The Korean Communications
in Statistics Vol. 7, No. 3, 2000
pp. 721-729

A Conditional Randomized Response Model for Detailed Survey

Gi Sung Lee[1], Ki Hak Hong[2]

Abstract

In this paper, we propose a new conditional randomized response model that has improved the Carr et al.'s model in view of the variance and the protection of privacy of respondents. We show that the suggested model is more effective and protective than the Loynes' model and Carr et al.'s model.

Keywords : conditional randomized response model, Loynes' model, Carr et al's model.

1. 서론

사회 여러 분야의 조사에는 응답자들이 응답을 회피하거나 정직하게 응답하지 않는 질문들이 종종 포함된다. 예를 들어, 불로소득, 탈세여부, 전과경험, 알코올중독, 환각제사용, 낙태경험, 동성연애 등과 같은 민감한 질문을 공개적으로 하게 되면 무응답이나 거짓응답 또는 응답을 회피함으로써 응답자들로부터 정확한 정보를 얻을 수 없게 된다. 따라서, 민감한 조사에서 응답자의 신분이나 비밀을 노출시키지 않고서 민감한 질문에 대한 정보를 이끌어 내기 위하여 Warner(1965)는 응답자들에게 민감한 질문과, 민감한 질문과 배반되는 질문으로 구성된 확률장치를 사용하여 간접적으로 응답하도록 하는 확률화응답모형(randomized response model ; RRM)을 처음으로 제시하였다. 그 후 수많은 학자들에 의해 이에 대한 연구가 확대되고 발전되었으며, Loynes(1976)는 Warner모형의 민감한 질문과 배반이 되는 질문 대신에 "예"라고 응답하도록 강요하는 강요질문모형을 제안하였다. 한편, Carr 등은 덜 민감한 속성 B와 강요질문으로 구성된 확률장치를 통해 "예"라고 응답한 사람들에게만 Loynes의 강요질문모형을 사용하도록 하여 민감한 속성 A에 대한 모비율을 추정해 내는 조건부 확률화응답모형을 제안하였다. Carr 등은 그들이 제안한 조건부 확률화응답모형에서 덜 민감한 속성 B에 대한 정보를 확률장치를 이용하여 얻고 있으며, 그 정보를 이용하여 민감한 속성 A에 대한 좀 더 심층적인 조사가 될 수 있도록 하였고, Loynes의 강요질문모형보다 적절한 조건하에서 늘 효율적임을 보였다. 하지만 Carr 등의 모형은 덜 민감한 속성에 대해서도 강요질문을 포함한 확률장치를 이용하고 있어서 그 모형의 사용절차가 복잡하므로

1) Associate Professor, Department of Computer Science & Statistics, Woosuk University, Wanju-gun, Chonbuk, 565-701, Korea.
E-mail : gisung@core.woosuk.ac.kr
2) Associate Professor, Department of Computer Science, Dongshin University, Daeho-dong, Naju, Chonnam, 520-714, Korea.
E-mail : khhong@blue.dongshinu.ac.kr

덜 민감한 속성에 대해서는 특별히 확률장치를 사용하지 않고 응답자들에게 직접질문을 함으로써 좀 더 단순화된 절차를 통해 정보를 얻을 수 있다고 생각된다.

따라서, 본 논문에서는 덜 민감한 속성 B에 대하여 응답자들에게 직접질문을 하여 "예"라고 응답한 사람들에게만 Loynes의 강요질문모형을 사용하도록 하여 민감한 속성 A에 대한 좀 더 심층적인 정보를 얻을 수 있는 조건부 확률화응답모형을 제안하고자 한다. 그리고, 제안된 조건부 확률화응답모형에서 덜 민감한 속성 B를 가지고 있는 응답자가 직접질문을 받음으로써 완전히 진실된 응답을 하지 않은 경우에 대해서도 살펴보고자 한다. 또한, 제안된 조건부 확률화응답모형과 Loynes의 모형 그리고 Carr 등의 모형과의 효율성을 분산과 비밀보장 측면에서 비교해 보고자 한다.

