The Korean Journal of Applied Statistics (2012)
25*(3), 447–456*
DOI: http://dx.doi.org/10.5351/KJAS.2012.25.3.447

An Additive Quantitative Randomized Response Model by Cluster Sampling

Gi-Sung Lee[1]

[1]Department of Children Welfare, Woosuk University

(Received March 27, 2012; Revised April 21, 2012; Accepted April 30, 2012)

Abstract

For a sensitive survey in which the population is comprised of several clusters with a quantitative attribute, we present an additive quantitative randomized response model by cluster sampling that adapts a two-stage cluster sampling instead of a simple random sample based on Himmelfarb-Edgell's additive quantitative attribute model and Gjestvang-Singh's one. We also derive optimum values for the number of 1st stage clusters and the optimum values of observation units in a 2nd stage cluster under the condition of minimizing the variance given constant cost. We can see that Himmelfarb-Edgell's model is more efficient than Gjestvang-Singh's model under the condition of cluster sampling.

Keywords: Additive randomized response model, quantitative attribute, cluster sampling, efficiency.

1. 서론

Warner (1965)가 응답자들의 민감한 속성을 정확하게 조사하기 위해 개발한 확률장치를 이용하는 확률화응답모형은 민감한 질적 속성뿐만 아니라 양적속성을 파악할 수 있는 모형으로 발전되어 왔다. 한편, 응답자들의 민감한 정보를 얻기 위한 한 방법으로 Himmelfarb와 Edgell (1980)은 응답자들의 민감한 양적 변수에 어떤 일정한 값을 더하여 응답하도록 하는 가법모형(additive model)을 제안하였고, Eichhorn과 Hayre (1983)는 응답자들의 민감한 양적 변수에 어떤 일정한 값을 곱하여 응답하도록 하는 승법모형(multiplicative model)을 제안하기도 하였다. 특히, Gjestvang와 Singh (2009)은 두 개의 알고 있는 양의 실수 값 α와 β를 변환된 변수 Z에 곱한 후 민감한 양적 변수 X에 더하거나 빼는 가법 양적속성 모형을 제안하였다. 이 가법 양적속성 모형은 단순임의추출된 i번째 응답자들에게 확률장치를 이용하여 $p = \beta/(\alpha + \beta)$의 확률로 $X_i + \alpha Z$라는 변환된 응답을 하도록 하고, $1 - p = \alpha/(\alpha + \beta)$의 확률로 $X_i - \beta Z$라는 변환된 응답을 하도록 하여 모형의 효율성을 높였다. 이 두 대표적인 가법모형인 Himmelfarb-Edgell의 가법 양적속성 모형과 Gjestvang-Singh의 가법 양적속성 모형은 응답자들을 추출하는 데 있어서 모두 단순임의추출법을 가정하고 있다. 만약 우리가 관심을 가지고 있는 모집단이 집락으로 구성되어 있을 경우에는 응답자들을 단순임의추출하는 것보다는 집락추출을 하는 것이 현실적으로 더 타당한 방법이라고 생각된다. 이 처럼 모집단이 집락으로 구성되어 있고 우리가 관심을 가지

This work was supported by Woosuk University(2012).
[1]Professor, Department of Children Welfare, Woosuk University, 490 Hujeong-ri, Wanju-gun, Jeonbuk, 565-701, Korea. E-mail: gisung@woosuk.ac.kr

고 있는 변수가 민감한 속성일 때, 집락추출법을 확률화응답모형에 적용하는 연구는 Ahn과 Lee (2002), Lee와 Hong (2003), Lee (2006), Lee 등 (2007) 등에 의해 발전되어 왔다.

본 논문에서는 매우 민감한 조사에서 모집단이 양적속성을 갖는 여러 개의 집락으로 구성되어 있을 때, Himmelfarb-Edgell의 가법 양적속성 모형과 Gjestvang-Singh의 가법 양적속성 모형에서 기존에 사용하고 있는 단순임의추출법 대신에 모집단으로부터 집락을 단순임의비복원 추출한 후, 추출된 각 집락에서 다시 조사단위의 표본을 단순임의복원 추출하는 2단계 집락추출법을 적용한 2단계 집락추출법에 의한 가법 양적속성 확률화응답모형을 제안하고자 한다. 그리고 제안한 두 2단계 집락추출법에 의한 가법 양적속성 모형으로부터 일정한 비용 하에서 분산을 최소로 하는 1단계 집락의 수와 2단계 집락에서 추출된 조사단위의 수의 최적값을 도출하고자 한다. 또한, 2단계 집락추출법에 의한 Himmelfarb-Edgell의 가법 양적속성 모형과 Gjestvang-Singh의 가법 양적속성 모형간의 효율성을 비교해 보고자 한다.

