Journal of the Korean Data & Information Science Society
2016, **27**(2), 363–372

http://dx.doi.org/10.7465/jkdi.2016.27.2.363
한국데이터정보과학회지

가구조사를 위한 이단추출 표본설계에서의 집락선택[†]

박인호[1]

[1]부경대학교 통계학과

접수 2016년 2월 17일, 수정 2016년 3월 23일, 게재확정 2016년 3월 24일

요 약

우리나라 가구조사는 흔히 통계청의 조사구를 집락으로 사용한 이단추출의 자체가중 표본설계의 형태로 진행된다. 집락구조는 모집단내 개체변동성을 집락간과 집락내 분산으로 분해되기 때문에 이와 연관된 표본집락수와 집락내 표본수의 결정은 표본추정에 영향을 미치게 된다. 하지만 조사구의 규모, 노후화, 가구명부 접근불가 등의 여러가지 이유로 집계구와 같은 대안적 집락선택이 고려되기도 한다. 또한 2015 인구주택총조사부터는 전통적 가구방문조사 방식에서 행정자료를 이용한 등록센서스 형태로 바뀜에 따라 기존 조사구의 형태나 규모의 변경되어 구축되는 것으로 알려져 있다. 본 논문에서는 집락추출을 반영한 설계효과식을 통해 계통적 혹은 내포적 구성을 갖는 집락들의 선택이 주는 분산식 차이를 유도하고, 주어진 표본크기에서 동일한 분산을 갖는 집락구조별 표본할당에 대해 살펴보았다. 미국 매릴랜드주 앤어룬델 카운티 자료를 사용하여 우리나라 조사구와 집계구와 다소 유사한 사례연구를 포함하였다. 조사변수별로 집락통합이 주는 동일성 계수의 변화는 같지 않으며 이에 따라 집락구조에 따른 표본할당이 집락표본수와 더불어 종합적으로 고려되어야 할 것이다.

주요용어: 계통적 구조, 동질성 계수, 설계효과, 집락효과, 초모집단모형.

1. 서론

조사연구를 위한 표본설계에서는 흔히 다단계추출 (multistage sampling)이 고려된다. 집락추출 (cluster sampling)을 이용하면 조사를 위해 방문해야 하는 대상이나 지역을 제한할 수 있어 시간과 비용을 절감할 수 있게 된다. 개체단위의 표본추출틀은 존재하지 않고 집락단위의 표본추출틀만 존재할 경우에도 집락추출은 현실적으로 고려할 수 있는 대안이 된다.

가구조사의 경우에 일부 행정구역이나 통계구역을 집락으로 선택하고 해당구역에 대해서 가구명부를 구축하여 일부만을 추출하는 지역추출 (area sampling)이 흔히 사용된다. 우리나라 가구조사의 대부분은 통계청 인구주택총조사 (이하 인총)의 조사구 (enumeration district; ED)를 표집틀로 사용하여 일부을 선택하고 조사구내 가구를 다시 추출하는 이단추출에 의한 표본설계를 따른다 (Yoon 등, 2004). 조사구는 일반가구, 아파트가구 및 기타특성에 따라 분류되는데 통상 60가구를 기준한 근접 지역으로 구성된다 (Lee 등, 2006). 조사구가 가구조사의 집락으로 선택되는 이유는 승인통계에 대해서는 표본조사구내 가구명부가 (요청시 요도와 함께) 매우 저렴한 비용으로 제공되고 있기 때문인 것으로 판단된다.

다른 지역단위들이 다양한 이유에 의해 대안적 집락으로 고려되기도 한다. 먼저, 조사구 구성이 매 5년 주기의 인총과 함께 이루어지므로 시간적으로 간격이 있는 조사의 경우에는 대안적 집락선택을 통해 표본틀 노후화에 따른 비포함률을 축소하기도 하며 (Park 등, 2010; Kang 등, 2009), 표본틀 구성의

[†] 이 논문은 부경대학교 자율창의학술연구비 (2015)에 의하여 연구되었음.
[1] (48513) 부산광역시 남구 용소로 45, 부경대학교 통계학과, 부교수. E-mail: ipark@pknu.ac.kr

신뢰성이나 집락별 목표응답수의 관리가 어려운 경우 등의 여러가지 이유로 새롭게 가구조사 표집틀을 구성하기도 한다 (Ku 등, 2014). 또한 (승인제한 등의 이유로) 조사구내 가구명부를 제공받을 수 없는 경우, 행정단위, 국가기초구역나 집계구 등이 대안적 집락으로 고려되기도 한다 (Park 등, 2015).

집계구는 조사구와 유사하게 인총을 기반으로 공간통계의 기본단위로 이용하기 위해 작성한 별도의 블럭 (block)이다. 집계구는 평균 1.1 km^2의 면적에 평균적으로 200가구를 포함하고 있어 조사구에 비해 약 3.3배의 크기를 갖는다고 할 수 있다. 집계구는 조사구와는 달리 일반가구와 아파트를 함께 포함할 수 있어서 집락내 이질성이 크고 집락간 차이가 작을 것으로 예상된다. 집계구에 대한 정보는 통계청의 통계지리정보 사이트 (http://sgis.kostat.go.kr)를 통해 지도와 함께 세대수, 가구수, 1인 가구수, 아파트 등의 정보가 제공된다.

