The Korean Communications
in Statistics Vol. 5, No. 1, 1998
pp. 99-105

# 2단계 집락추출법에 의한 확률화응답모형

이기성[1] 홍기학[2]

## 요 약

본 논문에서는 매우 민감한 조사에서 모집단이 여러 개의 집락으로 구성되어 있을 때, 모집단으로부터 집락을 단순임의추출한 후 추출된 각 집락에서 다시 조사단위의 표본을 추출하는 2단계 집락추출법에 확률화응답모형을 적용하였다. 그리고, 일정한 비용 하에서 분산을 최소로 하는 1단계 집락의 수와 2단계 집락에서 추출된 조사단위의 수의 최적값을 구하여 최소분산의 형태를 도출하였다.

## 1. 서 론

사회 여러 분야의 조사에서 응답자들은 민감하거나 개인적인 이해와 관계되는 질문을 받았을 경우 응답을 회피하거나 고의적인 거짓응답을 하게 된다.

이에 Warner(1965)는 확률장치를 통한 간접응답으로 응답자의 신분이나 비밀을 노출시키지 않고서 민감한 질문에 대해 정보를 이끌어 낼 수 있는 확률화응답모형(randomized response model ; RRM)을 처음으로 제시하였다. 그 후 수많은 학자들에 의해 이에 대한 연구가 확대되고 발전되었고, Fox와 Tracy(1986), Chaudhuri와 Mukerjee(1988), 그리고 류제복, 홍기학과 이기성(1993)들이 확률화응답기법들을 정리 요약하여 체계화시켰다. 그러나, 이러한 확률화응답기법들은 직접질문을 사용하는 경우보다 시간, 비용과 노력을 더 필요로 하는 단점을 가지고 있으며, 특히 모집단이 큰 경우에 단순임의추출법을 이용하여 응답자들을 추출하여 조사를 하는 데에는 여러 가지 어려움이 따르게 된다.

본 논문에서는 모집단이 여러 개의 집락으로 구성되어 있을 때 사용하는 집락추출법에 확률화응답모형을 적용해 보고자 한다. 2장에서는 모집단으로부터 집락을 단순임의비복원추출한 후, 추출된 각 집락에서 다시 조사단위의 표본을 단순임의복원추출하는 2단계 집락추출법에 확률화응답모형을 적용하였다. 3장에서는 일정한 비용 하에서 분산을 최소로 하는 1단계 집락의 수와 2단계 집락에서 추출된 조사단위의 수의 최적값을 구하였다.

## 2. 2단계 집락추출법에 의한 확률화응답모형

1) (565-701) 전북 완주군 삼례읍 후정리 490 우석대학교 전산통계학과 조교수
2) (520-714) 전남 나주시 대호동 252 동신대학교 컴퓨터학과 부교수

각 집락의 크기가 $M_i(i=1,2,\cdots,N)$인 $N$개의 집락으로 구성되어 있는 모집단으로부터 $n$개의 집락을 단순임의비복원추출한 후, 추출된 각 집락에서 다시 $m_i(i=1,2,\cdots,n)$개의 조사단위의 표본을 단순임의복원추출하는 2단계 집락추출법에 Warner모형을 적용해 보고자 한다.

2단계 집락추출법에 의해 추출된 응답자들은 다음과 같은 두 개의 설문으로 구성되어 있는 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.

$$\lambda_i = (2p-1)\pi_i + (1-p)\ . \tag{2.1}$$

여기서, $\pi_i$는 $i$번째 집락에서의 민감한 그룹에 속하는 비율이다.

$i$번째 집락에서의 $j$번째 조사단위를 $z_{ij}$라 하고, 다음과 같이 정의하자.

$$z_{ij} = \begin{cases} 1 : i\text{번째 집락에서 } j\text{번째 응답자가 "예"라고 응답하면} \\ 0 : i\text{번째 집락에서 } j\text{번째 응답자가 "아니오"라고 응답하면} \end{cases}$$

2단계 집락추출법에 의해 각 집락으로부터 표본으로 추출된 $m_i$명의 응답자 중에서 "예"라고 응답한 사람의 수를 $Z_i = \sum_{j=1}^{m_i} z_{ij}$라 하면, $Z_i$는 $b(m_i, \lambda_i)$를 따른다.