2. 2단계 집락추출법에 의한 Himmelfarb-Edgell의 가법 양적속성 모형

이 절에서는 Himmelfarb와 Edgell (1980)의 가법 양적속성 모형에 2단계 집락추출법을 적용한 2단계 집락 가법 양적속성 모형을 제안하고자 한다. 매우 민감한 조사에서 각 집락의 크기가 $M_i\,(i = 1, 2, \ldots, N)$인 N개의 집락으로 구성되어 있는 모집단으로부터 n개의 집락을 단순임의비복원 추출한 후, 추출된 각 집락에서 다시 $m_i\,(i = 1, 2, \ldots, n)$개의 조사단위의 표본을 단순임의복원 추출하는 2단계 집락추출법에 가법 양적속성 확률화응답모형을 적용해 보고자 한다.

i번째 집락의 j번째 응답자들은

$$Y_{ij} = X_{ij} + Z$$

라는 변환된 응답(scrambled response)을 기록하도록 요구받게 되는데, 여기서 $X_{ij}\,(\geq 0)$는 i번째 집락의 j번째 민감한 양적변수의 참값이며, Z는 $E(Z) = \mu_z$이고 $V(Z) = \sigma_z^2$인 분포를 따르는 변환변수로 양수값을 갖고 알고 있다고 가정한다.

단순임의복원 추출된 i번째 집락에서 j번째 응답자가 응답한 확률화응답의 관찰치를 Y_{ij}라 하면, i번째 집락에서의 민감한 변수 X의 평균 μ_{xi}의 추정량 $\hat{\mu}_{xi}$는 다음과 같이 표현된다.

$$\hat{\mu}_{xi} = \frac{1}{m_i}\sum_{j=1}^{m_i}(Y_{ij} - \mu_z).$$

이 때, $\hat{\mu}_{xi}$의 기대값과 분산은 다음과 같다.

$$E(\hat{\mu}_{xi}) = \mu_{xi},$$
$$V(\hat{\mu}_{xi}) = \frac{1}{m_i}\left(\sigma_z^2 + \sigma_{xi}^2\right).$$

한편, 민감한 변수 X의 조사단위당 모평균 μ_x는 다음과 같다.

$$\mu_x = \frac{1}{M_0}\sum_{i=1}^{N} M_i\mu_{xi}.$$

여기서 $M_0 = \sum_{i=1}^{N} M_i$이다.

응답자들이 $i\,(i=1,2,\ldots,n)$번째 집락으로부터 단순임의복원 추출되었을 때, 이러한 절차에 의해 얻어진 민감한 변수 X의 조사단위당 모평균 μ_x의 추정량 $\hat{\mu}_{x(H)}$는 다음과 같이 정의할 수 있다.

$$\hat{\mu}_{x(H)} = \frac{N}{nM_0}\sum_{i=1}^{n} M_i\hat{\mu}_{xi} = \frac{1}{n\bar{M}}\sum_{i=1}^{n} M_i\hat{\mu}_{xi}, \tag{2.1}$$

여기서 $\bar{M} = 1/N\sum_{i=1}^{N} M_i = M_0/N$이다.

정리 2.1 추정량 $\hat{\mu}_{x(H)}$는 모평균 μ_x의 비편향추정량이다.

증명:

$$\begin{aligned} E_1E_2(\hat{\mu}_{x(H)}) &= E_1E_2\left(\frac{N}{nM_0}\sum_{i=1}^{n} M_i\hat{\mu}_{xi}\right) = E_1\left(\frac{N}{nM_0}\sum_{i=1}^{n} M_i\mu_{xi}\right) \\ &= \frac{1}{M_0}\sum_{i=1}^{N} M_i\mu_{xi} = \mu_x. \end{aligned}$$

□

정리 2.2 각 집락의 크기가 M_i인 N개의 집락에서 n개의 집락을 단순임의비복원 추출하고, 추출된 집락에서 다시 m_i개의 조사단위의 표본을 단순임의복원 추출한다. 이러한 2단계 절차에 의해 얻어진 민감한 변수 X의 모평균 μ_x의 추정량 $\hat{\mu}_{x(H)}$의 분산은 다음과 같다.

$$V(\hat{\mu}_{x(H)}) = \frac{N-n}{nN(N-1)}\sum_{i=1}^{N}\left(\frac{M_i\mu_{xi}}{\bar{M}} - \mu_x\right)^2 + \frac{1}{nN\bar{M}^2}\sum_{i=1}^{N}\frac{M_i^2}{m_i}\left(\sigma_z^2 + \sigma_{xi}^2\right). \tag{2.2}$$