2015 인총부터는 전가구를 방문하여 조사하는 전통적 수행방식 대신 주민등록 및 건축물대장 등의 행정자료를 활용하고 약 20%의 가구만 표본조사하는 등록센서스로 변경되었다. 이에 따라 기존 조사구를 변경한 새로운 규모와 형태의 조사구를 구축하는 작업을 수행하고 있는 것으로 알려지고 있다 (Statistics Korea, 2014). 따라서 기존 2010 인총 조사구와 비교하여 대안으로 고려될 수 있는 집락에 대한 선택문제는 2015 인총 조사구를 사용할 수 있는 시점에서는 가구조사를 위한 표본설계의 중요한 고려사안이 될 수 있을 것이다.

본 논문에서는 가구조사를 위한 이단추출 표본설계에서 집락선택에 따른 정도수준의 변화에 대한 영향을 논의하고자 한다. 집락구조에 따라 집락내 개체가 얼마나 동질적인지가 결정되므로 어떠한 집락을 선택하는가는 표본설계의 효율성을 결정한다고 할 수 있다. 2절에서는 이단추출에 의한 표본설계를 간단히 소개하고 집락추출이 주는 분산에 대한 영향력을 집락효과의 측면에서 논의한다. 3절에서는 초모집단 모형의 가정을 통해 계층적 (hierachical) 혹은 내포적 (nested) 구성을 따르는 집락들의 선택이 주는 추정정도를 비교한다. 4절에서는 매릴랜드 지역모집단 사례를 통해 집락선택과 정도수준의 관계를 살펴본다. 5절에서는 조사구와 집계구의 선택에 따른 표본설계 효율성에 대해 논의하고 비용함수와 관련된 표본크기 결정에 대한 일반론과 더불어 향후 연구과제에 대해 간단히 언급한다.

2. 이단추출 표본설계 및 집락효과

2.1. 이단추출과 분산분해

M개의 개체로 이루어진 모집단 U로부터 크기 m의 표본 s를 이단추출에 의해 추출하는 표본설계를 고려하자. 이를 위해 모집단을 N개의 집락 U_i로 나누고 이중 n개 집락을 확률 p_i에 비례하여 추출하며, 표본집락내 총 M_i개의 개체중 m_i개를 단순확률로 추출한다고 하자. 단, $\sum_{i=1}^{N} p_i = 1$이다. 모평균 $\bar{Y} = Y/M$의 표본추정량은 다음과 같이 정의된다.

$$\bar{y}_{pps} = \frac{1}{M}\sum_{i=1}^{n}\frac{\hat{Y}_i}{np_i} = \sum_{i=1}^{n}\sum_{j=1}^{m_i} w_{ij}y_{ij}. \tag{2.1}$$

여기서 y_{ij}는 집락 i의 j번째 개체가 갖는 조사변수값이고 $Y_i = \sum_{j=1}^{M_i} y_{ij}$와 $Y = \sum_{i=1}^{N} Y_i$은 집락총합과 모총합을 각각 나타내며 집락총합은 $\hat{Y}_i = M_i m_i^{-1}\sum_{j=1}^{m_i} y_{ij}$으로 추정된다. 또한 $w_{ij} = (Mnp_im_i)^{-1}M_i$는 설계가중치 (design weight)가 된다.

논의를 단순화 하기 위해 표본설계에 대한 다음의 가정들을 추가로 고려한다.

(A1) 일차추출는 복원추출이고, 이차추출은 비복원추출

(A2) 동일한 집락표본크기 (모든 i에 대해, $m_i \equiv \bar{m}$)

(A3) 집락내 추출률은 매우 작음 (모든 i에 대해, $1 - m_i/M_i \approx 1$).

(A4) 집락추출확률은 집락크기에 비례함 (모든 i에 대해, $p_i \propto M_i$).

가정 (A1)과 (A2)은 평균추정량의 분산을 단순한 형태로 만들어 준다. 가정 (A2)와 (A4)을 만족하면, 식 (2.1)의 표본가중치는 $w_{ij} = (n\bar{m})^{-1}$으로 모든 개체에 대해 동일한 값을 갖는 자체가중 (self-weighting) 혹은 균등추출확률 (equal probability sampling method, *epsem*)의 표본설계가 되며 평균추정량 (2.1)은 $\bar{y}_{srs} = (n\bar{m})^{-1}\sum_{i=1}^{n}\sum_{j=1}^{\bar{m}} y_{ij}$이 된다.

가정 (A1)-(A4)하에서 표본추정량 $\bar{y}_{pps}$의 설계분산 (design variance)은 추출단계별 분산요소 (variance components)의 합으로 다음과 같이 유도된다.

$$V(\bar{y}_{pps}) = \frac{1}{n}S_B^2 + \frac{1}{n\bar{m}}S_W^2. \tag{2.2}$$

여기서 $S_B^2 = \sum_{i=1}^{N}(M_i/M)(\bar{Y}_i - \bar{Y})^2$는 집락(평균)간 분산을, $S_W^2 = (1/M^2)\sum_{i=1}^{N}(M_i/M)S_i^2$은 집락내 (개체)분산합을 각각 나타내며, $S_i^2 = (M_i-1)^{-1}\sum_{j=1}^{M_i}(y_{ij} - \bar{Y})^2$은 집락내 개체분산을 나타낸다 (Cochran, 1977). 분산분해식 (2.2)는 주어진 집락구조하에서 갖게 되는 분산요소을 이단추출에 따라 집락표본수 n과 집락평균표본개체수 $\bar{m}(= m/n)$을 어떻게 정하는가에 따라 추정량의 표본분산이 달라질 수 있음을 보여준다.

