The Korean Journal of Applied Statistics (2015) DOI: http://dx.doi.org/10.5351/KJAS.2015.28.6.1217
28(6), 1217–1225

Understanding Complex Design Features via Design Effect Models

Inho Park[a,1]

[a]Department of Statistics, Pukyong National University

(Received October 29, 2015; Revised November 13, 2015; Accepted November 13, 2015)

Abstract

Survey research, data is commonly collected through a sample design with complex design features that allow the relative efficiency on the precision of an estimator to be measured using the concept of the design effect compared to simple random sampling as a reference design. This concept is most useful when the design effect can be expressed as a function of various design features. We propose a design effect formula suitable under a stratified multistage sampling by generalizing Gabler *et al.* (1999, 2006)'s approaches for multistage sampling. Its use can either guide improvement in the design efficiency when in design stage or enable the evaluation of the adopted design features afterwards.

Keywords: stratified multistage sampling, haphazard weighting, intracluster correlation coefficient, mixed effect model, effective sample size

1. 서론

조사연구를 위한 자료수집은 집락화(clustering), 불균등가중(unequal weighting), 층화(stratificaiton)와 같은 복잡한 설계요소를 포함한 복잡표본설계(complex sample design)를 통해 이루어진다. 추정량에 대해 설계요소가 갖는 효율성은 주로 단순확률추출을 기준으로 비교한 설계효과(design effect)라는 개념을 통해 평가할 수 있다. 설계효과는 Cornfield (1951)에 의해 처음으로 소개되었고 Kish (1965)를 통해 널리 알려지기 시작한 후 그의 일련의 논문 (Kish, 1987, 1992)을 통해 지속적으로 연구발표되었다.

표본설계 $p(s)$ (혹은 p)에 의해 주어지는 개체 k의 조사값 y_k과 표본가중치 w_k로 계산되는 가중표본평균 $\bar{y}_p = \sum_{k\in s} w_k y_k / \sum_{k\in s} w_k$은 다음과 같이 정의되는 설계효과(design effect 혹은 deff)를 갖게 된다.

$$\mathrm{Deff}_p\left(\bar{y}_p\right) = \frac{V_p\left(\bar{y}_p\right)}{V_{srs}\left(\bar{y}\right)}. \tag{1.1}$$

여기서 $V_p(\bar{y}_p)$는 가중표본평균 $\bar{y}_p$의 분산을 나타내며, $V_{srs}(\bar{y})$는 단순표본평균 $\bar{y} = m^{-1}\sum_{k=1}^{m} y_k$이 가상적인 단순표본추출(srs)에서 갖게 되는 분산이다.

This work was supported by a research grant of Pukyong National University (2014).
[1]Department of Statistics, Pukyong National University, Yongso-ro 45, Nam-gu, Busan 48513, Korea.
E-mail: ipark@pknu.ac.kr

표본설계와 관련하여 또 하나의 유용한 개념은 유효표본크기(effective sample size)이다. 이는 표본설계 $p(s)$에서 (주로 응답수를 의미하는) 표본크기 m으로 얻게 되는 분산 $V_p(\bar{y}_p)$과 동일한 크기의 분산 $V_{srs}(\bar{y})$을 주는 단순확률추출 하에서의 표본크기 m_{eff}를 말하는데 다음과 같이 정의된다.

$$\frac{m}{m_{eff}} = \mathrm{Deff}_p(\bar{y}_p) \tag{1.2}$$

설계효과 혹은 유효표본크기의 개념을 이용하면 단순확률추출에서 갖는 간단한 분산식을 이용하여 복잡한 표본설계에서 필요한 표본크기를 결정할 수 있게 된다. 또한 단순확률추출에서 유도된 전통적 통계량을 설계효과를 통해 적절히 조정하므로 복잡표본설계에서도 타당성을 확보한 통계적 추론을 만들 수도 있게 된다 (예로, Rao와 Scott, 1987; Korn과 Graubard, 1999).

