Journal of Korean
Data & Information Science Society
2003, Vol. 14, No.2 pp. 275~278

Combined Procedure of Direct Question and Randomized Response Technique

Kyoung Ho Choi[1)]

Abstract

In this paper, a simple and obvious procedure is presented that allows to estimate π, the population proportion of a sensitive group. Suggested procedure is combined procedure of direct question and randomized response technique. It is found that the proposed procedure is more efficient than Warner's(1965).

Keywords : Sensitive issues, Direct question, Randomized response technique, Privacy protection

1. 서 론

사회조사시 발생되는 문제점 중, 응답자로부터 신뢰할만한 정보를 얻지 못하는 경우를 고려할 수 있다. 이는 조사의 내용이 법적이나 도덕적으로, 즉 사회통념상 인정되기 어려운 민감한 사안(sensitive issue)일 때 주로 발생된다. 그래서 민감한 사안에 대한 조사시 직접질문(direct question : DQ)을 하게되면 응답거절이나 거짓응답률이 높게 발생되어, 즉 비표본오차의 증대로 인하여 추정의 신뢰도가 떨어진다.

이에 대한 해결방안으로, 응답자의 신분보호(privacy protection)를 통하여 신뢰할만한 응답을 얻음으로써, 추정의 신뢰도를 높일 수 있는 간접질문방식인 확률화응답기법(randomized response technique : RRT)이 Warner(1965)에 의하여 개발되었고, Chudhuri와 Mukerjee(1988) 그리고 류제복 등(1993)은 이를 체계적으로 정리하였다. 한편 Bhargava와 Singh(2002)은 확률화응답기법의 신분보호 측면을 다루었다.

모집단 내의 모든 구성요소가 민감집단(A)과 비민감집단($\overline{A}$)으로 구성된 이지(dichotomous)모집단에서 민감집단의 모비율 π를 추정하는 문제를 고려해 보자. 모집

1) Associate Professor, Department of Information Statistics, Jeonju University, Wansan-Gu, 560-759, Korea
E-mail : ckh414@jeonju.ac.kr

단으로부터 단순임의복원 추출된 n명의 표본에 대하여, Warner(1965)의 확률화응답기법에 의한 추정량과 분산은 각각 다음과 같다. 단 n_1은 표본 중에서 확률장치를 통하여 '예'라고 응답한 응답자의 수이며, p는 확률장치에서 민감질문이 선택될 확률이다.

$$\hat{\pi}_W = \frac{n_1/n-(1-p)}{(2p-1)}, \quad p \neq \frac{1}{2} \tag{1.1}$$

$$Var(\hat{\pi}_W) = \frac{\pi}{n} - \frac{\pi^2}{n} + \frac{p(1-p)}{n(2p-1)^2} \tag{1.2}$$

전술한바와 같이 확률화응답기법은 본래 민감사안에 대해, 응답자의 신분보호를 통하여 정직한 응답을 얻기 위하여 고안된 방법이다. 즉, 확률화응답기법을 이용함으로써 직접질문에 비하여 추정량의 분산은 증가하지만 좀 더 신뢰성 높은 추정을 수행할 수 있게된다. 이러한 이유로 인하여 민감사안에 대한 조사 시, 조사비용과 시간이 더 요구됨에도 불구하고 직접질문 대신에 확률화응답기법을 사용한다.

일반적으로 추정하고자 하는 사안(issue)에 대한 모비율 π가 낮을수록 민감한 내용이라고 할 수 있다. 그런데, 실제 조사를 수행하다 보면 조사자 입장에서는 민감하다고 생각하는 사안에 대하여 응답자 입장에서는 전혀 민감하다고 생각하지 않는 경우도 있다. 이러한 응답자에 대해서는 굳이 수행절차가 복잡한 확률화응답기법을 적용하여 조사를 수행할 필요가 없다. 즉 이러한 응답자에 대해서는 직접질문을 이용함으로써 조사의 효율을 높일 수 있다. 이러한 측면에서 조사하고자 하는 사안에 대한 민감도가 크게 높다고 생각되지 않는 경우라면(예, $0.2 < \pi < 0.4$ 정도), 직접질문과 확률화응답기법을 결합한 조사절차를 이용하여 추정량의 신뢰성 확보와 더불어 추정의 효율을 높일 수 있을 것으로 생각된다. 이에 본 논문에서는 직접질문과 확률화응답기법의 결합을 통한 추정과정에 대해서 알아보고, 분산비교를 통하여 제안된 방법이 효율적임을 보이고자 한다.

