Review On Current Issues Of The Unrelated
Randomized Response Technique

Journal of Korean
Data & Information Science Society
2002, Vol. 13, No.1 pp. 79 86

Review On Current Issues Of The Unrelated Randomized Response Technique

Kyung Ho Choi[1)]

Abstract

Recently, it is shown that the unrelated question randomized response models proposed by Moors(1971), Folsom et al.(1973), Greenberg et al.(1971) are incapable of protecting the privacy of the respondent. Thus, in this paper, we review recent days research tendency. Also modification model of Mahmood et al.(1998) is proposed, and we show that this model is more efficient than Greenberg et al.(1969). Furthermore we treat the privacy protection based on Lanke's(1975) risk of suspicion measure.

Keywords : 무관질문기법, 신분보호 측도, 확률화응답기법

1. 서 론

개인의 사적인 문제(private issues)와 관련된 조사에서 응답자의 신분보호를 통하여 획득된 응답의 신뢰성을 높이고자 고안된 방법이, 1965년 Warner에 의하여 제안된 확률화응답기법(randomized response technique)이다. 그런데 Warner 기법에서 제시되는 두 질문은 모두 민감한 사안과 관련되어 있기에, 응답자가 확률장치를 통한 정직한 응답을 행함에 있어 여전히 부담을 느낄 수 있다. 나아가 응답자들의 신분보호를 위한 두 질문 모두가 반드시 민감한 사안과 관련을 가질 필요가 없으며, 오히려 민감한 질문과 전혀 관련이 없는 질문을 사용함으로써 응답자들의 정직한 응답에 대한 부담을 줄여줌과 동시에 신분을 더욱 보호해 줄 수 있다. 이러한 형태의 기법을 무관질문기법(unrelated question model)이라 하는데, 이는 Horvitz et al.(1967), Greenberg et al.(1969)로부터 Moors(1971), Folsom et al.(1973)에 의하여 확장·발전되어, 지난 20여 년 동안 여러 조사에서 사용되어 왔다.

그런데 최근 들어 Tracy와 Mangat(1996), Mahmood et al.(1998), Singh et al.(2000)에 의하여, 그 동안 사용되어온 질적(quality) 조사를 위한 무관질문기법에 문제가 있음이 제기되었다. 또한 Singh(1999), Chang과 Huang(2001)은 양적(quantity)인 경우로

1. 전주시 완산구 효자동 1200, 전주대학교 정보통계전공 부교수
E-mail : ckh414@jeonju.ac.kr

까지 이 문제를 확장하여 다루었다.

이에 본 연구에서는 그 동안 사용되어 온 무관질문기법의 문제점과 최근 개발된 해결방안들에 대한 연구동향에 대해서 고찰해 보고, 신분보호(privacy protection) 측면에서 Lanke(1975)의 노출위험(risk of suspicion)에 기초한 무관질문기법의 올바른 사용방안에 대해서 알아보고자 한다.

2. 무관질문기법의 문제점과 해결방안

2.1 무관질문기법의 문제점

Greenberg et al.(1969)에 의하여 이론적으로 정립된 질적 속성 추정을 위한 무관질문기법은 다음과 같다. 민감속성(A)의 비율 π를 추정하기 위하여 모집단으로부터 크기가 n_1과 n_2인 독립표본을 추출하고, 이에 무관질문을 각각 적용하여 다음의 추정량을 구한다.

$$\hat{\pi}_G = \frac{(1-P_2)\hat{\lambda}_1 - (1-P_1)\hat{\lambda}_2}{P_1 - P_2} \tag{2.1}$$

여기서 P_i는 i (i= 1, 2)번째 표본에서 민감질문이 선택될 확률이고, $\hat{\lambda}_1$, $\hat{\lambda}_2$는 각 표본에서 '예'라는 응답의 관측비율이다. 한편 위의 추정량에 대한 분산은 다음과 같다.

