응용통계연구 제 19권 1호, 2006년, pp. 69–80

확률비례추출법에 의한 확률화응답기법에 관한 연구*

이기성[1)]

요 약

본 연구에서는 매우 민감한 조사에서 모집단이 집락의 크기가 서로 다른 여러 개의 집락으로 구성되어 있을 때, 집락의 크기에 비례하게 추출확률을 부여하는 확률비례추출법(probability proportional to size : pps)을 이용한 확률화응답기법을 제안하고자 한다. 민감한 속성에 대한 모수의 추정치와 분산 및 분산추정량을 구하여 이론적 체계를 구축하고, 확률비례추출법에 의한 확률화응답기법과 등확률 2단계 추출법에 의한 확률화응답기법의 효율성을 비교해 보고자 한다. 또한, 실제조사를 통해 제안한 확률비례추출법에 의한 확률화응답기법에 대한 실용화의 타당성을 검토하고자 한다.

주요용어: 민감한 정보, 크기가 다른 집락, 확률비례추출, Warner기법, 실용화.

1. 서론

사회적으로나 개인적으로 매우 민감한 문제에 관한 조사에서 응답자들이 직접질문을 받았을 경우 응답을 회피하거나 거짓 응답으로 인해 비표본오차가 증가하게 된다. 예를 들어 음주운전, 낙태경험, 환각제사용, 동성연애 및 탈세여부 등과 같은 민감한 조사에서 기존의 직접질문방식을 그대로 사용할 경우 응답자들이 민감한 질문에 응답함으로써 불이익을 받거나 사생활이 보장되지 않는다고 생각하기 때문에 응답을 회피하거나 거짓 응답을 하게 된다. 이처럼 민감한 질문에 대한 조사에서 발생하는 비표본오차를 줄이기 위하여 1965년 Warner는 응답자들에게 직접적인 응답을 요구하는 것이 아니라 확률장치를 통한 간접적인 응답만을 요구함으로 응답자들의 신분을 보장해 주는 획기적인 확률화응답기법을 제시하였다. 그 후 미국, 캐나다, 영국, 호주 등 서구 여러 나라와 일본, 인도 등 몇몇 아시아 국가에서도 이 분야에 대한 연구가 활발히 진행되고 있다. 특히, Fox와 Tracy(1986), Chaudhuri와 Mukerjee(1988)는 확률화응답기법을 정리, 요약하여 체계화하였고, 최근에는 이들 이론들의 실제적 활용에 많은 관심이 집중되고 있으며, 사회학, 경영학, 의학 등 여러 학문분야에서의 조사활동에도 이의 활용이 적극 모색되고 있다. 국내에서는 류제복외 2인(1993)이 확률화응답기법에 관한 책을 출간하여 그 중요성을 강조하였고, 류제복외 2인(1995)은 확률화응답기법의 실용화 방안을 제시하였다. 이러한 확률화응답기법은 직접 질문을 사용하는 경우보다 시간, 비용과 노력을 더 필요로 하며, 특히 모집단이 큰 경우에 단순임의추출법

* 이 논문은 정부(교육인적자원부)의 재원으로 한국학술진흥재단의 지원을 받아 수행된 연구임 (R05-2004-000-10160-0).

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

을 이용하여 응답자들을 추출하여 조사를 하는 데에는 여러 가지 어려움이 따르게 된다. 최근, 이기성외 3인(2003)은 이러한 문제점을 해결하기 위하여 매우 민감한 조사에서 모집단이 여러 개의 집락으로 구성되어 있을 때, 3단계 집락추출법에 확률화응답기법을 적용하기 위한 이론적 체계를 구축하였으며, 3단계 집락 확률화응답기법을 실제조사에 적용해 봄으로써 나타나는 문제점을 파악하고 그 대안을 마련하였다. 이 연구에서 그들은 집락의 크기가 모두 동일하다고 가정하여 집락과 부차표본을 추출하는 데 임의추출법을 사용하고 있다. 하지만 실제조사에 있어서 많은 경우 집락의 크기가 같지 않은 것을 볼 수 있으며, 이때 집락을 추출하는 데 있어서 집락의 크기를 고려하지 않을 경우에 현실성이 떨어지는 문제점이 있게 된다.