하지만 앞서 언급한 설계효과의 유용성은 복잡설계요소의 함수 형태로 표현될 수 있을 때에 극대화될 수 있다. 예를들어, Rust와 Broene (2010)은 다음과 같이 지적하고 있다: "설계효과의 개념은 할 수만 있다면 모집단의 구조와 설계요소들이 갖는 개별적 영향을 명백히 나타낼 수 있는 함수적 형태로 표현될 때 그 유용성은 극대화 된다. 이렇게 될때 향후 적용시 설계효율을 증진시킬 수 있는 지침으로 사용될 수 있기 때문이다."

본 연구에서는 먼저 기존 연구들이 제시하고 있는 설계효과모형에 대해 간략히 살펴본 후, 층화다단추출의 표본설계에서 적용될 수 있는 설계효과모형을 제시하고자 한다. 제시된 설계효과모형과 기존 연구들에서 제시된 모형들간의 관계성을 논의하며 적용사례를 살펴본다. 2절에서는 집락과 불균등가중치를 반영한 Kish (1965, 1987, 1992)와 Gabler 등 (1999)의 기본적 연구를 정리한다. 또한 다수의 다단추출이 혼재된 표본설계에 적용될 수 있는 Gabler 등 (2006)의 확장된 설계효과모형을 살펴본다. 3절에서는 층화다단추출의 표본설계에서 적용이 가능한 일반화 설계효과모형을 제시하고 기존 연구의 모형들과 어떻게 비교될 수 있는 지 살펴본다. 더불어 제시된 설계효과모형이 활용된 설계사례를 간단히 소개한다. 4절에서는 설계효과모형과 관련된 추가적인 몇 가지 연구 이슈들에 대해 간단히 논하고자 한다.

2. 집락추출을 위한 설계효과모형

2.1. Kish 모형

다단집락추출(multistage cluster sampling)에 의한 표본설계 $p(s)$를 통해 i번째 집락내 k번째 개체에 대해서 관측치 y_{ik}와 표본가중치 w_{ik}가 주어지고, 가중표본평균이 $\bar{y}_p = \sum_{i=1}^{n}\sum_{k=1}^{m_{ik}} w_{ik}y_{ik} / \sum_{i=1}^{n}\sum_{k=1}^{m_{ik}} w_{ik}$와 같이 정의된다고 하자. Kish (1965)는 램덤가중(haphazard weighting), 즉 $\mathrm{cor}(w_{ik}, y_{ik}) = 0$으로 표본가중치와 관측치와 상관관계가 없을 경우, 집락효과(cluster effect)와 가중치효과(weight effect)를 반영하기 위해 다음의 설계효과모형(design effect model 혹은 design effect formula)을 제시하였다.

$$\mathrm{deff}_K\left(\bar{y}_p\right) = \left[1 + (\bar{m} - 1)\rho_K\right]\left[1 + \mathrm{cv}_w^2\right]. \tag{2.1}$$

여기서 $\bar{m} = \sum_{i=1}^{n} m_i/n$는 평균집락크기, ρ_K는 집락내 동질성 척도(rate of homogeneity, roh), cv_w^2는 표본가중치의 상대분산(relative variance)을 나타낸다.

Kish의 설계효과모형 (2.1)는 집락내 개체간 동질성과 평균집락크기가 클수록 집락효과가 크고 (단순확률추출에 비해 추정량의 분산이 증가하고), 표본가중치의 변동이 클수록 가중치효과가 크게 (단순확률추출에 비해 추정량의 분산이 증가하게) 나타남을 잘 표현해 주고 있다. Kish의 설계효과모형 (2.1)은 일단집락추출과 램덤가중의 설계요소를 각각 갖는 표본설계하에서 유도되는 설계효과식 $\mathrm{deff}_c(\bar{y}_p) =$

$1+(\bar{m}-1)\rho_K$와 $\text{deff}_w(\bar{y}_p)=1+cv_w^2$의 승법모형으로 나타낸 것이다. 이에 대한 상세한 논의는 Kish (1987), Lê 등 (2001), Park과 Lee (2004) 등을 참조할 수 있다.

