The Korean Communications
in Statistics Vol. 7, No. 2, 2000
pp. 427-433

Assessment for Efficiency of Two-Stage Randomized Response Technique

Kyung Ho Choi[1)]

Abstract

In this paper, we review several two-stage randomized response techniques for gathering self-report data when persons are asked sensitive question. Also efficiencies and privacy protections based on the two-stage randomized response procedures are compared. Finally, we find optimal parameter conditions.

Keywords : Two-stage RRT, Sensitive Issue, Privacy Protection

1. 서 론

많은 사회조사의 수행시 발생되는 오차 중, 비교적 통제가 어려운 부분은 거짓응답 등으로 인하여 유발되는 응답오차이다. 이러한 응답오차는 비표본오차 중에서 가장 취급하기 어려운 오차로 조사설문이 응답자의 명예나 사생활에 깊이 관련되어 있거나, 또는 개인 재산에 영향을 미치는 경우에 응답자가 응답을 회피하거나 거짓응답을 하게 됨으로써 발생된다. 부언하면 개인의 사생활과 관련된 민감한 사안 - 예컨대, A.I.D.S나 동성연애, 약물중독, 혼전성경험, 낙태여부 및 탈세 등 - 에 대한 조사시 직접질문(direct question)을 하게 되면 질문의 민감성 때문에 응답자는 응답을 회피하거나 거짓응답을 하는 경향이 있게되어, 결국 비표본오차의 증대를 가져와 추정의 신뢰성이 떨어진다. 따라서 이러한 민감한 사안에 대한 조사시에는 직접질문대신에 응답자의 조사에 대한 협력의 정도를 높일 수 있는 간접질문방식이 필요하다.

확률화응답기법(randomized response technique)이란 민감한 사안에 대한 조사시 거짓응답 등으로 인한 비표본오차를 줄이기 위하여 1965년에 Warner에 의하여 제안된 조사기법이다. 본질적으로 확률화응답기법이란 응답자에게 어떤 확률하에서 질문을 선택할 수 있는 기회를 부여하여 응답의 편의를 없애거나 줄일 수 있도록 고안된 간접질문방식이다.

그러나 확률화응답기법은 조사과정에서 응답자의 신분보호를 위하여 확률장치가 도입되는 간접질문방식이기 때문에 직접질문에 비하여 정보의 손실(information loss)이 있게 된다. 그래서 확률장치에 기인한 이러한 손실을 줄여서 추정의 효율을 높이고 얻어진 정보를 좀 더 효율적으로 이용할 수 있는 새로운 기법의 개발에 관한 연구가 지속되어 온 바, 최근 들어 발표되고 있는 2단계 확률화응답기법은 이에 대한 일환으로 고려할 수 있는 방법이다. 2단계 확률화응답기법의 대표적인 예로는 Mangat와 Singh(1990), 그리고 김종호등(1992)을 들 수 있다.

1) School of Information Technology, Jeonju University, Wansan-Gu, 560-759, Korea
E-mail : ckh414@www.jeonju.ac.kr

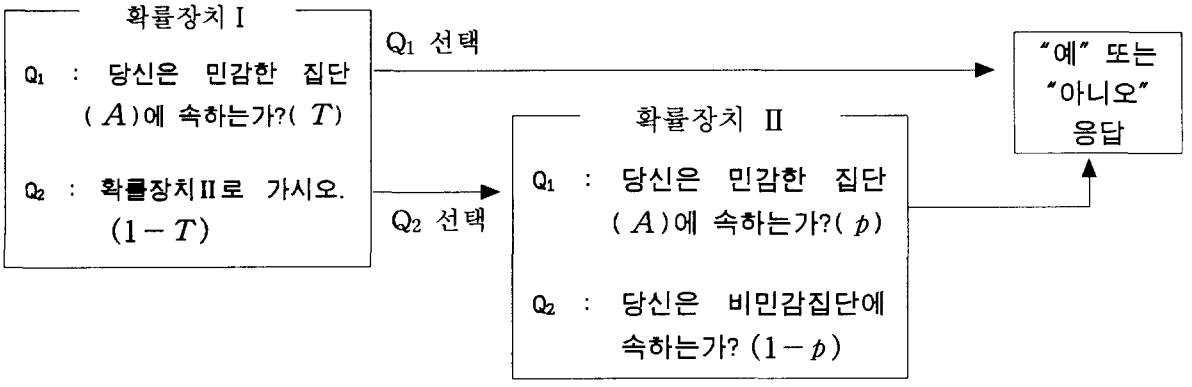
나아가 확률화응답기법을 이용한 조사시 추정의 효율성 문제와 더불어 간과해서는 안될 점 중의 하나로 신분보호(privacy protection)의 문제를 들 수 있다. 이는 확률화응답기법을 이용한 조사에 있어서 획득되는 정보의 양과 신분보호의 정도 사이에는 상충되는 면이 있기에 그러하다. 그래서 신분보호의 정도를 측정하기 위한 노력도 행하여지고 있는데 Warner와 Leysieffer(1976)등은 응답자가 질문에 응답을 함으로써 입게되는 사생활의 노출위험정도를 측정할 수 있는 위험함수(jeopardy function)를 정의하고 있으며, Lanke(1976)는 조건부 확률을 이용하여 신분보호의 정도를 측정할 수 있는 측도를 제시하고 있다.