본 연구에서는 첫째, 확률화응답기법의 실용화를 위한 선행연구들에 대하여 간략히 살펴보고 둘째, 민감한 조사에서 모집단이 집락의 크기가 서로 다른 여러 개의 집락으로 구성되어 있을 때, 집락의 크기에 비례하게 추출확률을 부여하는 확률비례추출법(probability proportional to size : pps)을 이용한 확률화응답기법을 제안하여, 실제조사에 사용할 수 있는 이론적 체계를 구축하고자 한다. 셋째, 민감한 속성에 대한 모수의 추정치와 분산 및 분산추정량을 구하고, 확률비례추출법에 의한 확률화응답기법과 등확률 2단계 추출법에 의한 확률화응답기법의 효율성을 비교해 보고자 한다. 넷째, 실제조사를 통해 제안한 확률비례추출법에 의한 확률화응답기법에 대한 실용화의 타당성을 검토하고자 한다.

2. 확률화응답기법의 실용화 선행 연구

확률화응답기법은 사회학, 교육심리, 의약 등 다양한 분야에서 민감한 정보를 얻는 데 실제 활용되고 있다. 확률화응답기법이 실제조사에 적용된 사례들과 최근 연구방향에 대하여 간략하게 살펴보고자 한다. Barth와 Sandler(1976)는 무관질문기법(unrelated question technique)을 이용하여 고등학교 학생들의 알콜 중독자 비율을 추정하였고, Begin, Boivin과 Bellerose(1979)는 성문제, 마약복용, 낙태 등 사회적으로 민감한 문제에 대하여 Morton기법을 적용하여 민감한 모수를 추정하였다. 그리고 Tracy와 Fox(1981)는 성인이 된 후 법적으로 구속된 경우의 수를 추정하기 위하여 Liu와 Chow(1976)의 확률화응답기법을 이용하였으며, Scheer와 Dayton(1987)는 무관질문기법을 이용하여 학생들의 부정행위 비율을 추정하였다. 또한 Danermark와 Swensson(1987)는 마리화나, 코카인, 약물복용 등에 대한 민감한 정보를 얻는 데 Warner기법을 이용하였으며, 류제복외 2인(1995)은 대학생들의 성경험과 책값의 유용 비율을 추정하기 위하여 Morton기법을 이용하기도 하였다. 이처럼 민감한 정보를 얻는데 있어서 다양한 분야에서 여러 형태의 확률화응답기법들이 활용되고 있으며, 실제조사에서 사용되는 확률장치로는 동전, 주사위, 공기돌, 회전판, 지폐 등 일상생활에서 응답자들에게 거부감을 주지 않는 것들이 있다. 이러한 확률화응답기법을 실제조사에 적용하는 경우에 있어서 응답자들을 추출하는 방법으로는 대개 단순임의추출법을 사용하고 있었다. 하지만, 최근에는 다양한 표본추출법들을 확률화응답기법에 적용하여 실용화를 가속화시키고 있다. 특히, 이기성과 홍기학(1988)은 민감한 조사에서 모집단이 집락으로 구성되어 있을 때, 사용 가능한 집락추출법을 확률화응답기법에 적용하였으며, 이기

성외 3인(2003)은 3단계 집락추출법을 Warner기법에 적용하여 대학생들의 성적충동에 대한 실제조사를 실시하였다. 또한, Kim과 Warde(2004)는 층화추출법을 Warner기법에 적용하는 층화 확률화응답기법을 제안하기도 하였다.

3. 확률비례추출법에 의한 확률화응답기법

사회적으로나 개인적으로 매우 민감한 조사에서 각 집락의 크기가 $M_i(i=1,2,\cdots,N)$인 N개의 집락으로 구성되어 있는 모집단으로부터 n개의 집락을 확률비례추출한 후, 추출된 각 집락에서 다시 $m_i(i=1,2,\cdots,n)$개의 조사단위를 단순임의추출하는 2단계 추출법에 확률화응답기법을 적용해 보고자 한다. 먼저, 집락을 확률비례복원추출하는 방법에 대하여 살펴보고, 다음으로 확률비례비복원추출하는 방법에 대하여 다루어 보고자 한다.

3.1. 확률비례복원추출에 의한 확률화응답기법

i번째 1차 추출단위의 추출확률 p_i에 의해서 n개의 1차 추출단위를 복원추출하고, 각 1차 추출단위에서 m_i개의 2차 추출단위를 단순임의비복원 추출한다고 가정하자. 이 때, n개의 1차 추출단위가 각 집락의 크기 M_i에 비례할 경우, 이를 확률비례추출(probability proportional to size : pps)이라 한다. 이러한 2단계 절차에 의해 추출된 응답자들은 다음과 같은 두 개의 설문으로 구성되어 있는 Warner기법의 확률장치를 사용하게 된다.

