The Korean Journal of Applied Statistics (2019) DOI: https://doi.org/10.5351/KJAS.2019.32.4.643
32(4), 643–652

Sample size determination using design effect formula for repeated surveys

Inho Park[a,1] · Hyeon Gil Hwang[a]

[a]Department of Statistics, Pukyong National University

(Received June 26, 2019; Revised July 21, 2019; Accepted July 26, 2019)

Abstract

We propose a method for sample size determination using design effect formulas when a sample is resigned for a repeated survey. The proposed method enables the determination of the sample size by incorporating the impact of various design components to the sampling error through design effect formulas that are applicable under multistage sampling design and stratified multistage sampling designs.

Keywords: repeated survey, coefficient of variation, design effect formula, stratified multistage sampling

1. 서론

반복조사에 있어 표본재설계는 주로 모집단 변화(표본틀 개편), 유효표본 감소, 목표오차 수정, 조사환경 및 비용 변동 등의 다양한 이유로 인해 고려된다. 표본수를 결정함에 있어 조사 예산은 우선적으로 고려하지만, 변동이 크지 않다면 주로 주요 항목별 추정오차값이 목표수준을 만족하도록 한다. 표본수 결정에 흔히 고려되는 방법은 기존조사의 표본수와 상대표준오차(coefficient of variation; CV)를 이용하여 재설계의 상대표준오차가 목표수준을 달성할 수 있게 하는 것이다. 기존시점 k에서 n_k는 표본수, N_k는 모집단 크기, $\bar{Y}_k$는 모평균, $\bar{y}_k$는 평균추정량, $V_k = V(\bar{y}_k)$와 $\mathrm{CV}_k = \sqrt{V_k}/\bar{Y}_k$는 각각 평균추정량의 분산과 상대표준오차를 나타낸다고 하자. 다음의 비례식은 $k+1$시점의 표본설계가 목표 상대표준오차 CV_{k+1}를 만족하는 표본수 n_{k+1}를 산정한다 (Park, 1989; Statistics Korea, 2007; Kim, 2012).

$$n_{k+1} = n_k \times \frac{\widehat{\mathrm{CV}}_k^2}{\mathrm{CV}_{k+1}^2}, \tag{1.1}$$

여기서 $\widehat{\mathrm{CV}}_k$는 k시점의 표본자료로 추정된 표본평균의 상대표준오차 추정량을 나타낸다. 복합표본설계(complex sampling design)를 통해 얻은 자료분석의 표본오차는 층 구분 및 규모, 층내 및 층간 조사특성의 분포, 표본할당, 집락내 개체간 동질성, 집락표본크기, 가중치 조정 등의 다양한 설계요소(design component)의 영향을 받게 된다.

본 연구는 반복조사를 위한 표본재설계에서 복합표본설계의 다양한 설계요소를 반영한 표본수 결정의 문제를 다룬다. 2절에서는 단순추출과 비교하여 복합표본설계가 갖는 표본오차 측면의 (비)효율성을

This work was supported by a Research Grant of Pukyong National University (2017 year).
[1]Corresponding author: Department of Statistics, Pukyong National University, 45 Yongso-ro, Nam-gu, Busan 48513, Korea. E-mail: ipark@pknu.ac.kr

측정하는 설계효과(design effect)에 대해 살펴보고, 조사연구에서 흔히 채택하는 다단추출(multistage sampling)과 층화다단추출(stratified multistage sampling) 하에서 설계요소별 영향력을 표현할 수 있는 설계효과모형식(design effect formula)을 소개한다. 3절에서는 모집단 크기변동을 반영한 표본수 결정을 다룬 Kim (2012)의 방법을 살펴보고, 다단추출과 층화다단추출의 설계효과모형식을 활용한 표본수 결정의 방법을 소개한다. 4절에서는 농업인 복지실태조사의 재설계에서 고려한 표본수 결정방법을 소개한다. 5절은 연구 내용을 정리하여 논한다.

2. 설계효과, 설계요소 및 설계효과모형

설계효과는 복합표본추정량의 (상대)분산을 동일 크기의 가상적 단순표본추정량이 갖는 (상대)분산으로 나눈 값으로 정의하는데, 비복원 단순추출 기준으로 다음과 같이 주어진다 (Kish, 1965).

$$\mathrm{DEFF}_k = \frac{V_k}{V_{k,\mathrm{srswor}}} = \frac{\mathrm{CV}_k^2}{\mathrm{CV}_{k,\mathrm{srswor}}^2}, \tag{2.1}$$

여기서 $V_{k,\mathrm{srswor}} = n_k^{-1} S_k^2 (1 - n_k/N_k)$와 $\mathrm{CV}_{k,\mathrm{srswor}}^2 = V_{k,\mathrm{srswor}}/\bar{Y}_k^2$는 각각 비복원 단순추출 하의 평균추정량 $\bar{y}_{\mathrm{srs}} = n_k^{-1}\sum_{i=1}^{n} y_{ki}$의 분산과 상대분산이고, $S_k^2 = (N_k - 1)^{-1}\sum_{i=1}^{N_k}(y_{ki} - \bar{Y}_k)^2$는 k시점의 조사특성 y의 모분산이고 y_{ki}는 k시점의 i번째 개체의 조사특성값을 나타낸다.

