Journal of the Korean Data & Information Science Society
2010, **21**(2), 271–277

한국데이터정보과학회지

2단계 창의 확률화응답기법

최경호[1]

[1]전주대학교 기초의과학과

접수 2010년 2월 2일, 수정 2010년 3월 10일, 게재확정 2010년 3월 19일

요 약

확률화응답기법은 조사과정에서 응답자의 신분보호를 위하여 확률장치가 도입되는 간접질문방식이기 때문에 직접질문에 비하여 정보의 손실이 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, 2단계 확률화응답기법은 이에 대한 일환으로 고려할 수 있는 방법이다. 한편 Chang 등 (2004) 또한 개선된 강요형기법을 제안하고 Warner (1965)에 비하여 효율적인 조건을 찾았다. 이에 본 연구에서는 Chang 등 (2004)의 기법을 2단계로 확장한 모형을 제안하고, Chang 등 (2004) 및 Mangat와 Singh (1990) 등의 2단계기법에 비하여 효율적인 조건을 찾아보았다.

주요용어: 2단계기법, 비표본오차, 확률화응답, 효율비교.

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 있게 되어, 결국 비표본오차의 증대를 가져와 추정의 신뢰성이 떨어진다 (최경호, 2000).

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 개발되었는데, 구체적인 내용은 다음과 같다.

모집단 내의 모든 구성요소가 민감집단 (A)과 비민감집단 ($BARA$)으로 구성된 이지 (dichotomous)모집단에서 민감집단의 모비율 π를 추정하는 문제를 고려해 보자. 모집단으로부터 단순임의복원추출된 n명의 표본에 대하여, Warner (1965)의 확률화응답기법에 의한 추정량과 분산은 각각 다음과 같다.

$$\widehat{\pi}_W = \frac{n_1/n - (1-p_1)}{(2p_1-1)}, \quad p_1 \neq \frac{1}{2} \tag{1.1}$$

$$Var_W = \frac{\pi(1-\pi)}{n} + \frac{p_1(1-p_1)}{n(2p_1-1)^2} \tag{1.2}$$

[1] (560-759) 전북 완산구 효자동 3가 1200, 전주대학교 기초의과학과 (통계학), 교수.
E-mail: ckh414@jj.ac.kr

단, n_1은 표본 중에서 확률장치를 통하여 '예'라고 응답한 응답자의 수이며, p_1은 확률장치에서 민감질문이 선택될 확률이다.

이후 Greenberg 등 (1969), Abernathy 등 (1970), 그리고 Horvitz 등 (1976) 등에 의하여 발전되어 왔으며, Chudhuri와 Mukerjee (1988) 그리고 류제복 등 (1993)은 이를 체계적으로 정리하였다. 이를 토대로 Lee 등 (2002)과 Ahn 등 (2003)은 무관질문기법으로 확장된 모형을 제안하였으며, Park 등 (2003)은 강요형기법을 온라인 조사에 적용해 보았다.

그러나 확률화응답기법은 조사과정에서 응답자의 신분보호를 위하여 확률장치가 도입되는 간접질문 방식이기 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, 2단계 확률화응답기법은 이에 대한 일환으로 고려할 수 있는 방법이다. 2단계 확률화응답기법의 대표적인 예로는 Mangat와 Singh (1990), 그리고 김종호 등 (1992)을 들 수 있다.

한편 Chang 등 (2004)은 개선된 강요형기법을 제안하고 Warner (1965)에 비하여 효율적인 조건을 찾은 바, 효율을 높이기 위하여 고안된 최근의 연구로 평가된다. 이에 본 연구에서는 Chang 등 (2004)의 기법을 2단계로 확장한 모형을 제안하고, Chang 등 (2004) 및 Mangat와 Singh (1990) 등의 2단계기법에 비하여 효율적인 조건을 찾고자 한다.

2. 선행연구 고찰

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.