설문 1 : 당신은 그룹 A(민감한 그룹)에 속합니까?
설문 2 : 당신은 그룹 A(민감한 그룹)에 속하지 않습니까?

여기서 설문 1이 선택될 확률은 $p(\neq 0.5)$이고, 설문 2가 선택될 확률은 $1-p$이다.

응답자들은 확률장치에 의해서 선택된 설문에 대해 "예" 또는 "아니오"라고 응답한다. 따라서, $i(i=1,2,\cdots,N)$번째 집락에서 응답자가 "예" 라고 응답할 확률을 구해 보면 다음과 같다.

$$\lambda_i = (2p-1)\theta_i + (1-p) \tag{3.1}$$

여기서 θ_i는 i번째 집락에서의 민감한 그룹에 속하는 비율이다.

i번째 집락에서의 j번째 조사단위를 z_{ij}라 하고, i번째 집락에서의 j번째 응답자가 "예" 라고 응답하면 $z_{ij}=1$, "아니오"라고 응답하면 $z_{ij}=0$이라고 정의하자. 각 집락으로부터 표본으로 추출된 m_i명의 응답자 중에서 "예" 라고 응답한 사람의 수를 $Z_i = \sum_{j=1}^{m_i} z_{ij}$라 하면, Z_i는 $B(m_i, \lambda_i)$를 따른다.

식 (3.1)에서 λ_i의 추정치는 $\hat{\lambda}_i = Z_i/m_i$이므로 θ_i의 추정량 $\hat{\theta}_i$는 다음과 같다.

$$\hat{\theta}_i = \frac{\hat{\lambda}_i - (1-p)}{2p-1}, \; p \neq \frac{1}{2} \tag{3.2}$$

한편, 민감한 그룹에 속하는 조사단위당 모비율 θ는 다음과 같다.

$$\theta = \frac{1}{M_0}\sum_{i=1}^{N} M_i\theta_i \tag{3.3}$$

여기서 $M_0 = \sum_{i=1}^{N} M_i$이다.

n개의 1차 추출단위를 확률비례복원추출한 후, 추출된 각 집락에서 다시 m_i개의 2차 추출단위를 단순임의비복원추출할 경우, 민감한 그룹에 속하는 모비율 θ의 추정량 $\hat{\theta}_{ppswr}$는 다음과 같다.

$$\hat{\theta}_{ppswr} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i \tag{3.4}$$

정리 3.1 추정량 $\hat{\theta}_{ppswr}$는 모비율 θ의 비편향추정량이다.

증명:

$$\begin{aligned} E_1E_2(\hat{\theta}_{ppswr}) &= E_1E_2\left[\frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i\right] \\ &= E_1\left(\frac{1}{n}\sum_{i=1}^{n}\theta_i\right) \\ &= \frac{1}{M_0}\sum_{i=1}^{N}M_i\theta_i \\ &= \theta \end{aligned}$$

□

정리 3.2 각 집락의 크기가 M_i인 N개의 집락에서 n개의 집락을 확률비례복원추출하고, 추출된 집락에서 다시 m_i개의 조사단위의 표본을 단순임의비복원추출하여 얻어진 민감한 그룹에 속하는 모비율 θ의 추정량 $\hat{\theta}_{ppswr}$의 분산은 다음과 같다.

$$\begin{aligned} V(\hat{\theta}_{ppswr}) &= \frac{1}{nM_0}\sum_{i=1}^{N}M_i(\theta_i-\theta)^2 \\ &\quad + \frac{1}{nM_0}\sum_{i=1}^{N}M_i\frac{1}{m_i}\left[\left(\frac{M_i-m_i}{M_i-1}\right)\theta_i(1-\theta_i) + \frac{p(1-p)}{(2p-1)^2}\right] \end{aligned} \tag{3.5}$$

증명:

$$V(\hat{\theta}_{ppswr}) = V_1E_2(\hat{\theta}_{ppswr}) + E_1V_2(\hat{\theta}_{ppswr})$$

에서

$$\begin{aligned} V_1E_2(\hat{\theta}_{ppswr}) &= V_1E_2\left[\frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i\right] \\ &= V_1\left(\frac{1}{n}\sum_{i=1}^{n}\theta_i\right) \\ &= \frac{1}{nM_0}\sum_{i=1}^{N}M_i(\theta_i-\theta)^2 \end{aligned}$$

