민감한 정보를 얻기 위한 대체 전략에 관한 연구

홍기학, 이기성, 손창균[1)]

요 약

Hansen과 Hurwitz(1946)는 우편조사에서의 무응답 문제를 처리하는 방법으로 표본을 응답결과에 따라 응답층과 무응답층으로 나눈 다음, 무응답층의 일부를 랜덤 추출하여 면대면 직접조사에 의해 무응답층의 정보를 얻는 방법을 제안하였다. 본 연구에서는 민감한 모집단에 대한 자료수집 방법으로 직접질문 방법인 Black-Box 방법과 간접질문 방법인 확률화응답기법(RRT)의 결합적 방법을 제시하였고, 층화이중 추출방법을 이용하여 모수를 추정하였다.

주요용어 : 무응답, 민감한 정보, Black-Box 방법, 확률화응답기법, 층화이중추출

1. 서론

사회가 복잡하고 다양해짐에 따라 신속하고 보다 정확한 정보가 요구되고 이를 충족시키기 위하여 표본조사의 필요성이 점차 증대되고 있다. 사회 여러 분야의 표본조사에서 발생하는 오차에는 표본오차와 비표본오차가 있으며, 최근에 연구의 관심은 비표본오차를 줄이는 데 있다. 이러한 비표본오차는 응답자들이 민감하거나 개인적인 이해와 관계되는 질문을 받았을 경우 더욱 증가하게 된다. 예를 들어 음주운전, 낙태경험, 환각제사용, 동성연애 및 탈세여부 등과 같은 사회적으로나 개인적으로 매우 민감한 문제에 관한 조사에서 기존의 직접질문방식을 그대로 사용할 경우 응답자들이 응답을 회피하거나 거짓으로 응답하는 경향이 뚜렷이 나타나게 된다. 이는 응답자들이 민감한 질문에 응답함으로써 불이익을 받거나 사생활이 보장되지 않는다고 생각하기 때문이다. 이 같은 문제점을 해결하고 사생활을 보장해주기 위한 대표적인 조사 방법으로 크게 두 가지를 들 수 있다. 첫째는 간접질문 조사 방법인 확률화응답기법(randomized response technique : RRT)이며, 둘째로는 직접질문 조사 방법인 무기명직접질문조사방법인데 일명 Black-Box(BB)방법이라고 불린다.

1965년 Warner에 의해 처음 제시된 확률화응답기법은 응답자의 신분이나 비밀을 노출시키지 않고 민감한 질문에 대한 정보를 이끌어 내기 위하여 응답자들에게 확률장치를 통한 간접 응답을 하게 함으로써 그들의 익명성을 보장해 주면서 조사자가 얻고자하는 민감한 정보를 최대한 얻을 수 있도록 한 방법이다.

확률화응답기법의 장점이 응답자의 익명성보장에 있다면, 단점은 추정량의 효율성이 익명성을 강조할수록 직접조사의 것에 비해 떨어진다는 것이다.

BB방법은 설문조사에서 응답자들이 설문에 직접 체크를 한다는 면에서 일반 설문조사 방법과 같으나, 응답자들의 신분을 밝히지 않고 무기명으로 조사된다는 점에서 다르다. 즉, BB방법은 응답자들이 주어진 설문에 무기명으로 답한 다음 그것을 조사자에게 주는 것이 아니고 따로 설치된 밀폐된 상자 속에 집어넣고, 조사자는 최종적으로 상자 속에 수집된 질문지만을 회수함으로써 설문의 답을 누가 했는지 알 수가 없다. 따라서 응답자는 자신의 프라이버시를 보장받

1) 동신대학교 컴퓨터학과
2) 우석대학교 전산통계학과
3) 동신대학교 교양교직학부

을 수 있는 조사방법이다.

한편 Hansen과 Hurwitz(1946)는 우편조사에서의 무응답 문제를 처리하는 방법으로 표본을 응답 결과에 따라 응답층과 무응답층으로 나눈 다음, 무응답층의 일부를 랜덤 추출하여 면대면 직접조사에 의해 무응답층의 정보를 얻는 방법을 제안하였다.

본 논문에서는 모집단의 민감한 모수를 추정하기 위한 자료수집 방법으로 직접질문 방법인 Black-Box 방법과 간접질문 방법인 확률화응답기법(RRT)의 결합적 방법을 제시하였고, 층화이중추출방법을 이용하여 모수를 추정하였다. 또한 주어진 추정량의 효율성을 2단계 확률화 응답기법인 Mangat과 Singh 모형의 추정량과 비교 분석하였다.