증명:

$$V(\hat{\mu}_{x(H)}) = V_1E_2(\hat{\mu}_{x(H)}) + E_1V_2(\hat{\mu}_{x(H)})$$

에서

$$\begin{aligned} V_1E_2(\hat{\mu}_{x(H)}) &= V_1E_2\left(\frac{1}{n\bar{M}}\sum_{i=1}^{n} M_i\hat{\mu}_{xi}\right) \\ &= V_1\left(\frac{1}{n\bar{M}}\sum_{i=1}^{n} M_i\mu_{xi}\right) \\ &= \frac{N-n}{nN(N-1)}\sum_{i=1}^{N}\left(\frac{M_i\mu_{xi}}{\bar{M}} - \mu_x\right)^2 \end{aligned}$$

이고

$$\begin{aligned} E_1V_2(\hat{\mu}_{x(H)}) &= E_1V_2\left(\frac{1}{n\bar{M}}\sum_{i=1}^{n} M_i\hat{\mu}_{xi}\right) \\ &= E_1\left[\frac{1}{(n\bar{M})^2}\sum_{i=1}^{n} M_i^2\frac{1}{m_i}\left(\sigma_z^2 + \sigma_{xi}^2\right)\right] \\ &= \frac{1}{nN\bar{M}^2}\sum_{i=1}^{N}\frac{M_i^2}{m_i}\left(\sigma_z^2 + \sigma_{xi}^2\right) \end{aligned}$$

이므로, 추정량 $\hat{\mu}_{x(H)}$의 분산은 식 (2.2)와 같다. □

또한, 집락의 크기가 $\bar{M}$로 일정한 N개의 집락에서 n개의 집락을 단순임의비복원 추출 후, 추출된 집락에서 다시 m_i개의 조사단위의 표본을 단순임의복원 추출할 때, 민감한 변수 X의 모평균 μ_x의 추정량 $\hat{\mu}_{x(H)}$의 분산은 다음과 같다.

$$V(\hat{\mu}_{x(H)}) = \frac{1}{nN}\sum_{i=1}^{N}(\mu_{xi}-\mu_x)^2 + \frac{1}{nN}\sum_{i=1}^{N}\frac{1}{m_i}\left(\sigma_z^2+\sigma_{xi}^2\right). \tag{2.3}$$

만약 식 (2.3)에서 각 집락으로부터 표본으로 추출된 m_i가 m으로 일정하다면, 식 (2.3)의 분산식은 다음과 같이 표현될 수 있다.

$$V(\hat{\mu}_{x(H)}) = \frac{1}{nN}\sum_{i=1}^{N}(\mu_{xi}-\mu_x)^2 + \frac{1}{nmN}\sum_{i=1}^{N}\left(\sigma_z^2+\sigma_{xi}^2\right). \tag{2.4}$$

식 (2.4)로부터 1단계 집락의 수 n과 각 집락에서 추출된 조사단위의 수 m을 증가시키면 분산은 감소하지만 n과 m의 증가에 따라 조사비용은 증가하게 된다. 표본의 최적배분을 위해 일정한 비용 하에서 표본의 정도를 최대로 하는 n과 m의 값을 결정해 보고자 한다.

먼저 비용함수를 고려해야 하는데, 2단계 추출의 경우 비용함수는 대개 다음과 같은 형태를 취한다.

$$C = c_0 + nc_1 + nmc_2, \tag{2.5}$$

여기서, C는 총비용이고, c_0는 고정비용으로 조사행정비, 표본설계비용 등을 포함하며 표본의 크기와는 관계없이 소요되는 비용이다. c_1은 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_2는 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.

일정한 비용 하에서 분산을 최소로 하는 n과 m의 값을 식 (2.5)의 비용함수와 분산 식 (2.4)를 이용하여 구해보기로 하자. Lagrange 승수법으로 m의 최적값 m_0를 구하면 다음과 같다.

$$m_0 = \sqrt{\frac{\frac{1}{N}\sum_{i=1}^{N}(\sigma_z^2+\sigma_{xi}^2)}{\frac{1}{N}\sum_{i=1}^{N}(\mu_{xi}-\mu_x)^2}\frac{c_1}{c_2}}. \tag{2.6}$$

또한, 비용함수 식 (2.5)를 n의 함수로 표현해 보면

$$n = \frac{C-c_0}{c_1+mc_2} \tag{2.7}$$

이므로, 식 (2.6)의 m_0값을 식 (2.7)에 대입하여 n의 최적값 n_0를 구하면 다음과 같다.

$$n_0 = (C-c_0)\frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mu_{xi}-\mu_x)^2/c_1}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mu_{xi}-\mu_x)^2c_1}+\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\sigma_z^2+\sigma_{xi}^2)c_2}}. \tag{2.8}$$