한편 Carr 등(1982)은 추정의 효율을 높이기 위한 일환으로 조건부 확률화응답기법을 제시하고 있는데, 이 방법 역시 응답획득의 수행절차는 2단계 확률화응답기법을 따르고 있다. 이에 본 논문에서는 이 방법과 Mangat와 Singh 그리고 김종호 등의 방법과의 비교를 통해 2단계 확률화응답기법의 효율성에 대한 평가를 수행하고자 한다. 이 과정에서 효율성비교와 함께 Lanke의 신분보호 측도를 이용한 신분보호의 정도를 비교하여 최적의 반복 확률화응답기법을 수행할 수 있는 모수의 선택에 대해서 언급하고자 한다.

2. 2단계 확률화응답기법

2.1 Mangat와 Singh의 2단계 기법

Mangat와 Singh(1990)에 의하여 제안된 기법으로, 이지모집단(dichotomous population)내의 민감집단(A)의 비율 π를 추정함에 있어 2단계에 걸친 조사를 통하여 응답자로부터 많은 정보를 얻을 수 있도록 고안된 기법이다. 단순임의복원추출된 n명의 응답자에 대하여 Mangat와 Singh의 기법을 이용하여 응답을 얻는 과정은 다음과 같다.



[그림 1. Mangat와 Singh기법의 응답과정]

이 과정을 통하여 얻어진 n명의 응답자에 대한 응답에서 "예"라고 응답한 응답자의 수를 n_1이라 하면, 고려되는 π의 불편추정량 $\hat{\pi}_1$와 이의 분산 $Var(\hat{\pi}_1)$은 다음과 같다.

$$\hat{\pi}_1 = \frac{n_1/n-(1-T)(1-p)}{(2p-1)+2T(1-p)} \tag{2.1}$$

$$Var(\hat{\pi}_1) = \frac{\pi(1-\pi)}{n} + \frac{(1-T)(1-p)[1-(1-T)(1-p)]}{n[(2p-1)+2T(1-p)]^2} \tag{2.2}$$

2.2 김종호 등의 2단계 기법

김종호 등(1992)은 Mangat와 Singh의 기법을 응용하여 그림 1의 확률장치Ⅱ에서 질문 Q_2 대신에 무조건 "예"라고 응답하도록 하는 강요모형을 사용한 2단계 기법을 제시하였다.

단순임의복원추출된 n명의 응답자에 대한 이 과정을 통하여 얻어진 응답에서 "예"라고 응답한 응답자의 수를 n_1이라 하면, 이지모집단내의 민감집단(A)의 비율 π의 불편추정량 $\hat{\pi}_2$와 이의 분산 $Var(\hat{\pi}_2)$은 다음과 같다.

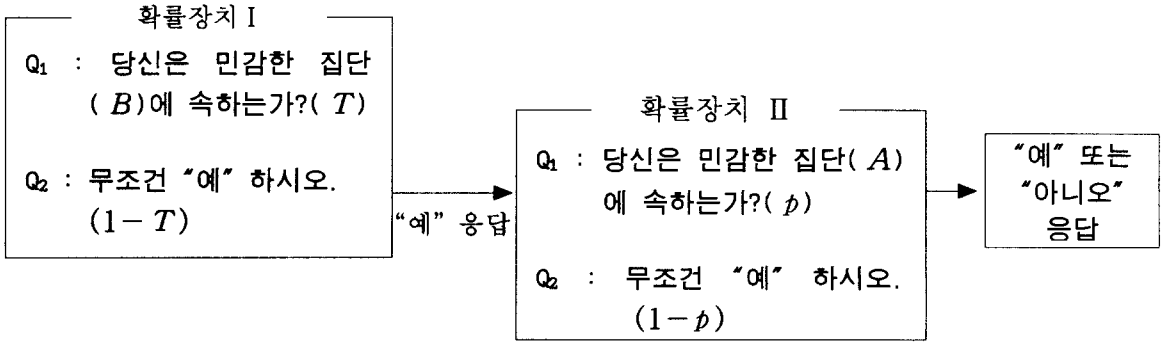
$$\hat{\pi}_2 = \frac{n_1/n-(1-T)(1-p)}{p+T(1-p)} \tag{2.3}$$

