응용통계연구 제 17권 2호, 2004년, pp. 303-310

층화이중추출을 이용한 결합 확률화응답기법 *

홍기학 1)

요 약

본 연구에서는 민감한 모집단에 대한 자료수집 방법으로 직접질문 방법인 Black-Box 방법과 간접질문 방법인 확률화응답기법(RRT)의 결합적 방법을 제시하였고, 층화이중추출방법을 이용하여 모수를 추정하였다. 또한, 주어진 추정량의 효율성을 Mangat과 Singh 추정량과 비교 분석하였다.

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

1. 서론

사회가 복잡하고 다양해짐에 따라 신속하고 보다 정확한 정보가 요구되고 있으며, 이를 충족시키기 위하여 표본조사의 필요성이 점차 증대되고 있다. 사회 여러 분야의 표본조사에서 발생하는 오차에는 표본오차와 비표본오차가 있으며, 최근에 연구의 관심은 비표본오차를 줄이는 데 있다. 응답자들이 민감하거나 개인적인 이해와 관계되는 질문을 받았을 경우 이러한 비표본오차는 더욱 증가하게 된다. 예를 들어 음주운전, 낙태경험, 환각제 사용, 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. 이같은 문제점을 해결하고 사생활을 보장해주기 위한 대표적인 조사 방법으로 크게 두 가지를 들 수 있다. 첫째는 간접질문 조사방법인 확률화응답기법(randomized response technique : RRT)이며, 둘째로는 직접질문 조사방법인 무기명 직접질문 조사방법인데 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 익명성 보장에 있다면, 단점은 추정량의 효율성이 익명성을 강조할수록 직접조사의 것에 비해 떨어진다는 것이다. 이러한 확률화응답기법은 Warner 이후 많은 학자들이 연구를 해왔으며, 특히 Mangat과 Singh(1990)은 Warner 모형의 효율성을 개선하기 위한 한 방법으로 두개의 확률장치를 통해 응답을 얻는 방법을 제시하였다. BB방법은 설문조사에서 응답

* 본 연구는 동신대학교 교내 학술연구지원에 의해 이루어 졌음.

1) (520-714) 전남 나주시 대호동 252, 동신대학교 컴퓨터학과, 교수
E-mail:khhong@dsu.ac.kr

자들이 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 보장받을 수 있는 조사방법이다. Mangat-Singh방법의 경우 Warner 방법에 비하여 효율성은 높지만, 확률장치를 두 번 사용함으로써 조사 대상자들에게 불편을 주고 조사시간이 길어지는 단점이 있다. 이를 보완하고자 하는 방법이 본 논문에서 제시하고자하는 BB 방법과 RRT를 결합시킨 조사 방법이다. 본 논문에서는 모집단의 민감한 모수에 대한 정보를 수집하는 방법으로 직접질문 방법인 BB방법과 간접질문 방법인 RRT의 결합적 방법을 제시하고, 층화이중추출방법을 이용하여 모수를 추정하고자 한다. 또한 주어진 추정량의 효율성을 Mangat과 Singh 추정량과 비교 분석하고자 한다.

2. Mangat-Singh 모형

Mangat과 Singh(1990)은 Warner 모형의 효율성을 개선하기 위한 한 방법으로 두개의 확률장치를 통해 응답을 얻는 방법을 제시하였다. 즉, 모집단으로부터 복원으로 뽑은 n'명의 조사대상자들은 다음과 같은 구조로 이루어진 두 개의 확률장치를 이용해서 조사자에게 "예" 또는 "아니오"로 응답을 한다.

	질 문	질문을 선택할 확률
확률장치 1	나는 민감한 그룹에 속한다	T
	확률장치 2로 가시오	$1-T$

	질 문	질문을 선택할 확률
확률장치 2	나는 민감한 그룹에 속한다.	p
	나는 민감한 그룹에 속하지 않는다.	$1-p$

조사에서 "예"라고 응답한 응답자의 수를 n'_r라고 하면, 민감한 모수에 대한 추정량 및 분산은 다음과 같이 구해진다.

$$\hat{\pi}_M = \frac{n'_r/n - (1-T)(1-p)}{2p-1+2T(1-p)}, \tag{2.1}$$

$$V(\hat{\pi}_M) = \frac{\pi_A(1-\pi_A)}{n'} + \frac{(1-T)(1-p)(1-(1-T)(1-p))}{n'(2p-1+2T(1-p))^2} \tag{2.2}$$

여기서, π_A는 민감한 그룹에 속하는 모비율이다.