2.2. 집락효과와 설계효과

집락내 분산에 대한 집락간 분산이 갖는 상대적 크기는 동질성 계수 (measure of homogeneity)로 다음과 같이 정의된다.

$$\rho_D = \frac{S_B^2}{S_B^2 + S_W^2}. \tag{2.3}$$

식 (2.3)의 동질성 계수는 (모집단) 개체분산이 집락구성에 따라 집락간 변동차이로 얼마나 많이 설명되는가를 나타낸다고 할 수 있다. 집락내 분산에 비해 집락간 분산이 크다면 동질성 계수는 큰 값을 갖게되며, 반대로 집락간 분산이 작다면 동질성 계수는 작은 값을 갖는다. 다시말해, 집락내 개체들이 유사할수록 동질성 계수값이 크고, 집락내 개체들이 이질적일수록 동질성 계수값은 작다. 예를들면, 집락내 개인소득은 차이가 많이 나고 집락간 평균소득은 서로 근소한 차이만 난다면 동질성 계수는 0에 가까운 값을 갖게 되지만, 집락내 개인소득은 매우 동질적인 반면에 집락간 평균소득이 매우 이질적이라면 동질성 계수는 1의 값을 갖게 된다. 동질성 계수는 집락규모에 영향을 받는 경향이 있는 것으로 알려져 있다. 집락규모가 작을수록 집락내 개체는 좀 더 동질적인 경향을 띤다 (Hansen 등, 1953a; Lohr, 2010; Valliant 등, 2013). 예를 들어, 시군구보다는 읍면동이, 읍면동보다는 조사구가 집락내 개체들이 좀 더 동질적이다.

분산식 (2.2)는 동질성 계수를 이용하면 다음과 같이 표현될 수 있다 (Valliant 등, 2013; Hansen 등, 1953b).

$$V(\bar{y}_{pps}) = \left(\frac{S_U^2}{n\bar{m}}\right)\kappa_D[1 + \rho_D(\bar{m} - 1)]. \tag{2.4}$$

여기서 $S_U^2 = (M-1)^{-1}\sum_{i=1}^{N}\sum_{j=1}^{M_i}(y_{ij} - \bar{Y})^2$는 개체분산이고 $\kappa_D = (S_B^2 + S_W^2)/S_U^2$이다. Valliant 등 (2013)은 집락크기가 모두 동일하고 집락수와 집락크기가 모두 큰 경우이거나 크기비례 집락추출인 경우에 식 (2.4)의 κ_D은 근사적으로 1의 값을 가짐을 각각 보여주고 있다. 정의상 개체분산이 집락간 분산과 집락내 분산의 합으로 분해될 때 $\kappa_D \doteq 1$을 만족한다.

만약 표본 s를 단순확률로 추출한다면 평균추정량은 $\bar{y}_{srs}$이 되고 분산식은 $V(\bar{y}_{srs}) = (n\bar{m})^{-1}S_U^2$이 된다. 따라서 식 (2.4)를 다음과 같이 표현한다면 이단추출의 표본분산이 단순추출의 표본분산 보다 얼

마나 더 증가 (혹은 감소)하였는가를 나타낼 수 있다.

$$V(\bar{y}_{pps}) = V(\bar{y}_{srs}) \times deff(\bar{y}_{pps}). \tag{2.5}$$

여기서 $deff(\bar{y}_{pps}) = V(\bar{y}_{pps})/V(\bar{y}_{srs}) = \kappa_D[1 + \rho_D(\bar{m} - 1)]$는 설계효과 (design effect, *deff*)에 해당하며 단순추출과 비교하여 집락구조 ρ_D, κ_D와 집락내 표본크기 $\bar{m}$의 선택이 주는 표본설계의 효율성을 나타낸다. $\kappa_D \doteq 1$인 경우, 설계효과는 일반적으로 알려져 있는 $deff(\bar{y}_{pps}) = 1 + \rho_D(\bar{m} - 1)$이 된다. 설계효과에 관한 좀 더 상세한 논의는 Park (2015), Park (2014), Heo (2013), Lee (2012), Kalton 등 (2005), Park과 Lee (2004), Kish (1965)를 참고할 수 있다.

초모집단 모형 (superpopulation model)을 통해서도 이단추출에 따른 집락효과의 평가가 가능하다. 주어진 집락구조에서 집락 i의 j번째 개체의 조사특성 y_{ij}는 공통평균 μ와 집락 i와 개체 (ij)가 갖는 상호 독립적인 램덤효과가 더해지는 일원랜덤모형 (one-ways random effects model)을 가정할 수 있다.

(C1) $y_{ij} = \mu + \alpha_i + \epsilon_{ij}$.