$$V(\hat{\pi}_G) = \frac{1}{(P_1 - P_2)^2}\left[\frac{(1-P_2)^2\lambda_1(1-\lambda_1)}{n_1} + \frac{(1-P_1)^2\lambda_2(1-\lambda_2)}{n_2}\right] \tag{2.2}$$

여기서 $\lambda_i = P_i\pi + (1-P_i)\pi_y$로 각 표본에서 '예'라는 응답이 얻어질 확률이며, π_y는 무관집단(Y)의 미지의 모비율이다. 한편 고정된 표본크기 $n(=n_1+n_2)$에 대한 최적 할당으로부터, 추정량의 최소분산은 다음과 같다.

$$Min\,V(\hat{\pi}_G) = \frac{\left[(1-P_2)\sqrt{\lambda_1(1-\lambda_1)} + (1-P_1)\sqrt{\lambda_2(1-\lambda_2)}\right]^2}{n\,(P_1 - P_2)^2} \tag{2.3}$$

한편 Moors(1971)는 P_2가 0에 가까울수록, 그래서 극단적으로 $P_2 = 0$일 때 Greenberg et al.(1969)기법에 의한 추정량의 분산이 최소가 되며 전체적인 조사비용을 줄일 수 있다고 하였다.

그런데 Moors의 주장대로 $P_2 = 0$을 선택하면 추정량의 분산을 최소화할 수는 있지만, 일부 응답자에 대해서는 신분노출이 되어 확률화응답기법 본래의 적용목적에 위배되는 상황이 발생한다. 예를 들어 첫 번째 표본에서 '예'라고 응답한 응답자가 두 번째 표본에도 추출되어 '아니오'라고 응답하면, 이 응답자는 민감속성(A)을 소유한 응답자라는 것이 밝혀지게 되어 확률화응답기법 본래의 취지에 어긋나게 된다. 임의의 응답자가 첫 번째와 두 번째 표본에 모두 포함될 가능성은 추정의 요구정도(required precision)에 따라 실제 발생가능한 일이다.

2.2 Moors기법의 해결방안

Mahmood et al.(1998)은 이러한 문제의 해결방안으로 다음과 같은 방법을 제안하고 있다. 표본 1에 적용되는 확률장치 구성 시, 민감질문과 무관질문 이외에 또 다른 선택 가능한 질문을 삽입함으로써 Moors기법이 갖는 문제점을 해결하고자 하였다. 한 예로 표본 1에 적용되는 확률장치 내의 질문을 다음과 같이 구성하였다. **Scheme 1** - (i) 당신은 민감집단(A)에 속하는가? (ii) 당신은 비 무관집단(Y^c)에 속하는가? (iii) 당신은 무관집단(Y)에 속하는가? - 이 때 각 질문이 선택될 확률은 각각 P_1, P_3 그리고 P_4이며 $P_1 + P_3 + P_4 = 1$이다. 이렇게 함으로써 Moors기법($P_2 = 0$)이 갖는 문제점을 해결함과 동시에, Greenberg et al.(1969)에 비하여 추정의 효율이 증대됨을 보였다. 또한 Mahmood et al.(1998)은 P_3의 확률로 선택되는 두 번째 질문이 **Scheme 2** - (ii)' 한번 더 반복하시오. 만약 반복에서도 이 질문이 선택되면 '예'라고 응답하시오. -, **Scheme 3** - (ii)'' 당신은 비 민감집단(A^c)에 속하는가? - 인 경우에 대해서도 추정량과 이의 분산을 유도하여 Greenberg et al.(1969)과 효율비교를 시도하였다.

나아가 Mahmood et al.(1998)은 수치계산을 통하여 자신들이 제안한 방법들이 Greenberg et al.(1969)보다 효율적일 수 있는 모수의 조건으로 $P_4 = 8(1 - P_1)/9$, $P_3 = 1 - P_1 - P_4$을 권고하고 있다.

