산분해요소와 동질성계수 .....	52

# 그림 차례

[그림 3.1]	집계구 정보의 가구수의 집락분포 .....	18
[그림 3.2]	단순임의/단순임의 & 비례확률/단순임의 분산구성요소 비교 .....	22
[그림 3.3]	각 지역별 이상점 확인을 위한 산점도 .....	41
[그림 3.4]	CCC를 이용한 각 구역에 따른 최적층수결정 .....	43
[그림 3.5]	R-squared를 이용한 각 구역에 따른 최적층수결정 .....	44
[그림 3.6]	Pseudo T-Squared를 이용한 각 구역에 따른 최적층수결정 .....	45
[그림 3.7]	Pseudo F-Squared를 이용한 각 구역에 따른 최적층수결정 .....	45
[그림 3.8]	K-평균 군집법을 이용한 65세 인구수와 아파트 비율의 층 배분 .....	46
[그림 3.9]	Ward법을 이용한 65세 인구수와 아파트 비율의 층 배분 .....	46



# Evaluating Stratified Multistage Sampling

Seong Jin Park

Department of Statistics, The Graduate School,  
Pukyong National University

## Abstract

A stratified multistage sample design is often considered for survey researches. In Korea, many household surveys are conducted by adopting stratified two-stage sample design utilizing so-called “enumeration district (ED for short)” as the primary sampling unit. However, some other alternative area units such as 'Output Area', 'apartment complexes', 'dong-eup-myeon' and so on may be suitable candidates for the primary sampling unit due to many reasons. In this thesis, we study methods to evaluate the efficiency of stratified two-stage sample design and apply them when the “Output Areas” of the Statistics Korea are used instead as area units in sample selection at the first or later stages through a variance decomposition. Use of Output Areas as clusters leads to attain intra-cluster homogeneity within them to some extent but their clustering effects due to both within-unit homogeneity and size-variation among the Output Areas may be reduced by applying the probability proportionate to size sampling.

However, we found out that stratification tends to enhance the precision of the total estimator by excluding between stratum-variations. Our study result shows that the Output Areas may be a plausible choice for designing the stratified multistage sampling in Korean household surveys.

keywords : measure of homogeneity; stratified Multistage sampling variance decomposition



# 제 1장 서론

조사연구(survey research)를 위한 표본설계에서는 흔히 집락추출(cluster sampling)과 층화(stratification)의 설계요소들을 혼합하여 사용한 복합표본추출(complex sampling)을 고려한다. 지역단위(area unit)를 집락으로 하는 집락추출을 사용하면 조사 경비와 시간을 줄일 수 있지만 집락 내 개체간 동질성으로 인해 표본추정의 정도수준은 낮아지게 된다. 반면 층화추출은 모집단을 서로 동질적인 개체들로 이루어진 층으로 구분하여 층별로 표본추출을 함으로 표본추정의 정도수준을 향상시킬 수 있다.

우리나라 가구조사의 대부분은 통계청의 인구주택총조사에서 작성한 조사구(enumeration district, ED)를 (지역단위의) 집락으로 하는 표본설계에 기반한다. 더불어 모집단을 특광역시·도와 더불어 아파트와 일반가구로 구분하는 조사구 특성으로 구분하는 층화의 기준으로 한다. 이러한 층화 및 집락추출의 선택은 우리나라 가구조사의 전형적 형태로 사용되고 있으며 대안적 선택을 고려한 표본조사는 그리 많지 않다. 하지만 통계청의 조사구 명부를 표본추출틀로 사용하기 위해서는 조사이전 공식통계의 자격을 획득해야 하므로 이러한 절차를 거치지 않거나 못하는 경우, 혹은 조사특성상 필요에 의해서 조사구는 물론 시도 및 주택유형적 고려가 아닌 대안적 선택을 해야 하는 경우도 발생할 수 있다. 예로, 조사비용의 제약이 있는 가구조사의 표본설계에서는 조사지역을 좀더 축소하는 소수의 지역적 선택이 필요할 수도 있다. 이러한 경우, 조사구 대신 동읍면을 먼저 집락단위로 선택하는 옵션을 고려할 수 있을

것이다. 또한 통계청에서 국가포털에서 무료로 제공하는 집계구는 조사구보다 다소 큰 규모의 지역적 가구집단으로 정의되는데 다양한 SGIS 정보를 제공받을 수 있으므로 조사구의 대안적 집락단위가 된다.

본 논문에서는 가구조사를 위한 층화다단추출의 설계요소들로 인한 표본추정량의 정도수준의 기여도를 분산분해의 관점에서 살펴보고 개별 설계요소의 효율성을 평가하는 방법에 대해 살펴본다. 2장에서는 층화다단 설계에서의 총합 추정량의 분산과 분산분해에 대한 이론을 정리한다. 3장에서는 층화다단설계에 의해 수행된 우리나라 가구조사 사례를 살펴본다. 더불어 조사구 대신 집계구를 대안적 집락으로 사용하는 층화다단추출을 고려하여 표본추정량 분산 및 분산 구성요소를 살펴보고 층화 및 집락추출에 따른 효율성에 대해 살펴본다. 4장에서는 우리나라 가구조사를 위한 복잡표본설계에서 기존 표본설계방식이 아닌 다양한 대안의 가능성에 대해 논의한다.

## 제 2장 복잡표본설계

### 2.1 층화이단표본설계

본 연구의 관심 대상인 모집단은  $H$ 개의 (표본설계) 층(design strata)으로 나뉘고,  $h$ 번째 층은  $N_h$ 개의 일차추출단위(primary sampling unit, PSU) 혹은 집락(cluster)으로 구성되어 총  $N = \sum_{h=1}^H N_h$ 개의 일차추출단위들로 이루어진다. 집락 ( $hi$ )는  $M_{hi}$ 개의 최종추출단위(ultimate sampling unit, USU)로 이루어지고,  $h$ 번째 층과 모집단은 각각  $M_h = \sum_{i=1}^{N_h} M_{hi}$ 와  $M = \sum_{h=1}^H M_h$ 개의 최종추출단위들로 이루어진다. 층별로 각각 독립적으로 일단계 추출에서  $n_h \geq 2$ 개의 집락을 불균등확률추출(unequal probability sampling) 혹은 단순확률추출(simple random sampling)로 선택한다. 실무(practice)에서는 비복원(without-replacement) 추출을 적용하지만 분산추정의 단순화를 위해 주로 복원(with-replacement) 추출인 것으로 가정된다. 이러한 가정에 따른 근사(approximation)는 분산의 과대추정(overestimation)을 초래하지만 일단계 추출의 표본추출률(sampling fraction)  $f_h = n_h/N_h$ 이 무시할 정도로 작다면 그 편향(bias)도 매우 작게 된다. 또한 선택된 표본집락 ( $hi$ )로부터 궁극적으로  $m_{hi}$ 개의 최종추출단위를 선택한다. 표본 내 총 최종추출단위는  $m = \sum_{h=1}^H \sum_{i=1}^{n_h} m_{hi}$ 가 된다. 집락 내 표본추출은 둘 이상의 단계로 구성할 수 있지만 본 연구에서는 논의의 단순화를 위해 일단계, 즉 집락 내 최종단위를 뽑는 이단추출을 고려한다.

모집단 내  $h$ 번째 층의 집락  $i$ 내  $k$ 번째 개체(즉, 최종추출단위)는 조사특성  $y_{hik}$ 를 갖는다면, 모총합(population total)은 다음 식 (2.1)과 같이 정의할 수 있다.

$$Y = \sum_{h=1}^H \sum_{i=1}^{N_h} \sum_{k=1}^{M_{hi}} y_{hik} \quad (2.1)$$

집락 ( $hi$ )와 층  $h$ 의  $y$ -총합을 각각  $Y_{hi} = \sum_{k=1}^{M_{hi}} y_{hik}$ 와  $Y_h = \sum_{i=1}^{N_h} Y_{hi}$ 로 정의하면, 식 (2.1)은 집락 총합 혹은 층 총합의 합으로 다음 식 (2.2)와 같이 표현될 수도 있다.

$$Y = \sum_{h=1}^H \sum_{i=1}^{N_h} Y_{hi} = \sum_{h=1}^H Y_h \quad (2.2)$$

본 연구에서 고려하는 층화이단추출의 표본설계는 다음을 가정한다.

- (D1) 층별로  $n_h$ 개 집락을 불균등확률  $p_{hi}$ 에 따라 복원추출한다( $\sum_{i=1}^{N_h} p_{hi} = 1$ ).
- (D2) 표본집락 내  $m_{hi}$ 개 최종추출단위를 균등확률에 따라 비복원추출한다.
- (D3) 집락 내 부표본추출은 일단계의 표본추출 결과에 영향을 받지 않으며<sup>1)</sup>, 서로 다른 집락 내 부표본추출과는 독립적으로 수행한다.

위의 세 가정 D1-D3 하에서 식 (2.1) 혹은 (2.2)의  $y$ -모총합에 대한 추출

---

1) 이러한 성격은 불변성(invariance)라고 칭한다 (Sarndal et al., 1992, 134).

확률을 고려한 표본설계에 기반한 추정량(design-based estimator)은 다음과 같이 정의된다.

$$\hat{Y} = \sum_{h=1}^H \hat{Y}_h = \sum_{h=1}^H \frac{1}{n_h} \sum_{i=1}^{n_h} \frac{\hat{Y}_{hi}}{p_{hi}} \quad (2.3)$$

여기서  $\hat{Y}_h = n_h^{-1} \sum_{i=1}^{n_h} (\hat{Y}_{hi}/p_{hi})$ 는 층 총합  $Y_h$ 의 표본추정량이고  $\hat{Y}_{hi}$ 는 집락 총합  $Y_{hi}$ 의 표본추정량으로 다음 식 (2.4)와 같이 정의된다.

$$\hat{Y}_{hi} = M_{hi} \bar{y}_{hi} = \frac{M_{hi}}{m_{hi}} \sum_{k=1}^{m_{hi}} y_{hik} \quad (2.4)$$

단,  $\bar{y}_{hi} = m_{hi}^{-1} \sum_{k=1}^{m_{hi}} y_{hik}$ 는 집락 내 단순평균을 나타낸다.

참고로 일차추출단위 내 부표본추출(subsampling)이 이단계 이상으로 구성될 때는 흔히 집락 총합  $Y_{hi}$ 에 대한 불편추정(unbiased estimation)이 가능한 것으로 가정하면 다단계추출로의 확장이 가능하다. 식 (2.3)의 모집단 총합 추정량  $\hat{Y}$ 과 층 총합 추정량  $\hat{Y}_h$ 은 각각 불편추정량이며 이들의 설계기반 분산은 각각 다음 식 (2.5), (2.6)과 같이 주어진다.

$$V(\hat{Y}) = \sum_{h=1}^H V(\hat{Y}_h) \quad (2.5)$$

$$V(\hat{Y}_h) = \frac{1}{n_h} \sum_{i=1}^{N_h} p_{hi} \left( \frac{Y_{hi}}{p_{hi}} - Y_h \right)^2 + \sum_{i=1}^{N_h} \frac{M_{hi}^2}{n_h p_{hi} m_{hi}} \left( 1 - \frac{m_{hi}}{M_{hi}} \right) S_{hi}^2 \quad (2.6)$$

식 (2.6)의 유도는 Sarndal et al. (1992, 결과 4.4.1)를 참고할 수 있으며  $S_{hi}^2$ 는 집락 내 단위 분산으로 다음과 같이 정의된다.

$$S_{hi}^2 = \frac{1}{M_{hi} - 1} \sum_{k=1}^{M_{hi}} (y_{hik} - \bar{Y}_{hi})^2.$$

여기서  $\bar{Y}_{hi} = M_{hi}^{-1} \sum_{k=1}^{M_{hi}} y_{hik}$ 는 집락 내  $y$ -평균이다

## 2.2 집락효과

가구, 학생, 환자 등의 단위는 지역단위, 학교, 병원 등과 같은 해당 단위 묶음인 집락을 표본추출단위로 사용하면 조사 경비, 시간 및 관리노력을 절감할 수 있는 경제적 효과를 얻는 반면, 집락 내 개체간 동질성으로 인해 표본추정의 정도수준이 낮아질 수 있는 단점도 갖게 된다. 본 절에서는 층화이단추출 하에서 표본설계가 갖는 집락효과의 평가 측면에서 총합추정치의 분산을 분해하고 재구성하였다. 특히 층화효과를 제외한 집락효과를 살펴보기 위해 모총합 추정량 보다는 층별 총합 추정에 대해 살펴본다.

논의를 단순화하기 위해 다음과 같은 가정을 먼저 고려한다.

(C1) 모든 집락에 대해서 동일한 수의 표본개체, 즉  $m_{hi} \equiv \bar{m}_h$ 을 추출한다.

(C2) 집락 내 표본추출률  $f_{hi} = m_{hi}/M_{hi}$ 은 무시해도 될 정도로 매우 작다.