2.2. Gabler-Häder-Lahiri 모형

Kish의 설계효과모형 (2.1)는 조사분야에서 많이 인용되고 활용되어 왔지만 2.1절에서 기술한대로 모형식의 유도가 분석적이라기 보다는 경험을 바탕으로한 직관에 의해 제시되었다. 이에 Gabler 등 (1999)은 전통적인 일원확률효과모형(one-way random effects model)에 근거하여 Kish의 설계효과모형 (2.1)에 대한 정당성을 부여하였다. 먼저 집락추출을 반영할 수 있는 모수모형 ψ을 다음과 같이 가정하고 가중표본평균 $\bar{y}_p$의 모형분산 $V_\psi(\bar{y}_p)$을 유도하였다.

$$y_{ik}=\mu+\alpha_i+\epsilon_{ik}. \tag{2.2}$$

또한 모수모형 ψ과 비교될 수 있는 단순확률추출을 반영한 모형 ψ_0인 $y_{ik}=\mu+\epsilon_{ik}^0$의 가정에서 단순표본평균 $\bar{y}$의 모형분산 $V_{\psi_0}(\bar{y})$을 구하여 다음의 설계효과모형을 유도하였다.

$$\text{deff}_{G1}(\bar{y}_p)=\frac{V_\psi(\bar{y}_p)}{V_{\psi_0}(\bar{y})}.$$

여기서 모수모형 ψ의 두 확률변수 α_i와 ϵ_{ik}는 서로 독립이며 평균 0, 분산 $V_\psi(\alpha_i)=\rho_y\sigma_y^2$과 $V_\psi(\epsilon_{ik})=(1-\rho_y)\sigma_y^2$을 갖는다. 또한 모수모형 ψ_0의 확률변수 ϵ_{ik}^0는 평균 0이고 분산 $V_{\psi_0}(\epsilon_{ik}^0)=\sigma_y^2$을 갖는다. 또한 Gabler 등 (1999)은 설계효과모형 $\text{deff}_{G1}(\bar{y}_p)$이 다음과 같은 상한값을 가짐을 보였다.

$$\text{deff}_{G1}^*(\bar{y}_p)=\left[1+(m^*-1)\rho_y\right]\left[1+\text{cv}_w^2\right]. \tag{2.3}$$

여기서 m^*은 일종의 평균집락표본크기로 다음과 같이 정의되는데, 이는 모든 집락표본이 동일한 크기일 때 식 (2.1)의 $\bar{m}$과 같게 된다.

$$m^*=\frac{\sum_{i=1}^n\left(\sum_{k=1}^{m_i}w_{ik}\right)^2}{\sum_{i=1}^n\sum_{k=1}^{m_i}w_{ik}^2}.$$

2.3. Gabler-Häder-Lynn 모형

Gabler 등 (2006)는 국가간 비교를 목적으로한 유럽사회조사(European social survey; ESS)의 표본설계 지침은 물론 조사결과를 통한 표본설계 요소의 효율성 평가를 위해 집락들로 이루어진 영역(domain)들로 이루어진 모집단에 대한 집락효과를 반영한 표본설계모형을 제시하였다. 표본은 총 n개의 집락들로 이루어져 있고 각각 H개의 영역 C_h에 속하고 개별영역은 n_h개 표본집락들로 구성된다고 가정한다. 영역 h내 i번째 집락이 m_{hi}개 개체들로 구성된다면 가중표본평균은 다음과 같이 정의될 수 있다.

$$\bar{y}_p=\frac{\sum_{i=1}^n\sum_{k=1}^{m_i}w_{ik}y_{ik}}{\sum_i^n\sum_{k=1}^{m_i}w_{ik}}=\frac{\sum_{h=1}^H\sum_{i\in C_h}\sum_{k=1}^{m_i}w_{ik}y_{ik}}{\sum_{h=1}^H\sum_{i\in C_h}\sum_{k=1}^{m_i}w_{ik}}. \tag{2.4}$$