$$Var(\hat{\pi}_2) = \frac{\pi(1-\pi)}{n} + \frac{(1-T)(1-p)(1-\pi)}{n[p+T(1-p)]^2} \tag{2.4}$$

한편 김종호 등은 식 (2.2)와 (2.4)의 비교를 통하여 식 (2.4)가 (2.2)에 비하여 통계적인 관점에서 효율적일 수 있는 조건을 구하였다.

2.3 Carr 등의 2단계 기법

Carr 등(1982)은 추정의 효율을 높이기 위한 일환으로 조건부 확률화응답기법을 제시하였다. 그런데 이 기법은 자료획득의 과정이 2단계에 걸쳐서 수행되는 바, 이를 2단계 확률화응답기법으로 간주하도록 하자. 단순임의복원추출된 n명의 응답자에 대해, Carr 등(1982)의 2단계 기법을 통하여 응답을 얻는 과정은 다음과 같다.



[그림 2. Carr 등의 2단계 기법의 응답과정]

여기서 확률장치 I 에서 질문되는 민감사안 B는 확률장치 II에서 질문되는 민감사안 A에 비하여 민감의 정도가 떨어지는 사안이며, 조사대상자 중에서 민감사안 A에 속하는 응답자는 민감사안 B에도 속하는 것으로 가정하자. 이 방법의 장점은 확률장치 I 을 통하여 n명의 응답자 중 추정을 원하는 집단(A)과 관계가 많을 것으로 판단되는 일부에 대하여 다시 2단계 조사를 실시함으로써 추정의 효율을 높이는데 있다.

그림 2의 응답과정을 통하여 얻어진 응답을 토대로 관심의 대상인 민감집단(A)의 모비율 π에 대한 불편추정량 $\hat{\pi}_3$과 이의 분산 $Var(\hat{\pi}_3)$은 Carr 등(1982)을 이용하면 다음과 같다.

$$\hat{\pi}_3 = \frac{n_2 - (1-p)\, n_1}{np} \tag{2.5}$$

$$Var(\hat{\pi}_3) = \frac{[(1-T) + T\pi_B](1-p) - \pi(1-2p+p\pi)}{np} \tag{2.6}$$

단, n_1은 확률장치 I 에서 "예"라고 응답한 응답자의 수이며 n_2는 확률장치 II에서 "예"라고 응답한 응답자의 수이다. 그리고 π_B는 민감집단 B의 모비율이다.

3. 효율비교와 신분보호

3.1 Lanke의 신분보호 측도

확률화응답기법을 이용하여 이지모집단내의 민감속성의 비율을 추정함에 있어, 통계적인 관점에서는 효율적인 추정량을 찾는데 목적을 두고 있다. 그러나 응답자의 입장에서는 자신이 정직하게 응답하여도 자신의 신분이 노출되지 않기를 바란다. 따라서 확률화응답기법을 이용한 조사시 간과해서는 안될 중요한 사항중의 하나는, 응답자가 응답을 함으로써 느끼는 신분보호의 정도이다.

조건부확률을 이용한 Lanke(1976)의 신분보호 측도는 다음과 같다. 모집단내의 민감집단을 A라 했을 때, 응답자의 " r"(r="예" 또는 "아니오")이라는 응답에 대하여 그 응답자가 민감집단에 속하는 것으로 여겨질 확률, 즉 Pr [응답자가 민감집단에 속한다 | 응답자가 " r"이라는 응답을 한다] $\equiv \Pr[A \mid r]$을 계산하여 $\max[\Pr(A \mid \text{예}), \Pr(A \mid \text{아니오})]$ 이 작은 방법일수록 더욱 신분보호의 정도가 잘되는 확률화응답기법이라 할 수 있다.