이고,

$$\begin{aligned} E_1V_2(\hat{\theta}_{ppswr}) &= E_1V_2\left[\frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_i\right] \\ &= E_1\left[\frac{1}{n^2}\sum_{i=1}^{n}\frac{1}{m_i}\left(\left(\frac{M_i-m_i}{M_i-1}\right)\theta_i(1-\theta_i)+\frac{p(1-p)}{(2p-1)^2}\right)\right] \\ &= \frac{1}{nM_0}\sum_{i=1}^{N}M_i\frac{1}{m_i}\left[\left(\frac{M_i-m_i}{M_i-1}\right)\theta_i(1-\theta_i)+\frac{p(1-p)}{(2p-1)^2}\right] \end{aligned}$$

이므로 추정량 $\hat{\theta}_{ppswr}$의 분산은 식 (3.5)와 같다. □

3.2. 확률비례비복원추출에 의한 확률화응답기법

각 집락의 크기가 M_i인 N개의 집락으로 구성되어 있는 모집단으로부터 n개의 집락을 확률비례비복원추출한 후, 추출된 각 집락에서 다시 m_i개의 조사단위를 단순임의비복원추출하는 2단계 추출법에 확률화응답기법을 적용해 보고자 한다. 이러한 2단계 비복원 추출절차에 의해 얻어진 민감한 그룹에 속하는 조사단위당 모비율 θ의 추정량 $\hat{\theta}_{ppswor}$는 다음과 같다.

$$\hat{\theta}_{ppswor}=\frac{1}{M_0}\sum_{i=1}^{n}\frac{M_i\hat{\theta}_i}{\pi_i} \tag{3.6}$$

여기서 π_i는 조사단위 i가 표본에 포함되는 확률이다.

그리고 $\hat{\theta}_{ppswor}$의 분산과 분산추정량은 각각 다음과 같다.

$$\begin{aligned} V(\hat{\theta}_{ppswor}) &= \frac{1}{M_0^2}\sum_{i=1}^{N}\sum_{j>i}^{N}(\pi_i\pi_j-\pi_{ij})\left(\frac{M_i\theta_i}{\pi_i}-\frac{M_j\theta_j}{\pi_j}\right)^2 \\ &\quad+\frac{1}{M_0^2}\sum_{i=1}^{N}\frac{M_i^2}{\pi_i}\frac{1}{m_i}\left[\left(\frac{M_i-m_i}{M_i-1}\right)\theta_i(1-\theta_i)+\frac{p(1-p)}{(2p-1)^2}\right] \end{aligned} \tag{3.7}$$

$$\begin{aligned} \hat{V}(\hat{\theta}_{ppswor}) &= \frac{1}{M_0^2}\sum_{i=1}^{n}\sum_{j>i}^{n}\frac{\pi_i\pi_j-\pi_{ij}}{\pi_{ij}}\left(\frac{M_i\hat{\theta}_i}{\pi_i}-\frac{M_j\hat{\theta}_j}{\pi_j}\right)^2 \\ &\quad+\frac{1}{M_0^2}\sum_{i=1}^{n}\frac{M_i^2}{\pi_i}\frac{1}{m_i-1}\left[\left(\frac{M_i-m_i}{M_i-1}\right)\hat{\theta}_i(1-\hat{\theta}_i)+\frac{p(1-p)}{(2p-1)^2}\right] \end{aligned} \tag{3.8}$$

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

4. 효율성 비교

이 절에서는 등확률 2단계 추출에 의한 확률화응답기법에 대하여 살펴본 후, 이 기법과 확률비례추출에 의한 확률화응답기법과의 효율성을 비교해 보고자 한다.

4.1. 등확률 2단계 추출에 의한 확률화응답기법

우선, 각 집락의 크기가 $M_i(i=1,2,\cdots,N)$인 N개의 집락으로 구성되어 있는 모집단으로부터 n개의 집락을 단순임의복원추출한 후, 추출된 각 집락에서 다시 $m_i(i=1,2,\cdots,n)$개의 조사단위를 단순임의비복원추출하는 등확률 복원 2단계 추출에 확률화응답기법을 적용해 보고자 한다. 이러한 등확률 복원 2단계 추출절차에 의해 얻어진 민감한 그룹에 속하는 조사단위당 모비율 θ의 추정량 $\hat{\theta}_{wr}$는 다음과 같다.