반면 설계효과를 복원 단순추출으로 비교대상으로 정의하면 다음과 같이 주어진다 (Kish, 1995).

$$\mathrm{DEFT}_k^2 = \frac{V_k}{V_{k,\mathrm{srswr}}} = \frac{\mathrm{CV}_k^2}{\mathrm{CV}_{k,\mathrm{srswr}}^2}. \tag{2.2}$$

여기서 $V_{k,\mathrm{srswr}} = n_k^{-1} S_k^2$이고 $\mathrm{CV}_{k,\mathrm{srswr}}^2 = V_{k,\mathrm{srswr}}/\bar{Y}_k^2$이다. 설계효과 정의 (2.2)는 복원추출도 설계요소의 한 가지로 설계효과에 표현되어야 하며 통계추론(예로, 신뢰구간)에서의 유용성이 고려될 수 있도록 제안되었다 (Kish, 1995).

새로운 조사를 기획할 때 일반적으로 유사사례를 참고하거나 대략적인 가정을 바탕으로 표본설계를 수행하며, 자료수집 이후에 평균, 분산, 설계효과 등을 추정하여 표본설계의 효율성을 평가하고, 반복조사인 경우에는 이를 바탕으로 표본재설계를 고려할 수 있다. Rust와 Broene (2010)은 표본 효율성 평가와 표본설계 개선의 관점에서 설계효과가 가져야 할 측면을 다음과 같이 논하고 있다. "설계효과의 개념은 할 수만 있다면 모집단 구조와 설계요소들이 갖는 표본오차에 대한 개별적 영향을 명백히 나타낼 수 있는 함수적 형태로 표현될 때 그 유용성은 극대화되며, 향후 적용시 설계효율을 증진시킬 수 있는 지침으로 사용될 수 있다."

다단추출이 평균추정량에 대해 갖는 설계효과는 다음의 설계효과모형식으로 표현될 수 있다 (Kish, 1987).

$$\mathrm{DEFT}_k^2 = L_k\left[1 + (\bar{n}_k - 1)\rho_k\right], \tag{2.3}$$

여기서 $L_k = 1 + \mathrm{CV}_{kw}^2$는 불균등가중치의 사용에 따른 분산증가분, $\mathrm{CV}_{kw}^2 = N_k^{-1}\sum_{i=1}^{N_k}(w_{ki} - \bar{W}_k)^2/\bar{W}_k^2$는 가중치 상대분산, $\bar{W}_k = \sum_{i=1}^{N_k} w_{ki}/N_k$는 가중치 평균값, w_{ki}는 시점 k의 i번째 개체의 가중치, $\bar{n}_k$는 집락크기, ρ_k는 집락내동질성계수를 나타낸다.

시점간 기본 모수 및 설계요소에 대해 다음의 가정을 고려할 수 있다.

(D1) $\bar{Y}_{k+1} = \bar{Y}_k$

(D2) $S_{k+1} = S_k$

(D3) $\mathrm{DEFF}_{k+1} = \mathrm{DEFF}_k$

(D4) $\mathrm{DEFT}^2_{k+1} = \mathrm{DEFT}^2_k$

(D5) $L_{k+1} = L_k$

(D6) $\rho_{k+1} = \rho_k$

먼저, D1, D2, D3, D4는 각각 시점간 조사특성 y의 모평균, 표준편차, 비복원추출과 복원추출 기준의 설계효과가 동일함을 가정한다. D5는 시점간 불균등가중치로 인한 분산증가분이 동일함을 가정한다. D6는 조사특성 y의 집락내동질성계수가 변함없음을 가정한다.

