The Korean Journal of Applied Statistics (2013) DOI: http://dx.doi.org/10.5351/KJAS.2013.26.2.209
26(2), 209–222

A Study on the Stratified Cluster Replicated Systematic Unrelated Question Model

Gi-Sung Lee[a,1]

[a]Department of Children Welfare, Woosuk University

(Received December 17, 2012; Revised February 6, 2013; Accepted February 7, 2013)

Abstract

We apply stratified cluster sampling to a replicated systematic unrelated question model for a large scale survey in which the population is comprised of several strata developed by several clusters and with sensitive parameters. We first present a replicated systematic unrelated question model using an unrelated question model to procure sensitive information from the population of clusters and then develop a suggested model to an unrelated question by a stratified cluster replicated systematic sampling that can be used in large population of strata. We cover the proportional and optimum allocation for the suggested model. Finally, we compare and analyze the efficiency of the suggested model with the replicated systematic unrelated question model.

Keywords: Stratified cluster sampling, replicated systematic sampling, unrelated question model, proportional allocation, optimal allocation.

1. 서론

Warner (1965)는 사회적으로 혹은 개인적으로 민감한 내용의 조사에서 정확한 정보를 얻기 위해 확률장치를 이용한 확률화응답모형(randomized response model; RRM)을 처음으로 제시하였으며, 이때의 확률장치는 두 개의 질문으로 두 질문 중 하나는 민감한 질문이고 다른 하나는 민감한 질문과 배반되는 질문으로 구성되었다. Greenberg 등 (1969)은 민감한 질문과 배반되는 질문 대신에 민감한 질문과 전혀 관계가 없는 질문을 사용하는 무관질문모형(unrelated question model)을 제안하여 그 이론적 체계를 구축하였다. 이후 새로운 모형의 개발과 다양한 확률장치가 고안되는 등 이에 대한 많은 연구가 이루어지고 있다. 최근에는 확률화응답모형의 실용화를 위해 모집단의 특성을 고려한 다양한 추출법들이 확률화응답모형에 적용되고 있다. 특히, Ahn과 Lee (2003, 2004)은 모집단이 여러 개의 층으로 구성되어 있고, 얻고자 하는 민감한 정보가 질적인 속성일 때 사용할 수 있는 층화 무관질문모형과 민감한 속성에 대한 정보가 이산인 양적속성일 때 사용할 수 있는 층화추출법에 의한 이산 양적 확률화응답모형을 제안하였다. 그리고 Kim과 Warde (2004)는 층화 Warner 모형을 제안하였고, Kim과 Elam (2005)은 최적할당을 이용한 2단계 층화 Warner 모형을 제안하기도 하였다. 또한 Lee와 Park (2007)은 모집단이

This research was supported by Basic Science Research Program through the National Research Foundation of Korea(NRF) funded by the Ministry of Education, Science and Technology(2011-0007800).
[1]Professor, Department of Children Welfare, Woosuk University, 490 Hujeong-ri, Wanju-gun, Jeonbuk 565-701, Korea. E-mail: gisung@woosuk.ac.kr

여러 개의 집락으로 구성되어 있을 때, 민감한 정보를 얻을 수 있는 확률비례추출법에 의한 무관질문모형을 제안하였고, Lee (2010, 2012)은 반복계통추출법을 Warner 모형과 무관질문모형에 적용하였으며, 이를 층화 반복계통 확률화응답모형으로 발전시켰다.

본 논문에서는 대규모 표본조사에서 많이 나타나는 조사하고자 하는 모집단이 층으로 형성되어 있고, 각 층들이 집락으로 구성되어 있을 때 사용 가능한 층화 집락추출법을 얻고자 하는 정보가 민감할 때 반복계통 무관질문모형에 적용하였다. 먼저 모집단이 집락으로 구성되어 있고, 추출된 집락으로부터 계통표본을 반복적으로 추출하여 민감한 정보를 얻는 데 무관질문모형을 사용한 집락 반복계통 무관질문모형을 제안하였다. 다음으로 제안한 모형을 층화된 모집단에서도 사용할 수 있도록 층화 집락 반복계통추출법에 의한 무관질문모형으로 발전시켰으며, 각 층의 집락을 확률비례복원추출 또는 확률비례비복원추출하는 층화 확률비례 반복계통 무관질문모형을 제안하였다. 또한 제안한 층화집락 반복계통추출법에 의한 무관질문모형에서 각 층의 표본배분하는 문제를 비례배분과 최적배분 측면에서 다루었다. 마지막으로 제안한 층화집락 반복계통추출법에 의한 무관질문모형과 집락 반복계통추출법에 의한 무관질문모형과의 효율성을 비교하였다.

2. 2단계 집락 반복계통 무관질문모형

이 절에서는 민감한 조사에서 각 집락의 크기가 $M_i\,(i=1,2,\ldots,N)$인 N개의 집락으로 구성되어 있는 모집단으로부터 n개의 집락을 단순임의추출한 후, 추출된 각 집락에서 다시 $m_i\,(i=1,2,\ldots,n)$개인 계통표본을 추출하는 데 있어서 두 개 이상의 부차표본을 독립적으로 추출하는 집락 반복계통추출법을 무관질문모형에 적용한 2단계 집락 반복계통 무관질문모형을 제안하고자 한다.

i번째 집락으로부터 m_i인 계통표본을 추출하는 데 있어서 크기가 m_i' 부차표본을 k_i개 추출할 경우 $m_i'=m_i/k_i$이고 k_i개의 표본의 단위 수는 모두 m_i'개가 된다.

i번째 집락으로부터 계통추출된 응답자들에게 다음과 같은 Greenberg 등 (1969)의 무관질문모형 확률장치 R_i를 사용하도록 한다.

설문 1 : 당신은 민감한 그룹 A에 속합니까?

설문 2 : 당신은 무관한 그룹 Y에 속합니까?

여기서, 설문 1이 선택될 확률은 p_i이고, 설문 2가 선택될 확률은 $1-p_i$이다. 응답자들은 확률장치에 의해서 선택된 설문에 대해 "예" 또는 "아니오"라고 응답한다.

따라서 $i\,(i=1,2,\ldots,N)$번째 집락의 $j\,(j=1,2,\ldots,M_i)$번째 부차표본 응답자가 "예"라고 응답할 확률은

$$\lambda_{ij}=p_i\pi_{ij}+(1-p_i)\pi_{yij} \tag{2.1}$$

이고, 이 때 π_{ij}는 i번째 집락의 j번째 부차표본에서의 민감한 속성의 모비율이고, π_{yij}는 i번째 집락의 j번째 부차표본에서의 무관한 속성의 모비율이며 알고 있다고 가정한다.

i번째 집락의 j번째 부차표본의 l번째 추출단위의 관찰치를 z_{ijl}이라 하고, 최종단위인 응답자가 "예"라고 응답하면 $z_{ijl}=1$, "아니오"라고 응답하면 $z_{ijl}=0$이라고 정의하자. 이 때, i번째 집락의 j번째 부차표본으로 추출된 응답자들 중에서 "예"라고 응답한 사람의 수를 $Z_{ij}=\sum_{l=1}^{m_i'}z_{ijl}$이라 하면, $\hat{\lambda}_{ij}=Z_{ij}/m_i'$이 되므로, 식 (2.1)로부터 π_{ij}의 추정량 $\hat{\pi}_{ij}$은 다음과 같다.