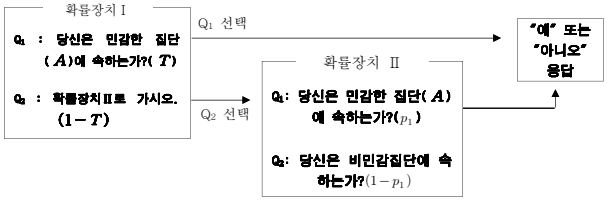

그림 **2.1** Mangat와 Singh기법의 응답과정

이 과정을 통하여 얻어진 n명의 응답자에 대한 응답에서 예 라고 응답한 응답자의 수를 n_1이라 하면, 고려되는 π의 불편추정량 $\widehat{\pi_{MS}}$와 이의 분산 Var_{MS}은 다음과 같다.

$$\widehat{\pi_{MS}} = \frac{n_1/n - (1-T)(1-p_1)}{(2p_1-1) + 2T(1-p_1)} \tag{2.1}$$

$$Var_{MS} = \frac{\pi(1-\pi)}{n} + \frac{(1-T)(1-p_1)[1-(1-T)(1-p_1)]}{n[(2p_1-1)+2T(1-p_1)]^2} \tag{2.2}$$

한편 Mangat와 Singh (1990)은 식 (1.2)와 식 (2.2)로부터 자신들의 2단계기법이 Warner (1965)기법에 비하여 효율적인 조건이 $T > \dfrac{(1-2p_1)}{(1-p_1)}$임을 밝혔다.

2.2. Chang 등의 기법

본 연구의 근간이 되는 Chang 등 (2004)의 기법은 Drane (1976)의 강요형기법의 확장으로 다음과 같다. 모집단으로부터 단순임의복원 추출된 n명의 표본에 대해, 확률장치를 통하여 '나는 민감집단에 속한다 (p_1)', '나는 비민감집단에 속한다 (p_2)', '무조건 예 라고 응답하시오 (p_3)', 그리고 '무조건 아니오 라고 응답하시오 (p_4)' 중 하나에 정직하게 응답하도록 한다. 단 $\sum_1^4 p_i = 1$이다. 응답자 중 예 라고 응답한 사람의 수를 n_1이라면, 식 (2.3)의 추정량과 식 (2.4)의 추정량의 분산을 얻게 된다.

$$\widehat{\pi}_C = \frac{n_1/n - (p_2 + p_3)}{(p_1 - p_2)} \tag{2.3}$$

$$Var_C = \frac{\pi(1-\pi)}{n} + \frac{\pi(1-p_1-p_2-2p_3)}{n(p_1-p_2)} + \frac{(p_2+p_3)(1-p_2-p_3)}{np_1^2} \tag{2.4}$$

3. 제안된 기법

이제 Chang 등 (2004)의 기법을 Mangat와 Singh (1990)이 제안한 2단계 형태로 확장해 보자. 즉 확률장치 R_1 과 R_2을 각각 다음과 같이 구성한다.

R_1: Q_1. 나는 민감한 집단에 속한다 (T) Q_2. 확률장치 R_2 로 가시오 ($1-T$)

R_2: Q_1. 나는 민감한 집단에 속한다 (p_1)

Q_2. 나는 비민감한 집단에 속한다 (p_2)

Q_3. 무조건 예 라고 응답하시오 (p_3)

Q_4. 무조건 아니오 라고 응답하시오 (p_4)

역시 모집단으로부터 단순임의복원 추출된 n명의 표본에 대해, 확률장치를 통하여 정직하게 응답하도록 했을 때 예 라고 응답한 수가 n_1이라 하자. 그러면 $n_1 b(n,\ \lambda)$, 단 λ는 임의의 응답자가 예 라고 응답할 확률로 $\lambda = piT + (1-T)[p_1\pi + p_2(1-\pi) + p_3]$이다. 따라서 식 (3.1)의 추정량을 얻게 된다.

$$\widehat{\pi_S} = \frac{n_1/n - (1-T)(p_2+p_3)}{[T + (1-T)(p_1-p_2)]} \tag{3.1}$$

한편 이항분포의 성질로부터 식 (3.1)은 $E(\widehat{\pi_s}) = \pi$ 가 되어 불편추정량임을 알 수 있다. 나아가 $Var(\widehat{\pi_S}) = \dfrac{\lambda(1-\lambda)}{n[T+(1-T)(p_1-p_2)]^2}$ 으로부터, 식 (3.2)를 유도할 수 있다.