한편 Singh et al.(2000)도 Moors기법이 갖는 문제점을 해결하기 위한 두 가지 방안을 제시하고 있는데, 이들의 해결방안은 다음과 같다. 두 독립표본에서 표본 1에 대해서는 Greenberg et al.(1969)에서와 같이 민감속성과 무관속성으로 구성된 확률장치를 이용하고, 표본 2에 선정된 응답자에 대해서는 표본 1과 중복된 경우에는 무관질문을 직접질문하고 중복되지 않은 응답자에 대해서는 민감속성과 무관속성으로 구성된 확률장치를 이용하여 응답획득을 수행하는데 이렇게 함으로써 응답자의 신분보호가 이루어지도록 하였다. 한편 Singh et al.(2000)의 두 번째 해결방안은 다음과 같다. 먼저 크기 n의 표본을 추출한 후, 이를 크기 n_1과 n_2인 부차표본(subsample)으로 나누어 표본 1에는 Greenberg et al.(1969)에서와 같은 방법으로, 표본 2에 대해서는 무관질문을 직접질문하는 방식으로 응답을 획득하여 추정을 수행한다. 이들 방법 역시도 Moors기법이 갖는 문제점을 해결함과 동시에 적절한 조건에서 Greenberg et al.(1969)보다 효율적인 방법이 되지만, Singh et al.(2000)방법의 단점은 추정량의 분산식 유도가 복잡하다는 것이다.

2.3 Folsom et al.기법의 해결방안

1973년 Folsom et al.은 무관질문기법에 있어, 추정의 효율을 높이고자 두 개의 무관속성(Y_1, Y_2)을 이용한 기법을 제시하였다. 그런데 이 경우에 있어서도 응답자의 신분보호가 이루어지지 않는 경우가 발생하게 된다. 예를 들어 두 표본에 모두 추출된 응답자가 표본 1의 확률장치를 통한 응답에서 '예'를 응답하고, 표본 2의 직접질문에서 '아니오'를 응답하게 되면 이 응답자는 민감속성(A)을 갖는 응답자임이 드러나게 되어 확률화응답기법 본래의 취지에 벗어나게 된다.

이에 대한 해결방안으로 Tracy와 Mangat(1996)는 Moors기법이 갖는 문제점에 대한

해결방안과 비슷한 방법을 제시하였다. 즉, 표본 1과 표본 2 모두에서 확률장치내에 또 다른 질문을 선택할 기회를 주어 응답자 신분보호를 이루었다. 그들이 제시한 확률장치의 구성은 다음과 같다.

< 표 1 > Tracy와 Mangat의 확률장치

응답획득 방법	표본 $i(i=1,2)$에 있는 응답자에게 주어지는 질문내용
확률장치	당신은 민감집단(A)에 속하는가? (w.p. P_1)
	당신은 무관집단(Y_i)에 속하는가? (w.p. P_2')
	당신은 비 무관집단(Y_i^c)에 속하는가? (w.p. P_3)
직접질문	당신은 무관집단(Y_j)에 속하는가? $j \neq i$

나아가 Tracy와 Mangat는 그들이 제안한 방법에 대한 분산식 유도를 통하여, 적절한 조건하에서 Greenberg et al.(1969)보다 항시 효율적임을 보였다.

2.4 양적속성 기법의 해결방안

Greenberg et al.(1971)은 무관질문모형을 양적속성으로 확장하였다. 즉 다음과 같은 질문으로 구성된 확률장치를 사용하여 민감속성(X)의 모평균 μ_X 추정에 대한 확률화응답기법을 제시하였다.

Q_1 : 당신의 민감속성(X)에 대한 값은 얼마입니까?

Q_2 : 당신의 무관속성(Y)에 대한 값은 얼마입니까?