$$\hat{\pi}_{ij}=\frac{\hat{\lambda}_{ij}-(1-p_i)\pi_{yij}}{p_i}.$$

그리고 i번째 집락의 민감한 속성의 모비율 π_i의 추정량 $\hat{\pi}_i$은 다음과 표현할 수 있다.

$$\hat{\pi}_i = \frac{1}{k_i}\sum_{j=1}^{k_i}\hat{\pi}_{ij}.$$

크기 $m_i'(=m_i/k_i)$인 독립적인 부차표본을 k_i개 추출한다는 것은 $g_i'(=k_ig_i)$개의 가능한 표본 중에서 k_i개를 계통추출한다는 뜻이 된다. 이 때, g_i는 계통추출에서의 추출간격을 의미한다. 그러면 추정량 $\hat{\pi}_i$의 분산은 비복원추출을 가정할 경우 다음과 같다.

$$V(\hat{\pi}_i) = \frac{k_ig_i - k_i}{k_ig_i}\frac{1}{k_i}\frac{1}{k_ig_i - 1}\sum_{j=1}^{k_ig_i}(\hat{\pi}_{ij} - \pi_i)^2,$$

여기서 $\hat{\pi}_{ij}$는 i번째 집락의 $j\,(i=1,2,\ldots,k_ig_i)$번째 부차표본에서 민감한 그룹에 속하는 표본비율이다. 또한 이러한 등확률 2단계 집락 반복계통 추출절차에 의해 얻어진 민감한 그룹에 속하는 조사단위당 모비율 π의 추정량 $\hat{\pi}$는 다음과 같다.

$$\hat{\pi} = \frac{N}{nM_0}\sum_{i=1}^{n}M_i\hat{\pi}_i,$$

여기서 $M_0 = \sum_{i=1}^{N}M_i$이다.

정리 2.1 추정량 $\hat{\pi}$는 모비율 π의 비편향추정량이다.

증명:

$$\begin{aligned}
E(\hat{\pi}) &= E_1E_2\left[\frac{N}{nM_0}\sum_{i=1}^{n}M_i\hat{\pi}_i\right] \\
&= E_1\left[\frac{N}{nM_0}\sum_{i=1}^{n}M_i\pi_i\right] \\
&= \frac{1}{M_0}\sum_{i=1}^{N}M_i\pi_i \\
&= \pi.
\end{aligned}$$

□

정리 2.2 N개의 집락에서 n개를 단순임의복원추출하고, 추출된 각 집락내의 M_i개 단위 중에서 m_i개인 계통표본을 추출하는 데 있어서 크기가 $m_i'(=m_i/k_i)$인 부차표본을 k_i개 추출할 경우 민감한 그룹에 속하는 모비율 π의 추정량 $\hat{\pi}$의 분산은 다음과 같다.

$$V(\hat{\pi}) = \frac{N}{nM_0^2}\sum_{i=1}^{N}\left(M_i\pi_i - \bar{M}\pi\right)^2 + \frac{N}{nM_0^2}\sum_{i=1}^{N}M_i^2\frac{k_ig_i - k_i}{k_ig_i}\frac{1}{k_i}\frac{1}{k_ig_i - 1}\sum_{j=1}^{k_ig_i}(\hat{\pi}_{ij} - \pi_i)^2, \tag{2.2}$$

여기서 $\bar{M} = M_0/N$이다.

증명:

$$V(\hat{\pi}) = E_1V_2(\hat{\pi}) + V_1E_2(\hat{\pi})$$

에서

$$\begin{aligned} V_1E_2(\hat{\pi}) &= V_1E_2\left[\frac{N}{nM_0}\sum_{i=1}^{n}M_i\hat{\pi}_i\right] \\ &= V_1\left[\frac{N}{nM_0}\sum_{i=1}^{n}M_i\pi_i\right] \\ &= \frac{N}{nM_0^2}\sum_{i=1}^{N}\left(M_i\pi_i-\bar{M}\pi\right)^2 \end{aligned}$$

이고,

$$\begin{aligned} E_1V_2(\hat{\pi}) &= E_1V_2\left[\frac{N}{nM_0}\sum_{i=1}^{n}M_i\hat{\pi}_i\right] \\ &= E_1\left[\frac{N^2}{(nM_0)^2}\sum_{i=1}^{n}M_i^2V_2(\hat{\pi}_i)\right] \\ &= E_1\left[\frac{N^2}{(nM_0)^2}\sum_{i=1}^{n}M_i^2\frac{k_ig_i-k_i}{k_ig_i}\frac{1}{k_i}\frac{1}{k_ig_i-1}\sum_{j=1}^{k_ig_i}(\hat{\pi}_{ij}-\pi_i)^2\right] \\ &= \frac{N}{nM_0^2}\sum_{i=1}^{N}M_i^2\frac{k_ig_i-k_i}{k_ig_i}\frac{1}{k_i}\frac{1}{k_ig_i-1}\sum_{j=1}^{k_ig_i}(\hat{\pi}_{ij}-\pi_i)^2 \end{aligned}$$

이므로 추정량 $\hat{\pi}$의 분산은 식 (2.2)와 같다. □

한편, N개의 집락에서 n개를 단순임의비복원추출하고, 추출된 각 집락내의 M_i개 단위 중에서 m_i개인 계통표본을 추출하는 데 있어서 크기가 $m_i'(=m_i/k_i)$인 부차표본을 k_i개 추출할 경우 민감한 그룹에 속하는 모비율 π의 추정량 $\hat{\pi}$의 분산은 다음과 같다.

$$V(\hat{\pi}) = \frac{N-n}{nM_0^2}\sum_{i=1}^{N}\left(M_i\pi_i-\bar{M}\pi\right)^2 + \frac{N}{nM_0^2}\sum_{i=1}^{N}M_i^2\frac{k_ig_i-k_i}{k_ig_i}\frac{1}{k_i}\frac{1}{k_ig_i-1}\sum_{j=1}^{k_ig_i}(\hat{\pi}_{ij}-\pi_i)^2. \quad (2.3)$$

3. 확률비례 반복계통 무관질문모형

각 집락의 크기가 $M_i\,(i=1,2,\ldots,N)$인 N개의 집락으로 구성되어 있는 모집단으로부터 n개의 집락을 확률비례추출한 후, 추출된 각 집락에서 다시 $m_i\,(i=1,2,\ldots,n)$개인 계통표본을 추출하는 데 있어서 두 개 이상의 부차표본을 독립적으로 추출하는 반복계통추출법을 무관질문모형에 적용한 확률비례 반복계통 무관질문모형을 제안하고자 한다. 먼저, 집락을 확률비례복원추출하는 방법에 대하여 살펴보고, 다음으로 확률비례비복원추출하는 방법에 대하여 다루어 보고자 한다.

n개의 1차 추출단위를 i번째 1차 추출단위의 추출확률 t_i에 의해서 복원추출하고, 각 1차 추출단위에서 m_i개의 2차 단위를 반복계통추출한다고 가정하자. i번째 집락으로부터 m_i인 계통표본을 추출하는 데 있어서 크기가 m_i'인 부차표본을 k_i개 추출할 경우 $m_i'=m_i/k_i$이고 k_i개의 표본의 단위 수는 모두 m_i'개가 된다. i번째 집락으로부터 계통추출된 응답자들에게 Greenberg 등의 무관질문모형 확률장치 R_i를 사용하도록 한다.