따라서 식 (2.6)과 식 (2.8)에서 구한 m_0와 n_0의 값을 식 (2.4)에 대입하여 최소분산 $V_{\min}(\hat{\mu}_{x(H)})$를 다음과 같이 얻을 수 있다.

$$\begin{aligned} V_{\min}(\hat{\mu}_{x(H)}) &= \frac{1}{n_0}\left[\frac{1}{N}\sum_{i=1}^{N}(\mu_{xi}-\mu_x)^2+\frac{1}{m_0N}\sum_{i=1}^{N}\left(\sigma_z^2+\sigma_{xi}^2\right)\right] \\ &= \frac{c_1+m_0c_2}{C-c_0}\left[\frac{1}{N}\sum_{i=1}^{N}(\mu_{xi}-\mu_x)^2+\frac{1}{m_0N}\sum_{i=1}^{N}\left(\sigma_z^2+\sigma_{xi}^2\right)\right] \\ &= \frac{1}{C-c_0}\left[\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mu_{xi}-\mu_x)^2c_1}+\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\sigma_z^2+\sigma_{xi}^2)c_2}\right]^2. \end{aligned} \tag{2.9}$$

3. 2단계 집락추출법에 의한 Gjestvang-Singh의 가법 양적속성 확률화응답모형

이 절에서는 Gjestvang와 Singh (2009)의 가법 양적속성 모형에 2단계 집락추출법을 적용한 2단계 집락 가법 양적속성 모형을 제안하고자 한다. 매우 민감한 조사에서 각 집락의 크기가 $M_i\,(i=1,2,\ldots,N)$인 N개의 집락으로 구성되어 있는 모집단으로부터 n개의 집락을 단순임의비복원 추출한 후, 추출된 각 집락에서 다시 $m_i\,(i=1,2,\ldots,n)$개의 조사단위의 표본을 단순임의복원 추출하는 2단계 집락추출법에 가법 양적속성 확률화응답모형을 적용해 보고자 한다.

i번째 집락의 j번째 응답자들은 $p=\beta_i/(\alpha_i+\beta_i)$의 확률로 민감한 변수 $X_{ij}+\alpha_iZ$라는 변환된 응답을 하도록 하고, $1-p=\alpha_i/(\alpha_i+\beta_i)$의 확률로 $X_{ij}-\beta_iZ$라는 변환된 응답을 하도록 한다. 이때, α_i,β_i는 i번째 집락의 양의 실수값으로 알고 있다고 가정한다. 즉, 응답자들의 응답의 분포는 다음과 같이 주어진다.

$$Y_{ij}=\begin{cases} X_{ij}+\alpha_iZ, & \text{선택확률}: p=\dfrac{\beta_i}{(\alpha_i+\beta_i)}, \\ X_{ij}-\beta_iZ, & \text{선택확률}: 1-p=\dfrac{\alpha_i}{(\alpha_i+\beta_i)}. \end{cases}$$

단순임의복원 추출된 i번째 집락에서 j번째 응답자가 응답한 확률화 응답의 관찰치를 Y_{ij}라 하면, i번째 집락에서의 민감한 변수 X의 평균 μ_{xi}의 추정량 $\hat{\mu}_{xi}$는 다음과 같이 표현된다.

$$\hat{\mu}_{xi}=\frac{1}{m_i}\sum_{j=1}^{m_i}Y_{ij}.$$

한편, 민감한 변수 X의 조사단위당 모평균 μ_x는 다음과 같다.

$$\mu_x=\frac{1}{M_0}\sum_{i=1}^{N}M_i\mu_{xi},$$

여기서 $M_0=\sum_{i=1}^{N}M_i$이다.

응답자들이 $i\,(i=1,2,\ldots,n)$번째 집락으로부터 단순임의복원 추출되었을 때, 이러한 절차에 의해 얻어진 민감한 변수 X의 조사단위당 모평균 μ_x의 추정량 $\hat{\mu}_{x(G)}$는 다음과 같이 정의할 수 있다.

$$\hat{\mu}_{x(G)}=\frac{N}{nM_0}\sum_{i=1}^{n}M_i\hat{\mu}_{xi}=\frac{1}{n\bar{M}}\sum_{i=1}^{n}M_i\hat{\mu}_{xi},$$

여기서 $\bar{M}=1/N\sum_{i=1}^{N}M_i=M_0/N$이다.

정리 3.1 추정량 $\hat{\mu}_{x(G)}$는 모평균 μ_x의 비편향추정량이다.