무관속성(Y)의 모평균 μ_Y가 미지일 때, 크기가 n_1과 n_2인 두 독립표본을 이용하여 $\hat{\mu}_X$과 이의 분산인 $V(\hat{\mu}_X)$를 유도하였다. 이 때 표본 i $(i=1,2)$에 적용되는 확률장치에서 질문 Q_1이 선택될 확률은 P_i이다. 한편 질적 무관질문에서와 마찬가지로, 이 경우에 있어서도 $V(\hat{\mu}_X)$을 최소로 하는 모수의 선택은 $P_2 = 0$인 경우이다.

그런데 $P_2 = 0$이면 두 표본에 모두 선택된 응답자의 경우, 표본 1에서의 응답과 표본 2에서의 응답이 다르면 민감속성에 대한 내용이 드러나 신분보호가 이루어지지 않는다. 이에 Singh(1999) 다음과 같이 구성된 확률장치를 표본 1에 적용하여 응답자의 신분보호문제를 해결하고자 하였다.

Q_1 : 당신의 민감속성(X)에 대한 값은 얼마입니까? (w.p. P_1)

Q_2 : 당신의 민감속성(X)과 무관속성(Y)에 대한 값의 평균은 얼마입니까? (w.p. $1-P_1$)

한편 Chang과 Huang(2001)은 Singh(1999)이 유도한 분산식에 오류가 있음을 지적하고, 같은 상황하에서 다음과 같은 올바른 분산식을 도출하였다.

$$V(\hat{\mu}_X) = \frac{\sigma_Z^2/n_1 + 0.25(1-P_1)\sigma_Y^2/n_2}{[P_1 + 0.5(1-P_1)]^2} \tag{2.4}$$

단, $\sigma_z^2 = 0.25[(1+3P_1)\sigma_X^2 + (1-P_1)\sigma_Y^2 + P_1(1-P_1)(\mu_X - \mu_Y)^2]$이다.

나아가 식 (2.4)를 이용하여 Greenberg et al.(1971)과 효율비교를 수행한 결과, Singh(1999)의 방법이 매우 효율적임을 보였다.

3. 무관질문기법의 올바른 사용에 대한 제언

본래 확률화응답기법은 응답자의 신분보호를 위하여 고안된 조사기법이다. 나아가 무관질문기법은 응답자의 응답에 따른 부담감을 더욱 줄여주고자 개발된 것이다. 그런데 그 동안 많이 활용되어 온 Moors(1971)와 Folsom et al.(1973) 등의 질적 무관질문기법과 Greenberg et al.(1971) 등의 양적 무관질문기법은 응답자의 신분보호 측면에서 문제점을 안고 있으므로, 질적인 경우에 대해서는 Tracy와 Mangat(1996), Mahmood et al.(1998) 그리고 Singh et al.(2000)의 방법을 나아가 양적인 경우에 대해서는 Singh(1999), Chang과 Huang(2001) 등이 제시한 방법을 사용하는 것이 적절하리라 여겨진다.

한편 여러 형태의 확률화응답기법에 대한 비교 시, 분산측면에서의 효율비교와 더불어 간과해서는 안 될 점이 신분보호의 정도이다. 많은 경우에 있어 이 두 측면은 상충되므로, 상황에 맞는 적절한 방법을 선택하는 것이 중요하다. 이에 여기에서는 Mahmood et al.(1998)이 제시한 방법(표본추출과 응답획득의 과정은 동일)에서 확률장치의 구성을 다음과 같이 변형하여, 효율비교 및 Lanke(1975)의 노출위험(risk of suspicion)에 기초한 무관질문기법들을 비교해 보도록 하겠다.