응답자들이 추출확률 t_i에 의해 복원추출된 $i\,(i=1,2,\ldots,n)$번째 집락으로부터 반복계통추출되었을

때, 이런 절차에 의해 얻어진 민감한 그룹에 속하는 조사단위당 모비율 π의 추정량 $\hat{\pi}_{ppz}$는 다음과 같다.

$$\hat{\pi}_{ppz} = \frac{1}{nM_0}\sum_{i=1}^{n}\frac{M_i}{t_i}\hat{\pi}_i,$$

여기서 $M_0 = \sum_{i=1}^{N} M_i$이다.

정리 3.1 추정량 $\hat{\pi}_{ppz}$는 모비율 π의 비편향추정량이다.

증명:

$$\begin{aligned} E(\hat{\pi}_{ppz}) &= E_1E_2\left[\frac{1}{nM_0}\sum_{i=1}^{n}\frac{M_i}{t_i}\hat{\pi}_i\right] \\ &= E_1\left[\frac{1}{nM_0}\sum_{i=1}^{n}\frac{M_i\pi_i}{t_i}\right] \\ &= \frac{1}{M_0}\sum_{i=1}^{N}t_i\frac{M_i\pi_i}{t_i} \\ &= \pi. \end{aligned}$$

□

정리 3.2 각 집락의 크기가 M_i인 N개의 집락에서 n개의 집락을 추출확률 t_i에 의해 복원추출하고, 추출된 각 집락내의 M_i개 단위 중에서 m_i개인 계통표본을 추출하는 데 있어서 크기가 $m_i'\,(= m_i/k_i)$인 부차표본을 k_i개 추출할 경우 민감한 그룹에 속하는 모비율 π의 추정량 $\hat{\pi}_{ppz}$의 분산은 다음과 같다.

$$V(\hat{\pi}_{ppz}) = \frac{1}{nM_0^2}\sum_{i=1}^{N}t_i\left(\frac{M_i\pi_i}{t_i} - M_0\pi\right)^2 + \frac{1}{nM_0^2}\sum_{i=1}^{N}\frac{M_i^2}{t_i}\frac{k_ig_i-k_i}{k_ig_i}\frac{1}{k_i}\frac{1}{k_ig_i-1}\sum_{j=1}^{k_ig_i}(\hat{\pi}_{ij}-\pi_i)^2. \tag{3.1}$$

증명:

$$V(\hat{\pi}_{ppz}) = E_1V_2(\hat{\pi}_{ppz}) + V_1E_2(\hat{\pi}_{ppz})$$

에서

$$\begin{aligned} V_1E_2(\hat{\pi}_{ppz}) &= V_1E_2\left[\frac{1}{nM_0}\sum_{i=1}^{n}\frac{M_i}{t_i}\hat{\pi}_i\right] \\ &= V_1\left[\frac{1}{nM_0}\sum_{i=1}^{n}\frac{M_i\pi_i}{t_i}\right] \\ &= \frac{1}{nM_0^2}\sum_{i=1}^{N}t_i\left(\frac{M_i\pi_i}{t_i} - M_0\pi\right)^2 \end{aligned}$$

이고,

$$E_1V_2(\hat{\pi}_{ppz}) = E_1V_2\left[\frac{1}{nM_0}\sum_{i=1}^{n}\frac{M_i}{t_i}\hat{\pi}_i\right]$$

$$= E_1\left[\frac{1}{(nM_0)^2}\sum_{i=1}^{n}\frac{M_i^2}{t_i^2}V_2(\hat{\pi}_i)\right]$$
$$= E_1\left[\frac{1}{(nM_0)^2}\sum_{i=1}^{n}\frac{M_i^2}{t_i^2}\frac{k_ig_i-k_i}{k_ig_i}\frac{1}{k_i}\frac{1}{k_ig_i-1}\sum_{j=1}^{k_ig_i}(\hat{\pi}_{ij}-\pi_i)^2\right]$$
$$= \frac{n}{(nM_0)^2}\sum_{i=1}^{N}t_i\frac{M_i^2}{t_i^2}\frac{k_ig_i-k_i}{k_ig_i}\frac{1}{k_i}\frac{1}{k_ig_i-1}\sum_{j=1}^{k_ig_i}(\hat{\pi}_{ij}-\pi_i)^2$$
$$= \frac{1}{nM_0^2}\sum_{i=1}^{N}\frac{M_i^2}{t_i}\frac{k_ig_i-k_i}{k_ig_i}\frac{1}{k_i}\frac{1}{k_ig_i-1}\sum_{j=1}^{k_ig_i}(\hat{\pi}_{ij}-\pi_i)^2$$

이므로 추정량 $\hat{\pi}_{ppz}$의 분산은 식 (3.1)과 같다. □

한편, n개의 1차 추출단위가 각 집락의 크기 M_i에 비례할 경우 $t_i = M_i/M_0$가 되며, 이를 확률비례추출(probability proportional to size; pps)이라 한다.

n개의 1차 추출단위를 확률비례복원추출한 후, 추출된 각 집락에서 다시 m_i개의 2차 단위를 반복계통추출할 경우, 이러한 2단계 절차에 의해 얻어진 민감한 그룹에 속하는 모비율 π의 추정량 $\hat{\pi}_{ppswr}$는 다음과 같다.

$$\hat{\pi}_{ppswr} = \frac{1}{n}\sum_{i=1}^{n}\hat{\pi}_i.$$

그리고 $\hat{\pi}_{ppswr}$의 분산은 다음과 같다.

$$V(\hat{\pi}_{ppswr}) = \frac{1}{nM_0}\sum_{i=1}^{N}M_i(\pi_i-\pi)^2 + \frac{1}{nM_0}\sum_{i=1}^{N}M_i\frac{k_ig_i-k_i}{k_ig_i}\frac{1}{k_i}\frac{1}{k_ig_i-1}\sum_{j=1}^{k_ig_i}(\hat{\pi}_{ij}-\pi_i)^2.$$

다음으로 각 집락의 크기가 M_i인 N개의 집락으로 구성되어 있는 모집단으로부터 n개의 집락을 확률비례비복원추출한 후, 추출된 각 집락에서 다시 m_i개의 조사단위를 반복계통추출한다고 하자.

이러한 확률비례비복원 추출절차에 의해 얻어진 민감한 그룹에 속하는 조사단위당 모비율 π의 추정량 $\hat{\pi}_{ppswor}$는 다음과 같다.

$$\hat{\pi}_{ppswor} = \frac{1}{M_0}\sum_{i=1}^{n}\frac{M_i}{\theta_i}\hat{\pi}_i,$$

여기서 θ_i는 조사단위 i가 표본에 포함되는 확률이다.

