응용통계연구 (2010)
23(3), 533–543

3단계 층화확률화응답모형

김종민[1] · 채성산[2]

[1]미네소타대학교 수리과학부 통계학과, [2]대전대학교 비즈니스정보통계학과

(2010년 2월 접수, 2010년 5월 채택)

요 약

직접면접으로 민감한 질문을 할때 발생하는 무응답이나 거짓응답의 문제를 개선하고자 Warner (1965)가 최초로 제안한 확률화응답모형에 관한 연구는 이후 많은 연구자들에 의해 개선, 발전되어 오고 있다. 본 연구에서 표본은 층화임의복원추출법에 의해 추출되었으며, 표본배분은 최적배분법에 의해 배분되었다. 한편, Kim과 Elam (2005)의 2단계 층화확률화응답모형을 확장한 3단계 층화확률화응답모형을 사용하였다. Kim과 Elam (2005)의 2단계 층화확률화응답모형과 상대효율을 비교한 결과 본 논문에서 제시한 3단계 층화확률화모형의 효율성이 상대적으로 높다는 결과가 도출되었다. 그러나 2단계확률화응답모형을 3단계로 확장함으로써 상대적으로 효율성은 증대되지만 반대로 조사과정의 어려움이 예상된다.

주요용어: 확률화 응답모형, 층화확률화응답모형.

1. 서론

사회 여러 분야의 조사에서 최근 연구의 관심은 응답을 회피하거나 고의적인 거짓응답으로 인한 비표본 오차를 줄이는 데 있다. 이러한 오차는 응답자들이 민감하거나 개인적인 이해와 관계되는 질문을 받았을 경우 증가하게 된다. Warner (1965)는 확률적장치를 이용하여 응답자의 신분이나 비밀을 노출시키지 않고 민감한 질문에 대해 정보를 이끌어낼 수 있는 확률화응답모형(randomized response model; RRM)을 제시하였다.

응답자의 신분이나 비밀을 보장함으로써 응답자로부터 민감한 질문에 보다 더 정확한 정보를 얻을 수 있는 확률화응답모형은 많은 학자들에 의해 연구, 발전되어 왔다. 특히 Abul-Ela 등 (1967)은 이지모집단에 대한 Warner의 관련질문기법(related question technique)을 다지모집단의 경우로 확장하였다.

Greenberg 등 (1969)은 무관질문기법(unrelated question technique)의 이론적 체계를 완성하였고, Moors (1971)와 Folsom 등 (1973)은 이를 개선, 보완하였다. Drane (1975)은 강요질문기법을 제시하였으며, Chaudhuri와 Mukerjee (1988)는 확률화응답모형에 대한 이론을 정리하여 체계화시켰다.

국내에서는 김종호 등 (1992)이 2단계 무관질문모형, 이기성과 홍기학 (1998, 2000)이 개선된 무관질문모형과 2단계 이표본 무관질문모형으로 확장하여 Greenberg 등 (1969)의 모형과 그 효율성을 비교하였고, 김종민과 채성산 (2010)는 2단계 확률화응답모형을 3단계 확률화응답모형으로 확장하여 Mangat와 Singh (1990)의 2단계 확률화응답모형에 대한 효율성을 비교하였다.

본 연구에서는 민감한 사항에 대한 조사를 위해 Kim과 Elam (2005)의 2단계 층화확률화응답모형과 김종민과 채성산 (2010)의 3단계 확률화응답모형을 결합하여 새로운 3단계 층화확률화응답모형을 제안하고, 제안된 모형의 효율성을 기존의 방법과 비교 분석한다.

[2]교신저자: (300-716) 대전시 동구 용운동 96-3, 대전대학교 비즈니스정보통계학과, 교수. E-mail: chae@dju.kr

2. 확률화응답모형

Warner (1965)는 응답자들에게 민감한 질문과, 민감한 질문에 배반되는 질문으로 구성된 확률적장치를 사용하여 민감한 속성에 대한 정보를 얻고자 하였다. 응답자들은 확률적장치, R에 의해 선택된 질문에 응답하고, 조사자는 응답자가 어떤 질문에 응답을 했는지를 알 수 없게 됨으로 응답자는 솔직하게 응답할 수 있다. Warner (1965)에 의해 제시된 관련질문기법의 확률화응답모형에 대하여 살펴보면, 다음과 같은 2개의 질문으로 구성되었다.

질문 1: "나는 민감한 집단에 속한다."

질문 2: "나는 민감한 집단에 속하지 않는다."

이 모형은 크기가 N인 모집단에서 단순임의복원추출된 n명의 응답자들이 확률적장치에 의해 선택된 질문에 대해 "예" 또는 "아니오" 라고 응답한다. 질문 1이 응답될 확률을 P, 질문 2가 응답될 확률을 $1-P$라 하고, 질문 1이 응답될 확률 P는 조사자가 확률적장치에서 사전에 조정할 수 있다.