< 표 2 > 변형된 확률장치

	표본 1에 있는 응답자에게 주어지는 질문내용	
확률장치	당신은 민감집단(A)에 속하는가?	(w.p. P_1)
	무조건 '예'라고 응답하시오.	(w.p. P_3)
	당신은 무관집단(Y)에 속하는가?	(w.p. P_4)

제안한 방법에 의한 응답획득 시, 표본 1에서 '예'라는 응답을 얻을 확률은 $\lambda = P_1\pi + P_3 + P_4\pi_Y$이며, 민감집단의 모비율 π에 대한 추정량과 이에 대한 분산식은 다음과 같다.

$$\hat{\pi} = \frac{\hat{\lambda} - P_3 - P_4\,\hat{\pi}_Y}{P_1} \tag{3.1}$$

$$V(\hat{\pi}) = \frac{\frac{\lambda(1-\lambda)}{n_1} + P_4^2\frac{\pi_Y(1-\pi_Y)}{n_2}}{P_1^2} \tag{3.2}$$

여기서 $\hat{\lambda}$는 표본 1에서 얻어진 응답 중 '예'라고 응답한 표본비율이며, $\hat{\pi}_Y$는 표본 2에서 얻어진 응답 중 '예'라고 응답한 표본비율이다. 한편, 고정된 표본크기 $n(=n_1+n_2)$에 대한 최적할당과 최소분산은 각각 다음과 같다.

$$\frac{n_1}{n_2} = \sqrt{\frac{\lambda(1-\lambda)}{P_4^2\pi_Y(1-\pi_Y)}} \tag{3.3}$$

$$Min\ V(\hat{\pi}) = \frac{[\sqrt{\lambda(1-\lambda)} + P_4\sqrt{\pi_Y(1-\pi_Y)}]^2}{nP_1^2} \tag{3.4}$$

이제 Mahmood et al.(1998)이 권고한 모수인 $P_4 = 8(1-P_1)/9$, $P_3 = 1-P_1-P_4$에 대해서, Greenberg et al.(1969)이 제안한 식 (2.3)과 본 논문에서 제안한 식 (3.4)의 효율비교를 하여보면 다음과 같다.

< 표 3 > Greenberg et al.(1969)과 제안한 방법의 효율비교

P_1	π_Y	π				
		0.1	0.3	0.5	0.7	0.9
0.7	0.1	159.037	186.552	201.677	215.489	233.513
	0.3	174.886	184.784	191.907	199.568	211.169
	0.5	186.374	188.624	191.044	194.707	202.397
	0.7	198.910	195.791	194.138	193.867	197.363
	0.9	223.250	213.835	206.914	200.794	195.475
0.8	0.1	111.804	124.458	130.887	137.471	148.200
	0.3	120.405	125.137	128.299	132.262	139.951
	0.5	126.139	127.045	128.213	130.483	136.421
	0.7	131.476	129.617	129.002	129.581	133.605
	0.9	140.630	135.461	132.421	130.405	130.353

표 3에서 볼 수 있듯이 제안한 방법은 Moors기법이 갖는 문제점을 해결함과 동시에 Greenberg et al.(1969)보다도 효율적임을 알 수 있다. 한편 Lanke(1975)는 확률화응답을 통한 응답과정에서 '예'라고 응답한 응답자가 민감집단(A)에 속할 확률, 즉 $P(A|\text{'예'})$을 계산하여 이 값이 작을수록 신분보호가 잘 된다고 하였다. 제안한 방법과 Mahmood et al.(1998)의 Scheme 1과의 노출위험을 비교해보면 거의 비슷하지만 제안한 방법의 신분보호가 약간 더 잘되는 것으로 나타났다. 그러나 Scheme 2와의 비교에서는 선택된 모수값에 따라 최고 27%까지 신분보호가 더 잘되는 것으로 나타났다. 따라서 전술했듯이 추정량의 분산에 대한 측면뿐만 아니라 노출위험에 기초한 응답자의 신분보호를 고려한다면, Mahmood et al.(1998)의 방법뿐만 아니라 본 논문에서 제안한 방법에 대해서도 고려해볼 가치가 있다고 여겨진다.