만약 개별 영역별로 조사변수의 모수모형이 식 (2.2)을 만족하고, 집락내 상관계수는 영역간에 다를 수 있지만 분산은 영역별 차이가 없음을 가정한다면 가중표본평균에 대한 설계효과모형은 다음과 같이 유도된다.

$$\text{deff}_{G2}(\bar{y}_p)=\sum_{h=1}^H\left(\frac{\hat{M}_h^2}{\hat{M}^2}\frac{m}{m_h}\right)\delta_{wh}\delta_{ch}. \tag{2.5}$$

여기서 $\hat{M}_h = \sum_{i\in C_h}\sum_{k=1}^{m_i} w_{ik}$는 영역 h의 크기추정량이고 $\hat{M} = \sum_{h=1}^{H}\hat{M}_h$는 모집단 전체크기의 추정량이다. 또한 $\delta_{wh} = 1 + cw_{wh}^2$와 $\delta_{ch} = 1 + (m_h^* - 1)\rho_{yh}$는 각각 영역별로 산출된 가중치효과와 집락효과의 설계요소모형을 나타낸다.

2.4. 기타 설계모형 및 논의

Verma 등 (1980)은 35개 개발도상국들에 대해 개별국가별로 실시한 국제출산률조사(World Fertility Survey)의 국가별 표본설계 특성 및 효율성 비교를 위한 연구에서 층화, 집락, 가중치, 영역 등의 설계요소별로 방대한 경험적 연구를 제시하고 있다. 이에 대한 논의에서 Holt (1980)는 조사변수 y에 대하여 식 (2.2)의 모수모형 ψ를 가정하여 가중표본평균 $\tilde{y}_p = \sum_i w_i\bar{y}_i$의 다음의 분산을 유도하여 설계효과모형에 대한 다양한 논의를 포함하였다.

$$v_\phi(\tilde{y}_p) = \sum_{i=1}^{n}\frac{w_i^2\sigma_i^2}{m_i}[1 + (m_i - 1)\rho_i].$$

여기서 w_i와 m_i는 각각 집락 i의 가중치와 표본크기를 나타낸다. Kalton 등 (2005)는 개발도상국가들의 가구표본설계를 위한 지침서에서 설계효과의 유용성 및 요소모형에 대한 상세한 설명을 포함하고 있다.

Park (2015)은 다단확률추출과 더불어 일차추출단위의 내재적 층화설계 요소를 추가한 표본설계에서 개별 설계요소효과들에 대해 승법모형을 고려하여 표본설계의 유용성을 평가하는 경험적 연구를 수행하였다. Park (2015)에서 제시된 설계효과모형은 3.2절에서 논의될 설계효과모형식 (3.6)와 유사한 층화, 집락, 가중치의 설계요소들이 갖는 설계요소모형을 곱한 승법모형 형태를 취한다.

3. 층화다단추출을 위한 설계효과모형

3.1. 제안모형

표본조사에서 일반적으로 많이 고려되는 층화다단추출(stratified multistage sampling)은 다음과 같이 기술될 수 있다. 먼저 총 M개의 개체로 구성된 모집단은 N개의 집락을 이루고 N_h개의 집락으로 구성된 총 H개 층으로 나뉜다. 층별로 집락표본 n_h개가 추출되며 추출된 집락내 M_{hi}의 개체 중 m_{hi}개가 추출된다고 하자. 또한 표본개체(hik)에 대해 조사값 y_{hik}와 표본가중치 w_{hik}가 주어진다고 가정하자. 여기서 $h = 1,\ldots,H$, $i = 1,\ldots,n_h$이고 $k = 1,\ldots,m_{hi}$이다. 모평균 $\bar{Y}$은 다음과 같이 정의되는 가중표본평균으로 추정할 수 있다.