이러한 확률화응답모형에서, 응답자가 "예"라고 응답할 확률 Y는 다음과 같다.

$$Y = P\pi_w + (1-P)(1-\pi_w). \tag{2.1}$$

위의 식 (2.1)에서, π_w는 응답자가 민감한 집단에 속하는 모비율이다. 이때, n명의 응답자 중에서 "예"라고 응답한 응답자의 수를 n_w라고 하면 n_w는 이항분포, $\mathrm{Bin}(n, Y)$를 따른다. Y의 추정량은, $\hat{Y} = (n_w/n)$ 임으로, π_w의 최우추정량, $\hat{\pi}_w$는,

$$\hat{\pi}_w = \frac{n_w}{(2P-1)n} + \frac{P-1}{2P-1}, \quad P \neq \frac{1}{2} \tag{2.2}$$

이고, $\hat{\pi}_w$의 분산은 다음과 같다.

$$\mathrm{Var}(\hat{\pi}_w) = \frac{\pi_w(1-\pi_w)}{n} + \frac{P(1-P)}{n(2P-1)^2}. \tag{2.3}$$

Mangat와 Singh (1990)는 Warner (1965)에 의하여 제안된 확률화응답모형에 대하여 효율성이 증가하는 2단계 확률화응답모형을 제안하였으며, 그 첫 번째 질문단계는 다음과 같다.

질문 1: "나는 민감한 집단에 속한다."

질문 2: "2번째 문항으로 가시오."

첫 번째 단계의 질문에서 질문 1에 응답한 응답자는 확률 M으로 민감한 질문으로 구성되는 확률적장치 R_1을, 질문 2에 응답한 응답자는 $1-M$의 확률로 두 번째 단계의 확률적장치 R_2로 진행하게 되며, 두 번째 단계의 질문은 다음과 같다.

질문 1: "나는 민감한 집단에 속한다"

질문 2: "나는 민감한 집단에 속하지 않는다."

두 번째 단계의 질문에서 질문 1에 응답할 확률 P로, 질문 2에 응답할 확률 $1-P$로 확률적장치 R_2를 이용하도록 하였다. 응답자가 민감한 집단에 속하는 모비율을 π_m이라고 하면, 확률화응답모형으로부터 응답자가 "예"라고 응답할 확률은,

$$Y = M\pi_m + (1-M)[P\pi_m + (1-P)(1-\pi_m)] \tag{2.4}$$

이며, π_m의 최우추정량 $\hat{\pi}_m$은 다음과 같다.

$$\hat{\pi}_m = \frac{\hat{Y} - (1-M)(1-P)}{2P - 1 + 2M(1-P)}. \tag{2.5}$$

위의 식 (2.5)에서 $\hat{Y}$는 표본에서 "예"라고 응답한 응답자 비율의 추정량이고, 응답자가 "예"라고 응답한 사람의 수를 n_m라 하면, n_m은 이항분포, $\text{Bin}(n, Y)$를 따르며, Y의 추정량은, $\hat{Y} = (n_m/n)$이다. 이때, $\hat{\pi}_m$의 분산은 다음과 같다.

$$\text{Var}(\hat{\pi}_m) = \frac{\pi_m(1-\pi_m)}{n} + \frac{(1-M)(1-P)[1-(1-M)(1-P)]}{n[2P-1+2M(1-P)]^2}. \tag{2.6}$$

김종민과 채성산 (2010)의 3단계 확률화응답모형은, Mangat와 Singh (1990)의 2단계 확률화응답모형을 3단계로 확장한 것으로, 그 첫 번째 질문단계는 2단계 확률화응답모형과 같으며, 두 번째 단계의 질문과 세 번째 단계의 질문은 다음과 같다.

두 번째 단계의 질문은

질문 1: "나는 민감한 집단에 속한다."

질문 2: "3번째 문항으로 가시오."

세 번째 단계의 질문은

질문 1: "나는 민감한 집단에 속한다."

질문 2: "나는 민감한 집단에 속하지 않는다."

세 번째 단계에서 질문 1에 응답할 확률은 L이고 질문 2에 응답할 확률은 $1-L$이다. 이러한 확률화응답모형에서가 응답자가 민감한 집단에 속하는 모비율이면, 응답자가 "예"라고 응답할 확률, Y는 다음과 같다.

$$Y = M\pi_t + (1-M)[P\pi_t + (1-P)\{L\pi_t + (1-L)(1-\pi_t)\}]. \tag{2.7}$$

위의 식 (2.7)에서 M은 첫 번째 단계에서 질문 1에 응답할 확률이고, 질문 2를 선택한 응답자는 $1-M$의 확률을 가지고 2번째 단계로 진행한다. P는 두 번째 단계에서 n명의 응답자 중에서 "예"라고 응답할 확률이며 질문 2를 선택한 응답자는 $1-P$의 확률을 가지고 세 번째 단계로 진행한다. 이때, π_t 의 최우추정량, $\hat{\pi}_t$는 다음과 같다.

$$\hat{\pi}_t = \frac{\hat{Y} - (1-L)(1-M)(1-P)}{2M - 1 + 2(1-M)L + 2(1-M)(1-L)P}. \tag{2.8}$$

위 식 (2.8)에서, $\hat{Y}$는 표본에서 "예"라고 응답할 확률의 추정량이고, n명 중에서 "예"라고 응답한 사람의 수를 n_t라 하면, n_t는 이항분포, $\text{Bin}(n, Y)$를 따르며, Y의 추정량, $\hat{Y} = (n_t/n)$이다. 여기서 추정량 $\hat{\pi_t}$의 분산 $\text{Var}(\hat{\pi}_t)$은 다음과 같다.

$$\text{Var}(\hat{\pi}_t) = \frac{\pi_t(1-\pi_t)}{n} + \frac{(1-M)(1-P)(1-L)[1-(1-M)(1-P)(1-L)]}{n[2M-1+2(1-M)L+2(1-M)(1-L)P]^2}. \tag{2.9}$$

Kim과 Elam (2005)이 제안한 2단계 층화확률화응답모형을 살펴보면, 표본은 각각의 층에서 단순임의 복원추출법을 이용하여 표본을 추출하였다. 첫 번째 단계의 질문에서 층 i에 속하는 각 응답자는 확률