$$\hat{\theta}_{wr} = \frac{N}{M_0 n}\sum_{i=1}^{n} M_i\hat{\theta}_i \tag{4.1}$$

그리고 $\hat{\theta}_{wr}$의 분산은 다음과 같다.

$$\begin{aligned} V(\hat{\theta}_{wr}) &= N^2\frac{1}{nM_0^2}\frac{1}{N-1}\sum_{i=1}^{N}(M_i\theta_i-\bar{M}\theta)^2 \\ &\quad +\frac{N}{nM_0^2}\sum_{i=1}^{N}M_i^2\frac{1}{m_i}\left[\left(\frac{M_i-m_i}{M_i-1}\right)\theta_i(1-\theta_i)+\frac{p(1-p)}{(2p-1)^2}\right] \end{aligned} \tag{4.2}$$

여기서 $\bar{M}=M_0/N$이다.

다음으로 n개의 1차 추출단위인 집락을 단순임의비복원추출한 후, 추출된 각 집락에서 다시 m_i개의 조사단위를 단순임의비복원추출하는 등확률 비복원 2단계 추출에 확률화응답기법을 적용해 보고자 한다.

이러한 등확률 비복원 2단계 추출절차에 의해 얻어진 민감한 그룹에 속하는 조사단위당 모비율 θ의 추정량 $\hat{\theta}_{wor}$은 다음과 같이 식 (4.1)과 동일한 형태를 취한다.

$$\hat{\theta}_{wor} = \frac{N}{M_0 n}\sum_{i=1}^{n} M_i\hat{\theta}_i \tag{4.3}$$

그리고 $\hat{\theta}_{wor}$의 분산은 다음과 같다.

$$\begin{aligned} V(\hat{\theta}_{\mathrm{wor}}) &= N^2\frac{N-n}{N}\frac{1}{nM_0^2}\frac{1}{N-1}\sum_{i=1}^{N}(M_i\theta_i-\bar{M}\theta)^2 \\ &\quad +\frac{N}{nM_0^2}\sum_{i=1}^{N}M_i^2\frac{1}{m_i}\left[\left(\frac{M_i-m_i}{M_i-1}\right)\theta_i(1-\theta_i)+\frac{p(1-p)}{(2p-1)^2}\right] \end{aligned} \tag{4.4}$$

4.2. 효율성 비교

등확률 복원 2단계 추출에 의한 분산 식 (4.2)와 확률비례복원추출에 의한 분산 식 (3.5)로부터 $N-1 \doteq N$일 때, 분산의 차이를 구해보면 다음과 같다.

$$\begin{aligned} V(\hat{\theta}_{wr}) - V(\hat{\theta}_{ppswr}) &= \frac{1}{nM_0\bar{M}} \left[\sum_{i=1}^{N} (M_i - \bar{M})^2 \theta_i^2 + \bar{M} \sum_{i=1}^{N} (M_i - \bar{M})(\theta_i^2 - \theta^2) \right. \\ &\quad + \sum_{i=1}^{N} (M_i - \bar{M})^2 \left[\left(\frac{M_i - m_i}{M_i - 1} \right) \theta_i(1-\theta_i) + \frac{p(1-p)}{(2p-1)^2} \right] \\ &\quad \left. + \bar{M} \left(\sum_{i=1}^{N} (M_i - \bar{M}) \left[\left(\frac{M_i - m_i}{M_i - 1} \right) \theta_i(1-\theta_i) + \frac{p(1-p)}{(2p-1)^2} \right] \right) \right] \end{aligned} \tag{4.5}$$