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

층화, 집락 및 불균등가중치의 세 가지 설계요소를 갖는 층화집락추출을 반영하기 위해 조사변수 y에 대해 다음의 혼합모형 ξ을 가정하자.

$$y_{hik} = \mu + \eta_h + \alpha_{hi} + \epsilon_{hik}. \tag{3.1}$$

여기서 η_h는 고정효과(fixed effect)로 $\sum_{h=1}^{H}\eta_h = 0$을 만족하고 α_{hi}와 ϵ_{hik}는 각각 램덤효과(random effect)로 서로 독립이고 분산 $V_\xi(\alpha_{hi}) = \rho_{yh}\sigma_{yh}^2$과 $V_\xi(\epsilon_{hik}) = (1 - \rho_{yh})\sigma_{yh}^2$을 갖는다. 이때 분산과 급내상관계수는 서로 다른 층에 대해서는 다른 값을 가질 수도 있음을, 즉 $\sigma_{yh}^2 \neq \sigma_{yh'}^2$이거나 $\rho_{yh} \neq \rho_{yh'}$을 허용한다.

Gabler 등 (2006)와 유사한 논리를 적용하면, 모수모형 (3.1)하에서 다음의 설계효과모형을 유도할 수 있다.

$$\text{deff}_\xi(\bar{y}_p) = \sum_{h=1}^{H} \delta_{sh}\delta_{ch}\delta_{wh}. \tag{3.2}$$

여기서 집락과 랜덤가중에 의한 설계효과 요소모형(design effect component model)은 각각 다음과 같이 정의된다.

$$\delta_{ch} = 1 + (m_h^* - 1)\rho_{yh}, \tag{3.3}$$

$$\delta_{wh} = 1 + cv_{wh}^2. \tag{3.4}$$

식 (3.3)과 (3.4)은 설계효과모형 (2.5)의 해당 요소모형과 동일하게 정의되며 m_h^*와 cv_{wh}^2는 식 (2.3)의 m^*과 cv_w^2을 층 h에 대해 적용한 형태를 취한다. 또한 층화설계에 따른 설계효과모형은 다음과 같이 정의된다.

$$\delta_{sh} = \frac{(\hat{M}_h/\hat{M})^2}{m_h/m}\frac{\sigma_{yh}^2}{\sigma_y^2} = \frac{\hat{A}_h^2}{a_h}\frac{\sigma_{yh}^2}{\sigma_y^2}. \tag{3.5}$$

여기서 $\hat{M}_h = \sum_{i=1}^{n_h}\sum_{k=1}^{m_{hi}} w_{hi}$와 $\hat{M} = \sum_{h=1}^{H}\hat{M}_h$는 층과 모집단의 크기 추정량, $m_h = \sum_{i=1}^{n_h} m_{hi}$과 $m = \sum_{h=1}^{H} m_h$는 층과 전체 표본크기를 각각 나타낸다. 또한 $\hat{A}_h = \hat{M}_h/\hat{M}$과 $a_h = m_h/m$는 층 h의 모집단과 표본의 상대적 크기를 뜻한다.

집락효과와 랜덤가중효과에 대한 설계요소모형 δ_{ch}와 δ_{wh}는 층별 집락효과와 가중치효과에 의한 설계효과을 나타낸다면, 층화에 대한 설계요소모형 δ_{sh}는 층별 분산, 크기 및 표본할당의 설계효과를 나타낸다. 여기서 주의할 것은 δ_{ch}와 δ_{sh}는 모두 관심특성 y와 연관되지만 δ_{wh}는 랜덤가중의 가정으로 인해 그렇지 않음을 알 수 있다.

3.2. 기존모형과의 비교논의

Gabler 등 (2006)의 설계효과모형 (2.5)은 집락들이 중복되지 않는 상호배반적 영역(domain)들로 나뉘어지는 "다중표본설계(multiple design samples)"에 적용시킬 것을 목적으로 개발되었다. 이들의 관심특성 y에 대한 모수모형은 식 (3.1)의 혼합모형 ξ와 유사한 일원확률효과모형이지만 모든 영역 h들에 대해 동질적 분산 $\sigma_{yh}^2 = \sigma_y^2$을 가정한다.