2. 결합추정 과정

모집단내의 민감집단의 비율 π를 추정함에 있어, 표본으로 추출된 응답자에 대하여 먼저 직접질문을 통하여 "예"나 "아니오"로 응답하게 한 후, "아니오"라고 응답한 응답자에 대해서만 Warner(1965)의 확률화응답기법을 적용하여 응답을 얻는다. 한편 모집단으로부터 단순임의복원 추출된 n명의 응답자에 대하여 다음을 가정한다. 응답자 중 비민감집단에 속하는 응답자는 직접질문에 대하여 거짓응답이 없으며, 확률장치를 이용하는 응답자도 Warner(1965)에서와 같이 정직하게 응답한다고 하자.

이제 T(기지라고 가정)를 민감집단에 속하는 응답자가 직접질문에 정직하게 응답할 확률이라고 하면, n표본에 대하여 "예"라고 응답할 확률은 다음과 같다.

$$\theta = \pi T + \pi(1-T)p + (1-\pi)(1-p) \tag{2.1}$$

따라서 민감집단의 모비율 π에 대한 결합추정량 $\hat{\pi}_c$은 다음과 같다.

$$\hat{\pi}_c = \frac{n_1/n - (1-p)}{[\,T(1-p)+(2p-1)]} \tag{2.2}$$

그리고 식 (2.2)의 $\hat{\pi}_c$은 다음과 같이 불편추정량임을 알 수 있다.

$$\begin{aligned} E(\hat{\pi}_c) &= \frac{\theta-(1-p)}{[\,T(1-p)+(2p-1)]} \\ &= \frac{\pi[\,T+(1-T)p-(1-p)]}{[\,T(1-p)+(2p-1)]} \\ &= \pi \end{aligned} \tag{2.3}$$

한편, $\theta = (1-p)+[\,T(1-p)+(2p-1)]\pi$ 이므로, $Var(\hat{\pi}_c)$은 다음과 같다.

$$\begin{aligned} Var(\hat{\pi}_c) &= \frac{\theta(1-\theta)}{n[\,T(1-p)+(2p-1)]^2} \\ &= \frac{\pi((2p-1)-[\,T(1-p)+(2p-1)]\pi)}{n[\,T(1-p)+(2p-1)]} + \frac{p(1-p)}{n[\,T(1-p)+(2p-1)]^2} \\ &= \frac{(2p-1)}{[\,T(1-p)+(2p-1)]}\frac{\pi}{n} - \frac{\pi^2}{n} + \frac{p(1-p)}{n[\,T(1-p)+(2p-1)]^2} \end{aligned} \tag{2.4}$$

식 (2.2)와 (2.4)에서 $T=0$이면, 이는 식 (1.1), (1.2)와 각각 같게 된다.

[정리 1] 제안된 결합추정절차에 위한 추정량은 $p>1/2$에 대하여, Warner(1965)에 비하여 항시 효율적이다.

$$Var(\hat{\pi}_c) \le Var(\hat{\pi}_W) \tag{2.5}$$

(증명) $[\,T(1-p)+(2p-1)] \ge (2p-1)$이므로, $p>1/2$에 대하여 식 (1.2)와 (2.4)로부터 식 (2.5)가 성립한다. ■