2. 대체 전략

본 장에서는 모집단의 민감한 정보에 대한 자료 수집 방법으로서 직접질문 조사 방법인 BB 방법과 간접 응답기법인 RRT를 결합한 방법을 제시하고, Hansen과 Hurwitz(1946)추정 방법을 이용하여 민감한 모수를 추정하고자 한다.

크기가 N인 모집단으로부터 단순임의복원추출된(SRSWR) n'명을 대상으로 먼저 BB방법을 이용한 직접조사를 실시하여 조사에 응한 그룹과 응하지 않은 그룹으로 나눈다. 이때 n_1' 와 n_2'를 각각 조사에 응한 그룹과 응하지 않은 그룹에 속하는 사람들의 수라고 놓으면 $n' = n_1' + n_2'$ 이 된다. 다음으로 조사에 응하지 않은 그룹에 속하는 사람들 중 n_2명을 $(n_2 = kn_2', 0 < k \leq 1)$을 단순임의복원추출하여 이들로부터 Warner의 확률화응답기법을 이용한 간접조사 방법에 의해 민감한 모수에 대한 정보를 얻는다.

n_{1r}'이 BB방법을 이용한 직접조사에서 설문에 "예"라고 응답한 사람들의 수이고, n_{2r}이 Warner의 확률화응답기법을 이용한 간접조사에서 "예"라고 응답한 사람들의 수라고 하면 Hansen과 Hurwitz방법을 이용하여 다음과 같은 모수에 관한 결합 추정량을 얻을 수 있다.

$$\hat{\pi}_d = w_1\hat{\pi}'_1 + w_2\hat{\pi}_2 \tag{1}$$

위 식에서

$$w_1 = \frac{n_1'}{n'}\ ,\ w_2 = \frac{n_2'}{n'}$$

이고

$$\hat{\pi}'_1 = \frac{n'_{1r}}{n_1'}\ ,\ \hat{\pi}_2 = \frac{n_{2r}}{n_2}$$

이다.

<정리 1> $\hat{\pi}_d = w_1\hat{\pi}'_1 + w_2\hat{\pi}_2$ 는 민감한 모수 π의 비편향 추정량이다.

(증명)

$$
\begin{aligned}
E(\hat{\pi}_d) &= E_1E_2(\hat{\pi}_d) \\
&= E[E(w_1\hat{\pi}'_1 + w_2\hat{\pi}_2 \mid w)] \\
&= E(w_1\pi_1 + w_2\pi_2) \\
&= W_1\pi_1 + W_2\pi_2 \\
&= \pi
\end{aligned}
$$

■

<정리 2> 모집단으로부터 표본을 모두 SRSWR로 뽑았다고 가정할 경우 추정량 $\hat{\pi}_d$의 분산은 다음과 같다.

$$V(\hat{\pi}_d) = \frac{\pi_A(1-\pi_A)}{n'}\left(1 + W_2(\frac{1}{k} - 1)\right) + W_2(\frac{1}{k} - 1)\frac{p(1-p)}{n'(2p-1)^2} \tag{2}$$

(증명)

제시한 결합 추정량은 Cochran(1977)의 층화이중추출에 의한 층화추정량의 성질을 이용하여 다음과 같이 표현할 수 있다.

$$
\begin{aligned}
\hat{\pi}_d &= w_1\hat{\pi}'_1 + w_2\hat{\pi}_2 = w_1\frac{n'_{1r}}{n'_1} + w_2\frac{n'_{2r}}{n'_2} + w_2\left(\frac{n_{2r}}{n_2} - \frac{n'_{2r}}{n'_2}\right) \\
&= \frac{n'_r}{n'} + w_2\left(\frac{n_{2r}}{n_2} - \frac{n'_{2r}}{n'_2}\right) \\
&= \hat{\pi} + w_2\left(\frac{n_{2r}}{n_2} - \frac{n'_{2r}}{n'_2}\right)
\end{aligned}
$$

위 식에서 $n'_r = n'_{1r} + n'_{2r}$는 표본으로 뽑힌 모든 사람들이 BB방법에 의한 조사에 응했다고 가정할 경우 민감한 질문이 적힌 설문에 "예"라고 표시한 사람들의 수이다. 따라서 $\hat{\pi}$에 대한 분산은 모집단의 민감한 속성에 대한 직접질문에 의한 분산과 같다.