식 (4.5)에서 $M_i = \bar{M} = M_0/N$이면 $V(\hat{\theta}_{wr}) = V(\hat{\theta}_{ppswr})$이 된다. 즉, 각 집락의 크기가 동일하면 확률비례복원추출에 의한 확률화응답기법은 추출확률이 다같이 $1/N$이 되므로 등확률 복원 2단계 추출에 의한 확률화응답기법과 효율이 같다. 각 집락의 크기 M_i가 서로 크게 차이가 있으면 식(4.5)의 첫 번째 항 $\sum_{i=1}^{N}(M_i - \bar{M})^2\theta_i^2$은 커지게 되고, 두 번째 항 $\sum_{i=1}^{N}(M_i - \bar{M})(\theta_i^2 - \theta^2)$은 비교적 작은 값이 된다. 따라서, 일반적인 경우 M_i가 서로 같지 않을 때는 확률비례복원추출에 의한 확률화응답기법이 등확률 복원 2단계 추출에 의한 확률화응답기법보다 효율이 높다. 하지만 식 (4.5)에서도 볼 수 있듯이, 등확률 2단계 추출에 의한 확률화응답기법과 확률비례추출에 의한 확률화응답기법간의 효율을 이론적으로 비교하기가 쉽지 않으므로 수치적인 비교를 통하여 효율성을 비교해 보고자 한다. 이 때, Yates와 Grundy(1953)가 가정한 세 개의 모집단을 이용하고자 한다. 먼저, 각 집락의 크기가 $M_1 = 100$, $M_2 = 200$, $M_3 = 300$, $M_4 = 400$인 $N = 4$개의 집락으로부터 $n = 2$개의 집락을 비복원추출한 후, 추출된 각 집락에서 다시 m_i개의 조사단위를 비복원추출한다고 가정하자. 이 때, m_i의 크기는 $m_1 = 10$, $m_2 = 20$, $m_3 = 30$, $m_4 = 40$이라고 하고, 모집단은 다음 예와 같다고 하자.

표 4.1: 모집단의 예

	A				B				C			
M_i/M_0	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4	0.1	0.2	0.3	0.4
θ_i	0.05	0.06	0.07	0.08	0.08	0.07	0.06	0.05	0.02	0.03	0.03	0.02

$N = 4$개의 집락으로부터 $n = 2$개의 집락을 확률비례비복원추출할 경우, 포함확률 π_i와 π_{ij}값들을 구해보면 다음과 같다.

$\pi_1 = 0.235$	$\pi_2 = 0.441$	$\pi_3 = 0.609$	$\pi_4 = 0.715$		
$\pi_{12} = 0.047$	$\pi_{13} = 0.077$	$\pi_{14} = 0.111$	$\pi_{23} = 0.161$	$\pi_{24} = 0.233$	$\pi_{34} = 0.371$

민감한 질문이 선택될 확률 p를 변화시켜가면서 등확률 비복원 2단계 추출에 의한 추정량 $\hat{\theta}_{wor}$의 분산과 확률비례비복원추출에 의한 추정량 $\hat{\theta}_{ppswor}$의 분산을 이용하여 분산비를 구해보면 다음과 같다.

표 4.2: 효율성 비교

	$V(\hat{\theta}_{wor})/V(\hat{\theta}_{ppswor})$							
p	0.1	0.2	0.3	0.4	0.6	0.7	0.8	0.9
A	1.169	1.088	1.055	1.040	1.040	1.055	1.088	1.169
B	1.047	1.040	1.037	1.036	1.036	1.037	1.040	1.047
C	1.046	1.039	1.036	1.035	1.035	1.036	1.039	1.046

표 4.2에서 1보다 큰 값은 추정량 $\hat{\theta}_{ppswor}$이 $\hat{\theta}_{wor}$보다 효율적임을 나타낸다. 각 집락의 크기 M_i가 서로 크게 차이가 있도록 가정을 하였기 때문에, 등확률 비복원 2단계 추출에 의한 확률화응답기법보다 확률비례비복원추출에 의한 확률화응답기법이 효율성이 높게 나타남을 알 수 있다. 그리고, p값이 0.5보다 작을수록(또는 0.5보다 클수록) 효율이 증가하고 있으며, 각 집락의 크기에 따라 θ_i의 크기가 비례적으로 커지는 모집단 A가 다른 모집단의 경우보다 효율성이 높게 나타났다.

5. 확률비례추출에 의한 확률화응답기법의 실용화

이 절에서는 확률비례추출법에 의한 확률화응답기법을 이용하여 민감한 조사에 대한 실제조사를 실시하고자 한다. 이를 토대로 확률비례추출법에 의한 확률화응답기법의 실용화 타당성을 검토해 보고자 한다.