그리고 $\hat{\pi}_{ppswor}$의 분산은 다음과 같다.

$$V(\hat{\pi}_{ppswor}) = \frac{1}{M_0^2}\sum_{i=1}^{N}\sum_{j>i}^{N}(\theta_i\theta_j-\theta_{ij})\left(\frac{M_i\pi_i}{\theta_i}-\frac{M_j\pi_j}{\theta_j}\right)^2 + \frac{1}{M_0^2}\sum_{i=1}^{N}\frac{M_i^2}{\theta_i}\frac{k_ig_i-k_i}{k_ig_i}\frac{1}{k_i}\frac{1}{k_ig_i-1}\sum_{j=1}^{k_ig_i}(\hat{\pi}_{ij}-\pi_i)^2,$$

여기서 θ_{ij}는 조사단위 i와 j가 동시에 표본에 포함되는 확률이다.

4. 층화집락 반복계통 무관질문모형

대규모 표본조사에서 많이 사용되는 층화 2단계 집락추출법은 모집단을 층화한 다음 각 층에서 집락을 추출한 후 추출된 집락에서 최종 조사단위를 추출하는 방법이다. 이 절에서는 매우 민감한 조사에서 조

사하고자 하는 모집단이 L개의 층으로 형성되어 있고, h층에는 $N_h\,(h=1,2,\ldots,L)$개의 1차 추출단위가 포함되고, 1차 단위는 $M_{hi}\,(i=1,2,\ldots,N_h)$개의 2차 단위를 포함하고 있다고 가정할 때, h층의 N_h개의 집락에서 n_h개를 단순임의추출하고, 추출된 각 집락내의 M_{hi}개 단위 중에서 m_{hi}개인 계통표본을 추출하는 데 있어서 두 개 이상의 부차표본을 독립적으로 추출하는 층화집락 반복계통추출법을 무관질문모형에 적용한 층화 집락 반복계통 무관질문모형을 제안하고자 한다. 그리고 각 층의 집락을 확률비례복원추출 또는 확률비례비복원추출하는 층화 확률비례 반복계통 무관질문모형을 제안하고자 한다. 또한, 제안한 층화집락 반복계통추출법에 의한 무관질문모형에서 각 층의 표본배분하는 문제를 비례배분과 최적배분 측면에서 다루고자 한다.

h층의 i번째 집락으로부터 m_{hi}인 계통표본을 추출하는 데 있어서 크기가 m'_{hi}인 부차표본을 k_{hi}개 추출할 경우 $m'_{hi}=m_{hi}/k_{hi}$이고 k_{hi}개의 표본의 단위 수는 모두 m'_{hi}개가 된다. h층의 i번째 집락으로부터 계통추출된 응답자들에게 다음과 같은 Greenberg 등의 무관질문모형 확률장치 R_{hi}를 사용하도록 한다.

설문 1 : 당신은 민감한 그룹 A에 속합니까?

설문 2 : 당신은 무관한 그룹 Y에 속합니까?

여기서, 설문 1이 선택될 확률은 p_{hi}이고, 설문 2가 선택될 확률은 $1-p_{hi}$이다. 응답자들은 확률장치에 의해서 선택된 설문에 대해 "예" 또는 "아니오"라고 응답한다.

따라서 $h\,(h=1,2,\ldots,L)$층의 $i\,(i=1,2,\ldots,N_h)$번째 집락의 $j\,(j=1,2,\ldots,M_{hi})$번째 부차표본 응답자가 "예"라고 응답할 확률은

$$\lambda_{hij}=p_{hi}\pi_{hij}+(1-p_{hi})\pi_{hyij} \tag{4.1}$$

이고, 이 때 π_{hij}는 h층의 i번째 집락의 j번째 부차표본에서의 민감한 속성의 모비율이고, π_{hyij}는 h층의 i번째 집락의 j번째 부차표본에서의 무관한 속성의 모비율이며 알고 있다고 가정한다.

h층의 i번째 집락의 j번째 부차표본의 l번째 추출단위의 관찰치를 z_{hijl}이라 하고, 최종단위인 응답자가 "예"라고 응답하면 $z_{hijl}=1$, "아니오"라고 응답하면 $z_{hijl}=0$이라고 정의하자. 이 때, h층의 i번째 집락의 j번째 부차표본으로 추출된 응답자들 중에서 "예"라고 응답한 사람의 수를 $Z_{hij}=\sum_{l=1}^{m'_{hi}} z_{hijl}$이라 하면, $\hat{\lambda}_{hij}=Z_{hij}/m'_{hi}$이 되므로, 식 (4.1)로부터 π_{hij}의 추정량 $\hat{\pi}_{hij}$은 다음과 같다.

$$\hat{\pi}_{hij}=\frac{\hat{\lambda}_{hij}-(1-p_{hi})\pi_{hyij}}{p_{hi}}.$$

그리고 h층의 i번째 집락의 민감한 속성의 모비율 π_{hi}의 추정량 $\hat{\pi}_{hi}$은 다음과 표현할 수 있다.

$$\hat{\pi}_{hi}=\frac{1}{k_{hi}}\sum_{j=1}^{k_{hi}}\hat{\pi}_{hij}.$$

크기 $m'_{hi}\,(=m_{hi}/k_{hi})$인 독립적인 부차표본을 k_{hi}개 추출한다는 것은 $g'_{hi}\,(=k_{hi}g_{hi})$개의 가능한 표본 중에서 k_{hi}개를 계통추출한다는 뜻이 된다. 이 때, g_{hi}는 h층의 계통추출에서의 추출간격을 의미한다. 그러면 추정량 $\hat{\pi}_{hi}$의 분산은 비복원추출을 가정할 경우 다음과 같다.

$$V(\hat{\pi}_{hi})=\frac{k_{hi}g_{hi}-k_{hi}}{k_{hi}g_{hi}}\frac{1}{k_{hi}}\frac{1}{k_{hi}g_{hi}-1}\sum_{j=1}^{k_{hi}g_{hi}}(\hat{\pi}_{hij}-\pi_{hi})^2,$$

여기서 $\hat{\pi}_{hij}$는 h층의 i번째 집락의 $j\,(i=1,2,\ldots,k_{hi}g_{hi})$번째 부차표본에서의 민감한 그룹에 속하는 표본비율이다.

또한 h층의 i번째 집락의 j번째 부차표본에서의 모비율 π_h의 추정량 $\hat{\pi}_h$는 다음과 같다.

$$\hat{\pi}_h = \frac{N_h}{n_h M_{h0}} \sum_{i=1}^{n_h} M_{hi}\hat{\pi}_{hi}.$$

그러므로 민감한 그룹에 속하는 모비율 π의 층화추정량 $\hat{\pi}_{stcl}$는 다음과 같다.

$$\hat{\pi}_{stcl} = \sum_{h=1}^{L} W_h \hat{\pi}_h = \sum_{h=1}^{L} W_h \frac{N_h}{n_h M_{h0}} \sum_{i=1}^{n_h} M_{hi}\hat{\pi}_{hi}, \quad W_h = \frac{N_h}{N}.$$

정리 4.1 층화추정량 $\hat{\pi}_{stcl}$는 모비율 π의 비편향추정량이다.