식(2.1)에서 $\lambda_i$의 추정치는 $\hat{\lambda}_i = \dfrac{Z_i}{m_i}$가 되므로, $\pi_i$의 추정량 $\hat{\pi}_i$는 다음과 같다.

$$\hat{\pi}_i = \frac{\hat{\lambda}_i - (1-p)}{2p-1}\ . \tag{2.2}$$

그리고, 민감한 그룹에 속하는 조사단위당 모비율 $\pi$는 다음과 같다.

$$\pi = \frac{1}{N}\sum_{i=1}^{N}\pi_i\ . \tag{2.3}$$

단순임의비복원추출된 $i(i=1,2,\cdots,n)$번째 집락으로부터 추출된 응답자들이 민감한 그룹에

속하는 조사단위당 모비율 $\pi$의 추정량 $\hat{\pi}_{cl}$는 다음과 같다.

$$\hat{\pi}_{cl} = \frac{1}{n}\sum_{i=1}^{n}\hat{\pi}_i \tag{2.4}$$

**<정리 1>** 추정량 $\hat{\pi}_{cl}$는 모비율 $\pi$의 불편추정량이다.

**(증명)**

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

■

**<정리 2>** 각 집락의 크기가 $M_i$인 $N$개의 집락에서 $n$개의 집락을 단순임의비복원추출하고, 추출된 집락에서 다시 $m_i$개의 조사단위의 표본을 단순임의복원추출한다. 이러한 2단계 집락추출법에 의한 확률화응답모형에서 민감한 그룹에 속하는 모비율 $\pi$의 추정량 $\hat{\pi}_{cl}$의 분산은 다음과 같다.

$$V(\hat{\pi}_{cl}) = \frac{N-n}{nN(N-1)}\sum_{i=1}^{N}(\pi_i-\pi)^2 + \frac{1}{nN}\sum_{i=1}^{N}\frac{1}{m_i}\left[\pi_i(1-\pi_i) + \frac{p(1-p)}{(2p-1)^2}\right]. \tag{2.5}$$

**(증명)**

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

에서

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

이고,

$$
\begin{aligned}
E_1 V_2(\hat{\pi}_{cl}) &= E_1 V_2\left[\frac{1}{n}\sum_{i=1}^{n}\hat{\pi}_i\right] \\
&= \frac{1}{n^2}\frac{n}{N}\sum_{i=1}^{N} V_2(\hat{\pi}_i) \\
&= \frac{1}{nN}\sum_{i=1}^{N}\frac{1}{m_i}\left[\pi_i(1-\pi_i) + \frac{p(1-p)}{(2p-1)^2}\right]
\end{aligned}
$$

이므로, 추정량 $\hat{\pi}_{cl}$의 분산은 식(2.5)와 같다.

■

또한, 집락의 크기가 $M$으로 일정한 $N$개의 집락에서 $n$개의 집락을 단순임의복원추출할 때, 모비율 $\pi$의 추정량 $\hat{\pi}_{cl}$의 분산은 식(2.6)과 같다.

$$
V(\hat{\pi}_{cl}) = \frac{1}{nN}\sum_{i=1}^{N}(\pi_i-\pi)^2 + \frac{1}{nN}\sum_{i=1}^{N}\frac{1}{m_i}\left[\pi_i(1-\pi_i) + \frac{p(1-p)}{(2p-1)^2}\right]. \tag{2.6}
$$

만약 각 집락으로부터 표본으로 추출된 $m_i$가 $m$으로 일정하다면, 식(2.6)의 분산식은 다음과 같이 표현될 수 있다.

$$
V(\hat{\pi}_{cl}) = \frac{1}{nN}\sum_{i=1}^{N}(\pi_i-\pi)^2 + \frac{1}{nmN}\sum_{i=1}^{N}\left[\pi_i(1-\pi_i) + \frac{p(1-p)}{(2p-1)^2}\right]. \tag{2.7}
$$

## 3. 2단계 집락추출법에 의한 확률화응답모형에서의 $n$과 $m$의 최적값

식(2.7)로부터 1단계 집락의 수 $n$과 각 집락에서 추출된 조사단위의 수 $m$을 증가시키면 분산은 감소하지만 $n$과 $m$의 증가에 따라 조사비용은 증가하게 된다. 표본의 최적배분을 위해 일정한 비용 하에서 표본의 정도를 최대로 하는 $n$과 $m$의 값을 결정해 보고자 한다.

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

$$
C = c_0 + nc_1 + nmc_2 . \tag{3.1}
$$

여기서, $C$는 총비용이고, $c_0$는 고정비용으로 조사행정비, 표본설계비용 등을 포함하며 표본의 크기와는 관계없이 소요되는 비용이다. $c_1$은 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. $c_2$는 표본 2차 추출단위 당 비용

으로 조사단위 당 소요비용을 의미하며, 표본 2차 추출단위의 추출 및 확인에 소요되는 비용, 확률장치를 이용한 면접 또는 실측비용, 조사자료의 집계분석비용 등을 포함한다.

분산 식 (2.7)에서

$$S_1^{\,2} = \frac{1}{N}\sum_{i=1}^{N}(\pi_i - \pi)^2 ,$$

$$S_2^{\,2} = \frac{1}{N}\sum_{i=1}^{N}\left[\pi_i(1-\pi_i) + \frac{p(1-p)}{(2p-1)^2}\right]$$

로 두면, $V(\hat{\pi}_{cl})$는 다음과 같이 표현된다.

$$V(\hat{\pi}_{cl}) = \frac{S_1^{\,2}}{n} + \frac{S_2^{\,2}}{nm} . \tag{3.2}$$

일정한 비용 하에서 분산을 최소로 하는 $n$과 $m$의 값을 식(3.1)의 비용함수와 분산 식(3.2)를 이용하여 구해 보기로 하자.