증명:

$$E_1E_2(\hat{\mu}_{x(G)}) = E_1E_2\left(\frac{N}{nM_0}\sum_{i=1}^{n} M_i\hat{\mu}_{xi}\right) = E_1\left[\frac{N}{nM_0}\sum_{i=1}^{n} M_iE_2(\hat{\mu}_{xi})\right]$$

에서 E_R을 확률장치에 대한 기대값이라 하면

$$\begin{aligned}
E_2(\hat{\mu}_{xi}) &= \frac{1}{m_i}\sum_{j=1}^{m_i} E_R(Y_{ij}) \\
&= \frac{1}{m_i}\sum_{j=1}^{m_i} [pE_R(X_{ij}+\alpha_iZ) + (1-p)E_R(X_{ij}-\beta_iZ)] \\
&= \frac{1}{m_i}\sum_{j=1}^{m_i} [pX_{ij} + (1-p)X_{ij} + \alpha_ip\mu_z - \beta_i(1-p)\mu_z] \\
&= \frac{1}{m_i}\sum_{j=1}^{m_i} \left(X_{ij} + \frac{\alpha_i\beta_i\mu_z}{\alpha_i+\beta_i} - \frac{\mu_z\alpha_i\beta_i}{\alpha_i+\beta_i}\right) \\
&= \frac{1}{m_i}\sum_{j=1}^{m_i} X_{ij}
\end{aligned}$$

이므로 구하고자 하는 $\hat{\mu}_{x(G)}$의 기대값은 다음과 같다.

$$\begin{aligned}
E_1E_2\left(\hat{\mu}_{x(G)}\right) &= E_1\left(\frac{N}{nM_0}\sum_{i=1}^{n} M_i\frac{1}{m_i}\sum_{j=1}^{m_i} X_{ij}\right) \\
&= E_1\left(\frac{N}{nM_0}\sum_{i=1}^{n} M_i\mu_{xi}\right) \\
&= \frac{1}{M_0}\sum_{i=1}^{N} M_i\mu_{xi} \\
&= \mu_x.
\end{aligned}$$

□

정리 3.2 각 집락의 크기가 M_i인 N개의 집락에서 n개의 집락을 단순임의비복원 추출하고, 추출된 집락에서 다시 m_i개의 조사단위의 표본을 단순임의복원 추출한다. 이러한 2단계 절차에 의해 얻어진 민감한 변수 X의 모평균 μ_x의 추정량 $\hat{\mu}_{x(G)}$의 분산은 다음과 같다.

$$V\left(\hat{\mu}_{x(G)}\right) = \frac{N-n}{nN(N-1)}\sum_{i=1}^{N}\left(\frac{M_i\mu_{xi}}{\bar{M}} - \mu_x\right)^2 + \frac{1}{nN\bar{M}^2}\sum_{i=1}^{N}\frac{M_i^2}{m_i}\left\{\sigma_{xi}^2 + \alpha_i\beta_i\left(\sigma_z^2+\mu_z^2\right)\right\}. \tag{3.1}$$

증명:

$$V(\hat{\mu}_{x(G)}) = V_1E_2(\hat{\mu}_{x(G)}) + E_1V_2(\hat{\mu}_{x(G)})$$

에서

$$\begin{aligned} V_1E_2(\hat{\mu}_{x(G)}) &= V_1E_2\left(\frac{1}{n\bar{M}}\sum_{i=1}^{n}M_i\hat{\mu}_{xi}\right) \\ &= V_1\left(\frac{1}{n\bar{M}}\sum_{i=1}^{n}M_i\mu_{xi}\right) \\ &= \frac{N-n}{nN(N-1)}\sum_{i=1}^{N}\left(\frac{M_i\mu_{xi}}{\bar{M}}-\mu_x\right)^2 \end{aligned} \tag{3.2}$$

이고,

$$\begin{aligned} E_1V_2(\hat{\mu}_{x(G)}) &= E_1V_2\left(\frac{1}{n\bar{M}}\sum_{i=1}^{n}M_i\hat{\mu}_{xi}\right) \\ &= E_1\left[\frac{1}{n\bar{M}}\sum_{i=1}^{n}M_iV_2(\hat{\mu}_{xi})\right] \end{aligned}$$

이며, 여기서 V_R을 확률장치에 대한 분산이라고 하면, $\hat{\mu}_{xi}$의 분산을 다음과 같이 표현할 수 있다.