$$Var_S = \frac{\pi(1-\pi)}{n} + \frac{\pi(1-T)(1-p_1-p_2-2p_3)}{n[T+(1-T)(p_1-p_2)]} + \frac{(1-T)(p_2+p_3)[1-(1-T)(p_2+p_3)]}{n[T+(1-T)(p_1-p_2)]^2} \tag{3.2}$$

또한 $E[\dfrac{\widehat{\lambda}(1-\widehat{\lambda})}{n}] = \dfrac{\lambda(1-\lambda)}{n} \cdot \dfrac{(n-1)}{n}$ 임을 이용하면 식 (3.3)과 같은 불편추정량을 얻을 수 있다.

$$\widehat{Var}(\widehat{\pi_S}) = \frac{1}{(n-1)} \frac{\widehat{\lambda}(1-\widehat{\lambda})}{[T+(1-T)(p_1-p_2)]^2} \tag{3.3}$$

4. 효율비교

Chang 등 (2004)은 효율적일 수 있는 모수들의 조건을 찾는 과정에서, 식 (2.4)에서 p_2에 대한 1차미분값이 항상 양수인 점에 착안하여 p_2=0일 것을 제안하였다. 그런데 본 연구에서 유도한 식 (3.2)의 분산식 또한 p_2에 대한 1차미분값이 항상 양수이므로, 효율비교 시 p_2=0으로 하겠다. 그러면 이와 같은 조건에서 식 (2.4)와 식 (3.2)로부터 식 (4.1)이 만족되면 제안된 기법이 Chang 등 (2004)에 비하여 항상 효율적임을 알 수 있다.

$$p_3 < \frac{T}{[1-(1-T)^2]} \tag{4.1}$$

본 연구의 일차적인 목적은 제안된 기법이 Chang 등 (2004)에 비하여 효율적인 조건인 식 (4.1)을 찾는 것이다. 그러나 이에 더하여 제안된 기법과 Warner (1965) 및 Mangat와 Singh (1990) 그리고 Chang 등 (2004)과의 효율비교를 수치적 해석을 통하여 수행해 보자.

일반적으로 확률화응답기법이 적용되는 경우는 모집단의 비율이 매우 낮은 경우이다. 따라서 효율비교 시 π=0.1인 경우에 대하여만 수행토록 하고, 편의상 $p_3 = p_4$로 하겠다. 그러면 $\sum_1^4 p_i = 1$이므로, $p_3 = (1-p_1)/2$으로 고정된다.

표 4.1부터 표 4.3에서 볼 수 있듯이 제안된 2단계기법에 의할 때 많은 경우에서 효율적임을 알 수 있다. 특히 p_1이 0.3이상인 경우에는 제안된 기법을 사용하는 것이 바람직하다고 하겠다.

표 4.1 π =0.1일 때Var_W/Var_S

	T=0.1	*T=0.2*	*T=0.3*	*T=0.4*	*T=0.5*	*T=0.6*	*T=0.7*	*T=0.8*	*T=0.9*
P 1 =0.1	0.034	0.076	0.138	0.226	0.346	0.512	0.746	1.089	1.625
P 1 =0.2	0.176	0.302	0.472	0.699	1.000	1.404	1.959	2.750	3.948
P 1 =0.3	0.842	1.240	1.751	2.405	3.249	4.352	5.830	7.878	10.868
P 1 =0.4	5.962	7.965	10.443	13.523	17.390	22.322	28.753	37.401	49.539
P 1 =0.6	13.523	15.999	18.898	22.322	26.407	31.342	37.401	44.988	54.733
P 1 =0.7	4.537	5.141	5.830	6.622	7.539	8.613	9.884	11.409	13.268
P 1 =0.8	2.523	2.750	3.002	3.282	3.596	3.948	4.347	4.803	5.328
P 1 =0.9	1.625	1.704	1.787	1.876	1.971	2.073	2.182	2.299	2.426