위의 두 가정 C1-C2 하에서 식 (2.6)의 오른쪽 두 번째 항의 유한모집단 수정계수는 생략되어 다음 식 (2.7)과 같이 표현된다.

$$V(\hat{Y}_h) = \frac{S_{Bh}^2}{n_h} + \frac{S_{Wh}^2}{n_h \bar{m}_h} \quad (2.7)$$

여기서  $S_{yBh}^2 = \sum_{i=1}^{N_h} p_{hi} (Y_{hi}/p_{hi} - Y_h)^2$  이고  $S_{yWh}^2 = \sum_{i=1}^{N_h} M_{hi}^2 S_{hyi}^2 / p_{hi}$  이다.  $V(\hat{Y}_h)$ 는 일차추출단위의 분산요소와 이차추출단위의 분산요소들의 합으로 각 분리요소로 집락 간 분산을 표본집락의 수로 나눈 형태와 집락 내 분산을 표본집락의 수와 표본개체의 수의 곱의 형태에서 나눈 형태로 두 구조에 대한 합으로 표현된다. 각 요소들을 통해 각 단계에 따른 추출단위별 분산크기에 영향을 주는 정도수준을 확인할 수 있다. 마찬가지로 두 가정 C1-C2 하에서 (2.7)의 식은 집락 내 개체 간 동질성정도를 표현하는 다음의 형태로도 유도될 수 있다.

$$V(\hat{Y}_h) \equiv M_h^2 \frac{S_h^2}{n_h \bar{m}_h} \kappa_h [1 + \delta_h (\bar{m}_h - 1)] \quad (2.8)$$

여기서  $S_h^2 = (M_h - 1)^{-1} \sum_{k=1}^{M_h} (y_k - \bar{Y}_h)^2$ 는 층내 개체분산을 나타내고,  $k_h$ 는 집락크기의 불균등성을 나타내는 측도로 다음 식 (2.9)와 같이 정의된다.

$$\kappa_h = (S_{Bh}^2 + S_{Wh}^2) / (M_h^2 S_h^2) \quad (2.9)$$

Valliant et al. (2015)는 집락크기가 모두 동일하며 집락수와 크기가 모두 클 때 크기비례로 집락추출을 한다면  $\kappa_h$ 는 근사적으로 1의 값을 가짐을 보여주었다. 또한  $\delta_h$ 는 동질성계수로 다음 식 (2.10)과 같이 정의된다.

$$\delta_h = \frac{S_{Bh}^2}{S_{Bh}^2 + S_{Wh}^2} \quad (2.10)$$

개체 특성치의 집락 간 분산보다 집락 내 분산이 큰 값을 가지면 상대적으로 작은 값을 가지게 되며 집락 간 분산이 집락 내 분산보다 큰 값을 가지면 동질성 계수 또한 큰 값을 갖는데  $0 \leq \delta_h \leq 1$ 의 범위를 갖는다. 동질성계수는 집락 크기가 불균등할 때 집락 간 분산이 달라지기 때문에  $\kappa_h$ 에 의해 영향을 받는다. 크기비례 확률추출에서 복원추출을 이용하여 집락을 선택할 때 집락  $i$ 의 가중치가  $N/(mN_i)$ 이고 각 집락에서 동일한 확률로 표본이 선택되면 자체가중 (self-weighting)으로도 표현 가능하다.

집락은 동질적인 개체로 집락 내부는 이질적으로 개체구성이 이루어지면 동질성계수가 작은 값을 가져 추정에 있어서 좋은 효과를 나타낸다. 예를 들면, 집락 간 가구의 소득이 비슷하고 집락 내 가구원의 소득의 차이가 많이 난다면 동질성계수는 0에 가까워 집락의 효율을 볼 수 있지만 집락 간 가구의 소득이 차이가 심하고 집락 내 가구원의 소득이 동질적이라면 집락을 뽑았을 때 집락의 효율이 떨어진다.

분산분해요소의 평가를 위해 설계효과(design effect,  $deff$ )를 이용한다. 설계

효과를 정의하면 단순임의추출 대비 하에 추정량이 어느정도의 효율성을 가지는 지 나타내는 지표이며, 동질성계수를 반영하고 있기에 집락효과의 평가 지표로 사용할 수 있다. 설계효과는 집락 내 개체의 표본크기  $m$ 의 선택에 따른 표본설계의 효율성을 나타낸다 (박인호, 2016).

설계효과는 층화와 집락, 불균등 확률의 선택 및 무응답, 가중치 조정과 같은 여러 구성 요소의 결합 된 효과를 나타낸다(Graham Kalton et al., 1994).

$$def(\hat{Y}) = V(\hat{Y}_{pps}) / V(\hat{Y}_{srs}) = \kappa_* [1 + \delta_* (\bar{m} - 1)] \quad (2.11)$$

여기서  $def(\hat{Y})$ 은  $\hat{Y}$ 에 대한 설계효과,  $\delta_*$ 는 비례확률추출의 동질성계수,  $\kappa_*$ 는 비례확률에서의 집락크기의 불균등성을 말한다.

### 2.3 층화효과

복잡표본설계에서 층화는 모집단을 서로 겹치지 않는 몇 개의 그룹으로 나누는 작업을 일컫는다. 층화추출은 실무에서 가장 널리 활용되는 기법 중의 하나이다. 예로, 지역표집(area sampling)에 의한 가구조사의 경우는 전국에 대한 추정량 뿐만 아니라 각 지역별 추정량이 요구되므로 시도구분을 통한 층화를 주로 고려하게 된다. 층화를 통해 분석영역별 추정량의 정도수준을 확보할 수 있음은 물론 조사 관리가 보다 편리하여 조사비용을 절감할 수도 있다.

층화추출에서는 층화변수(stratification variables), 즉 모집단을 층으로 구분

하는 보조정보의 선택에 따라 추정의 효율이 달라질 수 있다. 그러므로 효율적인 층화변수의 선택은 조사연구의 수행에 있어 매우 중요한 역할을 한다고 할 수 있다. 층화변수로는 질적변수(qualitative variable)와 양적변수(quantative variable)으로 구분할 수 있다. 질적변수는 조사연구에서 고려하는 분석영역의 수준에 따라 결정되는데 주로 지역적 특성, 주거형태, 성별, 연령대 등의 변수가 널리 사용된다. 반면, 양적변수는 연구특성과 상관관계가 높은 변수를 택한다. 예로, 매출액 특성을 연구하는 사업체조사에 근로자수를 고려할 수 있고, 농작물 생산량조사에서는 경지면적을 고려할 수 있다.

모집단의 층구분이 주어졌을 때 이에 대한 효과, 즉 층화효과(stratification effect)는 2.2절의 설계효과의 접근과 유사하게 층화추출에 따른 추정량 분산을 층화요소를 사용하지 않았을 때의 추정량 분산과 비교하여 다음 식 (2.12)과 같이 정의할 수 있다(Kalton et al., 1994).

$$Q(H, y) = \frac{V(\hat{Y})}{V^*(\hat{Y}^*)} \quad (2.12)$$

위의 층효과 평가식의  $V(\hat{Y})$ 는 식 (2.5)에 주어져 있고  $V^*(\hat{Y}^*)$ 은 층화없이 이단추출에 의한 표본설계에 따른 추정량  $\hat{Y}^* = n^{-1} \sum_{i=1}^n (\hat{Y}_i/p_i)$ 와 해당 표본설계 하에서 갖게 되는 분산을 나타내며 다음 식 (2.13)과 같이 정의될 수 있다.

$$V^*(\hat{Y}^*) = \frac{1}{n} \sum_{i=1}^N p_i \left( \frac{Y_i}{p_i} - Y \right)^2 + \sum_{i=1}^N \frac{M_i^2}{n p_i m_i} \left( 1 - \frac{m_i}{M_i} \right) S_i^2 \quad (2.13)$$

여기서  $p_i$ 는 층 구분 없이 결정되는 집락  $i$ 의 추출확률이며  $Y_i, m_i, M_i, S_i^2$  또한 각각 집락  $i$ 의 총합, 표본과 모집단 크기, 개체분산을 나타낸다.

일반적으로 모집단 층화는 조사연구의 주요한 관심변수와 연관성이 높은 보조변수들을 기준으로 층 내에서는 개체간에 서로 동질적일 수 있도록 하며 층 간에는 이질적으로 나누는 것이 효율적이다(Lohr, 2010).



## 제 3장 가구조사 사례적용 및 평가

본 장에서는 우리나라의 지역표집을 통한 가구조사의 대표적인 표본설계 사례들을 간단히 살펴보고, 통계청의 집계구를 지역 집락 단위로 사용하는 복잡표본설계의 효율성에 대해 논의하고자 한다.

### 3.1 복잡가구표본설계 사례

#### 3.1.1 농어업인 복지실태조사

농어업인 복지실태조사는 농업진흥청이 주관하는 조사로 농어촌의 복지실태, 교육여건, 보건의료, 기초생활여건, 사회안전망 및 복지서비스 등에 관련된 실태를 분석하여 농어촌 특성에 맞는 복지증진 및 지역개발시책의 수립과 시행 등에 필요한 기초자료를 제공하는 데에 목적을 두고 있다(농촌진흥청, 2016). 농어업인 복지실태조사는 5년 주기로 시행되어 있으며 첫 해는 도시와 농촌 간의 비교를 위해 전국을 조사대상으로 하며 2~5년차 조사에서는 농어촌 지역내 가구만을 대상으로 조사를 수행한다. 연차별 표본규모는 약 4,000가구로 층화다단추출 방식을 이용하는데, 2~5년차 조사를 기준으로 요약하면 전국의 읍면지역을 도별로 층화하고 층별로 읍면, 조사구 및 가구를 차례대로 추출한다. 지역층내 아파트비율과 65세비율의 두 특성을 이용한 다변량 군집분석을 통해 나눈 후 이를 내재적 층(implicit strata)으로 사용하였다. 여기서 내재적 층이란 표본추출 전

에 내재적 층을 나누는 특성에 따라 추출단위를 나열한 후 계통추출(systematic sampling)로 표본선택을 함으로 해당특성에 대해 균형을 맞출 수 있는 방법을 일컫는다. 이와 비교하여 지역층은 명시적 층화(explicit stratification)라 부른다.

### 3.1.2 경제활동인구조사

경제활동인구조사는 통계청이 수행하는 표본조사로 국민의 경제활동 특성을 조사함으로써 거시적 경제분석과 인력자원의 개발정책 수립에 필요한 기초 자료를 제공하는 것에 목표를 두고 있다. 경제활동인구조사는 매달 15일이 포함된 일주일을 조사기간으로 조사하며, 표본가구 내에서 만 15세 이상인 사람을 조사하여 경제활동참가율, 실업률, 고용률 등의 지표를 산출한다. 표본규모는 15세 이상 가구원 전체에 해당하는 1,629개의 조사구(약 3만2천 가구)로 다단층화집락추출방법을 이용하는데, 표본층은 시·도 와 동·읍면으로 나누어 구성하며, 층별로 광역조사구, 일반조사구, 가구를 차례로 추출하는 삼단추출방식을 이용하여 표집한다. 조사모집단은 2010년 인총 결과 중 일반과 아파트 조사구는 물론 2011년 신축아파트내 거주하는 만 15세 이상의 인구를 모두 포함한다.

### 3.1.3 우리나라 가구표본의 설계를 위한 추출단위 선택

통계청은 물론 우리나라 대부분의 가구조사에서 주로 조사구(enumeration district)를 집락으로 사용한다. 조사구는 인구주택총조사의 표집틀로 사용되며 조사구를 추출하고 추출된 조사구 내에서가구를 추출하는 이단추출방식의 표본

설계를 이룬다. 하지만 조사구는 5년 주기의 인구주택총조사를 통해 갱신되고 그 정보를 구성하기 때문에 주기별로 조사구의 노후화가 발생 할 수 있다(박진우 외, 2008;). 또한 2015년 등록센서스의 시작으로 기존 전수조사에서 제공되었던 조사구의 가구명부에 대한 접근이 불가능해지는 문제가 발생한다(한국통계학회, 농촌연구원 외, 2015;). 다음과 같은 문제들로 인해 동읍면이나 집계구와 같은 대안적집락을 고려할 수 있다. 집계구란 통계청에서 정의한 분할된 구획으로 읍면동 기준 1/24 정도의 크기를 가지며 평균적으로 200가구에서 250가구를 포함하며 기존 조사구(약 60가구)보다 상대적으로 큰 규모를 가진다. 집계구 자료는 인구주택총조사, 농림어업총조사, 전국사업체조사 등의 지표들을 기반으로 통계지리정보서비스(<http://sgis.kostat.go.kr>)에서 제공되고 있다. 본 논문에서는 조사구의 대안적 단위로 집계구를 이용하는 이단추출을 상정하여 우리나라 가구조사의 전형적 표본설계가 갖는 효율성에 대해 특히 집락효과와 층화효과 측면에서 평가하고자 한다.

## 3.2 집락 효율성 평가

### 3.2.1 이단추출

이단추출의 집락효과를 보기 위해서 2010년 기준 전국구 집계구 표본추출틀 내 정보들 중 1인 가구수, 자가가구수 대졸자인구수, 10대 인구수, 65세 이상의 노인인구수를 고려하였다. 표 3.1은 이들 5개 특성의 모비율 및 모총합을 나타내고 있다. 표 3.2는 각 변수와 인구특성 및 가구특성에 따른 기술통계량을 나

타낸다.