M_i로 민감한 질문으로 구성되는 확률적장치 R_{1i}를, "예"라고 응답하지 않은 응답자는 확률 $1-M_i$로 두 번째 단계의 확률적장치 R_{2i}를 이용하도록 지도되었다. 층 i에 속하는 두 번째 단계의 질문에서 응답자는 확률 P_i로 민감한 질문으로 그리고 확률 $1-P_i$로 부정적인 질문으로 구성되는 확률적장치 R_{2i}를 이용하도록 하였다.

n_i는 층 i에 속하는 표본수이고, n은 모든 층에 속하는 총 표본수이며, $n=\sum_{i=1}^{l} n_i$이다. "예"와 "아니오"에 대한 응답이 진실하고, M_i, P_i가 연구자에 의하여 설정되었다고 가정하면, 층 i에 속하는 표본 중에서 "예"라고 응답할 확률 Y_i는 다음과 같다.

$$Y_i = M_i\pi_{k_i} + (1-M_i)[P_i\pi_{k_i} + (1-P_i)(1-\pi_{k_i})], \quad i=1,2,\ldots,l. \tag{2.10}$$

위의 식 (2.10)에서 π_{k_i}는 층 i에서 응답자가 민감한 그룹에 속하는 모비율이다. 이때, 층 i에 속하는 응답자 n_i에서 "예"라고 응답한 응답자의 수를 n_{k_i}라고 하면, n_{k_i}는 이항분포, $\mathrm{Bin}(n_i, Y_i)$를 따른다. Y_i의 추정량은, $\hat{Y}_i = n_{k_i}/n_i$임으로, π_{k_i}의 최우추정량, $\hat{\pi}_{k_i}$는 다음과 같고,

$$\hat{\pi}_{k_i} = \frac{\hat{Y}_i - (1-M_i)(1-P_i)}{2P_i - 1 + 2M_i(1-P_i)}, \quad i=1,2,\ldots,l. \tag{2.11}$$

$\hat{\pi}_{k_i}$의 분산은

$$\mathrm{Var}(\hat{\pi}_{k_i}) = \frac{\pi_{k_i}(1-\pi_{k_i})}{n_i} + \frac{(1-M_i)(1-P_i)[1-(1-M_i)(1-P_i)]}{n_i[2P_i-1+2M_i(1-P_i)]^2}, \quad i=1,2,\ldots,l \tag{2.12}$$

이다.

이때, 각각의 층에서 단순임의복원추출법을 이용하여 표본이 독립적으로 추출되었기 때문에, 각 층에서 추정량, $\hat{\pi}_{k_i}$, $i=1,2,\ldots,l$의 합, $\sum_{i=1}^{l} w_i\hat{\pi}_{k_i}$를 구하여 전체 모집단에 대한 추정량을 얻을 수 있다. N을 전체 모집단에 속하는 개체수, N_i는 층 i에 속하는 모집단에서의 개체수라고 하고, $w_i = (N_i/N)$, $i=1,2,\ldots,l$이며, $w=\sum_{i=1}^{l} w_i = 1$이라고 하면, 전체 모집단에서 민감한 집단에 속하는 응답자들의 비율, π_k에 대한 최우추정량 $\hat{\pi}_k$는 다음과 같다.

$$\hat{\pi}_k = \sum_{i=1}^{l} w_i\hat{\pi}_{k_i} = \sum_{i=1}^{l} w_i\left[\frac{\hat{Y}_i - (1-M_i)(1-P_i)}{2P_i-1+2M_i(1-P_i)}\right], \quad i=1,2,\ldots,l. \tag{2.13}$$

이때, $\hat{\pi}_k$의 분산은

$$\mathrm{Var}(\hat{\pi}_k) = \sum_{i=1}^{l}\frac{{w_i}^2}{n_i}\left\{\pi_{k_i}(1-\pi_{k_i}) + \frac{(1-M_i)(1-P_i)[1-(1-M_i)(1-P_i)]}{[2P_i-1+2M_i(1-P_i)]^2}\right\}, \quad i=1,2,\ldots,l \tag{2.14}$$

이며, Kim과 Elam (2005)의 근사적 최적배분공식,

$$\frac{n_i}{n} = \frac{w_i\left[\pi_{k_i}(1-\pi_{k_i}) + \dfrac{(1-M_i)(1-P_i)\{1-(1-M_i)(1-P_i)\}}{\{2P_i-1+2M_i(1-P_i)\}^2}\right]^{\frac{1}{2}}}{\displaystyle\sum_{i=1}^{l} w_i\left[\pi_{k_i}(1-\pi_{k_i}) + \frac{(1-M_i)(1-P_i)\{1-(1-M_i)(1-P_i)\}}{\{2P_i-1+2M_i(1-P_i)\}^2}\right]^{\frac{1}{2}}}, \quad i=1,2,\ldots,l \tag{2.15}$$

을 이용하여 n_i를 구하고, 식 (2.14)에 대입하면, 추정량 $\hat{\pi}_k$의 최소분산은 다음과 같다.

$$\mathrm{Var}(\hat{\pi}_k) = \frac{1}{n}\left[\sum_{i=1}^{l} w_i\left\{\pi_{k_i}(1-\pi_{k_i}) + \frac{(1-M_i)(1-P_i)[1-(1-M_i)(1-P_i)]}{[2P_i-1+2M_i(1-P_i)]^2}\right\}^{\frac{1}{2}}\right]^2, \quad i=1,2,\ldots,l. \tag{2.16}$$

3. 3단계 층화확률화응답모형 제안

본 연구에서는 Kim과 Elam (2005)이 제안한 2단계 층화확률화응답모형과 김종민과 채성산 (2010)의 3단계 확률화응답모형을 결합하여 새로운 3단계 층화확률화응답모형을 제안하였다. 모집단은 각각의 층으로 분할되어 있고, 표본은 각각의 층에서 단순임의복원추출법을 이용하여 표본을 추출하였으며, 각 층의 표본수는 알려져 있다고 가정하였다.