식(2.2)에서 $T=1$이면,

$$V(\hat{\pi}_M) = \frac{\pi_A(1-\pi_A)}{n'} \tag{2.3}$$

이 된다.

또한

$$T > \frac{(1-2p)}{(1-p)} \tag{2.4}$$

일 때 Warner 모형에 비하여 분산 측면에서 더 효율적임을 보이고 있다.

3. 결합 확률화응답기법

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 제안한 바 있다. 본 장에서는 Hansen과 Hurwitz의 추정 방법을 응용하여, 모집단의 민감한 정보에 대한 자료 수집 방법으로서 직접질문 조사 방법인 BB방법과 간접 응답기법인 RRT를 결합한 방법을 제시하고, 민감한 모수를 추정하고자 한다. 크기가 N인 모집단으로부터 단순임의복원추출된(SRSWR) n'명을 대상으로 BB방법에 대하여 설명한 다음, BB방법에 동의하는 사람들 (n_1')은 BB방법에 의해 조사하고, 동의하지 않는 사람들(n_2')은 BB방법에 대한 무응답으로 간주하여 그들 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, 민감한 모수에 대한 추정량을 다음과 같이 결합 추정량으로 나타낼 수 있다. 이때, W_h와 w_h $(h = 1, 2)$는 각각 BB방법과 RRT를 선택하는 모집단비율과 그에 대응하는 1차 표본비율(first sample proportion)을 나타낸다.

$$\hat{\pi}_H = w_1\hat{\pi}_1' + w_2\hat{\pi}_2 \tag{3.1}$$

위 식(3.1)에서

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

이고,

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

이다.

정리 3.1 $\hat{\pi}_H = w_1\hat{\pi}_1' + w_2\hat{\pi}_2$는 민감한 모수 π의 비편향추정량이다.

증명:

$$\begin{aligned} E(\hat{\pi}_H) &= E_1 E_2(\hat{\pi}_H) \\ &= E\left[E(w_1\hat{\pi}_1' + w_2\hat{\pi}_2|w)\right] \\ &= E(w_1\pi_1 + w_2\pi_2) \\ &= W_1\pi_1 + W_2\pi_2 \\ &= \pi \end{aligned}$$

□

정리 3.2 모집단으로부터 표본을 모두 SRSWR로 뽑았다고 가정할 경우 추정량 $\hat{\pi}_H$의 분산은 다음과 같다.

$$V(\hat{\pi}_H) = \frac{\pi_A(1-\pi_A)}{n'} + W_2\left(\frac{1}{k}-1\right)\left(\frac{\pi_A(1-\pi_A)}{n'} + \frac{p(1-p)}{n'(2p-1)^2}\right) \tag{3.2}$$

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

$$\begin{aligned} \hat{\pi}_H &= w_1\hat{\pi}_1' + w_2\hat{\pi}_2 \\ &= w_1\frac{n_{1r}'}{n_1'} + w_2\frac{n_{2r}'}{n_2'} + w_2\left(\frac{n_{2r}}{n_2} - \frac{n_{2r}'}{n_2'}\right) \\ &= \frac{n_r'}{n'} + w_2\left(\frac{n_{2r}}{n_2} - \frac{n_{2r}'}{n_2'}\right) \\ &= \hat{\pi} + w_2\left(\frac{n_{2r}}{n_2} - \frac{n_{2r}'}{n_2'}\right) \end{aligned}$$

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.

$$V(\hat{\pi}) = \frac{\pi_A(1-\pi_A)}{n'} \tag{3.3}$$

고정된 w_2에 대하여

$$\begin{aligned} V_2\left(w_2(\frac{n_{2r}}{n_2} - \frac{n_{2r}'}{n_2'})\right) &= w_2^2\left(V_2(\frac{n_{2r}}{n_2}) - V_2(\frac{n_{2r}'}{n_2'})\right) \\ &= w_2^2\left(\frac{1}{n_2} - \frac{1}{n_2'}\right)\left(\pi_A(1-\pi_A) + \frac{p(1-p)}{(2p-1)^2}\right) \\ &= w_2\frac{1}{n'}\left(\frac{1}{k}-1\right)\left(\pi_A(1-\pi_A) + \frac{p(1-p)}{(2p-1)^2}\right) \end{aligned} \tag{3.4}$$

가 되고, 모든 w_2의 분포에 대한 평균값을 취하면

$$E_1 V_2\left(w_2(\frac{n_{2r}}{n_2} - \frac{n_{2r}'}{n_2'})\right) = W_2\frac{1}{n'}\left(\frac{1}{k}-1\right)\left(\pi_A(1-\pi_A) + \frac{p(1-p)}{(2p-1)^2}\right) \tag{3.5}$$

이 된다.