표 3.1 분석변수의 모비율과 모총합

변수	특성비율	모총합
1인 가구수	0.237	4,088,573
자가 가구수	0.542	9,369,843
대졸자 인구수	0.334	15,732,266
65세이상 인구수	0.106	5,015,821
10대 인구수	0.136	6,397,805

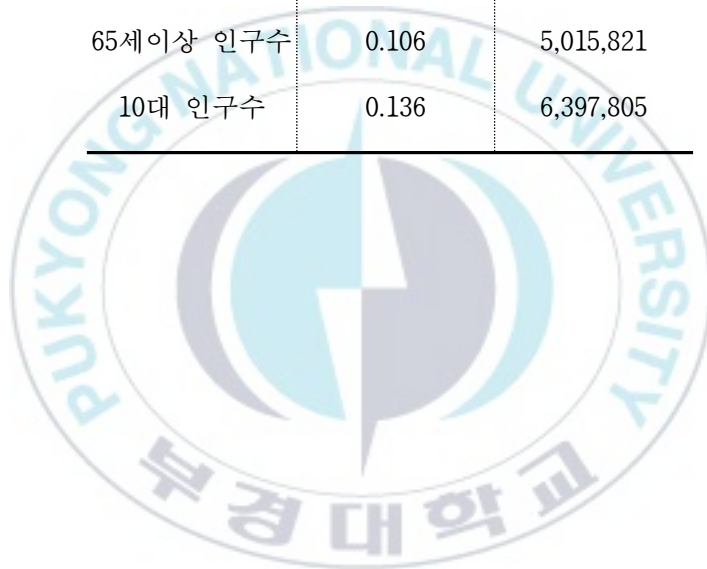


표 3.2 집계구 정보의 기술통계량

	인구수	가구수	1인가구수	자가가구수	대졸자수	65세인구수	10대인구수
최솟값	1	1	0	0	6	6	0
1사분위수	415	147	14	72	113	26	50
중위수	530	193	35	104	167	46	70
평균	571.07	209.43	49.55	113.56	190.77	64.22	77.54
3사분위수	694	254	66	147	242	80	97
최댓값	7613	3282	2444	798	4938	724	1778
변동계수	35.77	42.94	116.31	57.05	63.34	92.97	61.08

표 3.3 집계구정보 표본할당에 따른 상대분산과 설계효과

	$m = 65, \bar{n} = 30$		$m = 70, \bar{n} = 27$		$m = 94, \bar{n} = 20$		$m = 100, \bar{n} = 18$		$m = 188, \bar{n} = 10$	
	<i>RV</i>	<i>de ff</i>	<i>RV</i>	<i>de ff</i>	<i>RV</i>	<i>de ff</i>	<i>RV</i>	<i>de ff</i>	<i>RV</i>	<i>de ff</i>
1인 가구수	0.008	5.205	0.008	4.71	0.006	3.755	0.006	3.465	0.003	2.305
자가 가구수	0.004	8.984	0.003	8.159	0.002	6.231	0.002	5.681	0.001	3.478
대졸자 인구수	0.003	3.631	0.003	3.359	0.003	2.724	0.002	2.542	0.001	1.816
65세 이상 인구수	0.016	3.706	0.015	3.426	0.012	2.773	0.011	2.586	0.008	1.839
10대 인구수	0.006	1.789	0.005	1.707	0.005	1.517	0.005	1.462	0.004	1.244

※ 여기서  $RV = V(\hat{Y})/Y^2$

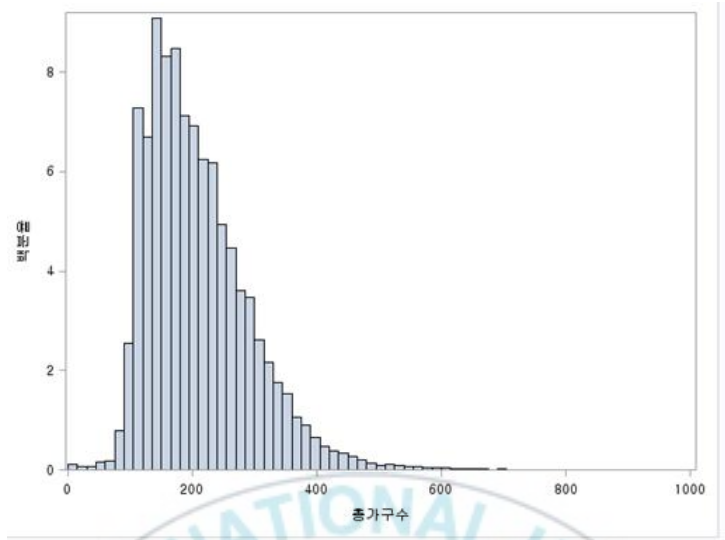


그림 3.1 집계구 정보의 가구수의 집락분포

표 3.3에서는 5가지 서로 다른 표본할당의 조합에 대한 모집단 특성의 상대 분산과 설계효과를 비교하고 있다. 위 표본할당에 있어서 집락의 표본크기를 늘리고 집락 내 개체 표본수를 작게 했을 때 상대적분산과 설계효과가 모두 작은 값을 가져 좋은 형태를 가진다. 그러나 집계구 정보는 집락 간 분산은 작으며 집락 내 분산이 큰 구조이기에 동질성계수가 작은 값을 가져 집락 내 개체들 간에 정보가 이질적인 정보를 가진다. 개체들 간의 정보가 이질적이기 때문에 집락 내 개체의 정보를 충분히 대표 할 만큼 집락 내 개체의 표본수를 늘려주는 것이 바람직하다.

표 3.4 단순임의추출/단순임의추출의 이단표본설계에서의 분산분해요소와 동질성계수

	$S_{yB}^2$	$S_{yW}^2$	$\delta_y$	$\kappa_y$	$RV$
1인 가구수	1.353	2.770	0.328	1.278	0.016
자가 가구수	0.325	0.614	0.346	1.113	0.004
대졸자 인구수	0.401	1.818	0.181	1.112	0.005
65세 이상 인구수	0.916	7.592	0.108	1.113	0.014
10대 인구수	0.373	6.209	0.057	1.034	0.008

**예시 1. 단순임의추출/단순임의추출에서 집락-간 그리고 집락-내 분산요소**

단순임의추출을 이용한 동질성 계수  $\delta_y$ 를 사용 했을 때 예상보다 표본이 효율적이지 않을 수가 있다. 이는 이단집락설계에 있어서 단순임의 추출을 할 경우  $\delta_y$ 의 값이 비효율적인 값으로 나타난다고 볼 수 있다. 표3.4와 같은 형태로 표본설계를 진행했을 때, 표본 배정에 있어서  $\delta_y$ 값이 높게 나타나며, 이는 집락의 크기가 다를 때 단순임의추출을 이용하면 집락크기의 변동을 반영하지 못하여 집락 간 분산이 크게 결정된다. 집락 크기를 반영하는 비례확률방법을 이용하면 이와 같은 문제를 해결할 수 있다.

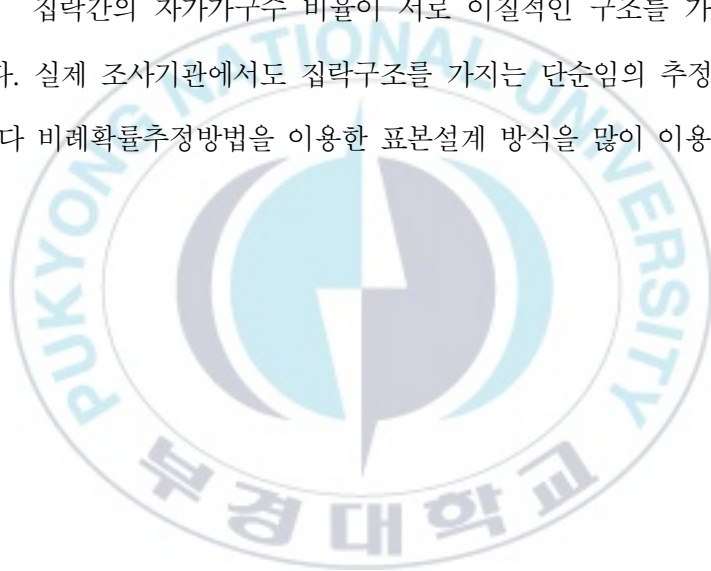
표 3.5 비례확률추출/단순임의추출의 이단표본설계에서의 분산분해요소와 동질성계수

	$S_{yB}^2$	$S_{yW}^2$	$\delta_y$	$\kappa_y$	$RV$
1인 가구수	0.469	2.77	0.145	1.004	0.006
자가 가구수	0.233	0.614	0.275	1.003	0.003
대졸자 인구수	0.181	1.817	0.091	1.001	0.003
65세 이상 인구수	0.781	7.585	0.093	0.996	0.012
10대 인구수	0.173	6.202	0.027	1.000	0.005

**예시 2. 비례확률추출/단순임의추출에서 집락-간 그리고 집락-내 분산 요소**

위의 표 3.5에서 앞의 예시1의  $\delta_y$ 의 값보다 상대적으로 작은 값을 가지는데 이는 집락의 정보가 집락 내에는 이질적이며 집락 간에는 동질적인 구조를 가지는 특징을 가지며, 집락크기의 변동성을 반영하여 집락 간 분산이 결정된다. 이는 비례확률추출방법을 이용하면 단순임의추출보다 동질성계수에 작은 값을 유도하며, 이단설계에 있어서  $\bar{n}$ 과  $m$ 을 고정했을 때 단순임의추출/단순임의추출의 설계보다 비례확률추출/단순임의추출의 설계방식이 더욱 효율적인 모습을 보여주는 것을  $\kappa_y$ 를 통해 확인할 수 있다. 예를 들어서 65세이상 인구수 변수에서  $\bar{n}=20$ 로 고정한 뒤 단순임의추출/단순임의추출 방법을 이용해 비례확률추출을 한다면  $\kappa[1 + \delta(\bar{n} - 1)] = 3.396$ 을 나타내며, 일차추출단위를 다음과 같은 비

비례확률추출/단순임의추출 방법을 이용한 추정값은  $\kappa_*[1 + \delta_*(\bar{n} - 1)] = 2.755$ 로 표현된다. 이는 상대적인 분산크기 비교할 때 비례확률추출/단순임의추출을 이용한 추정방법이 단순임의추출/단순임의추출 추정방법 보다 효율적인 모습을 나타낸다. 집락효과의 결과를 비교해보면 표 3.4에서 대졸자 인구수의  $\delta_y$ 의 값이 0에 근사하는 값을 가지는데 이는 집락 내에는 대졸자비율의 차이가 심한 반면에 집락 간에는 대졸자비율이 서로 비슷한 구조를 가진다고 볼 수 있다. 반면에 자가가구수의 경우  $\delta_y = 0.275$ 로 대졸자 비율에 비해 집락 내의 자가가구수의 차이가 적고 집락간의 자가가구수 비율이 서로 이질적인 구조를 가지는 것으로 볼 수 있다. 실제 조사기관에서도 집락구조를 가지는 단순임의 추정방법을 이용하는 것보다 비례확률추정방법을 이용한 표본설계 방식을 많이 이용한다.



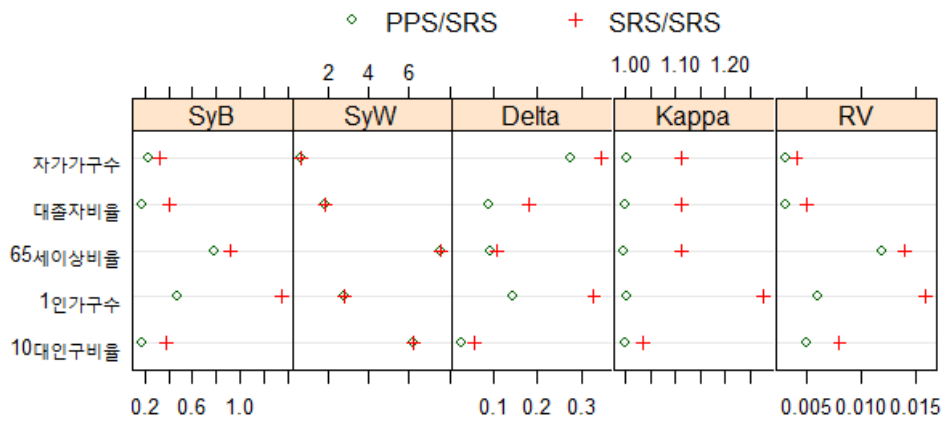


그림 3.2 단순임의/단순임의&비례확률/단순임의 분산구성요소 비교

### 3.3 층화 효율성 평가

층화효율성을 평가하기 위해 권역이나 지역 등 하나의 특징만을 고려한 단변량 층화 표본설계와 권역내의 동 읍면 또는 지역 내의 동 읍면을 이용한 다변량 층화 표본설계를 이용한다. 위의 방법과 같이 각 지역의 특성을 고려한 설계방식으로도 층을 구분할 수도 있지만, 대한민국의 고령화 진행 현상을 고려한 65세 이상 인구수의 정보와 밀집된 인구의 정보를 반영하기 위해 아파트비율을 이용하여 K-평균 군집법을 통해 다변량 층화 표본설계 또한 가능하다.

### 예시 3. 단변량 층화표본설계전략

단변량 층화표본설계를 하기 위하여 표 3.6과 같이 권역별로 층으로 구분하였으며, 각 권역별로 표본집계구  $m = 94$ 를 다음과 같이 할당한다.

표 3.6 권역별 표본집계구 설계방식

권역별	해당지역	표본집계구
수도권	서울/인천/경기	29
충청권	대전/충북/충남	12
호남권	광주/전북/전남/제주	16
대경권	대구/경북	14
동남권	부산/울산/경남	15
강원권	강원	8

분산을 층별로 분해했을 때 기존의 층을 나누지 않은 추정치의 분산보다 작은 추정치의 분산 값들을 기대할 수 있으며 이는 층화의 효율성을 나타낼 수 있다. 예를 들어서 1인가구수의 각 층에 해당되는  $V(\hat{Y}_h)$ 들의 총합을 표현하면 다음 식과 같이  $\sum_{h=1}^H V(\hat{Y}_h) = 1.345 * 10^{14}$ 로 표현이 되며 기존 층의 값  $V(\hat{Y}) = 2.948 * 10^{14}$ 과 비교 했을 때 층을 나눈 효율성을 비율로 나타낼 수 있다.  $Q(H, y) = 0.256$ 를 통해 층을 나눈 형태가 층을 구분하지 않은 형태에 비해 4배 더 적은 분산의 값을 제공하고 있으며 이는 분산분해의 효율성을 나타낸다.