5.1. 확률비례추출법에 의한 확률화응답기법을 이용한 실제조사

이 조사는 응답자들이 직접질문을 받았을 경우에 응답하기 곤란하거나 거짓응답의 가능성이 많은 민감한 질문들에 대하여 간접조사기법인 확률화응답기법을 사용하여 민감한 속성에 대한 모비율을 추정하는 데 그 목적이 있다. 조사모집단은 자연계열, 인문사회계열, 예체능계열, 기타계열 즉 4개의 집락으로 구성된 우석대학교 재학생 6,000명(인문사회계열 : 2,400명, 자연계열 : 1,800명, 예체능계열 : 1,200명, 기타계열 : 600명)으로 정하였다. $N = 4$ 개의 집락으로부터 $n = 2$개의 집락을 확률비례추출한 결과 인문사회계열과 자연계열이 추출되었으며, 추출된 각 집락으로부터 각각 72명과 54명의 재학생들을 임의로 최종 추출하였다. 대학생들에게 민감한 조사내용으로는 류제복 외 2인(1995)이 확률화응답기법의 실용화 방안에서 이미 사용한 적이 있는 "당신은 학과에 있는 이성에게서 성적 충동을 느낀 적이 있습니까?", "당신은 결혼대상이 아닌 사람과 성관계를 가진 적이 있습니까?" , "당신은 남의 물건을 훔친 적이 있습니까?", "당신은 책값이라는 명목으로 받은 돈을 유흥비로 사용한 적이 있습니까?" 라는 설문 중에서 사전조사를 통해 여전히 민감도가 높게 나타난 "당신은 학과에 있는 이성에게서 성적 충동을 느낀 적이 있습니까?", "당신은 결혼대

상이 아닌 사람과 성관계를 가진 적이 있습니까?"라는 설문을 이용하여 조사를 실시하였다.

본 조사는 2005.3.14 – 2005.3.18 기간에 실시되었고, 조사원이 응답자들을 직접 찾아가서 조사하는 자계식 면접방법을 원칙으로 하였다. 조사에 앞서 조사원이 응답자들에게 확률화응답기법을 충분히 설명하도록 하였다. 조사에서 확률장치는 붉은 구슬과 노란 구슬이 들어 있는 주머니를 사용하였으며, 응답자들은 확률장치(구슬)에 의해 선택된 질문에 대하여 응답을 함으로써 조사자는 어떠한 질문에 응답하였는지를 알 수 없도록 하였다. 확률장치에서 붉은 구슬 즉 민감한 설문이 선택될 확률은 0.6으로 하였다. "당신은 학과에 있는 이성에게서 성적 충동을 느낀 적이 있습니까?"라는 질문에 대하여 민감한 속성을 가지고 있는 사람들의 모비율을 추정해 본 결과 0.323으로 나타났으며, 분산추정치는 0.091이었다. 또한 "당신은 결혼대상이 아닌 사람과 성관계를 가진 적이 있습니까?" 라는 질문에 대하여 민감한 속성을 가지고 있는 사람들의 모비율을 추정해 본 결과 0.072로 나타났으며, 분산추정치는 0.065로 구해졌다.

5.2. 확률비례추출법에 의한 확률화응답기법의 실용화 타당성

4.2절의 효율성 비교에서 살펴보았듯이 확률비례추출법에 의한 확률화응답기법은 모집단의 집락의 크기가 서로 차이가 클 때 효율적이었다. 이러한 사실을 토대로 확률비례추출법에 의한 확률화응답기법을 실제조사에 적용하여 타당성을 높이기 위해서는 다음과 같은 사항들을 반드시 고려해야 한다.

첫째, 조사하고자 하는 내용이 민감한가를 사전조사를 통해 검토한다. 조사하는 내용이 민감하지 않을 경우에는 확률화응답기법보다는 직접질문기법이 효율적이므로 반드시 사전 검토가 있어야 한다. 둘째, 모집단이 확률비례추출법에 의한 확률화응답기법을 사용하기에 적합한가를 파악한다. 이를 위해서는 모집단의 집락의 크기가 서로 차이가 큰가를 살펴보아야 한다. 셋째, 연구자는 조사원과 응답자들이 확률화응답기법을 충분히 이해하고 있는가를 살펴본다. 연구자는 확률화응답기법에는 어떠한 속임수가 없으며, 단지 조사원이 알지 못하게 확률장치를 통해 나온 질문에만 응답을 함으로써 응답자의 신분이 보호가 됨을 정확하게 인식시켜야 한다. 넷째, 여러 가지 조사방법을 고려하여 확률장치를 적절하게 선택한다. 마지막으로, 조사과정에서 오차들이 발생하지 않도록 노력을 기울인다. 다른 직접조사들과는 달리 확률화응답기법은 확률장치를 이용해야 하므로 이로 인해 또 다른 오차가 발생하지 않도록 주의를 기울여야 한다.