여기서 $(\alpha_i, \epsilon_{ij})(1 \le i \le n_\alpha, 1 \le j \le \bar{m}_\alpha)$는 집락과 개체의 램덤효과들로 평균 0과 분산 σ_α^2와 σ_ϵ^2를 각각 갖게된다. 모형 (C1)에 의해 정의되는 동질성 계수는 다음과 같다 (Lohr, 2010).

$$\rho_\alpha = \frac{\sigma_\alpha^2}{\sigma_\alpha^2 + \sigma_\epsilon^2}. \tag{2.6}$$

모형 (C1)에 대한 표본설계는 표본집락수 n_α와 집락평균표본개체수 $\bar{m}_\alpha$을 가정하므로 평균추정량을 $\bar{y}_\alpha = (n_\alpha \bar{m}_\alpha)^{-1} \sum_{i=1}^{n_\alpha} \sum_{j=1}^{\bar{m}_\alpha} y_{ij}$으로 표기하고자 한다. 만약 표본추출이 단순표집이었다면 모형 (C1) 대신 $y_{ij} = \mu + \epsilon_{0ij}$이고 오차항 ϵ_{0ij}의 평균과 분산이 $(0, \sigma_y^2)$임을 가정할 수 있고 평균추정량의 갖는 모형기반 설계효과 (model-based design effect) 는 $deff(\bar{y}_\alpha) = 1 + \rho_\alpha(\bar{m}_\alpha - 1)$으로 유도된다 (Gabler 등, 2006; Skinner 등, 1989). 따라서 표본평균의 분산식은 식 (2.5) 혹은 (2.6)과 유사하게 다음과 같이 표현할 수 있다.

$$V(\bar{y}_\alpha) = \frac{\sigma_y^2}{n_\alpha \bar{m}_\alpha}[1 + \rho_\alpha(\bar{m}_\alpha - 1)]. \tag{2.7}$$

여기서 $\sigma_y^2 = \sigma_\alpha^2 + \sigma_\epsilon^2$이다. 설계기반 (design-based)의 분산식 (2.5)와는 달리 모형기반 (model-based)의 분산식 (2.7)은 κ_D가 생략되어 간단한 형태가 된다.

3. 집락선택에 따른 정도수준 비교

3.1. 집락효과 비교모형

고려 가능한 집락들이 서로 계통적 혹은 내포적 구조를 갖는다고 가정한다. 편의상 앞 절에서 정의한 집락을 기초집락 (elementary cluster)이라 칭하고 계통적 상위구조를 갖는 집락을 통합집락 (combined cluster)이라 부르고 다음과 같이 정의한다.

$$U_g = U_{i_1} \cup \cdots \cup U_{i_g}. \tag{3.1}$$

즉, 통합집락 g는 i_g개 기초집락으로 구성되고, $N = \sum_{g=1}^{G} i_g$와 $M_g = M_{i_1} + \cdots + M_{i_g}$을 만족한다. 통합집락을 사용한 표본설계에서는 n_β의 표본집락을 집락크기 M_g에 비례확률로 또한 집락내에서는 $\bar{m}_\beta$개 개체를 단순확률로 각각 추출한다고 가정하면 표본평균은 $\bar{y}_\beta = (n_\beta \bar{m}_\beta)^{-1} \sum_{g=1}^{n_\beta} \sum_{j=1}^{\bar{m}_\beta} y_{gj}$이 된다. 통합집락을 반영한 초모집단 모형은 모형 (C1)와 유사하게 다음과 같이 정의할 수 있다.

(C2) $y_{gj} = \mu + \beta_g + \eta_{gj}$.

여기서 $(\beta_g, \eta_{gj})(1 \le g \le n_\beta, 1 \le j \le \bar{m}_\beta)$는 통합집락과 개체의 램덤효과로 평균 0과 분산 σ_β^2와 σ_η^2를 각각 갖게 되며 동질성 계수는 다음과 같이 정의된다.

$$\rho_\beta = \frac{\sigma_\beta^2}{\sigma_\beta^2 + \sigma_\eta^2} \tag{3.2}$$

3.2. 정도수준 비교

조사특성 y의 모형분산은 집락선택과는 무관하게 동일하지만 규모가 큰 통합집락의 변동성은 기초집락에 비해 다소 큼으로 다음의 가정을 고려할 수 있다.

(C3) $\sigma_\beta^2 = \gamma\sigma_\alpha^2 \quad (0 < \gamma < 1)$.

결과 1 조건 (C1)-(C3)을 만족하면, $\sigma_\eta^2 = (1-\gamma)\sigma_\alpha^2 + \sigma_\epsilon^2$이고 $\rho_\beta = \gamma\rho_\alpha$이다.

만약 집락구조와 관계없이 전체표본크기가 m으로 정해져 있다면, 집락규모에 따라 집락표본수와 집락평균표본개체수를 조정하는 다음의 가정을 고려할 수 있다.

(C4) 주어진 $m > 0$에 대해, $\lambda > 0$이 존재하여 $n_\alpha = \lambda n_\beta$과 $\bar{m}_\beta = \lambda\bar{m}_\alpha$을 만족한다.