증명:

$$\begin{aligned}
E(\hat{\pi}_{stcl}) &= E_1 E_2 \left[\sum_{h=1}^{L} W_h \frac{N_h}{n_h M_{h0}} \sum_{i=1}^{n_h} M_{hi}\hat{\pi}_{hi} \right] \\
&= E_1 \left[\sum_{h=1}^{L} W_h \frac{N_h}{n_h M_{h0}} \sum_{i=1}^{n_h} M_{hi}\pi_{hi} \right] \\
&= \sum_{h=1}^{L} W_h \frac{1}{M_{h0}} \sum_{i=1}^{N_h} M_{hi}\pi_{hi} \\
&= \sum_{h=1}^{L} W_h \pi_h \\
&= \pi.
\end{aligned}$$

□

정리 4.2 h층의 N_h개의 집락에서 n_h개를 단순임의복원추출하고, 추출된 각 집락내의 M_{hi}개 단위 중에서 m_{hi}개인 계통표본을 추출하는 데 있어서 크기가 $m'_{hi}(= m_{hi}/k_{hi})$인 부차표본을 k_{hi}개 추출할 경우 민감한 그룹에 속하는 모비율 π의 층화추정량 $\hat{\pi}_{stcl}$의 분산은 다음과 같다.

$$\begin{aligned}
V(\hat{\pi}_{stcl}) = \sum_{h=1}^{L} W_h^2 \frac{N_h}{n_h M_{h0}^2} \Bigg[& \sum_{i=1}^{N_h} \left(M_{hi}\pi_{hi} - \bar{M}_h \pi_h \right)^2 \\
& + \sum_{i=1}^{N_h} M_{hi}^2 \frac{k_{hi} g_{hi} - k_{hi}}{k_{hi} g_{hi}} \frac{1}{k_{hi}} \frac{1}{k_{hi} g_{hi} - 1} \sum_{j=1}^{k_{hi} g_{hi}} (\hat{\pi}_{hij} - \pi_{hi})^2 \Bigg].
\end{aligned} \tag{4.2}$$

증명:

$$V(\hat{\pi}_{stcl}) = E_1 V_2(\hat{\pi}_{stcl}) + V_1 E_2(\hat{\pi}_{stcl})$$

에서

$$\begin{aligned}
V_1 E_2(\hat{\pi}_{stcl}) &= V_1 E_2 \left[\sum_{h=1}^{L} W_h \frac{N_h}{n_h M_{h0}} \sum_{i=1}^{n_h} M_{hi}\hat{\pi}_{hi} \right] \\
&= V_1 \left[\sum_{h=1}^{L} W_h \frac{N_h}{n_h M_{h0}} \sum_{i=1}^{n_h} M_{hi}\pi_{hi} \right] \\
&= \sum_{h=1}^{L} W_h^2 \frac{N_h}{n_h M_{h0}^2} \sum_{i=1}^{N_h} \left(M_{hi}\pi_{hi} - \bar{M}_h \pi_h \right)^2
\end{aligned}$$

이고,

$$\begin{aligned}
E_1V_2(\hat{\pi}_{stcl}) &= E_1V_2\left[\sum_{h=1}^{L} W_h \frac{N_h}{n_h M_{h0}} \sum_{i=1}^{n_h} M_{hi}\hat{\pi}_{hi}\right] \\
&= E_1\left[\sum_{h=1}^{L} W_h^2 \frac{N_h^2}{(n_h M_{h0})^2} \sum_{i=1}^{n_h} M_{hi}^2 V_2(\hat{\pi}_{hi})\right] \\
&= E_1\left[\sum_{h=1}^{L} W_h^2 \frac{N_h^2}{(n_h M_{h0})^2} \sum_{i=1}^{n_h} M_{hi}^2 \frac{k_{hi}g_{hi}-k_{hi}}{k_{hi}g_{hi}} \frac{1}{k_{hi}} \frac{1}{k_{hi}g_{hi}-1} \sum_{j=1}^{k_{hi}g_{hi}} (\hat{\pi}_{hij}-\pi_{hi})^2\right] \\
&= \sum_{h=1}^{L} W_h^2 \frac{N_h}{n_h M_{h0}^2} \sum_{i=1}^{N_h} M_{hi}^2 \frac{k_{hi}g_{hi}-k_{hi}}{k_{hi}g_{hi}} \frac{1}{k_{hi}} \frac{1}{k_{hi}g_{hi}-1} \sum_{j=1}^{k_{hi}g_{hi}} (\hat{\pi}_{hij}-\pi_{hi})^2
\end{aligned}$$

이므로 추정량 $\hat{\pi}_{stcl}$의 분산은 식 (4.2)와 같다. □

또한 h층의 N_h개의 집락에서 n_h개를 단순임의비복원추출하고, 추출된 각 집락내의 M_{hi}개 단위 중에서 m_{hi}개인 계통표본을 추출하는 데 있어서 크기가 $m'_{hi}(=m_{hi}/k_{hi})$인 부차표본을 k_{hi}개 추출할 경우 민감한 그룹에 속하는 모비율 π의 층화추정량 $\hat{\pi}_{stcl}$의 분산은 다음과 같다.

$$\begin{aligned}
V(\hat{\pi}_{stcl}) = \sum_{h=1}^{L} W_h^2 \Bigg[&\frac{N_h-n_h}{n_h M_{h0}^2} \sum_{i=1}^{N_h} \left(M_{hi}\pi_{hi} - \bar{M}_h\pi_h\right)^2 \\
&+ \frac{N_h}{n_h M_{h0}^2} \sum_{i=1}^{N_h} M_{hi}^2 \frac{k_{hi}g_{hi}-k_{hi}}{k_{hi}g_{hi}} \frac{1}{k_{hi}} \frac{1}{k_{hi}g_{hi}-1} \sum_{j=1}^{k_{hi}g_{hi}} (\hat{\pi}_{hij}-\pi_{hi})^2 \Bigg].
\end{aligned} \tag{4.3}$$

한편, h층의 응답자들이 추출확률 t_{hi}에 의해 복원추출된 $i\,(i=1,2,\ldots,n_h)$번째 집락으로부터 반복계통추출되었을 때, 이러한 절차에 의해 얻어진 민감한 그룹에 속하는 조사단위당 모비율 π의 층화추정량 $\hat{\pi}_{stppz}$는 다음과 같다.

$$\hat{\pi}_{stppz} = \sum_{h=1}^{L} \frac{W_h}{n_h M_{h0}} \sum_{i=1}^{n_h} \frac{M_{hi}}{t_{hi}} \hat{\pi}_{hi},$$

여기서 $W_h = N_h/N$, $M_{h0} = \sum_{i=1}^{N_h} M_{hi}$이다.

h층의 각 집락의 크기가 M_{hi}인 N_h개의 집락에서 n_h개의 집락을 추출확률 t_{hi}에 의해 복원추출하고, 추출된 각 집락내의 M_{hi}개 단위 중에서 m_{hi}개인 계통표본을 추출하는 데 있어서 크기가 $m'_{hi}(=m_{hi}/k_{hi})$인 부차표본을 k_{hi}개 추출할 경우 민감한 그룹에 속하는 모비율 π의 층화추정량 $\hat{\pi}_{stppz}$의 분산은 다음과 같다.