표 3.7 1인가구수 변수를 권역별로 층을 구분한 비례확률추출/단순임의추출의 이단표본설계에서의 분산분해요소와 동질성계수

	$S_B^2$	$S_W^2$	$\delta$	$\kappa$	$V(\hat{Y})$
수도권	0.572	3.042	0.158	1.003	$6.675 \cdot 10^{13}$
충청권	0.455	2.351	0.162	1.003	$1.612 \cdot 10^{12}$
호남권	0.324	2.486	0.115	1.004	$2.078 \cdot 10^{12}$
대경권	0.379	2.531	0.130	1.003	$1.636 \cdot 10^{12}$
동남권	0.401	2.879	0.122	1.004	$3.422 \cdot 10^{12}$
강원권	0.315	2.287	0.121	1.004	$8.276 \cdot 10^{10}$
전체	0.469	2.770	0.145	1.004	$2.948 \cdot 10^{14}$

표 3.8 자가가구수 변수를 권역별로 층을 구분한 비례확률추출/단순임의추출의 이단표본설계에서의 분산분해요소와 동질성계수

	$S_B^2$	$S_W^2$	$\delta$	$\kappa$	$V(\hat{Y})$
수도권	0.241	0.913	0.209	1.004	$1.212 \cdot 10^{14}$
충청권	0.234	0.484	0.326	1.003	$3.558 \cdot 10^{12}$
호남권	0.194	0.338	0.365	1.002	$6.019 \cdot 10^{12}$
대경권	0.193	0.439	0.306	1.003	$3.984 \cdot 10^{12}$
동남권	0.168	0.490	0.255	1.004	$8.033 \cdot 10^{12}$
강원권	0.193	0.505	0.277	1.003	$1.889 \cdot 10^{11}$
전체	0.233	0.614	0.275	1.003	$6.719 \cdot 10^{14}$

표 3.9 대졸학력인구수 변수를 권역별로 층을 구분한 비례확률추출/단순임의추출의 이단표본설계에서의 분산분해요소와 동질성계수

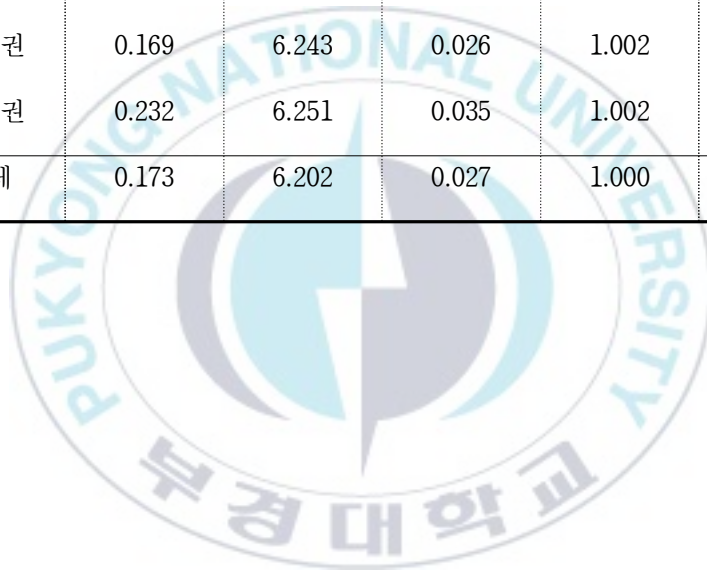
	$S_B^2$	$S_W^2$	$\delta$	$\kappa$	$V(\hat{Y})$
수도권	0.138	1.550	0.082	1.002	$4.596 \times 10^{14}$
충청권	0.261	2.056	0.113	1.002	$9.523 \times 10^{12}$
호남권	0.233	2.257	0.094	1.001	$1.330 \times 10^{12}$
대경권	0.198	2.121	0.085	1.002	$9.327 \times 10^{12}$
동남권	0.152	2.149	0.066	1.002	$1.991 \times 10^{13}$
강원권	0.284	2.379	0.106	1.001	$4.910 \times 10^{11}$
전체	0.181	1.817	0.091	1.001	$1.953 \times 10^{15}$

표 3.10 65세 이상 인구수 변수를 권역별로 층을 구분한 비례확률추출/단순임의추출의 이단표본설계에서의 분산분해요소와 동질성계수

	$S_B^2$	$S_W^2$	$\delta$	$\kappa$	$V(\hat{Y})$
수도권	0.395	10.472	0.036	0.999	$1.002 \times 10^{14}$
충청권	0.858	6.088	0.123	0.992	$5.216 \times 10^{12}$
호남권	0.835	4.825	0.147	0.993	$1.118 \times 10^{13}$
대경권	0.833	5.744	0.127	0.992	$6.449 \times 10^{12}$
동남권	0.789	7.507	0.095	0.995	$1.111 \times 10^{13}$
강원권	0.476	5.082	0.086	0.997	$2.758 \times 10^{11}$
전체	0.781	7.585	0.093	0.996	$8.462 \times 10^{14}$

표 3.11 10대인구수 변수를 권역별로 층을 구분한 비례확률추출/단순임의추출의 이단표본설계에서의 분산분해요소와 동질성계수

	$S_B^2$	$S_W^2$	$\delta$	$\kappa$	$V(\hat{Y})$
수도권	0.133	6.337	0.021	1.002	$1.244 \times 10^{14}$
충청권	0.225	5.953	0.036	1.002	$2.924 \times 10^{12}$
호남권	0.219	5.789	0.036	1.001	$4.847 \times 10^{12}$
대경권	0.233	6.244	0.035	1.002	$3.300 \times 10^{12}$
동남권	0.169	6.243	0.026	1.002	$7.318 \times 10^{12}$
강원권	0.232	6.251	0.035	1.002	$1.592 \times 10^{11}$
전체	0.173	6.202	0.027	1.000	$5.739 \times 10^{14}$



#### 예시 4. 다변량 층화표본설계전략 I

다변량 층화표본설계를 하기 위하여 다음 표 3.12 와 같이 권역별 동 읍면으로 층으로 구분하였으며, 표본집계구  $m = 94$ 를 다음과 같이 할당하였다.

표 3.12 권역당 동읍면별 표본집계구 설계방식

권역별	해당지역	표본집계구	
		동	읍면
수도권	서울/인천/경기	28	1
충청권	대전/충북/충남	8	4
호남권	광주/전북/전남/제주	11	5
대경권	대구/경북	10	4
동남권	부산/울산/경남	12	3
강원권	강원	5	3

권역에 이어 동과 읍면으로 세부적으로 분할하여 분산분해를 실시한 결과, 동 질성계수가 동과 읍면에 따라 차이를 나타내는 모습을 볼 수 있다. 그리고 예시 3과 같은 형태로 층화의 효과에 대한 평가를 할 수 있다. 예를 들어서 65세이상 인구수의 각 층에 해당되는  $V(\hat{Y}_h)$ 들의 총합을 표현하면 다음과 같이  $\sum_{h=1}^H V(\hat{Y}_h) = 1.066 * 10^{14}$ 로 표현이 되며 기존 층의 값  $V(\hat{Y}) = 4.025 * 10^{14}$ 과 비교 했을 때 층을 나눈 효율성을 비율로 나타낼 수 있다.

$Q(H, y) = 0.265$ 를 평가하면 층을 나눈 형태가 층을 구분하지 않은 형태에 비해 약 3.7배 분산이 적게 도출된다.

표 3.13 1인가구수 변수를 권역별 동 읍면으로 층을 구분한 비례확률추출/단순임의추출의 이단표본설계에서의 분산분해요소와 동질성계수

		$S_B^2$	$S_W^2$	$\delta$	$\kappa$	$V(\hat{Y})$
수도권	동	0.586	3.011	0.163	1.003	$2.860 \times 10^{13}$
	읍면	0.388	3.430	0.102	1.004	$1.012 \times 10^{10}$
충청권	동	0.581	2.327	0.199	1.003	$3.165 \times 10^{12}$
	읍면	0.257	2.378	0.097	1.004	$1.262 \times 10^{12}$
호남권	동	0.474	2.761	0.146	1.003	$2.408 \times 10^{12}$
	읍면	0.091	2.054	0.042	1.004	$6.746 \times 10^{11}$
대경권	동	0.473	2.688	0.149	1.003	$2.487 \times 10^{12}$
	읍면	0.197	2.218	0.081	1.004	$7.906 \times 10^{11}$
동남권	동	0.440	2.987	0.128	1.004	$5.006 \times 10^{12}$
	읍면	0.255	2.502	0.093	1.004	$1.031 \times 10^{12}$
강원권	동	0.421	2.162	0.163	1.003	$3.977 \times 10^{11}$
	읍면	0.135	2.497	0.051	1.003	$1.182 \times 10^{11}$
전체		0.469	2.770	0.145	1.004	$1.402 \times 10^{14}$

표 3.14 자가가구수 변수를 권역별 동 읍면으로 층을 구분한 비례확률추출/  
단순임의추출의 이단표본설계에서의 분산분해요소와 동질성계수

		$S_B^2$	$S_W^2$	$\delta$	$\kappa$	$V(\hat{Y})$
수도권	동	0.248	0.954	0.206	1.004	$4.944 \times 10^{13}$
	읍면	0.140	0.577	0.196	1.004	$9.594 \times 10^{12}$
충청권	동	0.268	0.651	0.291	1.003	$5.646 \times 10^{12}$
	읍면	0.152	0.296	0.339	1.003	$3.518 \times 10^{12}$
호남권	동	0.244	0.483	0.336	1.003	$6.326 \times 10^{12}$
	읍면	0.074	0.161	0.314	1.003	$1.846 \times 10^{12}$
대경권	동	0.221	0.582	0.275	1.003	$5.445 \times 10^{12}$
	읍면	0.099	0.231	0.302	1.003	$1.879 \times 10^{12}$
동남권	동	0.173	0.561	0.235	1.003	$1.107 \times 10^{13}$
	읍면	0.122	0.282	0.302	1.003	$2.628 \times 10^{12}$
강원권	동	0.239	0.672	0.263	1.003	$7.192 \times 10^{11}$
	읍면	0.104	0.331	0.239	1.003	$3.505 \times 10^{11}$
전체		0.233	0.614	0.275	1.003	$3.196 \times 10^{14}$

표 3.15 대졸학력인구수 변수를 권역별 동 읍면으로 층을 구분한 비례확률  
추출/단순임의추출의 이단표본설계에서의 분산분해요소와 동질성계수

		$S_B^2$	$S_W^2$	$\delta$	$\kappa$	$V(\hat{Y})$
수도권	동	0.125	1.488	0.078	1.001	$1.878 \times 10^{14}$
	읍면	0.184	2.834	0.061	1.001	$2.763 \times 10^{13}$
충청권	동	0.144	1.729	0.077	1.001	$1.432 \times 10^{13}$
	읍면	0.523	3.168	0.142	1.001	$8.887 \times 10^{12}$
호남권	동	0.108	1.835	0.055	1.001	$1.247 \times 10^{13}$
	읍면	0.479	5.049	0.087	1.000	$3.289 \times 10^{12}$
대경권	동	0.112	1.837	0.057	1.001	$1.199 \times 10^{13}$
	읍면	0.439	3.562	0.110	1.001	$4.395 \times 10^{12}$
동남권	동	0.122	2.005	0.057	1.001	$2.799 \times 10^{13}$
	읍면	0.269	3.112	0.080	1.001	$5.590 \times 10^{12}$
강원권	동	0.184	1.957	0.086	1.001	$1.848 \times 10^{12}$
	읍면	0.417	3.684	0.102	1.001	$8.665 \times 10^{12}$
전체		0.181	1.817	0.091	1.001	$9.289 \times 10^{14}$

표 3.16 65세 이상 인구수 변수를 권역별 동 읍면으로 층을 구분한 비례확률추출/단순임의추출의 이단표본설계에서의 분산분해요소와 동질성계수

		$S_B^2$	$S_W^2$	$\delta$	$\kappa$	$V(\hat{Y})$
수도권	동	0.305	11.226	0.026	0.998	$3.536 \times 10^{13}$
	읍면	0.521	5.841	0.082	0.994	$2.032 \times 10^{13}$
충청권	동	0.599	10.964	0.052	0.994	$3.695 \times 10^{12}$
	읍면	0.439	3.236	0.119	0.985	$7.691 \times 10^{12}$
호남권	동	0.654	9.447	0.065	0.995	$4.902 \times 10^{12}$
	읍면	0.242	2.097	0.104	0.990	$5.876 \times 10^{12}$
대경권	동	0.479	9.319	0.049	0.995	$4.111 \times 10^{12}$
	읍면	0.475	2.724	0.148	0.985	$6.017 \times 10^{12}$
동남권	동	0.484	9.660	0.048	0.995	$9.507 \times 10^{12}$
	읍면	0.703	3.650	0.162	0.988	$7.642 \times 10^{12}$
강원권	동	0.472	7.379	0.060	0.998	$6.918 \times 10^{11}$
	읍면	0.256	3.296	0.072	0.994	$7.547 \times 10^{11}$
전체		0.781	7.585	0.093	0.996	$4.025 \times 10^{14}$

표 3.17 10대인구수 변수를 권역별 동 읍면으로 층을 구분한 비례확률추출/  
단순임의추출의 이단표본설계에서의 분산분해요소와 동질성계수

		$S_B^2$	$S_W^2$	$\delta$	$\kappa$	$V(\hat{Y})$
수도권	동	0.127	6.307	0.020	1.002	$5.144 \times 10^{13}$
	읍면	0.214	6.709	0.031	1.002	$1.198 \times 10^{13}$
충청권	동	0.399	8.229	0.046	1.003	$9.772 \times 10^{12}$
	읍면	0.417	7.442	0.053	1.001	$2.887 \times 10^{12}$
호남권	동	0.122	5.195	0.023	1.001	$5.186 \times 10^{12}$
	읍면	0.542	8.105	0.063	1.002	$1.952 \times 10^{12}$
대경권	동	0.150	5.748	0.025	1.001	$4.721 \times 10^{12}$
	읍면	0.524	8.125	0.061	1.001	$1.779 \times 10^{12}$
동남권	동	0.139	6.070	0.022	1.001	$1.052 \times 10^{13}$
	읍면	0.321	7.139	0.043	1.001	$2.394 \times 10^{12}$
강원권	동	0.170	5.553	0.030	1.001	$6.418 \times 10^{11}$
	읍면	0.321	7.939	0.039	1.001	$3.144 \times 10^{11}$
전체		0.173	6.202	0.027	1.001	$2.729 \times 10^{14}$

예시 5. 다변량 층화표본설계전략 II

다변량 층화표본설계를 시도별 동 읍면으로 층으로 구분하였으며, 표본집계구  $m = 94$ 를 표 3.18과 같이 할당하였다.

표 3.18 시도당 동읍면별 표본집계구 설계방식

시도	표본집계구	
	동	읍면
서울	12	0
부산	5	2
인천	4	0
대구	6	0
광주	4	0
대전	4	0
울산	1	1
경기	10	2
강원	6	2
충북	3	1
충남	3	3
전북	3	2
전남	2	2
경북	4	3
경남	5	2
제주	1	1

예시4 와 동일한 방식으로 분산분해 요소를 살펴보면, 각 시도와 읍면에 따라 분산분해 결과를 다음 표 3.19부터 표 3.23에서 살펴 볼 수 있다. 층을 통해 층간 분산을 제거하여 분산분해의 효과를 볼 수 있는데, 예로써 표 3.13의 1인 가구 수의 각 해당 층에  $V(\hat{Y}_h)$ 들의 총합은  $\sum_{h=1}^H V(\hat{Y}_h) = 4.594 * 10^{13}$ 로 기존에 층을 나누지 않은  $V(\hat{Y}) = 1.40 * 10^{14}$ 와 비교 했을 때 분산이 줄어든 모습을  $Q(H, y) = 0.327$ 를 통해 확인할 수 있다.