$$V_2(\hat{\mu}_{xi}) = V_1E_R(\hat{\mu}_{xi}) + E_1V_R(\hat{\mu}_{xi})$$

이때

$$V_1E_R(\hat{\mu}_{xi}) = V_1\left[\frac{1}{m_i}\sum_{j=1}^{m_i}E_R(Y_{ij})\right] = V_1\left(\frac{1}{m_i}\sum_{j=1}^{m_i}X_{ij}\right) = \frac{\sigma_{xi}^2}{m_i},$$

$$\begin{aligned} E_1V_R(\hat{\mu}_{xi}) &= E_1\left[\frac{1}{m_i^2}\sum_{j=1}^{m_i}V_R(Y_{ij})\right] \\ &= E_1\left[\frac{1}{m_i^2}\sum_{j=1}^{m_i}\left\{E_R(Y_{ij}^2)-(E_R(Y_{ij}))^2\right\}\right] \\ &= E_1\left[\frac{1}{m_i^2}\sum_{j=1}^{m_i}\left\{pE_R(X_{ij}+\alpha_iZ)^2+(1-p)E_R(X_{ij}-\beta_iZ)^2-X_{ij}^2\right\}\right] \\ &= E_1\left[\frac{1}{m_i^2}\sum_{j=1}^{m_i}\left\{(\sigma_z^2+\mu_z^2)(p\alpha_i^2+(1-p)\beta_i^2)+2X_{ij}\mu_z(p\alpha_i-(1-p)\beta_i)\right\}\right] \\ &= E_1\left[\frac{1}{m_i^2}\sum_{j=1}^{m_i}\left\{(\sigma_z^2+\mu_z^2)\left(\frac{\alpha_i^2\beta_i}{\alpha_i+\beta_i}+\frac{\alpha_i\beta_i^2}{\alpha_i+\beta_i}\right)+2X_{ij}\mu_z\left(\frac{\alpha_i\beta_i}{\alpha_i+\beta_i}-\frac{\alpha_i\beta_i}{\alpha_i+\beta_i}\right)\right\}\right] \\ &= \frac{1}{m_i}\left\{\alpha_i\beta_i\left(\sigma_z^2+\mu_z^2\right)\right\} \end{aligned}$$

이므로

$$V_2(\hat{\mu}_{xi}) = \frac{1}{m_i}\left\{\sigma_{xi}^2+\alpha_i\beta_i(\sigma_z^2+\mu_z^2)\right\}$$

이 된다. 따라서

$$E_1 V_2(\hat{\mu}_{x(G)}) = E_1\left[\frac{1}{(n\bar{M})^2}\sum_{i=1}^{n} M_i^2 \frac{1}{m_i}\left\{\sigma_{xi}^2 + \alpha_i\beta_i(\sigma_z^2+\mu_z^2)\right\}\right]$$
$$= \frac{1}{nN\bar{M}^2}\sum_{i=1}^{N}\frac{M_i^2}{m_i}\left\{\sigma_{xi}^2 + \alpha_i\beta_i(\sigma_z^2+\mu_z^2)\right\} \tag{3.3}$$

이므로, 식 (3.2)와 식 (3.3)으로부터 $\hat{\mu}_{x(G)}$의 분산 식 (3.1)을 얻을 수 있다. □

또한, 집락의 크기가 $\bar{M}$로 일정한 N개의 집락에서 n개의 집락을 단순임의비복원 추출 후, 추출된 집락에서 다시 m_i개의 조사단위의 표본을 단순임의복원 추출할 때, 민감한 변수 X의 모평균 μ_x의 추정량 $\hat{\mu}_{x(G)}$의 분산은 다음과 같다.

$$V(\hat{\mu}_{x(G)}) = \frac{1}{nN}\sum_{i=1}^{N}(\mu_{xi}-\mu_x)^2 + \frac{1}{nN}\sum_{i=1}^{N}\frac{1}{m_i}\left\{\sigma_{xi}^2 + \alpha_i\beta_i(\sigma_z^2+\mu_z^2)\right\}. \tag{3.4}$$

만약 식 (3.4)에서 각 집락으로부터 표본으로 추출된 m_i가 m으로 일정하다면, 식 (3.4)의 분산식은 다음과 같이 표현될 수 있다.

$$V(\hat{\mu}_{x(G)}) = \frac{1}{nN}\sum_{i=1}^{N}(\mu_{xi}-\mu_x)^2 + \frac{1}{nmN}\sum_{i=1}^{N}\left\{\sigma_{xi}^2 + \alpha_i\beta_i(\sigma_z^2+\mu_z^2)\right\}. \tag{3.5}$$

식 (3.5)로부터 1단계 집락의 수 n과 각 집락에서 추출된 조사단위의 수 m을 증가시키면 분산은 감소하지만 n과 m의 증가에 따라 조사비용은 증가하게 된다. 표본의 최적배분을 위해 일정한 비용 하에서 표본의 정도를 최대로 하는 n과 m의 값을 결정해 보고자 한다. 먼저 비용함수를 고려해야 하는데, 2단계 추출의 경우 비용함수는 2절에서 정의한 비용함수 식 (2.5)를 사용하기로 하자.