$$\begin{aligned}
V(\hat{\pi}_{stppz}) = \sum_{h=1}^{L} W_h^2 \Bigg[&\frac{1}{n_h M_{h0}^2} \sum_{i=1}^{N_h} t_{hi} \left(\frac{M_{hi}\pi_{hi}}{t_{hi}} - M_{h0}\pi_h\right)^2 \\
&+ \frac{1}{n_h M_{h0}^2} \sum_{i=1}^{N_h} \frac{M_{hi}^2}{t_{hi}} \frac{k_{hi}g_{hi}-k_{hi}}{k_{hi}g_{hi}} \frac{1}{k_{hi}} \frac{1}{k_{hi}g_{hi}-1} \sum_{j=1}^{k_{hi}g_{hi}} (\hat{\pi}_{hij}-\pi_{hi})^2 \Bigg].
\end{aligned}$$

그리고 h층의 n_h개의 1차 추출단위가 각 집락의 크기 M_{hi}에 비례$(t_{hi} = M_{hi}/M_{h0})$가 되도록 확률비례복원추출한 후, 추출된 각 집락에서 다시 m_i개의 2차 단위를 반복계통추출할 경우, 이러한 층화 확률비

례복원 추출절차에 의해 얻어진 민감한 그룹에 속하는 모비율 π의 층화추정량 $\hat{\pi}_{stppswr}$과 그 분산은 다음과 같다.

$$\hat{\pi}_{stppswr} = \sum_{h=1}^{L} W_h \frac{1}{n_h} \sum_{i=1}^{n_h} \hat{\pi}_{hi}, \quad W_h = \frac{N_h}{N},$$

$$V(\hat{\pi}_{stppswr}) = \sum_{h=1}^{L} W_h^2 \left[\frac{1}{n_h M_{h0}} \sum_{i=1}^{N_h} M_{hi} (\pi_{hi} - \pi_h)^2 + \frac{1}{n_h M_{h0}} \sum_{i=1}^{N_h} M_{hi} \frac{k_{hi} g_{hi} - k_{hi}}{k_{hi} g_{hi}} \frac{1}{k_{hi}} \frac{1}{k_{hi} g_{hi} - 1} \sum_{j=1}^{k_{hi} g_{hi}} (\hat{\pi}_{hij} - \pi_{hi})^2 \right].$$

또한 h층의 각 집락의 크기가 M_{hi}인 N_h개의 집락으로 구성되어 있는 모집단으로부터 n_h개의 집락을 확률비례비복원추출한 후, 추출된 각 집락에서 다시 m_{hi}개의 조사단위를 반복계통추출할 경우, 이러한 층화 확률비례비복원 추출절차에 의해 얻어진 민감한 그룹에 속하는 조사단위당 모비율 π의 층화추정량 $\hat{\pi}_{stppswor}$과 그 분산은 다음과 같다.

$$\hat{\pi}_{stppswor} = \sum_{h=1}^{L} W_h \frac{1}{M_{h0}} \sum_{i=1}^{n_h} \frac{M_{hi}}{\theta_{hi}} \hat{\pi}_{hi},$$

$$V(\hat{\pi}_{stppswor}) = \sum_{h=1}^{L} W_h^2 \left[\frac{1}{M_{h0}^2} \sum_{i=1}^{N_h} \sum_{j>i}^{N_h} (\theta_{hi}\theta_{hj} - \theta_{hij}) \left(\frac{M_{hi}\pi_{hi}}{\theta_{hi}} - \frac{M_{hj}\pi_{hj}}{\theta_{hj}} \right)^2 + \frac{1}{M_{h0}^2} \sum_{i=1}^{N_h} \frac{M_{hi}^2}{\theta_{hi}} \frac{k_{hi} g_{hi} - k_{hi}}{k_{hi} g_{hi}} \frac{1}{k_{hi}} \frac{1}{k_{hi} g_{hi} - 1} \sum_{j=1}^{k_{hi} g_{hi}} (\hat{\pi}_{hij} - \pi_{hi})^2 \right],$$

여기서 θ_{hi}는 h층의 조사단위 i가 표본에 포함되는 확률이며, θ_{hij}는 h층의 조사단위 i와 j가 동시에 표본에 포함되는 확률이다.

다음으로 층화 집락 반복계통 확률화응답모형에서 n_h와 m_h의 최적값을 구해보고자 한다. 먼저 비용함수를 고려해야 하는데, 층화 2단계 추출의 경우 비용함수는 대개 다음과 같은 형태를 취한다.

$$C = c_0 + \sum_{h=1}^{L} c_{1h} n_h + \sum_{h=1}^{L} c_{2h} n_h m_h, \tag{4.4}$$

여기서 C는 총비용이고, c_0는 고정비용으로 조사행정비, 표본설계비용 등을 포함하며 표본의 크기와는 관계없이 소요되는 비용이다. c_{1h}은 h층의 표본 1차 추출단위 당 비용으로 집락 당 소요비용을 의미하며, 표본집락의 선정, 각 표본 1차 추출단위에서 2차 추출단위를 추출하기 위한 리스트 작성비와 1차 추출단위의 추출작업 등에 필요한 비용을 포함한다. c_{2h}는 h층의 표본 2차 추출단위 당 비용으로 조사단위 당 소요비용을 의미하며, 표본 2차 추출단위의 추출 및 확인에 소요되는 비용, 확률장치를 이용한 면접 또는 실측비용, 조사자료의 집계분석비용 등을 포함한다.

식 (4.3)에서 $k_{hi} = m_{hi}/m'_{hi}$이므로 $m_{hi} = m_h$라 가정하면, 분산 식을 다음과 같이 바꾸어 표현할 수 있다.

$$V(\hat{\pi}_{stcl}) = \sum_{h=1}^{L} W_h^2 \left[\frac{1}{n_h} S_{bh}^2 + \frac{1}{n_h m_h} S_{wh}^2 - \frac{1}{M_{h0}^2} \sum_{i=1}^{N_h} \left(M_{hi}\pi_{hi} - \bar{M}_h \pi_h \right)^2 \right], \tag{4.5}$$

여기서

$$S_{bh}^2 = \frac{N_h}{M_{h0}^2}\sum_{i=1}^{N_h}\left(M_{hi}\pi_{hi} - \bar{M}_h\pi_h\right)^2,$$

$$S_{wh}^2 = \frac{N_h m_h'}{M_{h0}^2}\sum_{i=1}^{N_h} M_{hi}^2\frac{k_{hi}g_{hi}-k_{hi}}{k_{hi}g_{hi}}\frac{1}{k_{hi}g_{hi}-1}\sum_{j=1}^{k_{hi}g_{hi}}(\hat{\pi}_{hij}-\pi_{hi})^2$$

이다.

일정한 비용 하에서 분산을 최소로 하는 n_h과 m_h의 값을 식 (4.4)의 비용함수와 분산 식 (4.5)를 이용하여 구해 보기로 하자. n_h과 m_h의 최적값을 구하기 위하여 라그랑즈 승수법을 이용하기로 하자. 이때, 최소화하는 함수 $\emptyset$는 다음과 같이 표현된다.

$$\emptyset = V(\hat{\pi}_{stcl}) + \lambda\left[\sum_{h=1}^{L} c_{1h}n_h + \sum_{h=1}^{L} c_{2h}n_h m_h - (C - c_0)\right], \tag{4.6}$$

여기서 λ는 라그랑즈 승수이다.