Lee (2012)는 설계효과모형 (2.5)이 층화다단추출의 표본설계에도 적용될 수 있음을 지적하였고 이에 따라 Gabler 등 (2006)의 가정들이 층화다단추출설계에 적용 가능하도록 다시 정리하여 설명하였다. Jabkowski (2013)은 Lee (2012)의 접근방식을 따라 Gabler 등 (2006)이 제시한 설계효과모형을 유럽사회조사의 폴란드 표본설계에 적용시킬 수 있는 여러가지의 변형된 모형들을 제시하였다. 하지만, Lee (2012)와 Jabkowski (2013)는 모두 동질적인 층분산을 가정하고 있어서 결과적으로 이들의 층화설계요소모형은 $\delta'_{sh} = \hat{A}_h^2/a_h$으로 표기되어 층분산이 다른 경우에는 적용될 수 없는 한계를 지닌다.

기존 설계효과모형은 제안된 모형 (3.1)의 특수한 형태가 되며 다양한 표본설계에 적용될 수 있음을 나타내기 위해 기존 연구에서 사용한 가정을 정리한다. 먼저 Lee (2012)의 기본 모형 가정을 정리하면 다음과 같다.

(L1) 모든 층 h에 대해 $\text{cor}(w_{hik}, y_{hik}) = 0$ (층별로 표본가중치는 관심특성과는 무관한 램덤성(haphazard)을 가짐).

(L2) $\sigma_{yh}^2 = \sigma_y^2$ (층화는 램덤성을 가져 층분산은 모든 층에 대해 동일).

(L3) $\rho_{yh} = \rho_y$ (층별 급내상관계수는 서로 다르지 않음).

(L4) $cv_{wh}^2 = cv_w^2$ (층별 가중치 상대분산은 모든 같음).

(L5) $m_h^* = m_0^*$ (층별 가중 집락표본수는 모든 층에 동일).

더불어 Gabler 등 (2006)의 추가적인 가정들을 나열하면 다음과 같다.

(G1) $w_{hik} = 1$ (모든 개체의 동일가중치).

(G2) $\hat{M}_h/\hat{M}_h = M_h/M$ (층 가중치합은 층 규모에 비례).

• Gabler 등 (2006) 설계효과모형

조건 (L1–L2)을 만족하면, 제안된 설계효과모형식 (3.1)은 다음의 Gabler 등 (2006)의 설계효과모형이 된다.

$$\text{deff}_G(\bar{y}_p) = \sum_{h=1}^{H} \left(\frac{\hat{A}_h^2}{a_h} \right) \delta_{ch}\delta_{wh}.$$

• Jabkowski (2013) 설계효과모형

조건 (L1–L4, G1–G2)을 만족한다면, 제안된 설계효과모형식 (3.1)은 다음의 Jabkowski (2013) 설계효과모형이 된다.

$$\text{deff}_J(\bar{y}_p) = \delta_w \sum_{h=1}^{H} a_h \delta_{ch}^{(J)}.$$

여기서 $\delta_w = 1 + cv_w^2$이고 $\delta_{ch}^{(J)}$는 집락추출이 없는 영역은 1이고 집락추출이 있는 영역은 $1 + (m_h^* - 1)\rho_{yh}$이 된다.

• Lee (2012) 설계효과모형

조건 (L1–L5)을 만족할 때, 설계모형식 (3.1)은 갖는 특수형태로 다음과 같이 층구분없이 세가지 요소효과모형식을 곱한 형태가 된다.

$$\text{deff}_L(\bar{y}_p) = \delta_s\delta_c\delta_w, \tag{3.6}$$

여기서 $\delta_s = \sum_{h=1}^{H}(\hat{A}_h^2/a_h)$이고 $\delta_c = 1 + (m_0^* - 1)\rho_y$, $\delta_w = 1 + cv_w^2$이다.