조사면접의 첫 번째 단계에서 i번째 층에 속하는 응답자는 확률 M_i로 민감한 질문으로 구성되는 확률적 장치 R_{1i}를, 그렇지 않은 경우의 응답자는 확률 $1-M_i$로 두 번째 단계의 확률적장치 R_{2i}를 이용하도록 지도되었다. 두 번째 단계의 질문에서 i번째 층에 속하는 응답자는 확률 P_i로 민감한 질문으로 구성되는 확률적장치 R_{2i}를, 확률 $1-P_i$로 세 번째 단계의 확률적장치 R_{3i}를 이용하도록 지도되었다. i번째 층에 속하는 세 번째 단계의 응답자는 확률 L_i로 민감한 질문으로, 그리고 확률 $1-L_i$로 부정적인 질문으로 구성되는 확률적장치 R_{3i}를 이용하도록 하였다.

n_i는 각 층 i에 속하는 개체수이고, n은 모든 층에 속하는 총 표본수이며, $n=\sum_{i=1}^{l} n_i$이다. "예"와 "아니오"의 응답이 진실하다고 가정하고 M_i와 P_i가 조사자에 의해 고정되어 있을 때, π_{s_i}는 층 i에 속하는 응답자가 민감한 집단에 속하는 모비율이다. 이때, 층화확률화응답모형에서 층 i에 속하는 응답자가 "예"라고 응답할 확률 Y_i는 다음과 같다.

$$Y_i = M_i\pi_{s_i} + (1-M_i)[P_i\pi_{s_i} + (1-P_i)\{L_i\pi_{s_i} + (1-L_i)(1-\pi_{s_i})\}], \quad i=1,2,\ldots,l. \tag{3.1}$$

위의 식 (3.1)에서 π_{s_i}는 층 i에서 응답자가 민감한 그룹에 속하는 모비율이다. 이때, 층 i에 속하는 응답자 n_i에서 "예"라고 응답한 응답자의 수를 n_{s_i}라고 하면, n_{s_i}는 이항분포, $\mathrm{Bin}(n_i, Y_i)$를 따르고, 층 i에 속하는 표본 중에서 "예"라고 응답한 비율 Y_i의 추정량은 $\hat{Y}_i = (n_{s_i}/n_i)$이며, π_{s_i}의 최우추정량, $\hat{\pi}_{s_i}$는 다음과 같다.

$$\hat{\pi}_{s_i} = \frac{\hat{Y}_i - (1-L_i)(1-M_i)(1-P_i)}{2M_i - 1 + 2(1-M_i)L_i + 2(1-M_i)(1-L_i)P_i}, \quad i=1,2,\ldots,l. \tag{3.2}$$

정리 3.1 추정량 $\hat{\pi}_{s_i}$은 모비율 π_{s_i}의 불편추정량이다.

증명:

$$\begin{aligned} E(\hat{\pi}_{s_i}) &= E\left[\frac{\hat{Y}_i - (1-L_i)(1-M_i)(1-P_i)}{2M_i - 1 + 2(1-M_i)L_i + 2(1-M_i)(1-L_i)P_i}\right] \\ &= \frac{Y_i - (1-L_i)(1-M_i)(1-P_i)}{2M_i - 1 + 2(1-M_i)L_i + 2(1-M_i)(1-L_i)P_i} \\ &= \pi_{s_i}, \quad i=1,2,\ldots,l. \end{aligned} \tag{3.3}$$

□

정리 3.2 추정량 $\hat{\pi}_{s_i}$의 분산은,

$$\mathrm{VAR}(\hat{\pi}_{s_i}) = \frac{\pi_{s_i}(1-\pi_{s_i})}{n_i} + \frac{(1-M_i)(1-P_i)(1-L_i)[1-(1-M_i)(1-P_i)(1-L_i)]}{n_i[2M_i - 1 + 2(1-M_i)L_i + 2(1-M_i)(1-L_i)P_i]^2}, \quad i=1,2,\ldots,l. \tag{3.4}$$

증명:

$$\mathrm{Var}(\hat{\pi}_{s_i}) = \frac{\mathrm{Var}(\hat{Y}_i)}{[2M_i - 1 + 2(1-M_i)L_i + 2(1-M_i)(1-L_i)P_i]^2}$$

$$= \frac{n_i Y_i(1-Y_i)}{n_i^2[2M_i - 1 + 2(1-M_i)L_i + 2(1-M_i)(1-L_i)P_i]^2}$$
$$= \frac{\pi_{s_i}(1-\pi_{s_i})}{n_i} + \frac{(1-M_i)(1-P_i)(1-L_i)[1-(1-M_i)(1-P_i)(1-L_i)]}{n_i[2M_i - 1 + 2(1-M_i)L_i + 2(1-M_i)(1-L_i)P_i]^2},$$
$$i = 1, 2, \ldots, l. \tag{3.5}$$