식 (4.6)은 n_h와 $n_h m_h$의 함수이므로 이를 각각에 관하여 미분하면 식 (4.7)과 식 (4.8)을 얻을 수 있다.

$$n_h\sqrt{\lambda} = W_h\frac{S_{bh}}{\sqrt{c_{1h}}}, \tag{4.7}$$

$$n_h m_h\sqrt{\lambda} = \frac{W_h S_{wh}}{\sqrt{c_{2h}}}. \tag{4.8}$$

식 (4.7)과 식 (4.8)에서 m_h의 최적값 $m_{h(opt)}$는 다음과 같다.

$$m_{h(opt)} = \frac{S_{wh}}{S_{bh}}\sqrt{\frac{c_{1h}}{c_{2h}}}\ .$$

식 (4.7)에서 $W_h \propto N_h M_h$이므로 n_h의 최적값은

$$n_h \propto \frac{N_h M_h S_{bh}}{\sqrt{c_{1h}}}$$

이므로 n_h의 최적값은 $N_h M_h$와 S_{bh}에 비례하고 $\sqrt{c_{1h}}$에 반비례한다.

5. 효율성 비교

이 절에서는 집락 반복계통추출법에 의한 무관질문모형과 층화집락 반복계통추출법에 의한 무관질문모형과의 효율성을 비교해 보고자 한다.

각 집락으로부터 표본으로 추출된 조사단위의 수가 동일하다는 가정 하에서 효율성을 비교하기 위하여 집락 반복계통추출법에 의한 분산 식 (2.3)과 층화집락 반복계통추출법에 의한 분산 식 (4.3)을 이용하였다. 상대효율은

$$\mathrm{RE} = \frac{V(\hat{\pi})}{V(\hat{\pi}_{stcl})}$$

와 같고, 상대효율이 1보다 큰 경우 층화집락 반복계통추출법에 의한 무관질문모형이 집락 반복계통추출법에 의한 무관질문모형보다 효율적임을 의미한다.

Table 5.1. Efficiency comparison

π_y	p								
	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$\hat{\lambda}_{12} = \hat{\lambda}_{112} = 0.45,\ \hat{\lambda}_{22} = \hat{\lambda}_{122} = 0.55,\ \hat{\lambda}_{32} = \hat{\lambda}_{132} = 0.65,\ \hat{\lambda}_{42} = \hat{\lambda}_{142} = 0.75$									
0.1	2.08173	2.04245	2.00512	1.96960	1.93578	1.90354	1.87280	1.84346	1.81543
0.2	2.54961	2.42424	2.31254	2.21268	2.12310	2.04245	1.96960	1.90354	1.84346
0.3	3.34182	3.02932	2.76785	2.54961	2.36680	2.21268	2.08174	1.96960	1.87280
0.4	4.55053	3.98299	3.45831	3.02932	2.69070	2.42424	2.21268	2.04245	1.90354
0.5	4.59152	4.95006	4.41227	3.70974	3.12748	2.69070	2.36680	2.12310	1.93578
0.6	1.92048	3.79076	4.95887	4.55053	3.70974	3.02932	2.54961	2.21268	1.96960
0.7	0.51960	1.45181	3.46720	4.95887	4.41227	3.45831	2.76785	2.31254	2.00512
0.8	0.17038	0.45016	1.45181	3.79076	4.95006	3.98299	3.02932	2.42424	2.04245
0.9	0.11673	0.17038	0.51960	1.92048	4.59152	4.55053	3.34182	2.54961	2.08173
$\hat{\lambda}_{12} = \hat{\lambda}_{112} = 0.4,\ \hat{\lambda}_{22} = \hat{\lambda}_{122} = 0.5,\ \hat{\lambda}_{32} = \hat{\lambda}_{132} = 0.6,\ \hat{\lambda}_{42} = \hat{\lambda}_{142} = 0.7$									
0.1	2.05733	2.01593	1.97675	1.93965	1.90447	1.87108	1.83934	1.80916	1.78042
0.2	2.56681	2.42722	2.30478	2.19683	2.10115	2.01592	1.93965	1.87108	1.80916
0.3	3.50907	3.12413	2.81563	2.56681	2.36403	2.19683	2.05733	1.93965	1.83934
0.4	5.20745	4.36289	3.65746	3.12413	2.72680	2.42722	2.19683	2.01593	1.87108
0.5	5.30274	5.87579	4.99254	3.98758	3.24306	2.72680	2.36403	2.10115	1.90447
0.6	1.72888	4.06959	5.90506	5.20745	3.98758	3.12413	2.56681	2.19683	1.93965
0.7	0.40911	1.24796	3.61226	5.90506	4.99254	3.65746	2.81563	2.30478	1.97675
0.8	0.15045	0.35375	1.24796	4.06958	5.87579	4.36289	3.12413	2.42722	2.01592
0.9	0.13102	0.15045	0.40911	1.72887	5.30274	5.20745	3.50908	2.56681	2.05733

상대효율을 계산하기 위하여 모집단에 2개의 층이 있다고 가정하고, 첫 번째 층에서는 각 집락의 크기가 $M_1 = M_{11} =$ 1,000, $M_2 = M_{12} =$ 1,000, $M_3 = M_{13} =$ 2,000, $M_4 = M_{14} =$ 3,000인 $N_1 = 4$개의 집락으로부터 $n_1 = 2$개의 집락을 비복원추출한다고 하고, 두 번째 층에서는 각 집락의 크기가 $M_5 = M_{21} = 500$, $M_6 = M_{22} = 500$, $M_7 = M_{23} =$ 1,000, $M_8 = M_{23} =$ 1,000인 $N_2 = 4$개의 집락으로부터 $n_2 = 2$개의 집락을 비복원추출한다고 하자. 이 때, 추출된 각 집락으로부터 동일한 크기의 50개의 계통표본을 추출하는데 있어서 크기가 25인 부차표본을 반복계통추출하게 되면, k_1에서 $k_4(k_{11}\sim k_{14})$까지 그리고 k_5에서 $k_8(k_{21}\sim k_{24})$까지 모두 2의 값을 갖게 되며, 추출간격은 $g_1 = g_{11} = 20$, $g_2 = g_{12} = 20$, $g_3 = g_{13} = 40$, $g_4 = g_{14} = 60$, $g_5 = g_{21} = 10$, $g_6 = g_{22} = 10$, $g_7 = g_{23} = 20$, $g_8 = g_{24} = 20$이 된다. 또한 $W_1 = 0.5$, $W_2 = 0.5$인 경우에 대하여 $\hat{\lambda}_{11} = \hat{\lambda}_{111} = 0.3$, $\hat{\lambda}_{21} = \hat{\lambda}_{121} = 0.4$, $\hat{\lambda}_{31} = \hat{\lambda}_{131} = 0.5$, $\hat{\lambda}_{41} = \hat{\lambda}_{141} = 0.6$, $\hat{\lambda}_{51} = \hat{\lambda}_{211} = 0.2$, $\hat{\lambda}_{52} = \hat{\lambda}_{212} = 0.25$, $\hat{\lambda}_{61} = \hat{\lambda}_{221} = 0.3$, $\hat{\lambda}_{62} = \hat{\lambda}_{222} = 0.35$, $\hat{\lambda}_{71} = \hat{\lambda}_{231} = 0.4$, $\hat{\lambda}_{72} = \hat{\lambda}_{232} = 0.45$, $\hat{\lambda}_{81} = \hat{\lambda}_{241} = 0.5$, $\hat{\lambda}_{82} = \hat{\lambda}_{242} = 0.55$라 두고, $\hat{\lambda}_{12} = \hat{\lambda}_{112} = 0.45$, $\hat{\lambda}_{22} = \hat{\lambda}_{122} = 0.55$, $\hat{\lambda}_{32} = \hat{\lambda}_{132} = 0.65$, $\hat{\lambda}_{42} = \hat{\lambda}_{142} = 0.75$인 경우와 $\hat{\lambda}_{12} = \hat{\lambda}_{112} = 0.4$, $\hat{\lambda}_{22} = \hat{\lambda}_{122} = 0.5$, $\hat{\lambda}_{32} = \hat{\lambda}_{132} = 0.6$, $\hat{\lambda}_{42} = \hat{\lambda}_{142} = 0.7$인 경우로 값을 변화시키고 $p_i = p_{hi}$와 $\pi_{yij} = \pi_{hyij}$값을 0.1에서 0.9까지 0.1씩 증가시켜가면서 상대효율을 구한 결과는 다음 Table 5.1과 같다.