이와 같은 결과를 통해 층을 구분하지 않은 형태보다 분산이 3배 더 줄어든 모습을 볼 수 있다. 표 3.16의 65세 이상인구수에서도 같은 방식으로 비교 했을 때  $Q(H, y) = 0.264$ 값으로 층을 나눈 구조가 층을 나누지 않은 구조에 비해 분산이 약 4배정도 감소한 모습을 확인 할 수 있다.

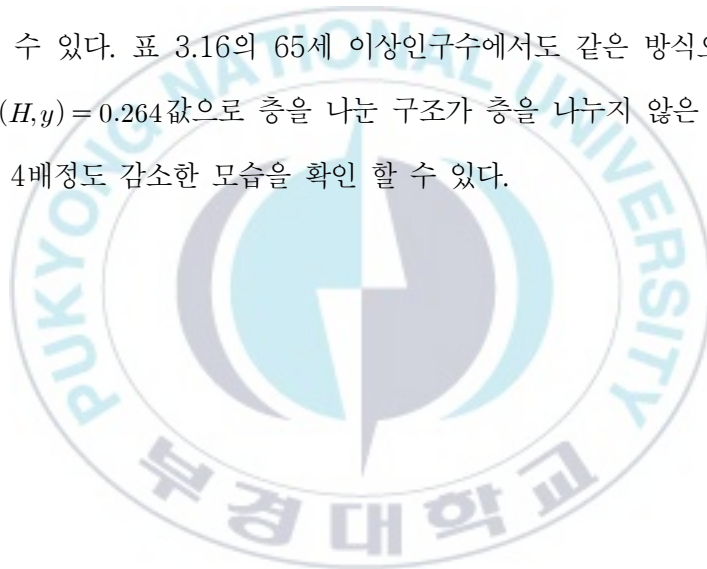


표 3.19 1인가구수 변수를 시도별 동 읍면으로 층을 구분한 비례확률추출/  
단순임의추출의 이단표본설계에서의 분산분해요소와 동질성계수

		$S_B^2$	$S_W^2$	$\delta$	$\kappa$	$V(\hat{Y})$
서울	동	0.477	2.661	0.152	1.003	$1.459*10^{13}$
부산	동	0.399	2.929	0.120	1.003	$3.404*10^{12}$
	읍면	0.362	3.002	0.108	1.004	$6.173*10^0$
대전	동	0.430	3.054	0.126	1.004	$1.925*10^{12}$
인천	동	0.535	3.427	0.135	1.003	$1.511*10^{12}$
광주	동	0.516	2.732	0.159	1.003	$9.655*10^{11}$
대전	동	0.611	2.385	0.204	1.003	$1.298*10^{12}$
울산	동	0.556	3.322	0.143	1.003	$1.169*10^{12}$
	읍면	0.366	3.542	0.094	1.004	$3.752*10^{10}$
경기	동	0.734	3.382	0.178	1.003	$1.401*10^{13}$
	읍면	0.404	3.461	0.105	1.004	$1.877*10^{12}$
강원	동	0.421	2.162	0.163	1.003	$3.314*10^{11}$
	읍면	0.135	2.497	0.051	1.004	$1.773*10^{11}$
충북	동	0.537	2.282	0.190	1.003	$6.841*10^{11}$
	읍면	0.225	2.388	0.086	1.003	$5.057*10^{11}$
충남	동	0.580	2.282	0.203	1.003	$5.797*10^{11}$
	읍면	0.260	2.407	0.098	1.004	$6.459*10^{11}$
전북	동	0.518	2.654	0.163	1.003	$1.015*10^{12}$
	읍면	0.075	2.098	0.034	1.004	$1.582*10^{11}$
전남	동	0.826	4.484	0.156	1.003	$9.608*10^{11}$
	읍면	0.097	1.972	0.047	1.003	$6.511*10^{11}$
경북	동	0.494	2.241	0.181	1.003	$1.159*10^{12}$
	읍면	0.165	2.093	0.073	1.004	$8.182*10^{11}$
경남	동	0.466	2.956	0.136	1.003	$1.283*10^{12}$
	읍면	0.228	2.367	0.088	1.003	$1.024*10^{12}$
제주	동	0.322	3.053	0.095	1.003	$1.881*10^{11}$
	읍면	0.063	2.672	0.023	1.004	$1.422*10^{10}$
전체		0.469	2.770	0.145	1.004	$1.402*10^{14}$

표 3.20 자가가구수 변수를 시도별 동 읍면으로 층을 구분한 비례확률추출/  
단순임의추출의 이단표본설계에서의 분산분해요소와 동질성계수

		$S_B^2$	$S_W^2$	$\delta$	$\kappa$	$V(\hat{Y})$
서울	동	0.260	1.181	0.181	1.004	$2.191*10^{13}$
부산	동	0.156	0.567	0.216	1.004	$7.258*10^{12}$
	읍면	0.271	0.690	0.282	1.003	$1.818*10^{10}$
대전	동	0.202	0.623	0.245	1.004	$4.641*10^{12}$
인천	동	0.148	0.670	0.181	1.004	$2.887*10^{12}$
광주	동	0.248	0.453	0.354	1.003	$2.493*10^{12}$
대전	동	0.279	0.692	0.287	1.003	$2.288*10^{12}$
울산	동	0.185	0.523	0.262	1.003	$2.784*10^{12}$
	읍면	0.125	0.418	0.230	1.003	$1.016*10^{11}$
경기	동	0.246	0.854	0.224	1.004	$2.654*10^{13}$
	읍면	0.143	0.597	0.193	1.004	$4.378*10^{12}$
강원	동	0.239	0.672	0.263	1.003	$5.993*10^{11}$
	읍면	0.104	0.331	0.240	1.003	$5.258*10^{11}$
충북	동	0.244	0.549	0.308	1.003	$1.293*10^{12}$
	읍면	0.130	0.281	0.317	1.002	$1.382*10^{12}$
충남	동	0.269	0.713	0.274	1.003	$9.651*10^{11}$
	읍면	0.159	0.291	0.353	1.003	$1.878*10^{12}$
전북	동	0.229	0.443	0.341	1.003	$2.433*10^{12}$
	읍면	0.049	0.144	0.255	1.003	$3.486*10^{11}$
전남	동	0.290	0.495	0.369	1.003	$1.622*10^{12}$
	읍면	0.087	0.146	0.372	1.002	$1.932*10^{12}$
경북	동	0.248	0.520	0.323	1.003	$2.334*10^{12}$
	읍면	0.092	0.216	0.300	1.003	$1.931*10^{12}$
경남	동	0.195	0.570	0.255	1.003	$2.921*10^{12}$
	읍면	0.107	0.249	0.300	1.003	$2.472*10^{12}$
제주	동	0.155	0.759	0.170	1.004	$3.988*10^{11}$
	읍면	0.040	0.437	0.084	1.004	$2.849*10^{10}$
전체		0.233	0.614	0.275	1.003	$3.196*10^{14}$

표 3.21 대졸학력인구수 변수를 시도별 동 읍면으로 층을 구분한 비례확률  
추출/단순임의추출의 이단표본설계에서의 분산분해요소와 동질성계수

		$S_B^2$	$S_W^2$	$\delta$	$\kappa$	$V(\hat{Y})$
서울	동	0.105	1.231	0.079	1.002	$9.175 \times 10^{13}$
부산	동	0.125	1.855	0.063	1.002	$2.064 \times 10^{13}$
	읍면	0.131	2.285	0.054	1.001	$3.571 \times 10^{10}$
대전	동	0.090	1.724	0.049	1.002	$1.103 \times 10^{13}$
인천	동	0.092	2.288	0.039	1.001	$7.660 \times 10^{12}$
광주	동	0.076	1.599	0.046	1.002	$4.533 \times 10^{12}$
대전	동	0.115	1.470	0.072	1.001	$5.997 \times 10^{12}$
울산	동	0.093	2.309	0.039	1.002	$5.491 \times 10^{12}$
	읍면	0.135	2.598	0.049	1.001	$2.567 \times 10^{11}$
경기	동	0.119	1.661	0.067	1.002	$8.860 \times 10^{13}$
	읍면	0.180	2.785	0.061	1.002	$1.286 \times 10^{13}$
강원	동	0.184	1.957	0.086	1.001	$1.540 \times 10^{12}$
	읍면	0.417	3.684	0.102	1.001	$1.299 \times 10^{12}$
충북	동	0.143	1.990	0.067	1.001	$2.762 \times 10^{12}$
	읍면	0.410	3.458	0.106	1.002	$2.981 \times 10^{12}$
충남	동	0.165	2.039	0.075	1.001	$2.387 \times 10^{12}$
	읍면	0.567	3.168	0.152	1.001	$4.797 \times 10^{12}$
전북	동	0.137	1.895	0.068	1.001	$5.218 \times 10^{12}$
	읍면	0.491	5.301	0.085	1.001	$8.067 \times 10^{11}$
전남	동	0.105	2.384	0.042	1.002	$2.263 \times 10^{12}$
	읍면	0.545	5.102	0.096	1.001	$3.061 \times 10^{12}$
경북	동	0.145	2.058	0.066	1.002	$4.420 \times 10^{12}$
	읍면	0.496	3.802	0.115	1.001	$4.457 \times 10^{12}$
경남	동	0.116	2.160	0.051	1.001	$6.450 \times 10^{12}$
	읍면	0.304	3.320	0.084	1.002	$5.327 \times 10^{12}$
제주	동	0.064	1.762	0.035	1.002	$1.162 \times 10^{12}$
	읍면	0.110	4.037	0.027	1.001	$8.184 \times 10^{10}$
전체		0.181	1.817	0.091	1.001	$9.289 \times 10^{14}$

표 3.22 65세 이상 인구수 변수를 시도별 동 읍면으로 층을 구분한 비례확률추출/단순임의추출의 이단표본설계에서의 분산분해요소와 동질성계수

		$S_B^2$	$S_W^2$	$\delta$	$\kappa$	$V(\hat{Y})$
서울	동	0.203	10.020	0.020	1.000	$1.675*10^{13}$
부산	동	0.301	7.747	0.037	1.000	$7.065*10^{12}$
	읍면	0.506	6.207	0.075	1.000	$2.323*10^{10}$
대전	동	0.326	9.012	0.035	0.998	$3.593*10^{12}$
인천	동	0.367	11.961	0.030	0.998	$2.292*10^{12}$
광주	동	0.649	10.328	0.059	0.995	$1.672*10^{12}$
대전	동	0.514	10.876	0.045	0.997	$1.456*10^{12}$
울산	동	0.839	17.274	0.046	0.994	$1.399*10^{12}$
	읍면	0.990	8.376	0.106	0.991	$1.739*10^{11}$
경기	동	0.416	12.575	0.032	0.997	$1.800*10^{13}$
	읍면	0.516	6.140	0.078	0.996	$8.810*10^{12}$
강원	동	0.472	7.379	0.060	0.998	$5.765*10^{11}$
	읍면	0.256	3.296	0.072	0.994	$1.132*10^{12}$
충북	동	0.599	10.225	0.055	0.995	$8.784*10^{11}$
	읍면	0.440	3.173	0.122	0.991	$3.357*10^{12}$
충남	동	0.773	12.116	0.060	0.993	$6.458*10^{11}$
	읍면	0.431	3.197	0.119	0.981	$3.963*10^{12}$
전북	동	0.651	8.437	0.072	0.996	$2.162*10^{12}$
	읍면	0.176	1.994	0.081	0.995	$1.341*10^{12}$
전남	동	0.749	9.607	0.072	0.991	$1.151*10^{12}$
	읍면	0.275	2.036	0.119	0.986	$5.690*10^{12}$
경북	동	0.804	9.865	0.075	0.990	$1.708*10^{12}$
	읍면	0.388	2.481	0.135	0.984	$6.078*10^{12}$
경남	동	0.746	12.250	0.057	0.992	$2.115*10^{12}$
	읍면	0.619	3.220	0.161	0.987	$7.888*10^{12}$
제주	동	0.422	9.743	0.042	0.998	$4.415*10^{11}$
	읍면	0.145	3.339	0.041	1.002	$1.078*10^{11}$
전체		0.781	7.585	0.093	0.996	$4.025*10^{14}$

표 3.23 10대 인구수 변수를 시도별 동 읍면으로 층을 구분한 비례확률추출/단순임의추출의 이단표본설계에서의 분산분해요소와 동질성계수

		$S_B^2$	$S_W^2$	$\delta$	$\kappa$	$V(\hat{Y})$
서울	동	0.133	7.134	0.018	1.001	$2.151*10^{13}$
부산	동	0.136	6.691	0.020	1.001	$6.472*10^{12}$
	읍면	0.308	6.860	0.043	1.001	$1.659*10^{10}$
대전	동	0.112	5.852	0.019	1.002	$4.149*10^{12}$
인천	동	0.079	5.886	0.013	1.001	$3.262*10^{12}$
광주	동	0.114	5.221	0.021	1.001	$1.944*10^{12}$
대전	동	0.131	5.615	0.023	1.001	$1.928*10^{12}$
울산	동	0.129	5.388	0.023	1.001	$2.855*10^{12}$
	읍면	0.170	5.662	0.029	1.001	$1.311*10^{11}$
경기	동	0.116	5.731	0.020	1.001	$2.914*10^{13}$
	읍면	0.209	6.631	0.031	1.001	$5.575*10^{12}$
강원	동	0.170	5.553	0.030	1.001	$5.349*10^{11}$
	읍면	0.321	7.939	0.039	1.001	$4.716*10^{11}$
충북	동	0.137	5.226	0.026	1.002	$1.100*10^{12}$
	읍면	0.370	7.789	0.045	1.001	$1.098*10^{12}$
충남	동	0.152	5.188	0.029	1.001	$9.394*10^{11}$
	읍면	0.435	7.354	0.056	1.001	$1.541*10^{12}$
전북	동	0.144	5.202	0.027	1.002	$2.074*10^{12}$
	읍면	0.518	8.958	0.055	1.001	$4.332*10^{11}$
전남	동	0.130	5.100	0.025	1.001	$1.220*10^{12}$
	읍면	0.631	7.943	0.074	1.000	$1.900*10^{12}$
경북	동	0.209	5.583	0.036	1.001	$1.938*10^{12}$
	읍면	0.594	8.548	0.065	1.001	$1.810*10^{12}$
경남	동	0.127	5.530	0.022	1.001	$2.819*10^{12}$
	읍면	0.348	7.490	0.044	1.002	$2.242*10^{12}$
제주	동	0.063	5.267	0.012	1.001	$4.965*10^{11}$
	읍면	0.101	6.649	0.015	1.001	$5.014*10^{10}$
전체		0.173	6.202	0.027	1.001	$2.729*10^{14}$