조사연구에서는 모집단을 서로 겹치지 않은 층(strata)으로 나눈 후 층별 다단계로 추출하는 층화다단추출이 종종 고려된다. 층별 모수와 표본통계량은 앞서 논의한 다단추출의 모수와 표본통계량에 층을 나타내는 첨자 h을 추가하여 표기하고자 한다. 예로, 시점 k의 층크기가 N_{kh}이고 층 표본수는 n_{kh}이며 층 모평균이 $\bar{Y}_{kh}$이라면, 전체 모집단 크기와 표본수는 $N_{k,st} = \sum_{h=1}^{H} N_{kh}$와 $n_{k,st} = \sum_{h=1}^{H} n_{kh}$이 된다. (층화) 모평균 $\bar{Y}_{k,st} = \sum_{h=1}^{H} A_{kh}\bar{Y}_{kh}$의 표본추정량은 $\bar{y}_{k,st} = \sum_{h=1}^{H} A_{kh}\bar{y}_{kh}$이고 분산은 $V_{st,k} = \sum_{h=1}^{H} A_{kh}^2 V_{kh}$이 된다. 여기서 $A_{kh} = N_{kh}/N_{k,st}$는 층 상대크기이고 $V_{kh} = V(\bar{y}_{kh})$는 층 평균추정량의 분산을 나타낸다. 층분산을 식 (2.2)의 설계효과로 표현하면, (층화)표본평균 $\bar{y}_{k,st}$의 상대분산은 다음과 같이 유도된다.

$$\mathrm{CV}^2_{k,st} = \frac{V_{k,st}}{\bar{Y}^2_{k,st}} = \frac{1}{\bar{Y}^2_{k,st}} \sum_{h=1}^{H} A_{kh}^2 V_{kh} = \frac{1}{\bar{Y}^2_{k,st}} \sum_{h=1}^{H} A_{kh}^2 \left(\frac{S^2_{kh}}{n_{kh}}\right) \mathrm{DEFT}^2_{kh}, \tag{2.4}$$

여기서 S^2_{hk}는 층 모분산을 나타내며 층 표본평균의 설계효과는 모형식 (2.3)을 적용하면 다음과 같이 나타낼 수 있다.

$$\mathrm{DEFT}^2_{kh} = L_{kh}\left[1 + (\bar{n}_{kh} - 1)\rho_{kh}\right], \tag{2.5}$$

여기서 L_{kh}, $\bar{n}_{kh}$, ρ_{kh}는 각각 층 h의 불균등가중치 사용에 따른 분산증가분, 층내 평균집락표본크기, 층내 집락내동질성개수를 나타낸다. 또한, 설계효과에 의해 표본평균 $\bar{y}_{k,st}$의 설계효과모형식을 유도하면 다음과 같다.

$$\mathrm{DEFT}^2_{k,st} = \left(\frac{S^2_k}{n_k \bar{Y}^2_k}\right)^{-1} \mathrm{CV}^2_{k,st} = \sum_{h=1}^{H} \left(\frac{A^2_{kh}}{a_{kh}} \frac{S^2_{kh}}{S^2_k}\right) \mathrm{DEFT}^2_{kh}, \tag{2.6}$$

여기서 $a_{kh} = n_{kh}/n_k$는 표본할당비를 나타낸다. 설계효과모형식 (2.6)의 유도는 Park (2015)와 Chen과 Rust (2017)를 참고할 수도 있다. Gabler 등 (2006)가 설계효과모형 (2.3)을 상이한 집락군별로 적용한 설계효과모형을 제시하였고 Lee (2012)는 이를 층화다단추출로 확대해석하여 설계효과모형을 제시하였는데, 이는 설계효과모형 (2.6)에서 $S^2_{kh}/S^2_k \approx 1$을 가정하는 특수한 형태에 해당한다 (Park, 2014).

3. 표본수 결정

3.1. Kim (2012)의 표본수 결정 방법

Kim (2012)은 D1, D2, D3의 가정 하에서 $(k+1)$ 시점의 표본평균의 상대분산을 다음과 같이 정의하였다.

$$\mathrm{CV}^2_{k+1} = \frac{1}{n_{k+1}} \left(\frac{S^2_k}{\bar{Y}^2_k}\right) \left(1 - \frac{n_{k+1}}{N_{k+1}}\right) \times \mathrm{DEFF}_k. \tag{3.1}$$

추정값 $\widehat{\mathrm{CV}}_k^2$을 식 (2.1)에 대입하면 식 (3.1)로부터 표본수 n_{k+1}는 다음과 같이 유도된다.

$$n_{k+1} = \left[\frac{1}{n_k} \left(\frac{\mathrm{CV}_{k+1}}{\widehat{\mathrm{CV}}_k} \right)^2 \left(1 - \frac{n_k}{N_k} \right) + \frac{1}{N_{k+1}} \right]^{-1}. \tag{3.2}$$

Kim (2012)는 복합표본설계를 특정하지는 않았지만 가상의 비복원 단순확률추출과 비교한 식 (2.1)을 이용함으로 조사시점에 따른 모집단 크기의 변동을 표본재설계에 반영하고자 하였다.