결과 2 조건 (C1)-(C4)을 만족하면, 표본평균 $\bar{y}_\beta$의 분산은 다음과 같이 표현된다.

$$V(\bar{y}_\beta) = \frac{\sigma_y^2}{n_\alpha\bar{m}_\alpha}[1 + \gamma\rho_\alpha(\lambda\bar{m}_\alpha - 1)]. \tag{3.3}$$

결과 3 조건 (C1)-(C4)을 만족하면, $\lambda \le \lambda_{\bar{m}_\alpha}(\gamma)$일때 $V(\bar{y}_\beta) \le V(\bar{y}_\alpha)$이고, $\lambda < \lambda_{\bar{m}_\alpha}(\gamma)$일때 $V(\bar{y}_\beta) > V(\bar{y}_\alpha)$이다. 여기서,

$$\lambda_{\bar{m}_\alpha}(\gamma) = \gamma^{-1}\left(1 - \bar{m}_\alpha^{-1}\right) + \bar{m}_\alpha^{-1}. \tag{3.4}$$

조건 (C1)-(C4)를 만족한다면, 결과 1과 결과2에 의해 통합집락과 기초집락을 사용한 표본평균의 분산차이는 $V(\bar{y}_\beta) - V(\bar{y}_\alpha) = (n_\alpha\bar{m}_\alpha)^{-1}\sigma_y^2\rho_\alpha[\gamma(\lambda-1) - (\bar{m}_\alpha - 1)]$이므로, 식 (3.4)이 성립함을 보일 수 있다.

결과 3은 표본크기가 m으로 주어지고 기초집락 대비 통합집락의 동질성 계수가 $\gamma(=\rho_\beta/\rho_\alpha)$의 비율로 감소한다면, 통합집락의 집락표본과 집락평균표본개체수는 $\lambda_{\bar{m}_\alpha}(\gamma)$만큼 조정되어야 함을 보여준다. 즉, $n_\alpha = \lambda_{\bar{m}_\alpha}(\gamma)n_\beta$이고 $\bar{m}_\beta = \lambda_{\bar{m}_\alpha}(\gamma)\bar{m}_\alpha$이다. 기초집락의 집락평균표본개체수 $\bar{m}_\alpha$가 충분히 크다면 표본조정계수는 $\lambda_\infty(\gamma) = \gamma^{-1}$으로 집락규모나 집락평균표본개체수에 관계없이 동질성 계수의 증감에 반비례하도록 결정된다. 예를들면, 기초집락 대비 통합집락의 동질성 계수가 50% ($\gamma = 0.5$) 정도 작다면 집락평균표본개체수는 2배 ($\lambda_\infty(.5) = 2$)로 늘일 때 집락구조에 관계없이 평균추정량의 분산이 같게 된다.

4. 사례연구

4.1. 매릴랜드 지역모집단 자료

계통적 구조를 갖는 집락단위간의 선택이 평균추정량의 정도에 주는 영향을 살펴보기 위해, Valliant 등 (2013)이 2000년 미국 센서스 자료에 근거하여 구성한 미국 매릴랜드주 앤어룬델 카운티 자료인 `MDarea.pop`를 고려하였다. 총 403,997 명을 포함하며 블럭그룹 (block group)과 트랙 (tracts)의 지역

단위 구분자를 포함하는데, 이는 각각 기초집락과 통합집락으로 간주할 수 있다. 트랙은 평균 3.2개의 블럭그룹으로 구성되는데 작게는 1개에서 많게는 6개의 블럭그룹을 포함한다. Table 4.1은 블럭그룹과 트랙의 크기분포를 나타내고 있다. 블럭그룹은 평균 1316명을 포함하며 최소 52명에서 최대 4744명을 포함하는 반면, 트랙은 평균 4253명을 포함하며 최소 86명이고 최대 13579명을 포함한다.

Table 4.1 Comparisons of cluster sizes

Clusters	Average	Standard Deviation	CV	Median	Min	Max
Block Group	1316.0	761.0	0.578	1240	52	4744
Tract	4252.6	2167.7	0.510	4132	86	13579

`MDarea.pop` 자료는 모수모형을 이용하여 발생한 5개 변수 `y1,y2,y3`, `ins.cov`, `hosp.stay`들을 포함하고 있다. `y1, y2, y3`은 연속형 변수들이고 `ins.cov`와 `hsp.stay`는 0-1의 값을 갖는 지시변수들이다. Table 4.2는 다섯개 변수의 모집단 평균, 표준편차, 상대표준편차, κ_D과 동질성 계수를 보여준다. 블럭그룹과 트랙의 κ_D값은 모두 근사적으로 0이므로 해당 ρ_D는 모형근거의 동질성 계수의 값으로 간주하였다. 두 동질성 계수의 비 γ는 0.4386에서 0.6432의 범위를 나타내는데 블럭그룹에 비해 트랙의 동질성 계수가 작아서 집락내 개체간 이질성이 큼을 알 수 있다.