$$V(\hat{\pi}) = \frac{\pi_A(1-\pi_A)}{n'} \tag{3}$$

고정된 w_2에 대하여

$$
\begin{aligned}
V_2\left(w_2(\frac{n_{2r}}{n_2} - \frac{n'_{2r}}{n'_2})\right) &= w_2^2\left(V_2(\frac{n_{2r}}{n_2}) - V_2(\frac{n'_{2r}}{n'_2})\right) \\
&= w_2^2\left(\frac{1}{n_2} - \frac{1}{n'_2}\right)\left(\pi_A(1-\pi_A) + \frac{p(1-p)}{(2p-1)^2}\right) \qquad (4) \\
&= w_2\frac{1}{n'}\left(\frac{1}{k} - 1\right)\left(\pi_A(1-\pi_A) + \frac{p(1-p)}{(2p-1)^2}\right)
\end{aligned}
$$

가 되고

모든 w_2의 분포에 대한 평균값을 취하면

$$E_1V_2\left(w_2(\frac{n_{2r}}{n_2} - \frac{n'_{2r}}{n'_2})\right) = W_2\frac{1}{n'}\left(\frac{1}{k} - 1\right)\left(\pi_A(1-\pi_A) + \frac{p(1-p)}{(2p-1)^2}\right) \tag{5}$$

이 된다.
따라서 식(3)과 (5)를 더해서 정리하면 식(2)를 얻을 수 있다.

■

식(2)에서 $k=1$이면,

$$V(\hat{\pi}_d) = \frac{\pi_A(1-\pi_A)}{n'} \tag{6}$$

이 된다.

3. 효율성 비교

본 장에서는 제시한 대체 전략과 Mangat-Singh 모형과의 효율성을 분산 측면에서 비교해 보고자 한다.
Mangat-Singh 모형의 추정량 및 분산은 다음과 같다.

$$\hat{\pi}_M = \frac{n'/n-(1-T)(1-p)}{2p-1+2T(1-p)} \tag{7}$$

$$V(\hat{\pi}_M) = \frac{\pi_A(1-\pi_A)}{n} + \frac{(1-T)(1-p)1-(1-T)(1-p)}{n(2p-1+2T(1-p))^2} \tag{8}$$

식(8)에서 $T=1$이면,

$$V(\hat{\pi}_M) = \frac{\pi_A(1-\pi_A)}{n} \tag{9}$$

이 된다.

$n=n'$ 이라고 가정할 때, 식(2)와 (8)로부터, $T=1$, $k=1$이면 두 방법의 분산은 일치한다. 이는 두 방법이 모두 민감한 모수에 대한 자료수집 방법으로 확률화응답기법과 같은 간접조사 방법이 아닌 직접조사 방법에 의해 자료를 수집함을 알 수 있다.
한편 $T \neq 1$ $k \neq 1$ 일 경우

$$\begin{aligned} V(\hat{\pi}_M) - V(\hat{\pi}_d) &= \frac{\pi_A(1-\pi_A)}{n} + \frac{(1-T)(1-p)(1-(1-T)(1-p))}{n(2p-1+2T(1-p))^2} - \\ &\quad \left[\frac{\pi_A(1-\pi_A)}{n'}\left(1+W_2(\frac{1}{k}-1)\right)+W_2(\frac{1}{k}-1)\frac{p(1-p)}{n'(2p-1)^2}\right] \\ &= \frac{1}{n}\left[\frac{(1-T)(1-p)(1-(1-T)(1-p))}{(2p-1+2T(1-p))^2}\right. \\ &\quad \left. - W_2(\frac{1}{k}-1)\left(\pi_A(1-\pi_A)+\frac{p(1-p)}{(2p-1)^2}\right)\right] \\ &\geqq 0 \end{aligned} \tag{10}$$

을 만족하는 W_2 와 T 그리고 k의 값을 구해야 한다.

식 (10)의 우변 두 번째 항의 값은 $k(0 < k \leq 1)$가 증가할수록, $W_2(0 \leq W_2 \leq 1)$가 감소할수록 작아짐을 알 수 있다.

k와 T값을 같게 놓고 효율성을 비교해본 결과 W_2의 값이 작을수록 Mangat-Singh 모형에 대하여 제시한 방법의 효율성이 높아짐을 알 수 있다.

참고문헌

[1] Cochran, W.G.(1977), *Sampling Technique*, 3rd edition, John Wiley& Sons, New York.

[2] Hansen, M.H. and Hurwitz, W.N. (1946), The problem of non-response in sample surveys, *Journal of the American Statistical Association*, vol.41, 517-529.

[3] Mangat, N.S. and Singh, R. (1990), An alternative randomized response procedure, *Biometrika*, vol.77, no.2, 439-442.