3.2 Lanke의 측도를 이용한 신분보호 비교

2.1절에 제시된 대표적인 2단계 기법인 Mangat와 Singh의 방법에 대한 Lanke의 측도에 따른 값을 τ_1이라 하면, 일반적인 손실없이 $p = T$ 에 대하여 이는 다음과 같다.

$$\tau_1 = \frac{\pi[p(2-p)]}{p\pi + (1-p)[p\pi + (1-p)(1-\pi)]} \tag{3.1}$$

한편 2.3절에 제시된 Carr 등의 2단계 기법에 대한 Lanke의 측도에 따른 값을 τ_2라 하면, $p = T$ 에 대하여 다음과 같다.

$$\tau_2 = \frac{[(1-p)+p\pi_B]\pi}{[(1-p)+p\pi_B](1-p)+p\pi} \tag{3.2}$$

식 (3.1)과 (3.2)로부터 $\pi < \pi_B$일 때 τ_2가 τ_1보다 작아지게 되는, 즉 Carr 등의 2단계 기법이 Mangat와 Singh의 2단계 기법보다 신분보호가 더 잘 될 수 있는 조건은 $p \geq 0.4$에 대하여 다음과 같다.

$$\pi_B < \frac{p-(p-1)^2}{p} \tag{3.3}$$

나아가 p와 π에 따른 τ_1과 τ_2, 즉 신분보호의 정도는 다음의 표 1과 같다.

<표 1. p와 π의 변화에 따른 신분보호정도. $\pi_B = \pi + 0.1$ >

π	0.1								
p	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
τ_1	0.02540	0.05882	0.10366	0.16495	0.25	0.36842	0.52907	0.72727	0.91667
τ_2	0.10979	0.12139	0.13523	0.15179	0.17143	0.19403	0.21782	0.23684	0.23729
π	0.2								
p	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
τ_1	0.05539	0.12329	0.20648	0.30769	0.42857	0.56757	0.71654	0.85714	0.96117
τ_2	0.21704	0.23626	0.25775	0.28125	0.30588	0.32955	0.34812	0.35484	0.34101
π	0.3								
p	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
τ_1	0.09135	0.19424	0.30847	0.43243	0.5625	0.69231	0.8125	0.91139	0.97697
τ_2	0.32192	0.34555	0.37048	0.39583	0.42	0.44037	0.45313	0.45349	0.43671

표 1로부터 알 수 있는 사실은 Carr 등의 2단계 기법에 의한 조사가 Mangat와 Singh의 기법에 의한 조사보다 0.4이상의 p에 대하여 신분보호가 더욱 잘되고 있음을 알 수 있다. 특히 Carr 등의 2단계 기법의 장점은 π및 p의 변화에 대해 신분보호의 정도가 크게 변하지 않음을 들 수 있다. 반면에 Mangat와 Singh의 기법의 경우는 예상했던 대로 p가 커짐에 따라 신분보호의 정도가 크게 떨어지게 됨을 확인할 수 있다.

3.3 효율비교

확률화응답기법을 사용함에 있어 확률장치에 기인한 정보의 손실을 줄여서 추정의 효율을 높이기 위하여 고안된 2단계 확률화응답기법들의 효율성에 대한 평가를 행하고, 이를 통하여 통계적인 관점에서 최적의 2단계 확률화응답기법을 수행할 수 있는 모수의 선택에 대하여 알아보자. 이는 2장에서 언급한 식 (2.2), (2.4), 그리고 식 (2.6)을 통하여 알 수 있다. 먼저 각 경우에 있어서 일반적인 손실없이 $p = T$ 인 경우에 대하여, π와 p의 변화에 따른 $Var(\widehat{\pi}_3)/Var(\widehat{\pi}_1)$ 과 $Var(\widehat{\pi}_3)/Var(\widehat{\pi}_2)$ 의 해석적 비교는 다음과 같다. 단, 식 (2.6)에서 π_B는 관심의 대상이 되는 민감집단 (A) 보다 민감의 정도가 떨어지는 집단에 대한 모비율로 $\pi_B = \pi + 0.1$로 하였다.

표 2로부터, π가 증가함에 따라 Carr 등의 2단계 기법이 Mangat와 Singh의 기법에 비하여 효율임을 알 수 있다. 나아가 π의 변화에 관계없이, $0.3 < p < 0.6$의 근방일 때는 $Var(\widehat{\pi}_3) < Var(\widehat{\pi}_1)$이 되어 Carr 등의 2단계 기법이 Mangat와 Singh의 기법에 비하여 통계적인 측면에서 우수함을 알 수 있다. 반면에 p가 작지 않은 한 Carr 등의 2단계 기법의 효율은 김종호 등의 기법에 비하여 효율이 떨어짐을 알 수 있다. 이는 김종호 등이 지적했듯이 $Var(\widehat{\pi}_2)$은 p가 증가할수록 작아지게 되는데, 이에 기인한 결과로 볼 수 있다.