Table 4.2 Population mean, standard deviation and measures of homogeneity by cluster-type

Variable	$\bar{Y}$	S_U	CV_Y	Block Group		Tract		γ
				κ_D	ρ_D	κ_D	ρ_D	
y1	69.71	84.3120	1.2094	1.0007	0.0109	1.0002	0.0063	0.5725
y2	7.66	7.7192	1.0081	1.0007	0.0173	1.0002	0.0106	0.6093
y3	87.52	29.4999	0.3371	1.0006	0.1857	1.0002	0.1194	0.6432
ins.cov	0.79	0.4052	0.5109	1.0007	0.0148	1.0002	0.0069	0.4700
hosp.stay	0.07	0.2584	3.5914	1.0008	0.0039	1.0002	0.0017	0.4386

4.2. 정도평가

집락선택에 따른 정도수준 비교를 위해 블럭그룹의 집락평균표본수를 $\bar{m}_\alpha = 10$으로 고정하고 트랙의 집락평균표본수 $\bar{m}_\beta$는 5개에서 40개로 5개씩 변경하여 상대효율을 계산하였다. 평균추정량의 상대효율 (relative efficiency)은 다음과 같이 정의된다.

$$RE(\bar{y}_\alpha, \bar{y}_\beta) = \frac{V(\bar{y}_\beta)}{V(\bar{y}_\alpha)} = \frac{1+\rho_\beta(\bar{m}_\beta - 1)}{1+\rho_\alpha(\bar{m}_\alpha - 1)}.$$

여기서 $V(\bar{y}_\alpha)$와 $V(\bar{y}_\beta)$는 식 (2.7)과 (3.3)에 각각 정의되었다.

Table 4.3은 다섯 변수가 갖는 동질성 계수의 상대크기 ($\gamma^{-1} = \rho_\alpha/\rho_\beta$)와 $\bar{m}_\beta$값에 대한 상대효율을 나타내고 있다. 동질성 계수의 상대크기는 1.555에서 2.280의 범위를 갖는다. 상대효율이 1의 값을 갖기 위해서는 변수 `y1`와 `y2`는 $\bar{m}_\beta = 15$과 20 사이의 값을 가질 때, 변수 `y3`은 $\bar{m}_\beta = 15$개, `ins.cov, hosp.stay`는 대략 $\bar{m}_\beta = 25$개이어야 함을 알 수 있다. 동일한 정도수준을 얻기 위해서는 집락평균표본개체수가 γ^{-1}와 유사한 비율로 정해질 때 가능함을 알 수 있다. Figure 4.1은 동질성 계수의 변동비 γ^{-1}에 대해 동일한 정도수준을 갖게하는 집락평균표본수의 배율에 대해 나타내고 있다. 집락평균표본수 $\bar{m}_\alpha$가 증가할수록 $\lambda_\infty(\gamma) = \gamma^{-1}$로 점진적으로 수렴함을 확인할 수 있다.

Table 4.4는 표본크기를 3000으로 할때 동일한 정도수준을 갖게하는 집락표본수와 집락평균표본개체수의 조합을 개별 변수별로 산출한 결과를 정리해 주고 있다. 먼저, 블럭그룹을 채택하였을때 표본집락수와 집락평균표본수를 각각 300 과 10으로 하였을 경우에 트랙 사용시 동일한 정도수준을 갖기 위해

서는 표본집락수는 15개에서 22개의 수준에서 정하되 해당 표본집락수에 따라 반비례하여 집락평균표본수를 200개에서 1390개의 수준에서 결정할 수 있다. 또한 블럭그룹의 표본집락수와 집락평균표본수를 각각 150과 20으로 하였을 때 트랙 사용시 동일한 정도수준을 갖기 위해서는 표본집락수는 68개에서 98개로 정하되 집락평균표본수를 역시 반비례적으로 44개에서 31개의 수준에서 결정할 수 있다.

Table 4.3 Comparisons of relative efficiency by cluster-type

Variable	γ^{-1}	$\bar{m}_\beta$							
		5	10	15	20	25	30	35	40
y1	1.747	0.933	0.962	0.990	1.019	1.047	1.076	1.104	1.133
y2	1.641	0.902	0.947	0.993	1.039	1.084	1.130	1.176	1.221
y3	1.555	0.553	0.777	1.000	1.224	1.447	1.671	1.895	2.118
ins.cov	2.128	0.907	0.938	0.968	0.999	1.030	1.060	1.091	1.122
hosp.stay	2.280	0.972	0.981	0.989	0.997	1.006	1.014	1.022	1.031

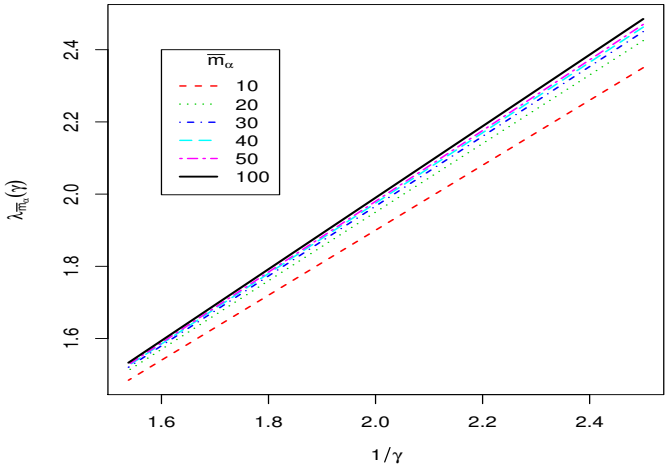

Figure 4.1 Sample adjustment factor for combined clusters to obtain the equal variance