□

이때, 다른 층에서의 표본추출이 독립적으로 이루어졌기 때문에, 각 층에서 추정량들의 합을 구하여 전체 모집단에 대한 추정량을 얻을 수 있다. 따라서 민감한 집단에 속하는 응답자들의 비율, π_s에 대한 최우추정량 $\hat{\pi}_s$는 다음과 같다.

$$\hat{\pi}_s = \sum_{i=1}^{l} w_i \hat{\pi}_{s_i} = \sum_{i=1}^{l} w_i \left[\frac{\hat{Y}_i - (1-L_i)(1-M_i)(1-P_i)}{2M_i - 1 + 2(1-M_i)L_i + 2(1-M_i)(1-L_i)P_i} \right], \quad i = 1, 2, \ldots, l. \tag{3.6}$$

위의 식 (3.6)에서, $w_i = (N_i/N)$, $i = 1, 2, \ldots, l$, $w = \sum_{i=1}^{l} w_i = 1$이며, N은 모집단에 속하는 개체수이고, N_i는 층 i에 속하는 개체수이다.

정리 3.3 추정량 $\hat{\pi}_s$는 모집단 비율 π_s의 불편추정량이다.

증명: 식 (3.6)에 기대값을 취하면 된다. □

정리 3.4 추정량 $\hat{\pi}_s$의 분산은 다음과 같다.

$$\text{VAR}(\hat{\pi}_s) = \sum_{i=1}^{l} \frac{w_i^2}{n_i} \left\{ \pi_{s_i}(1-\pi_{s_i}) + \frac{(1-L_i)(1-M_i)(1-P_i)[1-(1-L_i)(1-M_i)(1-P_i)]}{[2M_i - 1 + 2(1-M_i)L_i + 2(1-M_i)(1-L_i)P_i]^2} \right\},$$
$$i = 1, 2, \ldots, l. \tag{3.7}$$

증명: 식 (3.5), (3.6)과 Chochran (1977), Section 5.10의 보조정리 1을 이용하여 얻을 수 있다. □

위의 식 (3.7)에서, 일반적으로 π_{s_i}에 대한 정보는 알 수 없으나, 과거의 경험에서 π_{s_i}에 대한 사전정보를 얻을수 있기 때문에 다음과 같은 최적배분공식을 얻을 수 있다.

정리 3.5 $n = \sum_{i=1}^{l} n_i$의 표본에 대한 추정량 $\hat{\pi}_s$의 최소분산 유도하기 위하여 각각의 층에 $n_1, n_2, \ldots, n_l$로 배분하는 근사적 최적배분비율은 다음과 같다.

$$\frac{n_i}{n} = \frac{w_i \left[\pi_{s_i}(1-\pi_{s_i}) + \frac{(1-L_i)(1-M_i)(1-P_i)\{1-(1-L_i)(1-M_i)(1-P_i)\}}{\{2M_i-1+2(1-M_i)L_i+2(1-M_i)(1-L_i)P_i\}^2} \right]^{\frac{1}{2}}}{\sum_{i=1}^{l} w_i \left[\pi_{s_i}(1-\pi_{s_i}) + \frac{(1-L_i)(1-M_i)(1-P_i)\{1-(1-L_i)(1-M_i)(1-P_i)\}}{\{2M_i-1+2(1-M_i)L_i+2(1-M_i)(1-L_i)P_i\}^2} \right]^{\frac{1}{2}}}, \quad i = 1, 2, \ldots, l. \tag{3.8}$$

증명: Chochran (1977), Section 5.5의 Theorem 5.7에서 위의 결과를 얻을 수 있다. □

위의 식 (3.8)을 이용하여 n_i를 구하고 식 (3.7)에 대입하면, 추정량 $\hat{\pi}_s$의 최소분산은 다음과 같다.

$$\text{Var}(\hat{\pi}_s) = \frac{1}{n} \left[\sum_{i=1}^{l} w_i \left\{ \pi_{s_i}(1-\pi_{s_i}) + \frac{(1-L_i)(1-M_i)(1-P_i)[1-(1-L_i)(1-M_i)(1-P_i)]}{[2M_i - 1 + 2(1-M_i)L_i + 2(1-M_i)(1-L_i)P_i]^2} \right\}^{\frac{1}{2}} \right]^2,$$
$$i = 1, 2, \ldots, l. \tag{3.9}$$

4. Kim-Elam모형과의 효율성 비교

본 연구에서 제안된 3단계 층화확률화응답모형과 Kim과 Elam (2005)의 2단계 층화확률화모형의 비교를 위하여 상대효율(relative efficiency; RE)을 이용하였다. 여기서 상대효율은 응답자가 민감한 집단에 속하는 모비율의 추정량, $\hat{\pi}_k$, $\hat{\pi}_s$의 분산을 이용하여 계산하였으며, Kim과 Elam (2005)에 대하여 본 연구에서 제안된 3단계 층화확률화응답모형의 상대효율은 다음과 같이 정의하였다.

$$\mathrm{RE} = \frac{\mathrm{Var}(\hat{\pi}_k)}{\mathrm{Var}(\hat{\pi}_s)}$$

상대효율(RE)이 1.0 보다 큰 경우, 즉 3단계 층화확률화응답모형의 $\mathrm{Var}(\hat{\pi}_s)$이 Kim과 Elam (2005)의 2단계 층화확률화모형의 $\mathrm{Var}(\hat{\pi}_k)$ 보다 작으면, 상대효율이 좋다는 것을 의미한다.