일정한 비용 하에서 분산을 최소로 하는 n과 m의 값을 식 (2.5)의 비용함수와 분산 식 (3.5)를 이용하여 구해보기로 하자. 2절과 마찬가지로 Lagrange 승수법으로 m의 최적값 m_0와 n의 최적값 n_0를 구하면 다음과 같다.

$$m_0 = \sqrt{\frac{\frac{1}{N}\sum_{i=1}^{N}\left\{\sigma_{xi}^2 + \alpha_i\beta_i(\sigma_z^2+\mu_z^2)\right\}}{\frac{1}{N}\sum_{i=1}^{N}(\mu_{xi}-\mu_x)^2}\frac{c_1}{c_2}}, \tag{3.6}$$

$$n_0 = (C-c_0)\frac{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mu_{xi}-\mu_x)^2/c_1}}{\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mu_{xi}-\mu_x)^2 c_1} + \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left\{\sigma_{xi}^2 + \alpha_i\beta_i(\sigma_z^2+\mu_z^2)\right\}c_2}}. \tag{3.7}$$

따라서 식 (3.6)과 식 (3.7)에서 구한 m_0와 n_0의 값을 식 (3.5)에 대입하여 최소분산 $V_{\min}(\hat{\mu}_{x(G)})$를 다음과 같이 얻을 수 있다.

$$V_{\min}(\hat{\mu}_{x(G)}) = \frac{1}{n_0}\left[\frac{1}{N}\sum_{i=1}^{N}(\mu_{xi}-\mu_x)^2 + \frac{1}{m_0 N}\sum_{i=1}^{N}\left\{\sigma_{xi}^2 + \alpha_i\beta_i(\sigma_z^2+\mu_z^2)\right\}\right]$$

Table 4.1. Efficiency comparison

VR_i	C_z	$\alpha_i = 0.12,\ \beta_i = 0.08$	$\alpha_i = 0.09,\ \beta_i = 0.1$	$\alpha_i = 0.03,\ \beta_i = 0.09$
0.5	0.5	2.7372	2.7522	2.9211
	1.0	2.8890	2.8957	2.9679
	1.5	2.9190	2.9239	2.9767
	2.0	2.9296	2.9339	2.9798
1.0	0.5	1.9083	1.9138	1.9733
	1.0	1.9623	1.9646	1.9892
	1.5	1.9726	1.9743	1.9922
	2.0	1.9762	1.9777	1.9932
1.5	0.5	1.6149	1.6181	1.6518
	1.0	1.6456	1.6469	1.6606
	1.5	1.6514	1.6523	1.6623
	2.0	1.6534	1.6542	1.6629
2.0	0.5	1.4648	1.4669	1.4899
	1.0	1.4857	1.4866	1.4959
	1.5	1.4896	1.4903	1.4970
	2.0	1.4910	1.4916	1.4974

$$= \frac{c_1 + m_0 c_2}{C - c_0}\left[\frac{1}{N}\sum_{i=1}^{N}(\mu_{xi} - \mu_x)^2 + \frac{1}{m_0 N}\sum_{i=1}^{N}\left\{\sigma_{xi}^2 + \alpha_i\beta_i(\sigma_z^2 + \mu_z^2)\right\}\right]$$

$$= \frac{1}{C - c_0}\left[\sqrt{\frac{1}{N}\sum_{i=1}^{N}(\mu_{xi} - \mu_x)^2 c_1} + \sqrt{\frac{1}{N}\sum_{i=1}^{N}\{\sigma_{xi}^2 + \alpha_i\beta_i(\sigma_z^2 + \mu_z^2)\}c_2}\right]^2.$$

4. 효율성 비교

두 모형 간의 효율성을 비교하기 위하여 2단계 집락추출법에 의한 Himmelfarb-Edgell의 가법 양적속성 모형에 대한 2단계 집락추출법에 의한 Gjestvang-Singh의 가법 양적속성 모형의 상대효율을 구해보면 다음과 같다.

$$\mathrm{RE} = \frac{V(\hat{\mu}_{x(H)})}{V(\hat{\mu}_{x(G)})} = \frac{\sum_{i=1}^{N}(\sigma_z^2 + \sigma_{xi}^2)}{\sum_{i=1}^{N}\left[\sigma_{xi}^2 + \alpha_i\beta_i(\sigma_z^2 + \mu_z^2)\right]} = \frac{\sum_{i=1}^{N}(1 + \mathrm{VR}_i)}{\sum_{i=1}^{N}\left[\alpha_i\beta_i(1 + C_z^{-2}) + \mathrm{VR}_i\right]},$$

여기서 $\mathrm{VR}_i = \sigma_{xi}^2/\sigma_z^2$, $C_z = \sigma_z/\mu_z$이다.