Table 4.4 Measure of homogeneity and sample sizes for element and combined clusters by survey

Variable	y1	y2	y3	ins.cov	hosp.stay
γ	0.572	0.609	0.643	0.470	0.439
γ^{-1}	1.747	1.641	1.555	2.128	2.280
		($m = 3000$,	$n_\alpha = 300$	$\bar{m}_\alpha = 10$)	
$\lambda_{\bar{m}_\alpha}(\gamma)$	1.672	1.577	1.499	2.015	2.152
n_β	179	190	200	149	139
$\bar{m}_\beta$	17	16	15	20	22
		($m = 3000$,	$n_\alpha = 150$	$\bar{m}_\alpha = 20$)	
$\lambda_{\bar{m}_\alpha}(\gamma)$	1.709	1.609	1.527	2.071	2.216
n_β	88	93	98	72	68
$\bar{m}_\beta$	34	32	31	41	44

5. 논의

본 논문에서는 가구조사를 위한 이단추출의 표본설계에서 계통적 혹은 내포적 구조를 갖는 집락구조들 간의 선택에 따른 평균추정량의 정도수준에 대한 영향을 평가하였다. 특히 자체가중을 위해 흔히 사용하는 크기비례집락추출과 집락내 고정크기의 단순추출에 의한 표본설계를 고려하였다. 평균추정량의 분산은 설계기반이나 모형기반의 접근 모두에서 집락구조와 집락표본규모에 의해 결정되며 이는 단순확률추출과 비교하는 상대분산 개념의 설계효과인 $deff(\bar{y}) = 1 + \rho(\bar{m} - 1)$에 잘 반영되어 나타난다. 통합집락은 기초집락에 비해 집락내 개체간 이질성이 큼으로 동질성 계수가 작다. 따라서 집락선택에 관계없이 동일수준의 분산을 목표로 한다면 기초집락과 비교하여 집락내 표본개체수를 적절히 늘리되 집락표본수를 줄여야 함을 알 수 있었다. 물론 집락표본수는 단순계산 $m = n\bar{m}$에 의해서만 결정되는 것이 아닌 복잡조사 자료분석의 자유도 (degree of freedom, *df*)라는 측면에서 오차한계의 크기로 고려되어야 한다. 예를 들면, 95% 신뢰수준에서 갖는 평균추정량의 오차한계는 $t_{0.975,df}\sqrt{V(\bar{y})}$으로 근사되는데 이때 자유도 *df*는 집락표본수-1으로 오차한계의 크기에 영향을 주게 된다 (Korn과 Graubard, 1999).

우리나라 가구조사를 위해 고려되는 조사구와 집계구는 각각 평균 60가구와 200가구로 후자가 약 3.3정도 크다. 통계청 2015년 추계인구 기준으로 가구당 2.7명을 감안하면 조사구와 집계구는 평균 162명과 540명을 포함하므로 블럭그룹과 트랙에 비교할 때 1/10정도의 크기이고 상대적 크기는 유사함을 알 수 있다. 조사구와 집계구에 대한 세부자료가 주어진다면 두 집락구조의 선택이 주는 정도차이에 대한 상세한 연구가 가능할 것이다. 또한 2015 인총에서 새롭게 결정되어질 조사구 확정에 대해서도 실제 변수들을 활용한 정도평가의 진행도 가능할 것으로 판단된다.

하지만 앞선 논의에서는 조사비용을 전혀 고려되지 않았다. 집락선택에서 비용 및 분산을 동시에 고려할 수 있다면 보다 현실성을 반영한 가이드를 제시할 수 있을 것이다. 예로, 이단추출에서 고정된 조사비용 C_0와 표본집락당 조사비용 C_1 과 개체당 조사비용 C_2이 해당 표본수 만큼 곱하여 더해지는 매우 단순한 선형비용함수 (linear cost function)을 다음과 같이 가정할 수도 있다.

$$C_\psi = C_{\psi 0} + C_{\psi 1} n_\psi + C_{\psi 2} n_\psi \bar{m}_\psi .$$

여기서 ψ는 집락구조 ψ를 뜻하며 n_ψ와 $\bar{m}_\psi$는 해당 집락구조에 의한 표본집락수와 집락평균표본수를 각각 나타낸다. 조사비용함수를 근거로 결과 3의 정도비교에 더불어 동일 정도를 주는 표본설계가 갖는 총비용도 비교 가능윌옘. 더 나아가 집락선택별로 총 조사비용 C_ψ이 주어졌을때 표본추정량의 상대표준오차를 최소화하는 표본크기를 유도할 수도 있고 또한 주어진 정도수준을 만족하는 최소비용의 표본크기를 유도하여 비교할 수 있다. 이와 관련한 상세한 논의는 Valliant 등 (2013)와 Hansen 등 (1953a)을 참고할 수 있다. 하지만 위의 비용함수는 지나치게 단순한 형태일 수도 있으며 집락추출과 연계하여 좀 더 현실을 감안한 다양한 비용함수를 고려할 수 있을 것이다. 조사비용에 대한 상세한 논의는 Groves (1989)를 참고할 수 있다.