### 예시 6. 다변량 층화표본설계전략 III

다음 예시는 농어업인 복지실태조사를 참고하여 층화를 실시하는데 있어서 고려화문제를 보기 위한 변수로 65세 이상 노인인구비율과 도시화에 따른 인구특성을 고려하기 위한 변수로 아파트 비율을 사용하여 군집분석을 통한 층 할당을 실시한다. 표본설계에 있어서 이상점으로 나누어진 층화설계는 문제가 생기므로 이상점을 고려한 층화설계를 한다. 이상점을 식별하기 위해서 각 시도를 12개로 나누는 뒤 아파트비율과 65세이상 노인인구 비율을 통해 산점도 행렬로 표현하였으며, 다음 그림 1을 통해 파악되는 이상점으로 의심되는 자료로 강원도 철원군 근북면과 같은 점을 포함한 9개의 이상점을 파악하였는데, 해당 지역의 특징은 65세이상 비율이 그 지역에 비해 높거나, 아파트비율이 해당 지역보다 높거나 0에 가까운 값을 가진다. 이상점으로 확인된 9개의 지역은 층 결정을 위한 군집 분석의 대상에서 제외 시킨 뒤 층이 결정 되고 난 이후에 이상점을 가장 유사한 군집에 포함시킴으로써 이상점의 문제를 고려한다. 다변량 변수의 이상점에 대한 상세한 논의는 박진우 등 (2008)을 참고할 수 있다.

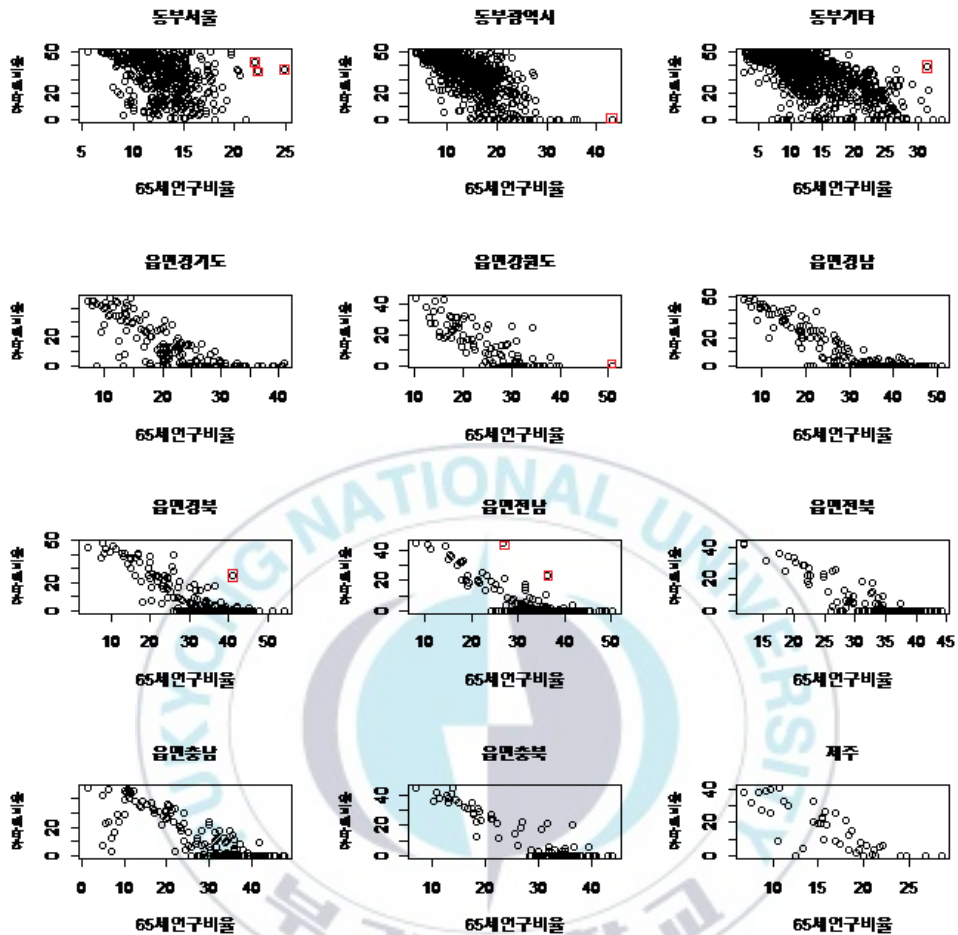


그림 3.3 각 지역별 이상점 확인을 위한 산점도

층화에 있어서 층의 수가 많아질수록 각 층 내부가 동질적인 성질을 가지지만 복잡하게 세분화 된 층을 가지는 것 보다 적절한 층의 수 확보가 중요하기 때문에 다음에 소개 할 네 가지 방법을 이용하여 최적의 층수를 결정하였다. 주로 *CCC*와 *R-Squared* 방법을 좀 더 고려하여 최적 층수를 결정하였다(최재혁, 2010).

### ①- Cubic Clustering Criterion (*CCC*)

군집의 최적 층수 결정을 위해 Sarle(1983)에 의해 처음 사용이 되었다. *CCC* 값이 2 혹은 3보다 큰 값을 가질 때 국부적 최고점(local peak)에 해당된다면, 적절한 군집개수의 후보라고 볼 수 있다.

### ②- *R-Squared*

군집들 간에 결정계수가 크다는 의미는 군집이 최적화 되어있다는 의미이며, 이는 군집의 개수가 증가할수록 *R-Squared*가 증가하는 경향을 가진다. 결정계수가 큰 군집수를 최적 군집수로 정하되 군집의 *R-Squared*가 급격히 증가하다가 증가분이 완만해지는 지점에서 군집개수의 적절한 후보군에 해당이 된다.

### ③- *Pseudo-F*

최적군집 개수를 찾기 위한 방법으로 *F*-통계량을 이용한 세 개 이상의 집단을 비교할 때 유용한 통계량이다. 일반적으로 *Pseudo-F*의 통계량이 커지면 집단 내부가 동질적이라고 볼 수 있으며, 이는 층화에 대한 효율이 높은 것을 의미한다. *Pseudo-F* 통계량은 군집들 간의 분리 정도를 측정하여 값이 크게 나오면 더 이상 군집을 합치는 것이 의미가 없음을 뜻한다(Calinski과 Harabasz, 1974).

### ④- *Pseudo-t<sup>2</sup>*

*Pseudo-F* 통계량과 마찬가지로 최적 군집개수를 결정하기 위한 방법이다. 두 집단을 비교하는 통계량으로 다변량 분산분석과 유사한 개념으로, *Pseudo-t<sup>2</sup>* 통계량의 최대값이 최적의 층수를 나타낸다. (Duda과 Hart, 1973).