상대효율(RE)을 계산하기 위해 모집단에 2개의 층이 있다고 가정하고, 다음과 같이 모수를 설정하였다.

1. 표본수 $n = 1000$이고, $n = n_1 + n_2$, $(n_1, n_2) = \{(300, 700), (500, 500), (700, 300)\}$;
2. 층의 표본수 (n_1, n_2)에 따라, $(w_1, w_2) = \{(.3, .7), (.5, .5), (.7, .3)\}$;
3. 층에서 응답자가 민감한 집단에 속하는 모비율 $(\pi_{s_1}, \pi_{s_2}) = \{(.08, .13), (.13, .18)\}$;
4. 응답자가 민감한 집단에 속하는 모비율, $\pi_s = w_1\pi_{s_1} + w_2\pi_{s_2}$;
5. 첫번째 단계에서 민감한 질문에 응답한 확률, $(M_1, M_2) = \{(.1, .15), (.2, .3), (.3, .45)\}$;
6. 두번째 단계에서 민감한 질문에 응답할 확률, $(P_1, P_2) = \{(.1, .2), (.2, .3), (.3, .4)\}$;
7. 세번째 단계에서 민감한 질문에 응답할 확률, $(L_1, L_2) = \{(.1, .15), (.2, .25), \ldots, (.9, .95)\}$.

설정된 모수들의 모든 조합에 대하여 $\mathrm{Var}(\hat{\pi}_k)$에 대한 $\mathrm{Var}(\hat{\pi}_s)$의 상대효율을 계산하였으며, 상대효율(RE)이 1.0보다 크면 3단계 층화확률화응답모형이 Kim과 Elam (2005)의 2단계 층화확률화응답모형보다 더 효율적임을 나타낸다. 설정된 모수의 모든 조합에 대하여 그 결과들을 살펴보았으나, 모수설정의 변화에 따른 상대효율의 경향이 서로 유사하여 대표적으로 $(\pi_{s_1}, \pi_{s_2}) = (.08, .13)$의 경우와 $(\pi_{s_1}, \pi_{s_2}) = (.13, .18)$의 경우를 표 4.1과 4.2에 요약하였다.

표 4.1에서 (w_1, w_2)의 설정에 관계없이 (M_1, M_2)의 값이 커지면 상대효율이 좋게 나타나는 결과를 얻을 수 있다. 또한, (M_1, M_2)의 고정된 값에 대하여 (P_1, P_2)의 값이 증가하면, $\mathrm{Var}(\hat{\pi}_k)$의 값이 증가하며, 반대로 $\mathrm{Var}(\hat{\pi}_s)$의 값은 감소하는 경향이 있다. 추가적으로 (L_1, L_2)의 값이 증가하면 $\mathrm{Var}(\hat{\pi}_s)$의 값은 감소하는 것으로 관찰되었으며, 제안된 3단계 층화확률화응답모형의 상대효율이 증가하였다.

다른 한편으로 $M_1 \le M_2 \le 0.2$이고 $P_1 \le P_2 \le 0.2$인 경우, $L_1 \le L_2 \le 0.65$이면, 3단계 층화확률화응답모형의 상대효율이 1.0보다 작은 경우도 관찰되었다. 그러나 제안된 3단계 층화확률화응답모형의 첫번째 단계에서 질문 1에 응답할 확률 (M_1, M_2)가 상당히 큰 경우$(M_2 \ge M_1 \ge 0.3)$, 두 번째 단계와 세번째 단계에서 질문 1에 응답할 확률인 (P_1, P_2)와 (L_1, L_2)의 값이 증가하면, 본 연구에서 제안된 3단계 층화확률화응답모형의 상대효율이 증가하는 경향이 있음을 알 수 있다.

표 4.1에 대하여 설명된 결과들은 $(\pi_{s_1}, \pi_{s_2}) = (.13, .18)$의 설정에 따른 표 4.2에서도 유사한 경향을 나타내고 있다. 여기서 표 4.2의 결과를 $(\pi_{s_1}, \pi_{s_2}) = (.08, .13)$로 설정된 표 4.1의 결과와 비교하여 살펴보면, $V(\hat{\pi}_k)$의 값이 증가하는 반면에 $V(\hat{\pi}_s)$의 값은 감소하여, 결과적으로 (π_{s_1}, π_{s_2})의 값이 증가하면 상대효율(RE)이 증가하는 경향이 관찰되었다.

각 층에서 단순임의 복원추출방법을 이용하여 표본을 추출한다는 전제하에, 표 4.1과 4.2의 결과를 전반적으로 살펴보았다. 결과적으로 $M_1 \le M_2 \le 0.2$이고 $P_1 \le P_2 \le 0.2$인 경우, $L_1 \le L_2 \le 0.65$이면 상

표 **4.1.** $(\pi_{s1}, \pi_{s2}) = (0.08, 0.13)$, $\pi_s = w_1\pi_{s1} + w_2\pi_{s2}$일 때, $\hat{\pi}_k$와 $\hat{\pi}_s$의 분산과 $\hat{\pi}_s$의 상대효율(RE)