표 4.2 π = 0.1일때 Var_{MS} /Var_S

	T=0.1	*T=0.2*	*T=0.3*	*T=0.4*	*T=0.5*	*T=0.6*	*T=0.7*	*T=0.8*	*T=0.9*
P 1 =0.1	0.072	0.374	2.124	38.085	37.272	6.725	3.304	2.126	1.493
P 1 =0.2	0.374	1.712	15.204	204.158	11.395	4.647	2.802	1.959	1.435
P 1 =0.3	2.124	15.204	779.911	16.472	6.064	3.511	2.424	1.810	1.379
P 1 =0.4	38.085	204.158	16.472	6.725	4.005	2.802	2.126	1.676	1.325
P 1 =0.5	37.272	11.395	6.064	4.005	2.953	2.317	1.883	1.552	1.270
P 1 =0.6	6.725	4.647	3.511	2.802	2.317	1.959	1.676	1.435	1.217
P 1 =0.7	3.304	2.802	2.424	2.126	1.883	1.676	1.493	1.325	1.163
P 1 =0.8	2.126	1.959	1.810	1.676	1.552	1.435	1.325	1.217	1.109
P 1 =0.9	1.493	1.435	1.379	1.325	1.270	1.217	1.163	1.109	1.055

5. 결론

통계적인 방법을 이용하는 많은 사회조사의 수행 시 질문의 내용이 민감한 사안인 경우 추정의 신뢰도를 높이기 위한 방안으로 Warner (1965)에 의하여 제안된 확률화응답기법이 종종 사용되고 있다. 그런

표 **4.3** π =0.1일때 Var_C /Var_S

	T=0.1	*T=0.2*	*T=0.3*	*T=0.4*	*T=0.5*	*T=0.6*	*T=0.7*	*T=0.8*	*T=0.9*
P 1 =0.1	3.708	8.422	15.703	26.571	42.935	68.683	112.946	203.198	476.916
P 1 =0.2	2.042	3.574	5.762	8.895	13.500	20.643	32.818	57.522	132.250
P 1 =0.3	1.604	2.427	3.554	5.126	7.397	10.884	16.787	28.718	64.731
P 1 =0.4	1.409	1.946	2.661	3.642	5.044	7.179	10.776	18.021	39.853
P 1 =0.5	1.301	1.688	2.195	2.882	3.857	5.333	7.811	12.789	27.769
P 1 =0.6	1.233	1.529	1.914	2.431	3.160	4.261	6.102	9.796	20.898
P 1 =0.7	1.187	1.423	1.728	2.136	2.710	3.573	5.014	7.901	16.570
P 1 =0.8	1.155	1.348	1.598	1.931	2.398	3.100	4.270	6.612	13.642
P 1 =0.9	1.130	1.293	1.502	1.781	2.171	2.757	3.734	5.689	11.553

데 이 기법은 사용과정에서 확률장치를 이용하게 되는데, 이에 기인하여 추정의 효율이 떨어지게 된다. 따라서 추정의 효율을 높이기 위한 다양한 방법이 강구된 바, Mangat와 Singh (1990)의 2단계 확률화응답기법도 이에 대한 일환이라 하겠다. 나아가 Chang 등 (2004)의 개선된 강요형기법 또한 추정의 효율을 높이기 위한 최근의 연구로 평가된다.

이런 상황에서 본 연구에서는 Chang 등 (2004)의 기법을 2단계로 확장한 모형을 제안하고, Chang 등 (2004) 및 Mangat와 Singh (1990) 등의 2단계기법에 비하여 효율적인 조건을 찾아보았다. 그 결과 표 4.1과 표 4.2에서 볼 수 있듯이 p_1이 0.3이상인 경우에는 제안된 기법을 사용하는 것이 효율적임을 알 수 있다. 특히 제한된 조건이지만 표 4.3을 그림으로 나타낸 그림 5.1에서 보듯이 제안된 기법은 Chang 등 (2004)에 비하여 항상 효율적이라고 할 수 있다.