<표 2. p와 π의 변화에 따른 효율비교>

π	0.1								
p	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$\frac{Var(\widehat{\pi}_3)}{Var(\widehat{\pi}_1)}$	15.23357	1.00701	0.00261	0.31696	0.70238	0.972	1.11290	1.14502	1.09662
$\frac{Var(\widehat{\pi}_3)}{Var(\widehat{\pi}_2)}$	0.36827	0.67263	0.91291	1.08965	1.20408	1.25815	1.25504	1.20097	1.10906
π	0.2								
p	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$\frac{Var(\widehat{\pi}_3)}{Var(\widehat{\pi}_1)}$	12.01005	0.90358	0.00246	0.30335	0.67033	0.91718	1.03923	1.07124	1.05038
$\frac{Var(\widehat{\pi}_3)}{Var(\widehat{\pi}_2)}$	0.37162	0.68116	0.92175	1.08907	1.18319	1.21068	1.18592	1.12981	1.06379
π	0.3								
p	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9
$\frac{Var(\widehat{\pi}_3)}{Var(\widehat{\pi}_1)}$	9.78105	0.80349	0.00228	0.28583	0.63542	0.87219	0.99457	1.03772	1.03391
$\frac{Var(\widehat{\pi}_3)}{Var(\widehat{\pi}_2)}$	0.37509	0.68998	0.93106	1.09060	1.17058	1.18424	1.15353	1.10241	1.04898

한편, 이상의 결과는 $\pi < \pi_B$를 만족하는 모든 π_B에 대해서 거의 동일하다. 따라서 지금까지 논

의한 결과를 토대로 2단계 확률화응답기법을 이용한 조사의 수행 시 통계적인 측면과 신분보호측면을 모두 고려했을 때, 합리적이라 생각되는 모수의 범위 및 2단계 기법으로는 0.3~0.4근방의 p를 이용하는 Carr 등의 2단계 기법을 고려하는 것이 바람직하다고 하겠다.

4. 결 론

통계적인 방법을 이용하는 많은 사회조사의 수행 시 질문의 내용이 민감한 사안인 경우 추정의 신뢰도를 높이기 위한 방안으로 확률화응답기법이 종종 사용되고 있다. 그런데 이 기법은 사용과정에서 확률장치를 이용하게 되는데, 이에 기인하여 추정의 효율이 떨어지게 된다. 따라서 추정의 효율을 높이기 위한 다양한 방법이 강구된 바, 2단계 확률화응답기법도 이에 대한 일환으로 고려될 수 있다.

본 논문에서는 Carr 등(1982)이 제안한 조건부 확률화응답기법이 2단계에 걸쳐서 수행되는 기법인 바, 이를 토대로 Mangat와 Singh(1990), 그리고 김종호 등(1992)의 2단계 기법과의 비교를 행하였다.

또한 확률화응답기법을 이용한 조사의 수행시 효율성증대와 더불어 간과해서는 안될 점의 중의 하나가 신분보호의 문제인데 본 논문에서는 이에 대해서도 다루었다. 그래서 2단계 확률화응답기법을 이용한 조사의 수행 시 효율성과 신분보호문제를 모두 고려했을 때, 0.3~0.4근방의 p를 이용하는 Carr 등의 2단계 기법을 이용하는 것이 바람직함을 알 수 있었다.

참 고 문 헌

[1] 김종호, 류제복, 이기성(1992). 새로운 2단계 확률화응답모형, 「응용통계연구」, 제5권 2호, 157-167.

[2] Carr, J. W. and Marascuilo, L. A. (1982). Optimal Randomized Response Models and Methods for Hypothesis Testing, *Journal of Educational Statistics*, Vol. 7, 295-310.

[3] Lanke, J.(1976). On the Degree of Protection in Randomized Interviews, *International Statistical Review*, Vol. 44, 197-203.

[4] Loynes, R. M. (1976). Asymptotically Optimal Randomized Response Procedures, *Journal of the American Statistical Association*, Vol. 71, 924-928.

[5] Mangat, N. S. and Singh, R. (1990). An Alternative Randomized Response Procedures, *Biometrika*, Vol. 77, 439-442.

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

[7] Warner, S. L. and Leysieffer, F. W. (1976). Respondent Jeopardy and Optimal Designs in Randomized Response Models, *Journal of the American Statistical Association*, Vol. 71, 649-656.