따라서 식(3.3)과 식(3.5)를 더해서 정리하면 식(3.2)를 얻을 수 있다. □

식(3.2)에서 $W_2 = 0$이면,

$$V(\hat{\pi}_H) = \frac{\pi_A(1-\pi_A)}{n'} \tag{3.6}$$

이 된다.

한편, N과 N_h가 커서 n_h/N_h 와 $1/N$의 값이 1과 비교하여 매우 작고, 또한 n_h가 n'에 비해서 상대적으로 작을 경우 $V(\hat{\pi}_H)$의 비편향 분산 추정량은 Des Raj(1968)의 방법을 이용하여 다음과 쓸 수 있다.

$$v(\hat{\pi}_H) = \frac{\hat{\pi}_A(1-\hat{\pi}_A)}{n'-1} + w_2\left(\frac{1}{k}-1\right)\frac{1}{n'-1}\left(\hat{\pi}_A(1-\hat{\pi}_A) + \frac{p(1-p)}{(2p-1)^2}\right) \tag{3.7}$$

4. 효율성 비교

본 장에서는 제시한 층화이중추출을 이용한 결합 확률화응답기법 과 Mangat-Singh 모형과의 효율성을 분산 측면에서 비교해 보고자 한다. 식(2.3)과 (3.2)로부터, $T = 1$, $W_2 = 0$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 $T \neq 1$, $W_2 \neq 0$일 경우

$$\begin{aligned} V(\hat{\pi}_M) - V(\hat{\pi}_H) &= \frac{\pi_A(1-\pi_A)}{n'} + \frac{(1-T)(1-p)(1-(1-T)(1-p))}{n'(2p-1+2T(1-p))^2} \\ &\quad - \left[\frac{\pi_A(1-\pi_A)}{n'}\left(1+W_2(\frac{1}{k}-1)\right) + W_2(\frac{1}{k}-1)\frac{p(1-p)}{n'(2p-1)^2}\right] \\ &= \frac{1}{n'}\left[\frac{(1-T)(1-p)(1-(1-T)(1-p))}{(2p-1+2T(1-p))^2}\right. \\ &\quad \left. - W_2(\frac{1}{k}-1)\left(\pi_A(1-\pi_A) + \frac{p(1-p)}{(2p-1)^2}\right)\right] \\ &\geq 0 \end{aligned} \tag{4.1}$$

을 만족하는 W_2와 T , k 그리고 p의 값을 구해야 한다. $W_2 = 1-T$ 라고 할 때, 식(4.1)의 우변을 보면 $k(0 < k \leq 1)$가 증가할수록 작아지고, p 값의 변화에 상당히 의존하고 있음을 알 수 있다. 다음 표4.1과 표4.2는 π 와 p값이 주어진 상황에서 $W_2 = 1-T$ 와 k값들에 따른 제시한 모형의 Mangat-Singh 모형에 대한 상대 효율($RE = V(\hat{\pi}_M)/V(\hat{\pi}_H)$) 값들을 나타낸 것이다.

T와 $1-W_2$의 값을 같게 놓고 효율성을 비교해본 결과, p값이 작고 , W_2의 값이 $0.5 \leq W_2 \leq 0.7$일 때, 제시한 방법의 효율성이 좋아지는 것을 알 수 있다. 특히 k의 값이 클수록 ($k \geq 0.7$) p값에 상관없이 제시한 모형의 효율성이 좋아지고 있음을 알 수 있었다. 이는 제