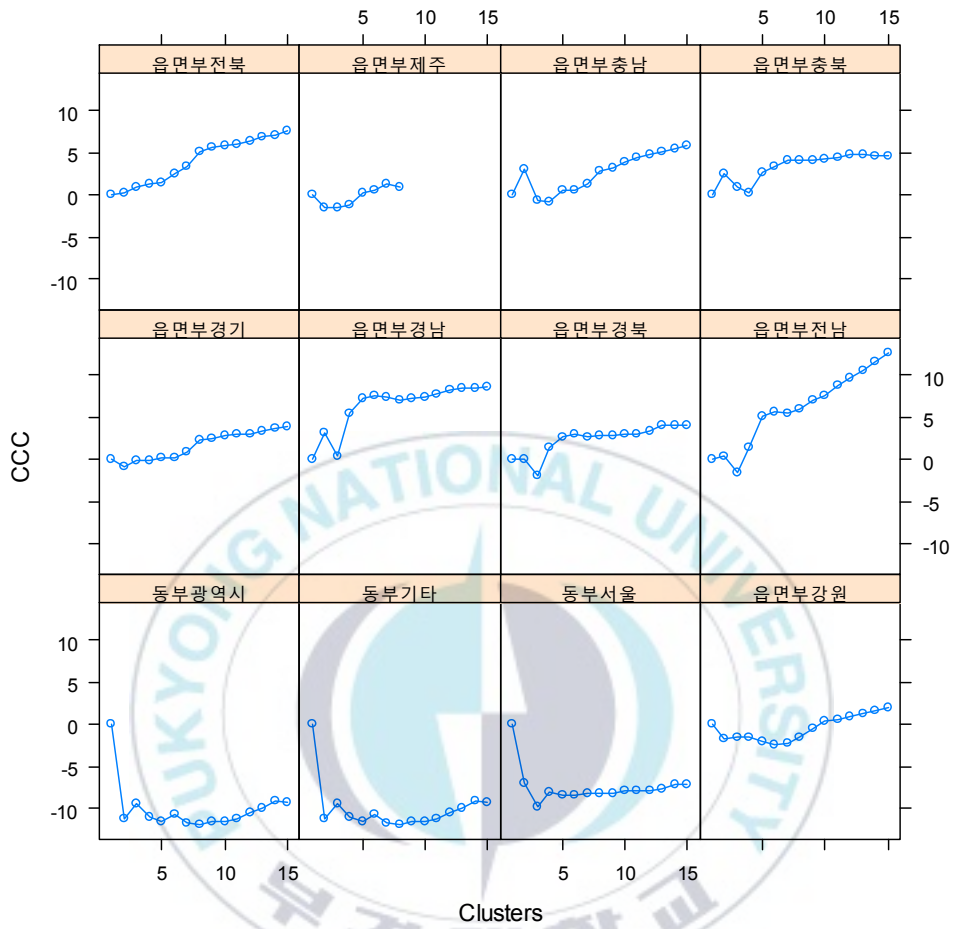


그림 3.4 CCC를 이용한 각 구역에 따른 최적층수결정

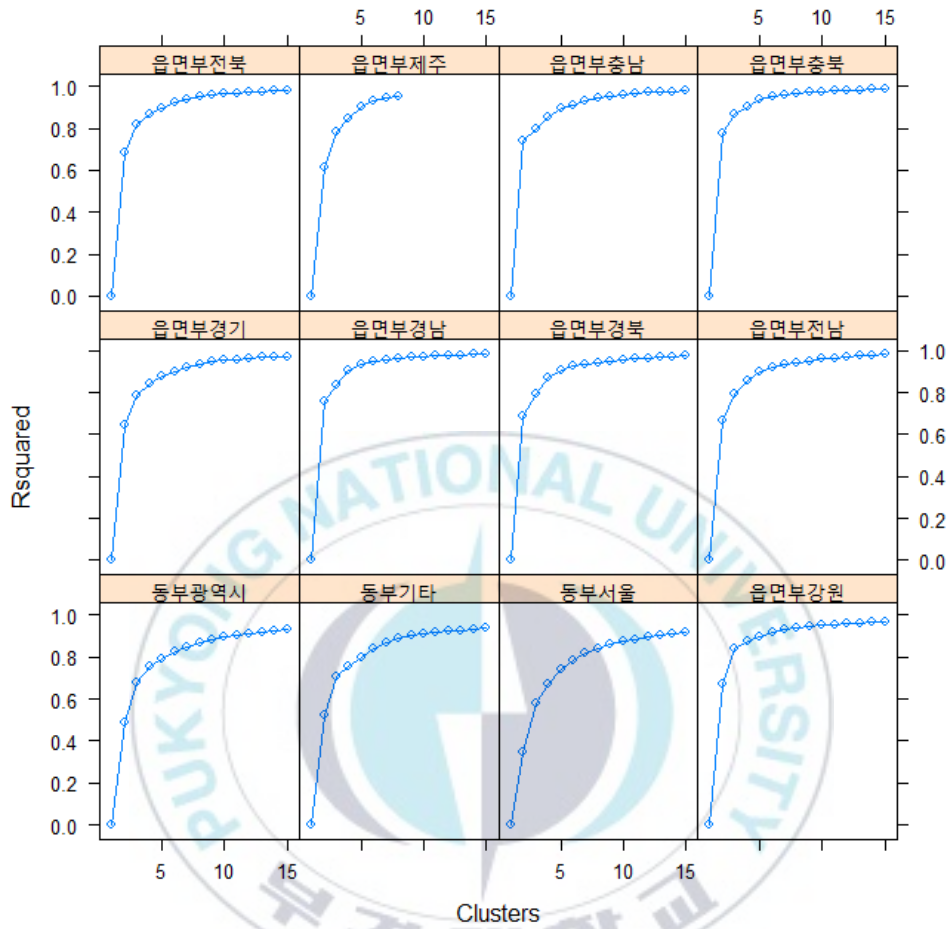


그림 3.5 R-squared를 이용한 각 구역에 따른 최적총수결정

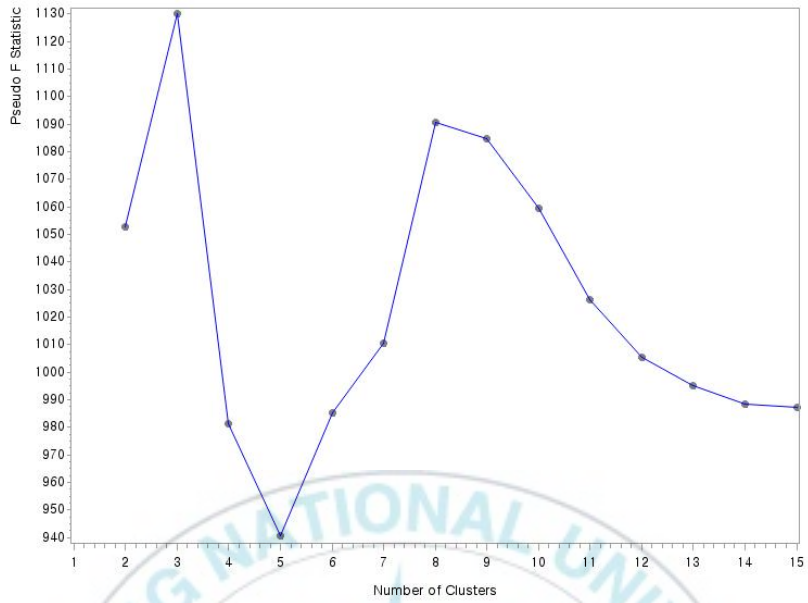


그림 3.6 Pseudo F-Squared를 이용한 각 구역에 따른 최적총수결정

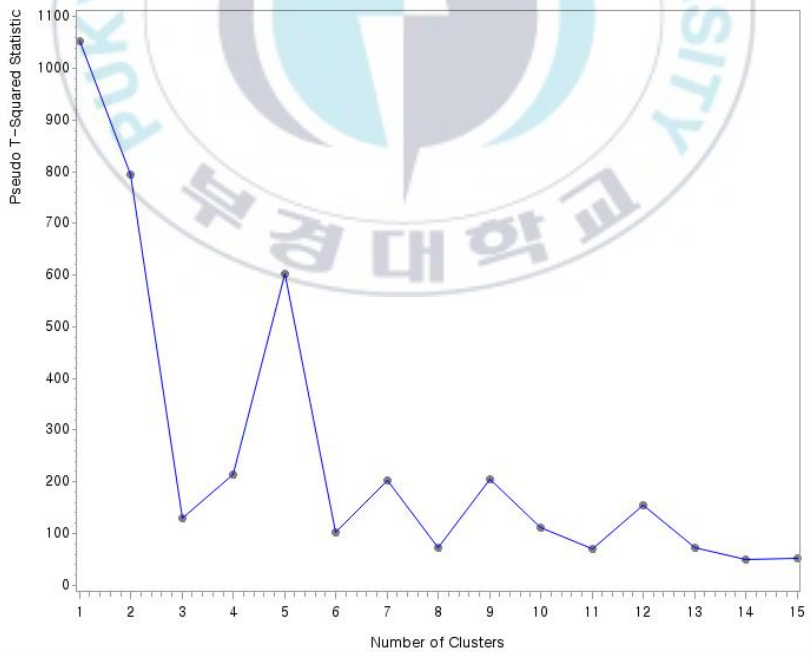


그림 3.7 Pseudo T-Squared를 이용한 각 구역에 따른 최적총수결정

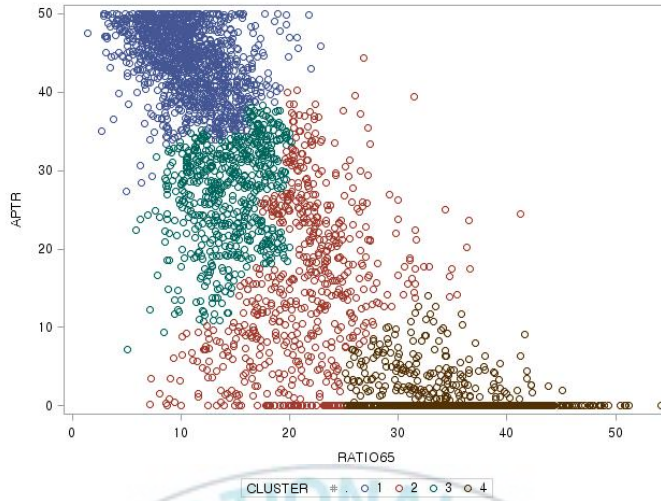


그림 3.8 Ward 법을 이용한 65세 이상 인구수와 아파트

비율의 층 배분

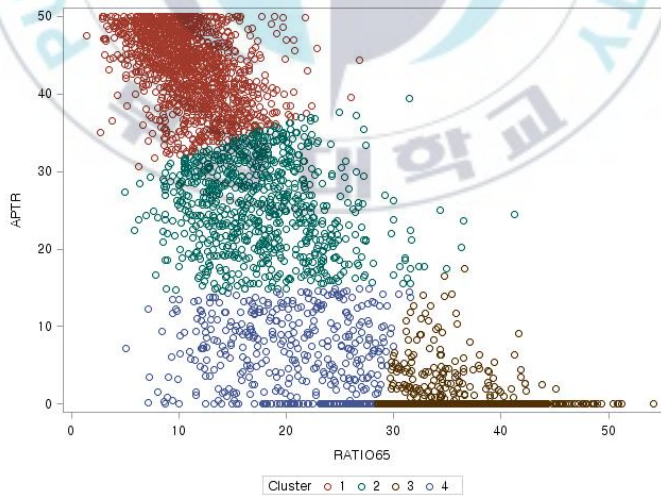


그림 3.9 K-평균 군집법을 이용한 65세 이상 인구수와

아파트 비율의 층 배분

K-평균 군집법과 Ward 방법의 최적 층수를 이용하여 군집을 그림 3.8과 그림 3.9와 같이 나눈 후 두 방법을 이용하여 서울특별시의 층4개에 적용하여 아파트비율과 65세이상 인구비율에 대한 층화효과를 다음표와 같이 확인한다.



표 3.24 1인가구수 변수를 K-평균 군집법으로 층을 구분한 비례확률추출/단순임의추출의 이단표본설계에서의 분산분해요소와 동질성계수

		$S_B^2$	$S_W^2$	$\delta$	<i>deff</i>	$V(\hat{Y})$
서울	층1	1.812	4.562	0.284	6.401	$4.493 \cdot 10^{12}$
	층2	1.286	4.979	0.205	4.900	$6.995 \cdot 10^{11}$
	층3	2.437	7.872	0.236	5.492	$4.618 \cdot 10^{12}$
	층4	1.537	4.026	0.276	6.249	$2.141 \cdot 10^{12}$
서울특별시		1.899	5.349	0.262	5.978	$3.239 \cdot 10^{13}$

표 3.25 자가가구수 변수를 K-평균 군집법으로 층을 구분한 비례확률추출/단순임의추출의 이단표본설계에서의 분산분해요소와 동질성계수

		$S_B^2$	$S_W^2$	$\delta$	<i>deff</i>	$V(\hat{Y})$
서울	층1	1.410	2.655	0.346	7.590	$6.780 \cdot 10^{12}$
	층2	1.172	3.685	0.241	5.585	$8.997 \cdot 10^{11}$
	층3	1.399	1.797	0.438	9.317	$1.465 \cdot 10^{13}$
	층4	1.328	3.366	0.283	6.375	$2.429 \cdot 10^{12}$
서울특별시		1.415	2.456	0.366	7.945	$6.586 \cdot 10^{13}$

표 3.26 대출학력인구수 변수를 K-평균 군집법으로 층을 구분한 비례확률  
추출/단순임의추출의 이단표본설계에서의 분산분해요소와 동질성계수

		$S_B^2$	$S_W^2$	$\delta$	<i>deff</i>	$V(\hat{Y})$
서울	층1	0.106	1.397	0.071	2.339	$9.429 \cdot 10^{12}$
	층2	0.082	1.739	0.045	1.856	$1.459 \cdot 10^{12}$
	층3	0.081	0.987	0.076	2.441	$2.008 \cdot 10^{13}$
	층4	0.113	1.499	0.070	2.332	$3.689 \cdot 10^{13}$
서울특별시		0.108	1.244	0.080	2.518	$5.682 \cdot 10^{13}$

표 3.27 65세 이상 인구수 변수를 K-평균 군집법으로 층을 구분한 비례확  
률추출/단순임의추출의 이단표본설계에서의 분산분해요소와 동질성계수

		$S_B^2$	$S_W^2$	$\delta$	<i>deff</i>	$V(\hat{Y})$
서울	층1	0.108	7.829	0.014	1.259	$3.218 \cdot 10^{12}$
	층2	0.076	7.674	0.010	1.186	$5.689 \cdot 10^{11}$
	층3	0.147	6.325	0.023	1.432	$7.819 \cdot 10^{12}$
	층4	0.106	8.028	0.013	1.248	$1.252 \cdot 10^{12}$
서울특별시		0.133	7.150	0.018	1.347	$1.321 \cdot 10^{13}$

표 3.28 10대 인구수 변수를 K-평균 군집법으로 층을 구분한 비례확률추출/단순임의추출의 이단표본설계에서의 분산분해요소와 동질성계수

		$S_B^2$	$S_W^2$	$\delta$	<i>deff</i>	$V(\hat{Y})$
서울	층1	0.108	7.829	0.014	1.259	$3.218 \times 10^{12}$
	층2	0.076	7.674	0.010	1.186	$5.689 \times 10^{11}$
	층3	0.147	6.325	0.022	1.432	$7.819 \times 10^{12}$
	층4	0.106	8.028	0.013	1.248	$1.252 \times 10^{12}$
서울특별시		0.133	7.150	0.018	1.347	$1.321 \times 10^{13}$

표 3.29 1인가구수 변수를 Ward법으로 층을 구분한 비례확률추출/단순임의추출의 이단표본설계에서의 분산분해요소와 동질성계수

		$S_B^2$	$S_W^2$	$\delta$	<i>deff</i>	$V(\hat{Y})$
서울	층1	2.321	6.222	0.272	6.162	$1.821 \times 10^{13}$
	층2	1.572	5.001	0.239	5.544	$1.022 \times 10^{12}$
	층3	1.502	4.092	0.269	6.102	$2.340 \times 10^{12}$
	층4	1.232	4.957	0.199	4.782	$2.571 \times 10^{11}$
서울특별시		1.899	5.349	0.262	5.978	$3.241 \times 10^{13}$

표 3.30 자가가구수 변수를 Ward법으로 층을 구분한 비례확률추출/단순임의추출의 이단표본설계에서의 분산분해요소와 동질성계수

		$S_B^2$	$S_W^2$	$\delta$	<i>deff</i>	$V(\hat{Y})$
서울	층1	1.437	2.034	0.414	8.866	$3.011 \cdot 10^{13}$
	층2	1.310	2.609	0.334	7.351	$1.721 \cdot 10^{12}$
	층3	1.318	4.065	0.245	5.652	$2.310 \cdot 10^{12}$
	층4	1.071	2.791	0.277	6.269	$4.126 \cdot 10^{11}$
서울특별시		1.415	2.456	0.365	7.945	$6.593 \cdot 10^{13}$

표 3.31 대졸학력인구수 변수를 Ward법으로 층을 구분한 비례확률추출/단순임의추출의 이단표본설계에서의 분산분해요소와 동질성계수

		$S_B^2$	$S_W^2$	$\delta$	<i>deff</i>	$V(\hat{Y})$
서울	층1	0.089	1.059	0.077	2.472	$4.171 \cdot 10^{13}$
	층2	0.106	1.650	0.060	2.147	$2.273 \cdot 10^{12}$
	층3	0.102	1.484	0.064	2.222	$4.013 \cdot 10^{12}$
	층4	0.090	1.946	0.044	1.839	$5.293 \cdot 10^{11}$
서울특별시		0.108	1.244	0.079	2.518	$5.686 \cdot 10^{13}$

표 3.32 65세 이상 인구수 변수를 Ward법으로 층을 구분한 비례확률추출/단순임의추출의 이단표본설계에서의 분산분해요소와 동질성계수

		$S_B^2$	$S_W^2$	$\delta$	<i>deff</i>	$V(\hat{Y})$
서울	층1	0.251	11.193	0.022	1.415	$1.033 \cdot 10^{13}$
	층2	0.118	7.470	0.016	1.295	$8.636 \cdot 10^{11}$
	층3	0.103	10.574	0.009	1.183	$1.102 \cdot 10^{12}$
	층4	0.088	6.529	0.013	1.253	$2.457 \cdot 10^{11}$
서울특별시		0.210	9.847	0.021	1.397	$1.060 \cdot 10^{13}$

표 3.33 10대 인구수 변수를 Ward법으로 층을 구분한 비례확률추출/단순임의추출의 이단표본설계에서의 분산분해요소와 동질성계수

		$S_B^2$	$S_W^2$	$\delta$	<i>deff</i>	$V(\hat{Y})$
서울	층1	0.146	6.705	0.021	1.405	$1.542 \cdot 10^{13}$
	층2	0.087	7.887	0.011	1.207	$8.172 \cdot 10^{11}$
	층3	0.104	7.927	0.013	1.246	$1.402 \cdot 10^{12}$
	층4	0.077	7.846	0.009	1.185	$2.113 \cdot 10^{11}$
서울특별시		0.133	7.150	0.018	1.347	$1.323 \cdot 10^{13}$

비계층적방법에 해당하는 K-평균 군집법과 계층적 방법인 Ward법을 이용하여 내재적 층화에 대한 층 효과를 비교하였다. 표 3.24의과 표 3.29는 각각 1인가구수 특성을 기준으로 평가한 Ward법과 K-평균법에 따른 층화효과를 비교하고 있다. K-평균법은  $Q(H,y)=0.368$ 이며, Ward법은  $Q(H,y)=0.506$ 으로 K-평균 군집법을 통해 층을 구분하였을 때 층화효과가 더 좋았다.

### 3.4 종합적 효율성 평가

층을 구분하지 않은 동질성계수와 층이 결정된 후 계산된 동질성계수와와의 차이를 통해 층화다단계추출의 집락효율성과 층화효율성을 동시에 고려할 수 있다. 예를 들어서 표 3.24의 1인가구수의  $\delta_y$ 의 값을 비교했을 때 층2와 3의 동질성계수가 줄어들어 집락효율성 측면에 있어서 개선된 부분을 볼 수 있으며 표 3.27의 설계효과에서 기존 층에 비해 층1, 층2, 층4의 설계효과가 작아져 층화를 통해 집락효율성이 기존에 비해 좋은 결과를 나타낸다. 또한 층을 나눈 이후에 집락간 분산과 집락내 분산을 통해서도 종합적인 효율성을 평가할 수 있는데 다음 예로 표 3.15의 충청권의 동과 읍면에 따라 비교할 수 있는데 동지역의 집락 간 분산  $S_{yB}^2 = 0.144$ 으로 기존의 층을 나누지 않은 집락 간 분산  $S_{yB}^2 = 0.181$ 에 비해 작아진 모습을 볼 수 있으며, 읍면지역의 집락 간 분산  $S_{yB}^2 = 0.523$ 으로 커진 값을 보인다. 이는 충청권 동지역의 대졸자인구수가 집락들 간에는 큰 차이를 이루지 않으며, 읍면 지역의 대졸자인구수는 집락들 간에 많은 차이를 보이는 것을 의미한다. 분산구조에서도 복잡표본설계의 효과를 표현할 수 있는데 다음 식을 통해 분산구조를 참고 할 수 있다.