두 모형간의 효율성을 수치적으로 비교하기 위하여 모집단이 $N = 4$개의 집락으로 구성되어 있을 때, $\alpha_i (i = 1, 2, 3, 4) = 0.12$, $\beta_i\,(i = 1, 2, 3, 4) = 0.08$, $\alpha_i = 0.09$, $\beta_i = 0.1$, $\alpha_i = 0.03$, $\beta_i = 0.09$이라 가정하였다. 그리고 분산비(VR_i)와 C_z를 0.5에서 2.0까지 변화시켜가면서 상대효율을 구해보면 다음 Table 4.1과 같다.

Table 4.1에서 1보다 큰 값은 2단계 집락추출법에 의한 Gjestvang-Singh의 가법 양적속성 모형이 2단계 집락추출법에 의한 Himmelfarb-Edgell의 가법 양적속성 모형보다 효율성이 좋음을 나타낸다. Table 4.1로부터 VR_i값이 작을수록 그리고 C_z값이 클수록 2단계 집락추출법에 의한 Gjestvang-Singh의 가법 양적속성 모형이 2단계 집락추출법에 의한 Himmelfarb-Edgell의 가법 양적속성 모형보다 효율

적인 것으로 나타났다. 또한 식 (4.1)에서 알 수 있듯이 값이 $\alpha_i\beta_i$작을수록 2단계 집락추출법에 의한 Gjestvang-Singh의 가법 양적속성 모형의 효율성이 증대됨을 알 수 있었다.

5. 결론

본 논문에서는 매우 민감한 조사에서 모집단이 양적속성을 갖는 여러 개의 집락으로 구성되어 있을 때, Himmelfarb-Edgell의 가법 양적속성 모형과 Gjestvang-Singh의 가법 양적속성 모형에서 기존에 사용하고 있는 단순임의추출법 대신에 모집단으로부터 집락을 단순임의비복원 추출한 후, 추출된 각 집락에서 다시 조사단위의 표본을 단순임의복원 추출하는 2단계 집락추출법을 적용한 2단계 집락추출법에 의한 가법 양적속성 확률화응답모형을 제안하였다. 그리고 제안한 두 2단계 집락추출법에 의한 가법 양적속성 모형으로부터 일정한 비용 하에서 분산을 최소로 하는 1단계 집락의 수와 2단계 집락에서 추출된 조사단위의 수의 최적값을 도출하였다. 또한, 2단계 집락추출법에 의한 Himmelfarb-Edgell의 가법 양적속성 모형과 Gjestvang-Singh의 가법 양적속성 모형간의 효율성을 비교해 본 결과, 2단계 집락추출법에 의한 Gjestvang-Singh의 가법 양적속성 모형이 Himmelfarb-Edgell의 가법 양적속성 모형보다 효율적임을 알 수 있었다. 특히 VR_i값이 작을수록 그리고 C_z값이 클수록 2단계 집락추출법에 의한 Gjestvang-Singh의 가법 양적속성 모형이 2단계 집락추출법에 의한 Himmelfarb-Edgell의 가법 양적속성 모형보다 효율적이었고, $\alpha_i\beta_i$값이 작을수록 2단계 집락추출법에 의한 Gjestvang-Singh의 가법 양적속성 모형의 효율성이 증대됨을 알 수 있었다.

References

Ahn, S. C. and Lee, G. S. (2002). The best choice of subsample size for the two-stage cluster unrelated question technique, *Journal of The Korean Data Analysis Society*, **4(4B)**, 397–407.

Eichhorn, B. H. and Hayre, L. S. (1983). Scrambled randomized response methods for obtaining sensitive quantitative data, *Journal of Statistical Planning and Inference*, **7**, 307–316.

Gjestvang, C. R. and Singh, S. (2009). An improved randomized response model: Estimation of mean, *Journal of Applied Statistics*, **36**, 1361–1367.

Himmelfarb, S. and Edgell, S. E. (1980). Additive constant model: A randomized response technique for eliminating evasiveness to quantitative response questions, *Psychological Bulletin*, **87**, 525–530.

Lee, G. S. (2006). A study on the randomized response technique by PPS sampling, *The Korean Journal of Applied Statistics*, **19**, 69–80.

Lee, G. S. and Hong, K. H. (2003). Unrelated question model with quantitative attribute by three-stage cluster sampling, *Journal of The Korean Data Analysis Society*, **5(1B)**, 85–99.

Lee, G. S., Ryu, J. B., Hong, K. H. and Son, C. K. (2007). A study on two-stage cluster Mangat-Singh model, *Journal of The Korean Data Analysis Society*, **9(4B)**, 1801–1810.

Warner, S. L. (1965). Randomized response; A survey technique for eliminating Evasive Answer Bias, *Journal of the American Statistical Association*, **60**, 63–69.