$n$과 $m$의 최적값을 구하기 위하여 Lagrange 승수법을 이용하기로 하자. 이 때, 최소화하는 함수 $\varnothing$는 다음과 같이 표현된다.

$$\varnothing = \frac{S_1^{\,2}}{n} + \frac{S_2^{\,2}}{nm} + \lambda(nc_1 + nmc_2 - c_0) . \tag{3.3}$$

식(3.3)을 $n$과 $m$에 대하여 편미분하면 식(3.4)와 식(3.5)를 얻을 수 있다.

$$\frac{\partial \varnothing}{\partial n} = -\frac{S_1^{\,2}}{n^2} - \frac{S_2^{\,2}}{n^2 m} + \lambda(c_1 + mc_2) . \tag{3.4}$$

$$\frac{\partial \varnothing}{\partial m} = -\frac{S_2^{\,2}}{nm^2} + \lambda nc_2 . \tag{3.5}$$

식(3.4)와 식(3.5)를 0으로 놓고, 두 식으로부터 $\lambda$를 구하면 다음과 같다.

$$\lambda = \frac{S_1^{\,2}}{n^2 c_1} . \tag{3.6}$$

그리고, 식(3.5)를 0으로 놓고 $\lambda$를 구하면

$$\lambda = \frac{S_2^{\,2}}{n^2 m^2 c_2} \tag{3.7}$$

이므로 식(3.6)과 식(3.7)으로부터 다음을 구할 수 있다.

$$m^2 = \frac{S_2^{\ 2}}{S_1^{\ 2}} \frac{c_1}{c_2} .$$

따라서, 구하고자 하는 $m$의 최적값 $m_0$는 다음과 같다.

$$m_0 = \sqrt{\frac{S_2^{\ 2}}{S_1^{\ 2}} \frac{c_1}{c_2}} . \tag{3.8}$$

또한, 비용함수 식(3.1)을 $n$의 함수로 표현해 보면

$$n = \frac{C - c_0}{c_1 + mc_2} . \tag{3.9}$$

이므로, 식(3.8)의 $m_0$값을 식(3.9)에 대입하여 $n$의 최적값 $n_0$를 구하면 다음과 같다.

$$\begin{aligned} n_0 &= \frac{C - c_0}{c_1 + m_0 c_2} \\ &= (C - c_0) \frac{\sqrt{\dfrac{S_1^{\ 2}}{c_1}}}{\sqrt{S_1^{\ 2} c_1} + \sqrt{S_2^{\ 2} c_2}} . \end{aligned} \tag{3.10}$$

따라서, 식(3.8)과 식(3.10)에서 구한 $m_0$와 $n_0$의 값을 식(3.2)에 대입하여 최소분산 $V_{\min}(\hat{\pi}_{cl})$를 다음과 같이 얻을 수 있다.

$$\begin{aligned} V_{\min}(\hat{\pi}_{cl}) &= \frac{1}{n_0}\left( S_1^{\ 2} + \frac{S_2^{\ 2}}{m_0} \right) \\ &= \frac{c_1 + m_0 c_2}{C - c_0}\left( S_1^{\ 2} + \frac{S_2^{\ 2}}{m_0} \right) \\ &= \frac{1}{C - c_0}\left( \sqrt{S_1^{\ 2} c_1} + \sqrt{S_2^{\ 2} c_2} \right)^2 . \end{aligned} \tag{3.11}$$

## 4. 결 론

본 논문에서는 모집단이 여러 개의 집락으로 구성되어 있을 때, 모집단으로부터 집락을 단순임의비복원추출한 후, 추출된 각 집락에서 다시 조사단위의 표본을 단순임의복원추출하는 2단계 집락추출법에 확률화응답모형을 새로이 적용하였다. 또한, 일정한 비용 하에서 분산을 최소로 하는 1단계 집락의 수와 2단계 집락에서 추출된 조사단위의 수의 최적값을 구하여 최소분산의 형태를 도출하였다.

2단계 집락추출법에 의한 확률화응답모형은 단순임의추출법에 의한 확률화응답모형과 단순집락추출법에 의한 확률화응답모형보다 분산이 증가할 수 있지만, 실제조사에서 많이 이용될 수 있는 장점을 가지고 있다. 또한, 이러한 2단계 집락추출법에 의한 확률화응답모형은 다단계 집락추출법에 의한 확률화응답모형으로 확장가능 하기 때문에 더욱 실용적이라 할 수 있다.

## 참고문헌

[1] 류 제복, 홍 기학, 이 기성(1993). 「확률화응답모형」, 자유아카데미, 서울.

[2] 박 홍래(1989). 「통계조사론」, 영지문화사, 서울.

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

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

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

[6] Warner, S. L.(1965). Randomized Response ; A Survey Technique for Eliminating Evasive Answer Bias, *Journal of the American Statistical Association*, 60, 63-69.