Table 5.1로부터 전반적으로 층화집락 반복계통추출법에 의한 무관질문모형이 집락 반복계통추출법에 의한 무관질문모형보다 효율성이 높게 나타남을 알 수 있었다. 또한 "예"라고 응답한 사람의 비율에 따라 효율성에 약간의 차이를 나타내기는 했지만 p값이 작고 π_y값이 큰 경우를 제외하고는 층화집락 반복계통추출법에 의한 무관질문모형이 집락 반복계통추출법에 의한 무관질문모형보다 효율적인 것으로 나타났다.

6. 결론

본 논문에서는 대규모 표본조사에서 많이 나타나는 조사하고자 하는 모집단이 층으로 형성되어 있고, 각 층들이 집락으로 구성되어 있을 때 사용 가능한 층화 집락추출법을 얻고자 하는 정보가 민감할 때 반복계통 무관질문모형에 적용하였다. 우선 모집단이 집락으로 구성되어 있고, 추출된 집락으로부터 계통표본을 반복적으로 추출하여 민감한 정보를 얻는 데 무관질문모형을 사용한 집락 반복계통 무관질문모형을 제안하여 그 이론적 체계를 구축하였다. 다음으로 제안한 모형을 층화된 모집단에서도 사용할 수 있도록 층화집락 반복계통추출법에 의한 무관질문모형으로 발전시켰으며, 각 층의 집락을 확률비례복원추출 또는 확률비례비복원추출하는 층화 확률비례 반복계통 무관질문모형을 제안하였다. 그리고 제안한 층화집락 반복계통추출법에 의한 무관질문모형에서 각 층의 표본배분하는 문제를 비례배분과 최적배분 측면에서 다루었다. 또한 제안한 층화집락 반복계통추출법에 의한 무관질문모형과 집락 반복계통추출법에 의한 무관질문모형과의 효율성을 비교한 결과 전반적으로 층화집락 반복계통추출법에 의한 무관질문모형이 집락 반복계통추출법에 의한 무관질문모형보다 효율성이 높게 나타남을 알 수 있었다. 또한 "예"라고 응답한 사람의 비율에 따라 효율성에 약간의 차이를 나타내기는 했지만 p값이 작고 π_y값이 큰 경우를 제외하고는 층화집락 반복계통추출법에 의한 무관질문모형이 집락 반복계통추출법에 의한 무관질문모형보다 효율적인 것으로 나타났다.

References

Ahn, S. C. and Lee, G. S. (2003). A stratified unrelated question model, *Journal of The Korean Data Analysis Society*, **5(4B)**, 853–864.

Ahn, S. C. and Lee, G. S. (2004). A stratified discrete quantitative randomized response model, *Journal of The Korean Data Analysis Society*, **6(1B)**, 181–191.

Greenberg, B. G., Abul-Ela, Abdel-Latif A., Simmons, W. R. and Horvitz, D. G. (1969). The unrelated question randomized response model: Theoretical framework, *Journal of The American Statistical Association*, **64**, 520–539.

Kim, J. M. and Elam, M. E. (2005). A two-stage stratified Warner's randomized response model using optimal allocation, *Metrika*, **61**, 1–7.

Kim, J. M. and Warde, W. D. (2004). A stratified Warner's randomized response model, *Journal of Statistical Planning and Inference*, **120**, 155–165.

Lee, G. S. (2010). A replicated systematic randomized response model, *Journal of The Korean Data Analysis Society*, **12(3B)**, 1549–1558.

Lee, G. S. (2012). Unrelated question randomized response model by stratified replicated systematic sampling, *Journal of The Korean Data Analysis Society*, **14(2B)**, 781–791.

Lee, G. S. and Park, C. I. (2007). An unrelated question model by PPS sampling, *Journal of The Korean Data Analysis Society*, **9(3B)**, 1471–1483.

Warner, S. L. (1965). Randomized response: A survey technique for eliminating evasive answer bias, *Journal of The American Statistical Association*, **60**, 63–69.

층화 집락 반복계통 무관질문모형에 관한 연구

이기성[a,1]

[a]우석대학교 아동복지학과

(2012년 12월 17일 접수, 2013년 2월 6일 수정, 2013년 2월 7일 채택)

요 약

본 논문에서는 대규모 표본조사에서 많이 나타나는 모집단이 층으로 형성되어 있고, 각 층들이 집락으로 구성되어 있을 때 사용 가능한 층화 집락추출법을 얻고자 하는 정보가 민감할 때 반복계통 무관질문모형에 적용하였다. 먼저 모집단이 집락으로 구성되어 있고, 추출된 집락으로부터 계통표본을 반복적으로 추출하여 민감한 정보를 얻는 데 무관질문모형을 사용한 집락 반복계통 무관질문모형을 제안하였다. 다음으로 제안한 모형을 층화된 모집단에서도 사용할 수 있도록 층화집락 반복계통추출법에 의한 무관질문모형으로 발전시켰으며, 각 층의 집락을 확률비례복원추출 또는 확률비례비복원추출하는 층화 확률비례 반복계통 무관질문모형을 제안하였다. 또한 제안한 층화집락 반복계통추출법에 의한 무관질문모형에서 각 층의 표본배분하는 문제를 비례배분과 최적배분 측면에서 다루었다. 마지막으로 제안한 층화집락 반복계통추출법에 의한 무관질문모형과 집락 반복계통추출법에 의한 무관질문모형과의 효율성을 비교하였다.

주요용어: 층화 집락추출법, 반복계통추출법, 무관질문모형, 확률비례추출, 표본배분.

이 논문은 2011년도 정부(교육과학기술부)의 재원으로 한국연구재단의 지원을 받아 수행된 기초연구사업임 (2011-0007800).

[1](565-701) 전북 완주군 삼례읍 후정리 490, 우석대학교 아동복지학과, 교수. E-mail: gisung@woosuk.ac.kr