만약 설계효과의 비교 기준을 복원단순추출로 하거나 시점에 관계없이 모집단이 표본에 비해 월등히 크다면 (즉, $1 - n_k/N_k \approx 1$) D1, D2, D4의 가정 하에서 식 (2.2)로부터 다음의 관계식을 도출할 수 있다.

$$\mathrm{CV}_{k+1}^2 = \frac{1}{n_{k+1}} \left(\frac{S_k^2}{\bar{Y}_k^2} \right) \times \mathrm{DEFT}_k^2 = \frac{n_k}{n_{k+1}} \mathrm{CV}_k^2.$$

위의 식에서 시점 k의 상대표준오차 추정값을 대입하면 식 (1.1)이 된다. Kim (2012)이 지적한 것과 같이 표본수가 동일한 경우에 현행 표본설계의 상대표준오차와 재설계의 상대표준오차가 근사적으로 같다는 가정이 성립되는 경우에는 식 (1.1)이 정당성을 갖게 된다. 시점간 설계효과에 차이가 없음(D3 혹은 D4)을 가정하면 식 (3.1)에서처럼 세부적 설계요소의 변동이 반복조사의 표본수 결정에서 반영될 수 있는 여지가 없게 된다. 표본오차에 영향을 줄 수 있는 다양한 설계요소, 예를 들면, 집락내동질성계수, 집락평균표본수, 불균등가중치, 층별 분산분해 및 표본할당 등을 반영하여 표본수를 결정할 수 있다면 좀 더 효율적일 수 있을 것이다. 다음 소절에서는 다단추출과 층화다단추출 하에서 다양한 설계요소를 고려할 수 있는 표본수의 결정 방법에 대해 논의하고자 한다.

3.2. 다단추출의 표본수 결정 방법

다단추출 하에서 표본평균의 상대분산은 식 (2.2)와 (2.3)으로부터 다음과 같이 표현할 수 있다.

$$\mathrm{CV}_{k+1}^2 = \frac{1}{\bar{Y}_{k+1}^2} \left(\frac{S_{k+1}^2}{n_{k+1}} \right) L_{k+1} \left[1 + (\bar{n}_{k+1} - 1)\rho_{k+1} \right]. \tag{3.3}$$

만약 D1, D2, D5, D6이고 시점 k의 조사자료를 이용한 표본추정값 $\bar{y}_k$, s_k^2, $\hat{L}_k$, $\hat{\rho}_k$를 식 (3.3)에 대입하 다음의 근사식을 유도할 수 있다.

$$\mathrm{CV}_{k+1}^2 \doteq \frac{1}{\bar{y}_k^2} \left(\frac{s_k^2}{n_{k+1}} \right) \hat{L}_k \left[1 + (\bar{n}_{k+1} - 1)\hat{\rho}_k \right], \tag{3.4}$$

여기서 $\hat{L}_k = 1 + cv_{kw}^2 = n_k^{-1} \sum_{i=1}^{n_k} (w_{ki} - \bar{w}_k)^2 / \bar{w}_k^2$이다. 따라서 표본평균의 목표상대표준오차가 CV_{k+1}이 되도록 하기 위해서는 식 (3.4)를 집락평균표본크기 $\bar{n}_{k+1}$에 대해 유도하면 된다. 물론 $\bar{n}_{k+1} = n_{k+1}/m_{k+1}$이므로 표본집락수 m_{k+1}와 표본개체수는 n_{k+1}는 식 (3.4)로부터 동시적으로 결정해야 할 필요가 있다. 하지만, 집락별로 조사부담을 균등하게 하기 위한 이유로 흔히 $\bar{n}_{k+1}$을 특정한 값 η (예, $10, 20$ 등)으로 정하게 된다 (Statistics Korea, 2007). 즉,

(D7) $n_{k+1} = \eta\, m_{k+1}$ (주어진 상수 η).

따라서 D7을 가정하면 n_{k+1} 혹은 m_{k+1}의 결정 문제로 귀결되고, $(k+1)$ 시점의 표본개체수 n_{k+1}는 다음과 같이 유도된다.

$$n_{k+1} \doteq \left(\bar{y}_k^2 \mathrm{CV}_{k+1}^2 \right)^{-1} s_k^2 \hat{L}_k \left[1 + (\eta - 1)\hat{\rho}_k \right]. \tag{3.5}$$

만약 시점 $(k+1)$에서 대안적 집락(예를 들면, 가구조사에서 조사구를 병합조사구 혹은 집계구)을 사용하면, 위의 식을 이용하여 상대분산을 예측할 수 있다. 예로, 집락 변경에 따른 가중치 영향력, 평균집락표본수, 집락내동질성계수를 적절히 추정 혹은 예측할 수 있다면 $(k+1)$ 시점에서의 표본추정량의 상대표준오차를 추정할 수 있을 것이다 (Park, 2016).

3.3. 층화다단추출

층화다단추출의 표본수 결정은 모집단 전체는 물론 층별 정도수준를 모두 고려할 수 있다. 만약 층평균의 적절한 정도수준이 요구된다면 집락추출의 표본수 결정식 (3.5)를 이용하여 층별로 (잠정적인) 표본수 $n_{k+1,h}$를 정하고 전체 표본에 대해 조율할 수 있다.