References

Cochran, W. G. (1977). *Sampling techniques*, 3rd ed., John Wiley & Sons, New York.

Gabler, S., Häder, S. and Lynn, P. (2006). Design effects for multiple design samples. *Survey Methodology*, **32**, 115-120.

Groves, R. M. (1989). *Survey errors and survey costs*, John Wiley & Sons, New York.

Hansen, M. H., Hurwitz, W. N. and Madow, W. G. (1953a). *Sample survey methods and theory. Volume 1: Methods and applications*, Wiley, New York.

Hansen, M. H., Hurwitz, W. N. and Madow, W. G. (1953b). *Sample survey methods and theory. Volume 2: Methods and applications*, Wiley, New York.

Heo, S. (2013). Error cause analysis of Pearson test statistics for k-population homogeneity test. *Journal of the Korean Data & Information Science Society*, **24**, 815-824.

Kalton, G., Brick, J. M. and Le, T. (2005). *Estimating components of design effects for use in sample design. In household sample surveys in developing and transition countries*, (Sales No. E.05.XVII.6). Department of Economic and Social Affairs, Statistics Division, United Nations, New York.

Kang, H., Park, S., Kim, J., Kim, I., Lee, D., Hwang, J. and Park, M. (2009). A case study on the construction of the sampling frame and sampling design for 2008 Seoul survey. *Survey Research*, **10**, 157-172.

Kish, L. (1965). *Survey sampling*, John Wiley & Sons, New York.

Korn, E. L. and Graubard, B. I. (1999). *Analysis of health surveys*, Wiley, New York.

Ku, M. J, Kim, S., Kim, H. Y. and Kim, J. (2014). Creating the household sampling frame: The Korean general social survey(KGSS). *Survey Research*, **15**, 153-174.

Lee, H. (2012). How should one find out the contributions to the design effect (variance) made by each of the design components (stratification, clustering, weighting) of a complex sample design? *Survey Statistician*, **66**, 16-20.

Lee, K., Lee, M. J., Seo, U. S. and Byun, M. R. (2006). Methods used in determining enumeration districts in the 2005 population and housing census. *Survey Research*, **7**, 109-129.

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

Park, I. (2014). A study on design effect models for complex sample survey. *Journal of the Korean Data & Information Science Society*, **24**, 523-531.

Park, I. (2015). Understanding complex design features via design effect models. *The Korean Journal of Applied Statistics*, **28**, 1-9.

Park, I. and Lee, H. (2004). Design effecs for the weighted mean and total estimators under complex survey sampling. *Survey Methodology,* **30**, 183-193.

Park, I., Son, C. K. and Shin, J. (2015). *Sampling design and weighting for the 2015 consumer behavior survey for food*, The Korean Statistical Society, Seoul.

Park, J., Byun, J. and Park, M. (2010). Construction of samplng frames for the 5th Korean national health and nutrition examination survey. *The Korean Journal of Applied Statistics,* **23**, 923-932.

Skinner, C. J., Holt, D. and Smith, T. M. F. (Eds.) (1989). *Analysis of complex survey*, Wiley, New York.

Statistics Korea (2014). *Launching the household and housing survey for successful register-census*, Press Release, Daejeon.

Valliant, R., Dever, J. A. and Kreuter, F. (2013). *Practical tools for designing and weighting survey samples*, Springer, New York.

Yoon, Y. O., Kim, K. Y. and Lee, M. H. (2004). Redesigning KNSO's household survey sample. *Survey Research*, **5**, 103-130.

Journal of the Korean Data & Information Science Society
2016, **27**(2), 363–372

http://dx.doi.org/10.7465/jkdi.2016.27.2.363
한국데이터정보과학회지

Choosing clusters for two-stage household surveys[†]

Inho Park[1]

[1]Department of Statistics, Pukyong National University

Received 17 February 2016, revised 23 March 2016, accepted 24 March 2016

Abstract

Two-stage sample designs are commonly used for household surveys in Korea using as clusters the enumeration districts (EDs). Since clustering decomposes the population variation into within- and between-cluster variations, the sample sizes allocated in stages can affect the overall precision. Alternative clusters are often considered due to diverse reasons such as the EDs' limitation in size, being out-of-date, and in-assessibility to their household lists. In addition, the EDs are currently under development by the Statistics Korea as an joint effort toward their transition from the traditional practice to the register census from 2015. We present an approach for evaluating the difference in the precision of the mean estimators of the sets of the cluster units in between a hierachical and nested form, where the design effect is used to reflect the effect of the clustering and the sample allocation. We also demonstrate our approach using the U.S. Census counts from the year 2000 for Anne Arundel County in Maryland. Our research shows that the within-cluster variance can be significantly different for survey variables and thus the choice of cluster units and the associated sample allocation scheme should reflect the corresponding variance decomposition due to clustering.

Keywords: Clustering effect, design effect, hierachical structure, measure of homogeneity, superpopulation model.

[†] This work was supported by a research grant of Pukyong National University (2015).
[1] Associate professor, Department of Statistics, Pukyong National University, Busan 48513, Korea. E-mail: ipark@pknu.ac.kr