3.3. 모형적용사례

제안된 설계효과모형 (3.1)은 한국농촌경제연구원이 주관하는 식품소비행태조사를 위한 표본설계에 이용되었다. 식품소비행태조사는 우리나라 소비자들의 전반적인 식품소비행태 및 트렌드분석을 수행하는 표본조사로 한국농촌경제연구원이 주관하며 2013년 처음 실시된 이후 지금까지 매년 수행되어오고 있다. 조사대상은 전국 1인 및 혈연가구내 식품 주구입자, 성인 및 만 13–18세의 청소년 구성원으로 이루어져 있고 표본대상은 16개 광역시도를 표본층으로 나누고 층별로 조사구와 조사구 내에서 가구를 각각 확률추출하였다. 주요 분석영역은 권역이고 세부영역인 표본층은 16개 광역시도로 이루어지며, 약 $m = 3000$ 가구가 층별로 절충비례에 의해 배분되었다.

조사에 포함되는 많은 조사특성치를 함께 반영하기 위해 $p = 0.5$의 크기를 갖는 임의의 비율 특성치에 대한 표본추정량의 상대표준오차가 전국수준에서 2.5%와 권역수준에서 8.8%가 넘지 않도록 배분규모를 정하였다. 추출확률, 무응답조정, 래이킹 비 조정 등 일련의 가중치 조정과정을 통해 얻게 되는 (불균등) 가중치의 상대분산이 20% 수준일 것을 가정하여 가중치효과 요소모형으로 $\delta_{wh} = 1 + 0.2^2 = 1.04$을 계산하였다. 층내 조사구 급내상관계수 ρ_h는 통계청의 사회조사를 기준으로 0.096임을 가정하고 집락평균 목표응답가구수로 $m_0 \equiv 5$을 설정하여 집락추출효과 요소모형 $\delta_{ch} = 1 + 4 \times 0.096 = 1.384$로 예측하였다. 또한 특성치의 층별 분산은 층과 관계없이 모두 동일한 $\sigma_{yh}^2 \equiv \sigma_y^2$이고 층별 모집단 크기 추정량 $\hat{M}_h$는 표본추출틀로 사용한 2010년 인구주택총조사의 층규모 M_h와 같다고 가정하에 층화효과 요소모형 $\delta_{sh} = (A_h^2/a_h)(\sigma_{yh}^2/\sigma_h^2) = A_h^2/a_h$을 층별로 계산하여 전체 설계효과모형식 (3.1)에 적용하여 표본설계를 수행하였다. 상세한 표본설계내역과 조사결과를 바탕으로 수행한 설계효과모형의 평가내용은 Park (2014)를 참고할 수 있다.

4. 논의

제안된 설계효과모형식 (3.1)은 층화다단추출 하에서 층화표본추정량에 대한 표본설계와 관련된 모집단 층화구조가 주는 정도수준에 대한 영향을 설계효과라는 개념을 통해 명쾌히 나타내 주고 있다. 따라서 다음과 같은 두 가지 측면의 유용성을 제공한다고 할 수 있다. 먼저, 사전에 기술된 추정정도를 얻기 위해 설정한 표본크기가 줄 수 있는 설계효과를 조사수행 전에 예측하는데 활용될 수 있다. 둘째, 조사수행 후 표본설계의 개별 설계요소들의 효율성을 평가하는데 활용될 수 있다. 물론, 제안된 설계효과모형식 (3.1)의 타당성은 이를 유도하기 위해 가정된 모수모형 ξ이 얼마나 적절하였는가에 따라 달라질 수 있을 것이다.

조사변수 y와 표본가중치 w_{hik} 사이에 0이 아닌 상관관계가 존재한다면 제안된 설계효과모형식 (3.1)은 적절치 않을 수도 있다. 이러한 경우, 일단추출, 즉 집락추출을 사용하지 않고 조사개체를 직접 추출하는 표본설계에서는 Spencer (2000)나 Henry (2011)의 연구내용을 참고할 수 있을 것이다. 향후 제안된 설계효과모형식이 이를 위해 가정된 설계상황에서 벗어날때 얼마나 강건한가를 경험적 연구를 통해 평가한다면 매우 유용할 것이다. 최근 Gabler 등 (2014)은 일단 집락추출하에서 설계효과모형식의 모수들을 설계근거분석 방식하에서 추정할 수 있는 가를 연구하였다. Gabler 등 (2014)의 연구를 층화다단추출의 상황으로 확대할 수 있다면 매우 유용한 결과를 얻을 수 있을 것으로 기대된다.