총 표본수가 먼저 주어지는 경우라면 층별 표본수 결정은 표본할당의 문제가 된다. 만약 표본할당 기준이 상대분산 (2.4)를 최소화하는 것이라면 표본층별로 다음의 최적할당이 유도된다.

$$n_{k+1,h} \propto N_{k+1,h} s_{k,h+1} \sqrt{\hat{L}_k \left[1+(\eta-1)\hat{\rho}_{kh}\right]}. \tag{3.6}$$

식 (3.6)은 네이만 할당의 일반화된 형태로 층화단순추출과 비교하면 집락효과와 불균등가중치 효과가 추가로 고려된 것임을 알 수 있다. 이 외에도 비례할당 $(n_{k+1,h} \propto N_{k+1,h})$, Kish 할당$(n_{k+1,h} \propto \sqrt{1/H^2 + A_{k+1,h}^2})$, 제곱근할당 $(n_{k+1,h} \propto \sqrt{N_{k+1,h}})$ 등 고려할 수 있다.

4. 사례분석

농업인 복지실태조사는 농업진흥청 국립농업과학원이 주관하며 5년 주기로 매년 약 4,000여 개의 읍면지역 가구를 대상으로 농촌 특성에 맞는 복지증진 및 지역 개발시책의 수립과 시행 등에 필요한 기초자료의 제공을 목적으로 한다. 매년 특정 부문별 특성을 조사하여 공표함으로써 농촌의 복지실태 상황 등에 관한 변화 추이를 파악하는데, 대부분의 조사항목은 5점 리커트 척도의 만족도로 측정된다. 주요 특성으로는 지역생활여건, 문화여건, 경제활동여건, 보건의료, 복지여건, 안전도 등의 만족도 등이 포함된다. 주기 내 첫 해는 읍면지역은 물론 동지역을 포함한 도농 종합조사로 진행하며, 이후 4년간 읍면지역만을 대상으로 매년조사로 진행한다 (Rural Development Administration, 2018).

표본추출은 읍면동 지역구분과 동 지역내 서울, 광역시, 기타 지역과 읍면 지역내 9개 도지역으로 층화한 뒤, 층별로 읍면동, 조사구, 가구 순으로 선택하며 마지막으로 가구를 대표하는 19세 이상의 가구주 혹은 배우자 중 한 명을 가구에서 자율적으로 선택하게 하는 층화다단확률추출의 방식을 택하고 있다. 2013년에 수행된 1주기 5년 조사의 표본설계에 대하여 2017년에 효율성 평가를 실시하였고, 2018년에는 2주기 5년 조사의 표본설계를 진행하였다.

표본재설계를 위해 지역생활여건, 경제활동여건, 보건의료, 복지여건, 안전도 등의 다섯 가지 주요 만족도 변수와 연간가구소득 변수의 설계효과를 평가하고 세부 설계요소에 대한 분석을 실시하였다. 논리의 단순화를 위해 층화이단추출을 가정하여 설계효과모형식 (2.6)을 적용한 다음의 표본추정량을 고려하였다.

$$\begin{aligned} \widehat{\mathrm{DEFT}}_k^2 &= \sum_{h=1}^{H} \left(\frac{A_{kh}^2}{a_{kh}} \frac{s_{kh}^2}{s_k^2} \right) \widehat{\mathrm{DEFT}}_{kh}^2, \\ &= \sum_{h=1}^{H} \left(\frac{A_{kh}^2}{a_{kh}} \frac{s_{kh}^2}{s_k^2} \right) \hat{L}_{kh} \left[1+(\bar{n}_{kh}-1)\hat{\rho}_{kh}\right], \end{aligned} \tag{4.1}$$

Table 4.1. Sampling error of Life Satisfaction in Rural Area

Stratum	n_{kh}	$\bar{y}_{kh}$	$\widehat{\mathrm{SE}}_{kh}$	$\widehat{\mathrm{CV}}_{kh}$	$\widehat{\mathrm{DEFT}}^2_{kh}$
Gyeonggi	477	58.6	2.06	3.51	4.00
Gangwon	340	52.9	1.54	2.91	2.00
Chungbuk	420	64.5	1.90	2.95	4.23
Chungnam	420	57.3	1.94	3.38	3.77
Jeonbuk	400	57.9	1.13	1.95	1.88
Jeonnam	500	63.7	1.56	2.46	5.39
Gyeongbuk	500	51.7	2.42	4.67	7.58
Gyeongnam	489	60.9	1.18	1.93	1.98
Jeju	269	59.4	1.45	2.44	1.68
Total	3,995	58.3	0.73	1.26	5.12