<표 1> T와 p에 따른 $\frac{Var(\hat{\pi}_W)}{Var(\hat{\pi}_c)}$ (단 $\pi=0.3$)

p \ T	0.1	0.2	0.3	0.4
0.6	1.43529	1.94945	2.54378	3.22009
0.7	1.14921	1.30877	1.47883	1.65959
0.8	1.06145	1.12472	1.18982	1.25679
0.9	1.02083	1.04188	1.06315	1.08464

한편 직접조사에 소요되는 단위당 조사비용을 C_1이라 하고, 확률화응답기법에 소요되는 단위당 조사비용을 C_2라 하자. 그러면 제안한 방법에서 먼저 n명에 대해서 조사비용이 C_1인 직접조사가 수행되고, 이들 중에서 "아니오"라는 응답자의 평균수

인 $(1-\pi)n+(1-T)\pi n$명에 대해서는 조사비용이 C_2인 확률화응답기법을 이용한 조사가 수행되므로 평균 총조사비용은 $E(C_c)=nC_1+n(1-\pi T)C_2$이며, Warner(1965)에 대한 총조사비용은 $C_W=nC_2$이다. 따라서 고정된 표본에 대하여 조사비용의 비(ratio)에 대한 평균은 $E(C_c/C_W)=C_1/C_2+(1-\pi T)$이므로, $C_1/C_2<\pi T$이면 제안된 절차에 의할 때 같은 크기의 표본에 대해서 Warner(1965)에 비하여 평균조사비용이 절감됨을 알 수 있다.

3. 결 론

확률화응답기법은 민감사안에 대한 조사 시, 거짓응답이나 무응답 등으로 인하여 발생하는 비표본오차의 증가 등과 같은 직접질문이 갖는 단점을 보완하여 신뢰성 있는 추정을 하기 위하여 고안된 방법이다. 그런데 확률화응답기법은 추정량의 분산측면과 시간과 비용 등의 추정절차 측면에서 직접질문에 비하여 효율성이 떨어진다. 즉, 효율성의 측면에서만 본다면 직접질문이 더욱 바람직하다고 하겠다.

한편 조사하고자 하는 민감사안의 민감정도가 크게 높지 않은 경우, 일부 응답자는 직접질문을 하여도 정직하게 응답하므로 표본으로 추출된 모든 응답자에 대해서 확률화응답기법을 적용할 필요가 없는 경우가 발생하기도 한다. 이에 본 논문에서는 직접질문을 통한 여과(filtering)과정을 이용하여 일부의 표본에 대해서만 확률화응답기법을 적용하는 결합추정절차에 대해서 알아보았다. 그 결과 식 (2.5)로부터 $p>1/2$에 대하여 제안된 방법이 효율적임을 알 수 있었다.

부언하면, 제안된 방법에서 $T>0$면 여과과정을 통하여 확률화응답기법이 적용되는 표본이 줄어들게 되어 전체적인 조사비용과 시간이 절약되는 효과를 거둘 수 있어, 조사사안의 민감도가 아주 높은 경우가 아니라면 제안된 방법을 사용하는 것이 효율적이라고 생각된다.

한편 제안된 방법에서 T를 기지(known)라고 가정하였는데, 이는 예비조사나 직접질문으로 수행된 비슷한 성격의 선행조사로부터 알아낼 수 있다. 그러나 만약 T를 아는 것이 불가능하거나 좀더 정확한 T의 추정을 원한다면 2표본을 이용하는 문제를 고려해 볼 수 있다.

참고문헌

1. 류제복, 홍기학, 이기성(1993). <확률화응답모형>, 자유아카데미, 서울.
2. Chaudhuri, A., and Mukerjee, R.(1988). *Randomized Response - Theory and Technique*, Marcel Dekker, Inc., New York.
3. Bhargava, M., and Singh, R.(2002). On the efficiency of certain randomized response strategies, *Metrika*, Vol. 55, 191-197.
4. Warner, S.L.(1965). Randomized response : A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*. Vol. 60, 63-69.

[2003년 2월 접수, 2003년 5월 채택]