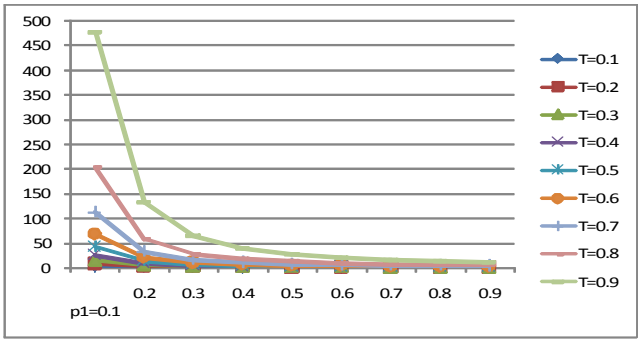

그림 **5.1** $\pi = 0.1$ 일 때Var_H/Var_S

다만 확률화응답기법을 이용한 조사의 수행 시 효율성증대와 더불어 간과해서는 안 될 점의 중의 하나가 신분보호 (Lanke, 1976)의 문제인데 본 논문에서는 이에 대해서는 다루지 못했다. 따라서 향후에는 제안된 2단계 확률화응답기법을 이용한 조사의 수행 시, 효율성과 신분보호문제를 모두 고려한 최적모수의 조건을 찾는 연구가 계속되어야 할 것으로 사료된다.

참고문헌

김종호, 류제복, 이기성 (1992). 새로운 2단계 확률화응답모형. <응용통계연구>, **5**, 157-167.
류제복, 홍기학, 이기성 (1993). <확률화응답모형>, 자유아카데미, 한국.

최경호 (2000). 2단계 확률화응답기법의 효율성 평가. <한국통계학회논문집>, **7**, 427-433.

Abernathy, J. R., Greenberg, B. G. and Horvitz, D. G. (1970). Estimates of induced abortion in urban North Carolina. *Demography*, **7**, 19-29.

Ahn, S. C., Lee, K. S. (2003). The three-stage cluster unrelated question model. *Journal of the Korean Data & Information Science Society*, **14**, 55-65.

Chang, H. J., Wang, C. L. and Huang, K. C. (2004). On estimating the proportion of a qualitative sensitive character using randomized response sampling. *Quality and Quantity*, **38**, 675-680.

Drane, W. (1976). On the theory of randomized response to sensitive questions. *Communications in Statistics - Theory and Methods*, **5**, 565-574.

Greenberg, B. G., Abul-Ela, Abdel-Latif 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.

Horvitz, D. G., Greenberg, B. G. and Abernathy, J. R. (1976). Randomized response: A data-gathering device for sensitive questions. *International Statistical Review*, **44**, 181-196.

Lanke, J. (1976). On the degree of protection in randomized interviews. *International Statistical Review*, **44**, 197-203.

Lee, K. S., Hong, K. H. (2002). A study on the multi trial of unrelated question models. *Journal of the Korean Data & Information Science Society*, **13**, 25-34.

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

Park, H. C., Ryu, J. H. and Lee, S. Y. (2003). Implementation of forced answer model for sensitive information at on-line survey. *Journal of the Korean Data & Information Science Society*, **14**, 489-499.

Warner, S. L. (1965). A survey technique for eliminating evasive answer bias. *Journal of the American Statistical Association*, **60**, 63-69.

Journal of the Korean Data & Information Science Society
2010, **21**(2), 271–277

한국데이터정보과학회지

Two stage Chang's randomized response technique

Kyoung Ho Choi[1]

[1]Department of Basic Medical Science, Jeonju University

Received 2 February 2010, revised 10 March 2010, accepted 19 March 2010

Abstract

The randomized response technique is an indirect question that employs a randomizing device to protect respondents' privacy. The technique is now considered the most efficient of the newly developed techniques. In this technique, Chang *et al.* (2004) suggests an improved forced-answer technique and finds more efficient conditions than Warner did in 1965. But it is the weakness of the technique to lose more information than a direct response technique does. Therefore, a lot of researches have developed new techniques to reduce loss of information, to enhance estimated efficiency, and to efficiently use collected information. Considering this tendency, this paper also tries to improve Chang's technique. It suggests the technique that is extended from Chang's and finds more efficient conditions than Chang's technique and Mangat and Singh's (1990) did.

Keywords: Efficiency comparison, non-sampling error, randomized response, two-stage technique.

[1] Professor, Department of Basic Medical Science, Jeonju University, Jeonju 560-759, Korea. E-mail: ckh414@jj.ac.kr