여기서 층별 설계효과는 $\widehat{\mathrm{DEFT}}^2_{kh} = \widehat{\mathrm{CV}}^2_{kh}/\widehat{\mathrm{CV}}^2_{kh,\mathrm{srswr}} = \hat{L}_{kh}[1 + (\bar{n}_{kh} - 1)\hat{\rho}_{kh}]$이고, 집락내동질성계수는 다음과 같이 추정하였다 (Kish, 1995).

$$\hat{\rho}_{kh} = \frac{\widehat{\mathrm{DEFT}}^2_{kh}/L_{kh} - 1}{\bar{n}_{kh} - 1}.$$

층별로 m_{kh}개의 표본읍면동와 $m_{khi} \equiv 2$개의 표본조사구, 그리고 평균 $n_{khij} \equiv 10$개의 표본 가구로 구성되었다. 따라서 층별 표본개체수는 총 $n_{kh} = \sum_{i=1}^{m_{kh}} \sum_{j=1}^{2} n_{khij} = 20m_{kh}$이다. 조사예산을 감안하여 전 주기와 동일하게 총 $n_{k+1} = 4{,}000$개의 응답가구를 목표로 하였고, 표본할당 전략으로는 고정 CV 할당, 최적할당, 제곱근할당, 키쉬할당 등을 고려하여 최종적으로 층별 및 전국 기준의 평균추정량의 기대 상대표준오차를 산출하였다.

설계요소를 반영한 표본수 결정에 대한 사례 소개를 위해 2017년의 조사항목 중 지역생활만족도(체감)를 예를 들어 살펴보았다. 먼저, 설계요소 중 모평균, 모분산, 층별 불균등가중치에 따른 분산증가분, 층별 집락내동질성계수가 시점간 변화가 없다는 D1, D2, D5, D6을 가정하였다. Table 4.1은 시점 k의 기본 통계량으로 표본가구수 n_{kh}, 평균추정치 $\bar{y}_{kh}$, 표준오차 $\widehat{\mathrm{SE}}_{kh}$, 상대표준오차 $\widehat{\mathrm{CV}}_{kh}$와 설계효과 $\widehat{\mathrm{DEFT}}^2_k$를 정리하고 있다. 총 3,995개의 응답가구 중 전남과 경북이 각각 500개 가구로 가장 많고, 제주와 강원은 각각 269개와 340개의 가구로 구성되었다. 전국기준의 지역생활만족도(체감) 점수는 58.3점이고, 지역별로는 충북이 64.5점, 전남이 63.7점으로 가장 높았고, 강원과 경북이 각각 52.9와 51.7로 가장 낮았다. 상대표준오차는 경남이 1.93% (경남)로 가장 낮고 경북이 4.67%이며, 전국수준은 1.26%이다.

Table 4.2는 표본층별 설계요소의 구성현황을 정리하고 있다. 상대규모 A_{kh}는 경기가 0.23으로 가장 크고 제주가 0.02로 가장 작으며 약 0.21의 범위를 갖는다. 해당 표본층의 표본할당비 a_{kh}는 각각 0.12와 0.07으로 약 0.05의 범위를 갖는데 이는 기존 표본설계에서 표본층별 상대규모에 대한 절충적 표본할당이 이루어졌음을 나타낸다. 표본층별 집락표본평균크기 $\bar{n}_{kh}$는 대부분 20개 가구 정도이지만 제주도만 29.9이었다. 불균등 가중치로 인한 분산증가분 $\hat{L}_{kh}$는 경남이 1.25으로 가장 작지만 충북과 전북은 각각 8.89와 22.06로 매우 높았다. 집락내동질성계수 $\hat{\rho}_{kh}$는 전북이 가장 작은 -0.0485이고 경기가 가장 큰 0.05286의 값을 갖는다. 또한 분산추정치 $\hat{S}^2_{kh}$는 전국수준 408.12이 비해 전북은 269.65로 낮은 반면 경기가 504.94로 매우 높다.

Table 4.3은 표본수 결정의 전략별로 예상되는 상대표준오차, 설계효과 및 표본할당을 비교하고 있다. 우선 집락평균표본수는 $\bar{n}_{k+1,h} \equiv 20$으로 모든 층에서 동일하다고 가정하였다. 표본층별로 표본추정량의 목표상대표준오차를 3.10%으로 정한다면, 식 (3.5)를 적용하여 표본수를 정할 수 있다. 집락내