4. 결 론

확률화응답기법을 사용하는 근본 목적은 응답자의 신분보호를 통한 신뢰성 있는 응답을 획득하는데 있다. 나아가 Greenberg et al.(1969), Moors(1971) 그리고 Folsom et al.(1973)의 무관질문기법은 응답자의 응답부담을 더욱 줄이고 분산의 측면에서 추정의 효율을 더욱 높이고자 개발된 방법들이다. 그런데 이들 방법들은 일부 응답자에 대해서 신분보호에 실패함으로써, 확률화응답기법 본래의 취지에서 벗어나는 약점을 갖게된다. 이에 본 논문에서는 이에 대한 해결방안인 Tracy와 Mangat(1996), Mahmood et al.(1998), Singh et al.(2000), Singh(1999), Chang과 Huang(2001) 등의 최근연구동향에 대해서 고찰해 보았다.

한편 확률화응답기법을 사용함에 있어 추정량의 분산을 줄이려는 노력과 더불어 간과해서는 안될 또 한가지 측면은 신분보호의 정도이다. 이에 여기에서는 Mahmood et al.(1998) 방법의 변형을 통한 효율증대와 더불어, Lanke(1975)의 측도에 기초하여 신

분보호를 높일 수 있는 방안에 대해서 알아보았다.

결론적으로 지난 20여 년 동안 사용되어 온 Moors(1971), Folsom et al.(1973), Greenberg et al.(1971)등의 방법은 문제점이 노출된 만큼 사용상에 주의가 요구되며, 나아가 해결방안으로 본 연구에서 소개되고 제안된 방법들을 적극 활용할 것을 권고하는 바이다.

참 고 문 헌

1. Chang, H. J., Huang, K. C.(2001). A note on an addendum to the confidentiality guaranteed under randomized response sampling by Mahmood, Singh, and Horn, *Biometrical Journal*, 43(4), 497-500.
2. Folsom, R. E., Greenberg, B. G., Horvitz, D. G. and Abernathy, J. R.(1973). The two alternate questions randomized response model for human surveys, *Journal of the American Statistical Association*, 68, 525-530.
3. Greenberg, B. G., Abul-Ela, A. L. A., Simmons, W. R. and Horvitz, D. G.(1969). The unrelated question randomized response model-theoretical framework, *Journal of the American Statistical Association*, 64, 520-539.
4. Greenberg, B. G., Kuebler, R. R., Abernathy, J. R. Jr. and Horvitz, D. G.(1971). Applications of the randomized response technique in obtaining quantitative data, *Journal of the American Statistical Association*, 66, 243-250.
5. Horvitz, D. G., Shah, B. V. and Simmons, W. R.(1967). The unrelated question randomized model, *Proceedings of the Social Statistics Section American Statistical Association*, 65-72.
6. Lanke, J.(1975). On the choice of unrelated question in Simmons' version of randomized response, *Journal of the American Statistical Association*, 70, 80-83.
7. Mahmood, M., Singh, S. and Horn, S.(1998). On the confidentiality guaranteed under randomized response sampling : A comparison with several new techniques, *Biometrical Journal*, 40(2), 237-242.
8. Moors, J. J. A.(1971). Optimization of the unrelated question randomized response model, *Journal of the American Statistical Association*, 66, 627-629.
9. Singh, S.(1999). An addendum to the confidentiality guaranteed under randomized response sampling by Mahmood, Singh, and Horn, *Biometrical Journal*, 41(8), 955-966.
10. Singh, S., Singh, R. and Mangat, N. S.(2000). Some alternative strategies to Moor's model in randomized response sampling, *Journal of Statistical Planning and Inference*, 83, 243-255.
11. Tracy, D. S.(1996). On response's jeopardy in two alternate questions randomized response model, *Journal of Statistical Planning and Inference*, 55, 107-114.
12. Warner, S. L.(1965). Randomized response : A survey technique for

eliminating evasive answer bias, *Journal of the American Statistical Association*, 60, 63-69.