6. 결론

매우 민감한 조사에서 모집단이 집락의 크기가 서로 다른 여러 개의 집락으로 구성되어 있을 때, 집락의 크기에 비례하게 추출확률을 부여하는 확률비례추출법을 확률화응답기법에 적용하여, 민감한 속성에 대한 모비율의 추정량과 분산 및 분산추정량을 도출하였다. 그리고 제안한 확률비례추출에 의한 확률화응답기법과 등확률 2단계 추출에 의한 확률화응답기법과의 효율성을 민감한 속성이 선택될 확률 p의 변화에 따라 수치적으로 비교하였다.

그 결과 각 집락의 크기가 차이가 날 때 확률비례추출에 의한 확률화응답기법이 등확률 2단계 추출에 의한 확률화응답기법보다 더 효율적임을 알 수 있었다. 또한 p값이 0.5보다 작을수록(또는 0.5 보다 클수록) 확률비례비복원추출에 의한 확률화응답기법이 등확률 2단계 추출에 의한 확률화응답기법보다 더 효율이 좋게 나타났다. 제안한 확률비례비복원추출에 의한 확률화응답기법은 π_i와 π_{ij}를 계산하는데 다소 복잡하다라는 단점을 가지고 있기는 하지만, 매우 민감한 조사에서 집락의 크기가 서로 차이가 있을 경우 기존의 등확률 추출에 비하여 효율적이라는 장점을 가지고 있으므로 사회학, 경제학, 의학, 경영학 등 여러 분야의 연구조사에 제시한 방법을 유효 적절히 사용하여 관련분야의 연구발전에 활용할 수 있다.

참고문헌

류제복, 이계오, 이기성 (1995). 확률화응답기법의 실용화 방안, <응용통계연구>, **8**, 9-26.

류제복, 홍기학, 이기성 (1993). <확률화응답모형>, 자유아카데미, 서울.

박홍래 (1989). <통계조사론>, 영지문화사, 서울.

이기성, 홍기학 (1998). 2단계 집락추출법에 의한 확률화응답기법, <한국통계학회논문집>, **5**, 99-105.

이기성, 홍기학, 손창균, 정영미 (2003). The three-stage cluster randomized response model for obtaining sensitive information, <한국통계학회논문집>, **10**, 247-256.

Barth, J. T. and Sandler, H. M. (1976). Evaluation of the randomized response technique in a drinking survey, *Journal of Studies on Alcohol*, **37**, 690-693.

Begin, G., Boivin, M., and Bellerose, J. (1979). Sensitive data collection through the randomized response technique : some improvements, *The Journal of Psychology*, **101**, 53-65.

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

Cochran, W. G. (1977). *Sampling Techniques*, 3rd ed. John Wiley and Sons, New York.

Danermark, B. and Swensson, B. (1987). Measuring drug use among swedish adolescents : randomized response versus anonymous questionnaires, *Journal of Official Statistics*, **3**, 439-448.

Fox, J. A. and Tracy, P. E. (1986). *Randomized Response : A Method for Sensitive Survey*, Sage Publications.

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 Warde, W. D. (2004). A stratified warner's randomized response model, *Journal of Statistical Planning and Inference*, **120**, 155-165.

Liu, P. T. and Chow, L. P. (1976). A new discrete quantitative randomized model, *Journal of the American Statistical Association*, **71**, 72-73.

Scheer, N. J. and Dayton, C. M. (1987). Improved estimation of academic cheating behavior using the randomized response technique, *Research in Higher Education*, **26**, 61-69.

Tracy, P. E. and Fox, J. A. (1981). The validity of randomized response for sensitive measurements, *American Sociological Review*, **46**, 187-200.

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

[2005년 5월 접수, 2005년 8월 채택]

A Study on the Randomized Response Technique by PPS Sampling*

GiSung, Lee[1)]

ABSTRACT

In this study, we make an effort to find a method to acquire sensitive information when sensitive populations are consisted of several clusters that vary in size. We suggest and systemize the theoretical validity for applying RRT(Randomized Response Technique) to PPS(Probability Proportional to Size) sampling method and derive the estimate and it's variance of the proportion of sensitive characteristic of population by using the suggested method. We compare the efficiency of the suggested technique by two-stage equal probability sampling. We examine practical aspects of the suggested method of RRT by PPS sampling through field survey.

Keywords: Sensitive information, Unequal size clusters, PPS, Warner's technique, Field survey.

* This work was supported by Korea Research Foundation Grant funded by the Korean Government (MOEHRD)(R05-2004-000-10160-0).

1) Professor, School of Computer, Woosuk University, 490 Hujeong-ri, Wanju-gun, Jeonbuk, 565-701, Korea.
E-mail : gisung@woosuk.ac.kr