$(w_1,\ w_2)$	$(M_1,\ M_2)$	$(P_1,\ P_2)$	$(L_1,\ L_2)$	$V(\hat{\pi}_k)$	$V(\hat{\pi}_s)$	RE
(0.3, 0.7)	(0.1, 0.15)	(0.1, 0.2)	(0.5, 0.55)		.003549	0.3689
			(0.6, 0.65)		.001139	1.1480
			(0.7, 0.75)	.001308	.000546	2.3935
			(0.8, 0.85)		.000294	4.4524
			(0.9, 0.95)		.000160	8.1789
		(0.3, 0.4)	(0.5, 0.55)	.	.001419	229.8111
			(0.6, 0.65)		.000679	480.2317
			(0.7, 0.75)	326203	.000389	838.2590
			(0.8, 0.85)		.000237	1375.8954
			(0.9, 0.95)		.000145	2247.0692
	(0.3, 0.45)	(0.1, 0.2)	(0.5, 0.55)		.000888	13.5479
			(0.6, 0.65)		.000525	22.9212
			(0.7, 0.75)	.012028	.000334	35.9876
			(0.8, 0.85)		.000218	55.0986
			(0.9, 0.95)		.000141	85.1832
		(0.3, 0.4)	(0.5, 0.55)	.	.000569	126.9400
			(0.6, 0.65)		.000379	190.3996
			(0.7, 0.75)	072169	.000264	273.4410
			(0.8, 0.85)		.000187	386.5641
			(0.9, 0.95)		.000131	549.2099
(0.7, 0.3)	(0.1, 0.15)	(0.1, 0.2)	(0.5, 0.55)		.005770	0.1034
			(0.6, 0.65)		.001737	0.4495
			(0.7, 0.75)	.000781	.000711	1.0974
			(0.8, 0.85)		.000346	2.2591
			(0.9, 0.95)		.000172	4.5303
		(0.3, 0.4)	(0.5, 0.65)		.002725	28.5029
			(0.6, 0.55)		.001022	75.9620
			(0.7, 0.75)	.077666	.000503	154.5121
			(0.8, 0.85)		.000273	284.1251
			(0.9, 0.95)		.000151	513.5133
	(0.3, 0.45)	(0.1, 0.2)	(0.5, 0.55)		.001467	4.4629
			(0.6, 0.65)		.000744	8.7954
			(0.7, 0.75)	.006547	.000424	15.4483
			(0.8, 0.85)		.000252	26.0285
			(0.9, 0.95)		.000147	44.4177
		(0.3, 0.4)	(0.5, 0.55)		.000901	356.4027
			(0.6, 0.65)		.000524	613.4899
			(0.7, 0.75)	.321288	.000326	985.4993
			(0.8, 0.85)		.000209	1540.1681
			(0.9, 0.95)		.000133	2421.5510

대효율이 1.0 보다 작은 경우도 관찰되었으나, 제안된 3단계 층화확률화응답모형의 상대효율이 증가하여 Kim과 Elam (2005)의 2단계 층화확률화응답모형보다 효율성이 좋은 것으로 판단된다. 또한, 첫 번째 단계와 두 번째 단계 그리고 세 번째 단계에서 민감한 질문에 응답할 확률이 증가하면 상대효율이 증가한다는 것이 관찰되었다. 현실적인 측면에서 조사자는 설문조사의 과정에서 응답자의 신분이나 비밀

표 **4.2.** $(\pi_{s1}, \pi_{s2}) = (0.13, 0.18)$, $\pi_s = w_1\pi_{s1} + w_2\pi_{s2}$일 때, $\hat{\pi}_k$와 $\hat{\pi}_s$의 분산과 $\hat{\pi}_s$의 상대효율(RE)

$(w_1,\ w_2)$	$(M_1,\ M_2)$	$(P_1,\ P_2)$	$(L_1,\ L_2)$	$V(\hat{\pi}_k)$	$V(\hat{\pi}_s)$	RE
(0.3, 0.7)	(0.1, 0.15)	(0.1, 0.2)	(0.5, 0.55)	.001348	.003589	0.3756
			(0.6, 0.65)		.001177	1.1453
			(0.7, 0.75)		.000583	2.3125
			(0.8, 0.85)		.000330	3.0868
			(0.9, 0.95)		.000196	6.8819
		(0.3, 0.4)	(0.5, 0.55)	.326334	.001460	223.4610
			(0.6, 0.65)		.000717	455.2845
			(0.7, 0.75)		.000426	766.8114
			(0.8, 0.85)		.000273	1194.7875
			(0.9, 0.95)		.000181	1801.6135
	(0.3, 0.45)	(0.1, 0.2)	(0.5, 0.55)	.012070	.000926	13.0297
			(0.6, 0.65)		.000562	21.4862
			(0.7, 0.75)		.000371	32.5717
			(0.8, 0.85)		.000254	47.4554
			(0.9, 0.95)		.000177	68.1280
		(0.3, 0.4)	(0.5, 0.55)	.072317	.000607	119.2286
			(0.6, 0.65)		.000416	173.9043
			(0.7, 0.75)		.000300	240.9248
			(0.8, 0.85)		.000223	324.7397
			(0.9, 0.95)		.000167	432.0292
(0.7, 0.3)	(0.1, 0.15)	(0.1, 0.2)	(0.5, 0.55)	.000823	.007595	0.1083
			(0.6, 0.65)		.001776	0.4632
			(0.7, 0.75)		.000750	1.0973
			(0.8, 0.85)		.000384	2.1446
			(0.9, 0.95)		.000210	3.9126
		(0.3, 0.4)	(0.5, 0.55)	.077799	.002768	28.1090
			(0.6, 0.65)		.001062	73.2614
			(0.7, 0.75)		.000541	143.7843
			(0.8, 0.85)		.000311	249.8378
			(0.9, 0.95)		.000189	411.1780
	(0.3, 0.45)	(0.1, 0.2)	(0.5, 0.55)	.006591	.001508	4.3722
			(0.6, 0.65)		.000783	8.4137
			(0.7, 0.75)		.000462	14.2619
			(0.8, 0.85)		.000290	22.7609
			(0.9, 0.95)		.000185	35.5578
		(0.3, 0.4)	(0.5, 0.55)	.321438	.000941	341.4159
			(0.6, 0.65)		.000563	571.4374
			(0.7, 0.75)		.000364	882.4662
			(0.8, 0.85)		.000247	1303.4599
			(0.9, 0.95)		.000171	1883.4773

이 보장됨을 주지시켜 응답자가 민감한 질문에 응답할 확률을 높이는 노력이 필요하며 상대효율도 증가하게 될 것이다.