Table 4.2. Population and Sample Compositions and Design Components

Stratum	A_{kh}	a_{kh}	$\bar{n}_{kh}$	$\hat{L}_{kh}$	$\hat{\rho}_{kh}$	$\hat{S}^2_{kh}$
Gyeonggi	0.23	0.12	20.74	1.96	0.05286	504.94
Gangwon	0.07	0.09	20.00	2.66	−0.01305	402.66
Chungbuk	0.08	0.11	20.00	8.89	−0.02757	356.49
Chungnam	0.12	0.11	20.00	4.34	−0.00687	416.94
Jeonbuk	0.06	0.10	20.00	22.06	−0.04815	269.65
Jeonnam	0.11	0.17	20.00	5.39	0.00004	308.23
Gyeongbuk	0.16	0.13	20.00	6.01	0.01381	384.23
Gyeongnam	0.16	0.12	19.56	1.25	0.03106	341.43
Jeju	0.02	0.07	29.89	3.02	−0.01542	336.98
Total	1.00	1.00	20.49	-	-	408.12

Table 4.3. Comparisons of Sample sizes, Expected CVs and Design Effects by Strategy

Stratum	$\bar{n}$	Fixed CV			Proportional allocation			Optimum allocation		
		CV	DEFT^2	n	CV	DEFT^2	n	CV	DEFT^2	n
Gyeonggi	20	3.10	3.92	600	2.55	3.92	884	2.39	3.92	1,011
Gangwon	20	3.10	2.00	299	3.34	2.00	258	3.91	2.00	188
Chungbuk	20	3.10	4.23	378	3.42	4.23	310	3.43	4.23	309
Chungnam	20	3.10	3.77	497	3.15	3.77	482	3.12	3.77	491
Jeonbuk	20	3.10	1.88	157	2.55	1.88	232	3.36	1.88	134
Jeonnam	20	3.10	5.39	426	3.10	5.39	426	3.03	5.39	446
Gyeonbuk	20	3.10	7.58	1,135	4.21	7.58	615	3.57	7.58	853
Gyeonnam	20	3.10	1.99	191	1.72	1.99	621	2.10	1.99	416
Jeju	20	3.10	2.14	212	5.49	2.14	68	6.62	2.14	47
Total	20	1.17	4.48	3,895	1.10	3.95	3,895	1.07	3.71	3,895

동질성계수와 불균등가중치로 의한 분산증가분이 상대적으로 큰 경북은 약 2배인 1,135개의 응답가구가 있어야 동일한 CV값을 갖게 됨을 보여준다. 고정 CV 기준으로 층별 표본수를 정하였을 때 전국은 총 3,895개 응답수가 필요하고, 표본추정량의 상대표준오차는 식 (3.4)에 의해 약 1.17%이 된다. 만약, 동일한 목표 표본수 3,895개 응답가구를 기준으로 표본층의 가구수에 비례한 표본할당을 고려하면, 상대규모가 가장 큰 경기도에 884개의 응답가구가 할당되어 고정 CV에 의한 할당결과와는 상이함을 알 수 있다. 비례할당에 의한 전국수준의 표본추정량이 갖는 상대표준오차는 고정 CV 보다 다소 작은 1.10%의 값을 갖게 된다. 반면, 전국수준의 평균추정량의 표본오차를 가장 작게하는 최적할당 식 (3.6)은 표본층의 상대크기, 특성분산, 층별 설계효과에 비례하여 표본수를 정한다. 경기도의 경우에는 층 크기와 특성분산의 크기에 비례하여 다른 두 가지의 표본할당에 비해 가장 많은 1,011개 응답가구가 된다. 전국수준의 상대표준오차는 1.07로 가장 낮은 값을 갖는다.

Table 4.4는 설계요소별로 갖는 영향력을 반영한 설계효과모형을 통해 산출한 식 (3.5)의 상대표준오차 $\mathrm{CV}_{k+1,h}$과 단순비례식 (1.1)에 근거해 산출한 다음의 비례산출 상대표준오차를 비교하고 있다.

$$\mathrm{CV}^*_{k+1,h} = \sqrt{\frac{n_k}{n_{k+1}}\widehat{\mathrm{CV}}^2_k}$$

더불어 두 상대표준오차 간의 상대차이(%) (즉, $100 \times (\mathrm{CV}^*_{k+1,h} - \mathrm{CV}_{k+1,h})/\mathrm{CV}_{k+1,h}$)를 정리하고 있다. 작게는 0.1%에서 크기는 −49.9%의 큰 상대차이를 보이고 있다. 이는 설계요소들의 비선형적 영향력으로 인해 층별 표본수에 단순히 비례하는 방식은 실질적 상대표준오차를 예측하는데는 적절하지 않

Table 4.4. Comparisons of CVs between Design Effect Formula and Relative Variance Ratio Approaches