복잡표본설계를 고려하지 않은 총합추정치의 분산형태는  $V_1(\hat{Y}) = M^2 \frac{S_y^2}{nm}$  과 같이 표현이 되며, 표 3.5의 1인가구수 예로  $V_1(\hat{Y}) = 7.82 * 10^{13}$  을 가진다. 복잡표본설계에서 집락의 형태만을 고려한 총합추정치의 분산을 다음과 같이 정의하며  $V_2(\hat{Y}) = M^2 \frac{S_y^2}{nm} \kappa_y [1 + \delta_y (\bar{m} - 1)]$  동일한 예로써  $V_2(\hat{Y}) = 2.95 * 10^{14}$  의 값을 가지며, 복잡표본설계의 층화와 집락의 형태를 고려한 총합추정치의 분산형태는  $V_3(\hat{Y}) = \sum_{h=1}^H V(\hat{Y}_h) = \sum_{h=1}^H M_h^2 \frac{S_{yh}^2}{n_h m_h} \kappa_{yh} [1 + \delta_{yh} (\bar{m}_h - 1)]$  와 같이 표현되며, 마찬가지로 동일한 예의 분산값을 구하면  $V_3(\hat{Y}) = 7.56 * 10^{13}$  의 값을 가진다. 각 분산  $V_1, V_2, V_3$  를 통해  $V_2$  는  $V_1$  에 비해 총합추정량의 분산이 더 커지는 단점을 가지지만,  $V_3$  의 층 구조를 통해 총합추정량의 분산을 줄여 정도수준을 향상시킨다. 또한 층 세분화를 통해 층별 설계요소들에 대한 평가를 더욱 명확하게 확인할 수 있다.

## 제 4장 결론

본 논문에서는 대안적 추출단위인 집계구를 이용한 층화 이단 표본설계에서 가구와 개체의 표본선택에 따른 추정량의 정도수준을 평가하였다. 집락 내 크기를 고정하는 단순확률추출의 표본설계와 자체가중을 반영하는 비례확률추출의 표본설계를 비교하였다. 실무를 반영하여 복원 추출을 이용하였으며, 위 두 방법을 고려하여 표본설계의 평가를 집락의 관점에서는  $def(\hat{Y})$ 를 통해 층화의 관점에서는 식 (2.12)의  $Q(H, y)$ 를 이용하여 확인 할 수 있었다. 비례확률추출의 설계방식은 집락크기의 불균등한 문제를 보정하는 역할을 통해  $\kappa$ 가 안정적인 모습을 가진다. 층화 이단 추출에 있어서 가구와 인구특성에 따른 표본설계효과의 차이가 있었으며 일차추출단위가 추정량의 분산에 주는 영향력이 더 큰 모습을 볼 수 있었다. 층화의 효과는 권역별 동읍면 분류가 층화에 있어서 층 효율이 적절히 이루어진 모습을 확인할 수 있었다. 표본 개체 표본수의 할당 또한 개체 간 정보가 이질적이라면 이후 표본설계에 있어 정보를 반영하여 개체 표본수를 늘리는 방안도 모색할 수 있다. 기존의 가구조사에서 조사구를 통한 표본설계는 조사구의 노후화, 등록센서스에 의한 정보제공의 불편성을 고려하여 대안적 집락을 통한 표본설계를 확인해 보았다. 통계청 경제활동인구조사와 농촌진흥청의 농어업인복지실태조사에서는 조사구를 이용하여 층화삼단추출을 고려한 표본설계를 사용하고 있기에 추가적으로 연구해야 할 향후 과제로 집계구와 읍면동 정보에 층화삼단추출 적용에 대한 연구를 논의할 수 있다.

## 참고문헌

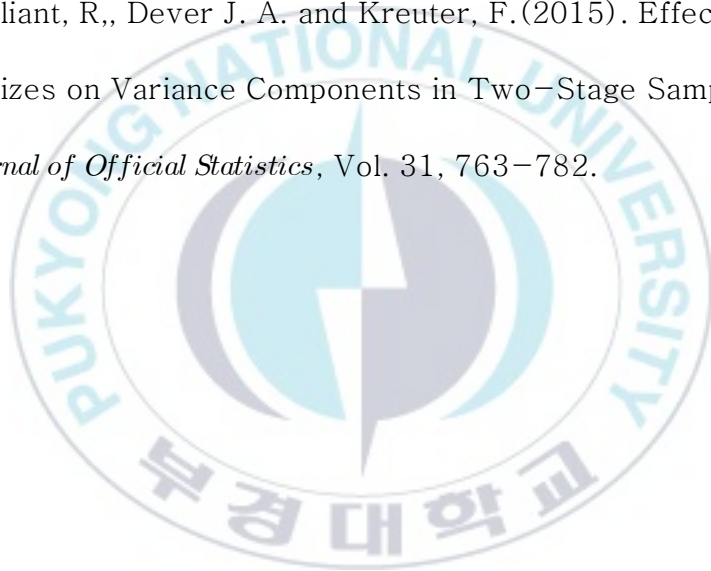
- [1] 박인호 (2016). 가구조사를 위한 이단추출 표본설계에서의 집락선택. 한국데이터정보과학회지, 27권 2호, 363-372.
- [2] 박진우, 윤석훈 (2008). 이상점을 고려한 다변량 층화. 한국통계학회. 응용통계연구, 21권 3호, 377-385.
- [3] 박진우, 변중석, 박민규 (2010). 국민건강영양조사 표본설계를 위한 추출틀 구축. 응용통계연구, 23권 5호, 923-932.
- [4] 최재혁 (2010). 다변량 층화표본설계. 성균관대학교학위논문.
- [5] 한국농촌경제연구원 (2015). 2015 식품소비행태조사 표본설계 및 가중치 산출방법 연구. 한국통계학회 위탁연구보고서.
- [6] Calinski, R. B. and Harabasz, J. (1974). "A dendrite method for cluster analysis". *Communications in statistics*, Vol. 3, 1-27.
- [7] Duda, R. O. and Hart, P. E. (1974). "Pattern Classification and Scene Analysis". *Journal of the Royal Statistical Society*, Vol. 137, 4 42-443.
- [8] Graham, K., Michael, B. J., Thanh, L. (1994). Estimating components of design effects for use in sample design.

*Household Surveys in Developing and Transition Countries*, 95–121.

[9] Lohr, S. L. (2010). *Sampling: design and analysis 2nd edition.*, Brooks/Cole, Boston.

[10] Sarndal, Carl-Erik., Swensson B., Wretman J. (1992).  
*Model Assisted Survey Sampling.* Springer Series in statistics, New York.

[11] Valliant, R., Dever J. A. and Kreuter, F. (2015). Effect of Cluster sizes on Variance Components in Two-Stage Sampling,  
*Journal of Official Statistics*, Vol. 31, 763–782.



## 부 록

### A.1 R코드

```
#층화에서 최적층수 결정
#CCC Rsquared Pseudo_F Pseudo_T
library(lattice)
a2<- read.csv("C:/Users/PSJ/Desktop/RCCC/CCCR11.csv", header=
T)
xyplot(CCC~Clusters| city,data=a2,layout=c(4,3),xlim=c(1,16),type
="o")
xyplot(Rsquared~Clusters| city,data=a2,layout=c(4,3),type="o")
xyplot(Pseudo_F~Clusters| city,data=a2,layout=c(4,3),type="o")
xyplot(Pseudo_T~Clusters| city,data=a2,layout=c(4,3),type="o")

#그림1의 r코드
n <- length(levels(cvse$method))
trellis.device(new = FALSE, theme = col.whitebg())
dotplot(stratum ~ value | characteristic,
        data = cvse, groups = method,
        layout = c(5, 1), aspect = 1.5,
        xlab = "분산구성 요소비교",
        key = list(points = Rows(trellis.par.get("superpose.symbol"),1:n),
        text = list(levels(cvse$method)), columns = n))
key = list(points = Rows(trellis.par.get("superpose.symbol"),1:
n)
```

## A.2 집계구 표본 설계 SAS코드

*/\*A<sub>hi</sub> <여기서 i는 집계구> n개의 (권역 또는 읍면동별) 1인가구 특성비율  
\*/*

```
%MACRO a;
```

```
    %DO i=1 %TO k;
```

```
    data frame5_&i.;
```

```
    set frame5_&i.;
```

```
    one_gagu_Ai=ga_shh/hh;
```

```
run;
```

```
%END;
```

```
%MEND;
```

```
%a
```

*/\*t<sub>hi</sub> 총계\*/*

```
%MACRO a;
```

```
    %DO i=1 %TO k;
```

```
    data frame5_&i.;
```

```
    set frame5_&i.;
```

```
    one_gagu_ti=hh*one_gagu_ai;
```

```
run;
```

```
    %END;
```

```
%MEND;
```

```
%a
```

*/\*각 (권역 또는 읍면동별) 전체 가구수 혹은 인구수\*/*

```
%MACRO a;
```

```
    %DO i=1 %TO k;
```

```
proc means data=frame5_&i. sum;
```

```

var hh(또는 pop);
output out=b&i. sum=N;

run;
%END;
%MEND;
%a

data table1;
    set b1-bn;

run;

/*  $T_{hi}$  모총합 */
%MACRO a;
    %DO i=1 %TO k;
proc means data=frame5_&i. sum;
    var one_gagu_ti;
    output out=c&i. sum=TU;

run;
    %END;
%MEND;
%a

data table2;
    set c1 c2 c3 c4 c5 c6;

run;

/* vtild 구하기 */
data tablenew;
    set tablenew;

```

```

vtild=Au*(1-Au)/Au**2;
run;

/*  $s_{u2i}^2$  구하기 */
%MACRO a;
    %DO i=1 %TO k;
data frame5_&i.;
    set frame5_&i.;
    one_gagu_s2=(hh/(hh-1))*one_gagu_Ai*(1-one_gagu_Ai);
run;
    %END;
%MEND;
%a

/* 추출확률  $p_i = N(i)/N$  구하기 */
%MACRO a;
    %DO i=1 %TO k;
data frame5_&i.;
    set frame5_&i.;
    N_1=집락1의 수;
    ...
    N_k=집락k의 수;
    m_1=표본집계구1의 수;
    ...
    m_k=표본집계구k의 수;
    one_gagu_pi=hh/N&i.;
run;
    %END;

```

```
%MEND;
```

```
%a
```

```
/*자체 가중  $w_{ik}$  */
```

```
%MACRO a;
```

```
    %DO i=1 %TO n;
```

```
data frame5_&i.;
```

```
    set frame5_&i.;
```

```
    one_gagu_wik=N&i./(m&i.*20);
```

```
run;
```

```
    %END;
```

```
%MEND;
```

```
%a;
```

```
/* $t_{pwr}$  구하기 */
```

```
%MACRO a;
```

```
    %DO i=1 %TO k;
```

```
data frame5_&i.;
```

```
    set frame5_&i.;
```

```
    one_gagu_t_pwrhat=1/m&i.*(one_gagu_ti/one_gagu_pi);
```

```
run;
```

```
proc means data=frame5_&i. sum;
```

```
    var one_gagu_t_pwrhat;
```

```
    output out=d&i. sum=sum2;
```

```
run;
```

```
    %END;
```

```
%MEND;
```

```

%a;

/* V(tpwr) 구하기 */
%MACRO a;
    %DO i=1 %TO k;
data frame5_&i.;
    set frame5_&i.;
    TU_1=총합추정치1의 값;
    ...
    TU_k=총합추정치k의 값;
one_gagu_t1=1/m&i.*(one_gagu_pi)*((one_gagu_ti/one_gagu_pi)-T
U&i.)**2;
run;
proc means data=frame5_&i. sum;
    var one_gagu_t1;
    output out=e&i. sum=sum2;
run;
%END;
%MEND;
%a;
%MACRO a;
    %DO i=1 %TO 6;
data frame5_&i.;
    set frame5_&i.;
one_gagu_t2=(hh**2)/(m&i.*one_gagu_pi*20)*(1-(20/hh))*one_gag
u_s2;
run;
proc means data=frame5_&i. sum;
    var one_gagu_t2;

```

```

        output out=ee&i. sum=sum3;
run;
%END;
%MEND;
%a;
data var1;
    set e1 e2 e3 e4 e5 e6;
run;
data var2;
    set ee1 ee2 ee3 ee4 ee5 ee6;
run;

data var_pwrhat;
    merge var1 var2;
run;
/*  $V(t_{pwr})$ 의 최종값*/
data var_pwrhat;
    set var_pwrhat;
    var_pwr=sum2+sum3;
run;
%MACRO a;
    %DO i=1 %TO 6;
/*  $SB^2$  구하기*/
data frame5_&i.;
    set frame5_&i.;
    BSTAR=(one_gagu_pi)*((one_gagu_ti/one_gagu_pi)-TU&i.)**2/TU&i
    .**2;
run;
proc means data=frame5_&i. sum;

```

```

var BSTAR;
output out=f&i. sum=BSTAR;
run;
/* SW*2 구하기*/
data frame5_&i.;
set frame5_&i.;
WSTAR=((hh**2)*one_gagu_s2/one_gagu_pi)/TU&i.**2;
run;
proc means data=frame5_&i. sum;
var WSTAR;
output out=g&i. sum=WSTAR;
run;
%END;
%MEND;
%a;
data B;
set f1 f2 f3 f4 f5 f6;
run;
data W;
set g1 g2 g3 g4 g5 g6;
run;
data BW;
merge B W;
run;
data all;
merge tablenew t_pwrhat var_pwrhat BW;
run;

```