층화를 통하여 응답자 개개인의 특성을 최대한 얻을 수 있는 방법이 제안된 모형의 장점이며, 특정한 조건을 제외하고는 본 연구에서 제안된 모형의 효율성이 증가함을 알 수 있다. 그러나, 2단계 층화확률화

응답모형을 3단계로 확장함으로써 상대적으로 효율성은 증대되지만 응답절차가 한 단계 증가함으로써 응답과정을 응답자에게 설명할 때 발생하는 조사과정의 어려움이 예상된다.

5. 결론

본 연구에서는 Kim과 Elam (2005)의 2단계 층화확률화모형과 김종민과 채성산 (2010)의 3단계 확률화응답모형을 결합하여 새로운 3단계 층화확률화응답모형을 제안하였다. 표본은 각각의 층에서 단순임의복원추출방법에 의해 추출하였으며, 제안된 3단계 층화확률화응답모형과 Kim과 Elam (2005)의 2단계 층화확률화응답모형의 상대효율을 비교하였다.

특정한 조건을 제외하고는 본 연구에서 제안된 모형의 효율성이 증가함을 알 수 있으며, 제안된 모형의 장점인 층화를 통하여 응답자 개개인에 대한 상세정보를 얻을 수 있으므로 연구자들은 3단계 층화확률화응답모형을 의학 또는 범죄와 관련된 연구에 적용하여 이점을 얻을 수 있을 것이다. 그러나, 2단계 층화확률화응답모형을 3단계로 확장함으로써 상대적으로 효율성은 증대되지만 응답절차가 한 단계 증가함으로써 응답과정을 응답자에게 설명할 때 발생하는 조사과정의 번거로움이 예상된다.

참고문헌

김종민, 채성산 (2010). 3단계 확률화응답모형, <대전대 사회과학논문집>, **28**, 77–87.

김종호, 류제복, 이기성 (1992). 새로운 2단계 확률화응답모형, <응용통계연구>, **5**, 157–167.

이기성, 홍기학 (1998). 단순집락추출법에 의한 양적속성의 무관질문모형, <응용통계연구>, **11**, 141–150.

이기성, 홍기학 (2000). 2단계 이표본 무관질문모형, <응용통계연구>, **13**, 575–590.

Abul-Ela, A. A., Greenberg, B. G. and Horvitz, D. G. (1967). A multiproportions randomized response model, *Journal of the American Statistical Association*, **62**, 990–1008.

Chaudhuri, A. and Mukerjee, R. (1988). *Randomized Response: Theory and Techniques*, Marcel Dekker, New York.

Chochran, W. G. (1977). *Sampling Techniques*, John Wiley & Sons, New York.

Drane, W. (1975). Randomized response to more than one question, *American Statistical Association Proceedings of the Social Statistics*, 395–397.

Folsom, R. E., Greenberg, B. G., Horvitz, D. G. and Abernathy, J. R. (1973). The two alternative questions randomized response model for human surveys, *Journal of the American Statistical Association*, **68**, 525–530.

Greenberg, B. G., Abul-Ela, A., Simmons, W. R. and Horvitz, D. G. (1969). The unrelated question randomized response: Theoretical framework, *Journal of the American Statistical Association*, **64**, 529–539.

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

Mangat, N. S. and Singh, R. (1990). An alternative randomized response procedure, *Biometrika*, **77**, 439–442.

Moors, J. J. A. (1971). Optimization of the unrelated question randomized response model, *Journal of the American Statistical Association*, **66**, 627–629.

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

Three-Stage Stratified Randomized Response Model

Jong-Min Kim[1] · Seong S. Chae[2]

[1]Statistics Discipline, University of Minnesota
[2]Department of Business Information Statistics, Daejeon University

(Received February 2010; accepted May 2010)

Abstract

Asking sensitive questions by a direct survey method causes non-response bias and response bias. Non-response bias arises from interviewees' refusal to respond and response bias arises from giving incorrect responses. To rectify these biases, Warner (1965) introduced a randomized response model which is an alternative survey method for socially undesirable or incriminating behavior questions. The randomized response model is a procedure for collecting the information on sensitive characteristics without exposing the identity of the respondent. Many survey researchers have proposed diverse variants of the Warner randomized response model and applied their model to collect the information of sensitive questions. Using an optimal allocation, we proposed three-stage stratified randomized response technique which is an extension of the Kim and Elam (2005) two-stage stratified randomized response technique. In this study, we showed that the estimator based on the proposed response model is more efficient than Kim and Elam (2005). But by adding one more survey step to the Kim and Elam (2005), our proposed model may have relatively less privacy protection compared to the Kim and Elam (2005) model.

Keywords: Randomized response model, stratified random sample.

[2]Corresponding author: Professor, Department of Business Information Statistics, Daejeon University, Yongun-Dong, Dong-Gu, Daejeon 300-716, Korea. E-mail: chae@dju.kr