표 4.1: $n' = 1,000$, $\pi = 0.1$, $p = 0.2$ 일때의 상대 효율

	$W_2 = 1 - T$								
k	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	0.340	0.362	0.499	0.878	2.441	52.450	4.976	0.769	0.256
0.2	0.640	0.735	1.045	1.872	5.255	113.71	10.843	1.682	0.562
0.3	0.905	1.122	1.647	3.004	8.535	186.22	17.863	2.785	0.933
0.4	1.142	1.521	2.313	4.308	12.407	273.36	26.416	4.141	1.394
0.5	1.355	1.933	3.054	5.823	17.048	380.09	37.062	5.852	1.981
0.6	1.547	2.361	3.883	7.608	22.711	513.83	50.680	8.076	2.755
0.7	1.721	2.803	4.817	9.740	29.776	686.33	68.713	11.085	3.820
0.8	1.880	3.261	5.877	12.332	38.838	917.28	93.725	15.383	5.381
0.9	2.025	3.736	7.091	15.551	50.881	1242.5	130.74	22.026	7.887

표 4.2: $n' = 1,000$, $\pi = 0.1$, $p = 0.8$ 일때의 상대 효율

	$W_2 = 1 - T$								
k	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
0.1	0.195	0.129	0.106	0.096	0.092	0.092	0.093	0.097	0.102
0.2	0.366	0.262	0.223	0.206	0.199	0.199	0.203	0.211	0.224
0.3	0.518	0.399	0.351	0.330	0.323	0.325	0.335	0.350	0.371
0.4	0.654	0.541	0.493	0.473	0.470	0.478	0.495	0.520	0.555
0.5	0.776	0.688	0.650	0.640	0.646	0.664	0.694	0.735	0.789
0.6	0.886	0.839	0.827	0.836	0.860	0.898	0.949	1.015	1.097
0.7	0.985	0.997	1.026	1.070	1.128	1.200	1.287	1.393	1.521
0.8	1.076	1.160	1.252	1.355	1.471	1.603	1.756	1.933	2.142
0.9	1.160	1.329	1.510	1.708	1.927	2.172	2.449	2.768	3.140

시한 모형의 효율성이 W_2와 T, k 그리고 p의 값 모두에 의존하지만 k 의 값이 클수록 안정화돼 간다는 것을 말해준다. 따라서 실제 조사에서 BB방법에 동의하지 않는 사람들 (n_2') 중 적절한 $k(\geq 0.7)$값에 의한 $n_2(n_2 = kn_2', 0 < k \leq 1)$ 명에게 Warner의 확률화응답기법을 적용할 경우 Mangat-Singh 모형보다 분산 측면에서 더 효율적인 결과를 얻을 수 있다. Mangat-Singh 모형의 경우 형식은 간접응답 방법을 취하고 있지만 직접질문의 선택 비율을 조사자가 사전에 결정함으로써 직접조사 또는 확률화응답기법에 협조하고자 하는 조사대상자의 의견 반영을 원천적으로 봉쇄하는 강제적인 면이 있다. 반면에 본 논문에서 제시하고 있는 층화이중추출을 이용한 결합 확률화응답기법의 경우 민감한 질문에 대한 직접질문의 비율을 실제적으로 조사대상자들이 정하게 함으로써 보다 유연한 조사가 가능하도

록 하였다. 민감한 모집단에 대한 조사에서 Mangat-Singh 모형과 같은 간접응답 방법만을 시용하는 것보다 조사하기가 쉬운 BB방법을 통해 정보를 얻고, 나머지 조사 대상자들을 상대로 간접응답방법을 사용함으로써 보다 수월하게 민감한 정보를 수집할 수 있을 것으로 기대된다.

참고문헌

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

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

Des Raj (1968). *Sampling Theory*, McGraw-Hill Book Company, New York.

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

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

[2003년 11월 접수, 2004년 2월 채택]

A Combined Randomized Response Technique Using Stratified Two-Phase Sampling *

Kihak, Hong [1)]

ABSTRACT

We suggest a method to procure information from the sensitive population which combine a direct survey method, BB and an indirect survey one, RRT, and a combined estimator that uses the stratified double sampling to estimate the sensitive parameter. We compare the efficiency of our estimator with that of Mangat and Singh model.

Keywords: Non-response; Sensitive information; Black-Box method; Randomized response technique; Staratified two-phase sampling.

* This research was supported by the Dongshin Unversity research grants in 2003.

1) Professor, Dept. of Computer Science, Dongshin University, Daeho-Dong 252, Naju, Chonnam, 520-714, Korea.
E-mail : khhong@dsu.ac.kr