Stratum	Design effect formula			$CV^*_{k+1,h}$			Relative difference		
	Fixed CV	Prop	Optimum	Fixed CV	Prop	Optimum	Fixed CV	Prop	Optimum
Gyeonggi	3.10	2.55	2.39	3.13	3.13	2.41	1.1%	22.7%	1.1%
Gangwon	3.10	3.34	3.91	3.11	3.11	3.92	0.2%	−7.0%	0.1%
Chungbuk	3.10	3.42	3.43	3.10	3.10	3.43	0.1%	−9.3%	0.1%
Chungnam	3.10	3.15	3.12	3.11	3.11	3.13	0.2%	−1.4%	0.1%
Jeonbuk	3.10	2.55	3.36	3.11	3.11	3.36	0.2%	21.8%	0.1%
Jeonnam	3.10	3.10	3.03	3.10	3.10	3.03	0.1%	0.1%	0.1%
Gyeongbuk	3.10	4.21	3.57	3.10	3.10	3.58	0.1%	−26.3%	0.1%
Gyeongnam	3.10	1.72	2.10	3.09	3.09	2.09	−0.3%	79.7%	−0.3%
Jeju	3.10	5.49	6.62	2.75	2.75	5.87	−11.2%	−49.9%	−11.3%

을 수도 있음을 나타낸다.

2주기 복지실태조사의 상세한 표본설계 내역은 Rural Development Administration (2018)을 참고할 수 있다. 최종적 표본수 결정은 Table 4.4에 명시된 표본수와는 다소 차이가 있는데 실질적인 표본재설계는 1주기 5년 조사에 포함한 주요 만족도 변수 5개(지역생활여건만족도, 경제활동여건만족도, 보건의료만족도, 복지여건만족도, 안전만족도)를 중심으로 표본오차가 안정적이도록 고려하였기 때문이다. 변수별로 5년간 개별 설계효과요소의 조화평균을 계산하고 이의 $a = 1/3$을 승수로 하는 멱배분(power allocation)을 적용한 후, 주요 분석영역인 표본층별 상대표준오차가 적절한 범위를 갖도록 추가적인 조정을 실시하였다.

5. 논의

본 연구에서는 반복조사에서 설계요소를 반영한 표본수 결정에 관하여 살펴보았다. 흔히 고려되는 기존 조사의 상대표준오차와 표본수에 대비하여 재설계의 목표상대표준오차와 예상표본수의 비례적 관계를 이용한 표본수 결정식 (1.1)에 비하여 표본층 구성에 따른 모집단 구조 및 특성, 표본할당, 집락효과, 가중치 변동 등의 세부적 내역을 상세히 반영할 수 있는 장점이 있다.

반복조사의 표본수 결정에 있어서 설계효과모형을 적용한다면 매우 전략적인 표본재설계가 될 수 있지만, 설계효과모형식의 타당성과 설계요소 모수의 추정량에 대한 추가적인 연구(예, Hwang, 2019)는 향후 과제로 남긴다.

References

Chen, S. and Rust, K. (2017). An extension of Kish's formula for design effects to two- and three-stage designs with stratification, *Journal of Survey Statistics and Methodology,* **5**, 111–130.

Gabler, S., Häder, S., and Lynn, P. (2006). Design Effects for Multiple Design Samples, *Survey Methodology*, **32**, 115–120.

Hwang, H. G. (2019). *Evaluation of design effect models for complex surveys* (Master's Thesis), Pukyong National University, Busan.

Kim, K. S. (2012). Sample size determination in repeated surveys with varying population sizes, *Survey Research*, **11**, 159–174.

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

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

Kish, L. (1995). Methods for design effects, *Journal of Official Statistics,* **11**, 55–77.

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, H. N. (1989). *Statistical Survey* (2nd ed.), Youngji Publishers, Seoul.

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

Park, I. (2015). Understanding complex design features via design effect models, *The Korean Journal of Applied Statistics*, **28**, 1217–1225.

Park, I. (2016). Choosing clusters for two-stage household surveys, *Journal of the Korean Data and Information Science Society*, **27**, 363–372.

Rural Development Administration (2018). *2018 Survey on Rural Well-Being*.

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

Statistics Korea (2007). *Review of the Sample Redesign of Household Surveys*.

반복조사에서 설계요소를 반영한 표본수 결정

박인호[a,1] · 황현길[a]

[a]부경대학교 통계학과

(Received June 26, 2019; Revised July 21, 2019; Accepted July 26, 2019)

요 약

본 연구에서는 반복조사의 표본재설계에서 설계요소를 반영한 표본수 결정 방법을 제안하였다. 제안된 방법은 다단추출과 층화다단추출 등에 적용할 수 있으며 시점간 모집단 구성 변화, 집락효과, 표본할당 등의 주된 설계요소가 갖는 표본오차에 대한 영향력을 구분하여 반영하므로 보다 전략적인 표본수 결정이 가능할 수 있다.

주요용어: 반복조사, 상대표준오차, 설계효과모형식, 층화다단추출

이 논문은 부경대학교 자율창의학술연구비(2017년)에 의하여 연구되었음.

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