References

Cornfield, J. (1951). Modern methods in the sampling of human populations, *American Journal of Public Health*, **41**, 654–661.

Gabler, S., Ganninger, M. and Lahiri, P. (2014). A new approximation to the true randomization-based design effect, Submitted to a journal.

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

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

Henry, K. A. (2011). Weighting adjustment methods and their impact on sample-based inference (Ph.D. thesis), University of Maryland, College Park, MD.

Holt, D. (1980). Discussion of the paper by Verma, Scott, and O'Muircheartaigh, *Journal of Royal Statistical Society-A*, **143**, 468–469.

Jabkowski, P. (2013). How (not) to estimate the design effect of a complex sampling scheme: a case study of the Polish section of the European social survey, round 5, *Ask Research & Methods*, **22**, 55–77.

Kalton, G., Brick, J. M. and Lê, 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, New York, Statistics Division, United Nations.

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

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

Kish, L. (1992). Weighting for unequal p_i, *Journal of Official Statistics*, **8**, 183–200.

Korn, E. L. and Graubard, B. I. (1999). *Analysis of Health Surveys*, John Wiley & Sons, New York.

Lê, T., Brick, J. M. and Kalton, G. (2001). Decomposing design effects, In *Proceedings of the Joint Statistical Meetings, Section on Survey Research Methods*, American Statistician Association, CD-Rom.

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.

Park, I. (2014). A study on design effect models for complex sample survey, *Journal of the Korean Data & Science Society*, **25**, 523–531.

Park, I. (2015). Assessing complex sample designs via design effect decompositions, Submitted to a journal.

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

Rao, J. N. K. and Scott, A. J. (1987). On simple adjustments to chi-square tests with sample survey data, *Annals of Statistics*, **15**, 385–397.

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

Spencer, B. D. (2000). An approximate design effect for unequal weighting when measurements may correlate with selection probabilitie, *Survey Methodology*, **26**, 137–138.

Verma, V., Scott, C. and O'Muircheartaigh, C. (1980). Sample designs and sampling errors for the World Fertility Survey, *Journal of Royal Statistical Society-A*, **143**, 431–473.

설계효과모형을 통한 설계요소의 유용성 이해

박인호[a,1]

[a]부경대학교 통계학과

(2015년 10월 29일 접수, 2015년 11월 13일 수정, 2015년 11월 13일 채택)

요 약

조사자료분석에 있어서 표본추정량에 대해 설계요소가 갖는 효율성은 단순확률추출과 비교한 복잡표본설계의 의한 표본추출이 주는 분산의 상대적 크기인 설계효과를 통해 평가할 수 있다. 설계효과의 유용성은 복잡설계요소의 함수형태로 표현될 수 있을때 극대화될 수 있다. 본 연구에서는 층화다단추출의 표본설계에서 적용될 수 있는 설계효과모형을 제시하였다. 제시된 설계효과모형은 기존 다단추출을 위한 Gabler 등 (1999, 2006)의 모형을 일반화한 것으로 층구조, 표본할당, 집락추출 및 불균등가중치 등의 설계요소들이 정도수준에 갖는 영향력을 함수식으로 명확히 나타내주고 있다. 이를 활용하면 사전에 기술된 추정정도를 얻기 위해 설정한 표본크기가 줄 수 있는 설계효과를 예측하는데 활용할 수 있다. 또한 사후적으로 표본설계의 개별 설계요소들이 표본추정량에 대해 갖는 효율성을 평가하는데 활용될 수 있다.

주요용어: 층화다단추출, 램던가중, 집락내상관계수, 혼합효과모형, 유효표본크기

이 논문은 부경대학교 자율창의학술연구비(2014년)에 의하여 연구되었음.

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