2. 심층조사를 위한 조건부 확률화응답모형

2.1 제안한 모형

Carr 등(1982)은 심층조사를 위해 첫 번째 단계에서 덜 민감한 속성 B와 강요질문으로 구성된 확률장치를 통해 "예"라고 응답한 사람들만을 대상으로 두 번째 단계에서 민감한 속성 A와 강요질문으로 구성된 확률장치를 이용하여 선택된 질문에 응답하도록 하여 민감한 속성 A에 대한 모비율을 추정하는 조건부 확률화응답모형을 제안하였다. 이 때, 덜 민감한 속성 B는 민감한 속성 A보다 민감의 정도가 떨어지는 속성이다. 예를 들어, 덜 민감한 속성 B를 흡연이라고 하고, 민감한 속성 A를 대마초 흡연이라고 하면, 흡연자들이나 강요질문에 의해 "예"라고 응답한 사람들에게만 대마초 흡연에 대하여 질문을 하게 된다. 이와 같이 Carr 등은 응답자들 중에서 민감한 속성 A와 별로 관계가 없을 것이라고 생각되는 일부를 덜 민감한 속성 B를 이용하여 "아니오"라고 응답한 사람들을 제외함으로써 좀 더 심층적인 조사가 될 수 있도록 하였다. Carr 등이 첫 번째 단계에서 덜 민감한 속성 B와 강요질문으로 구성된 확률장치를 사용하고 있는 데, 이 장에서는 덜 민감한 속성이라면 특별히 확률장치를 사용하지 않고 응답자들에게 직접질문을 하는 방법을 택하는 방법을 제안하고자 한다. 예를 들어 흡연여부는 직접질문을 통해 응답하게 하고, 흡연자중에서 확률장치를 이용하여 대마초 흡연율을 조사하고자 하는 것이다.

제안하고자 하는 조건부 확률화응답모형의 절차는 두 단계로 나누어지며, 첫 번째 단계에서는 직접질문을 두 번째 단계에서는 강요질문 확률화응답모형을 사용하게 된다.

크기가 N인 모집단은 덜 민감한 속성을 가지고 있는 사람들의 수 N_1과 덜 민감한 속성을 가지고 있지 않은 사람들의 수 N_2의 합으로 구성되어 있다고 하자. 이 때, 덜 민감한 속성을 가지고 있는 사람들의 수 N_1은 민감한 속성 A를 가지고 있는 사람들의 수 N_A를 포함하고 있다.

모집단으로부터 단순임의추출된 표본의 크기를 n이라 하고, 첫 번째 단계에서 n명의 응답자 중 "예"라고 응답한 사람들의 수를 n_1이라 하자. 그리고, 두 번째 단계에서 n_1명의 응답자 중 "예"라고 응답한 사람들의 수를 n_2라 하자.

첫 번째 단계에서 n명의 응답자들은

당신은 덜 민감한 속성 B를 가지고 있습니까?

라는 직접질문에 "예"와 "아니오"로 응답하게 된다.

첫 번째 단계에서 응답자가 완전히 진실된 응답을 하는 경우 "예"라고 응답할 확률을 λ_1이라 하면, 이는 덜 민감한 속성 B의 모비율 $\pi_1 = \frac{N_1}{N}$ 과 일치하게 된다. 이 때, "예"라고 응답한 사람들의 수를 n_1명이라고 하였으므로, $\hat{\lambda}_1 = \frac{n_1}{n}$ 이 되어 덜 민감한 속성 B에 대한 모비율 π_1의 추정량 $\hat{\pi}_1$는 다음과 같다.

$$\hat{\pi}_1 = \frac{n_1}{n} \ . \tag{2.1}$$

두 번째 단계에서는 첫 번째 단계에서 "예"라고 응답한 n_1명의 응답자들을 대상으로 다음과 같은 2개의 질문으로 구성된 Loynes의 강요질문모형의 확률장치를 통해 선택된 질문에 대하여 응답하도록 한다.

질문 1 : 당신은 민감한 속성 A를 가지고 있습니까?
질문 2 : "예"라고 응답하시오.

여기서, 질문 1이 선택될 확률은 p이고, 질문 2가 선택될 확률은 $1-p$이다. 이 때, 응답자들은 확률장치에 의해서 선택된 질문에 대해 "예" 또는 "아니오"라고 응답한다.

따라서, 응답자가 첫 번째 단계에서 "예"라고 응답했다는 조건하에서 두 번째 단계에서 "예"라고 응답할 확률을 λ_2라 하면, 그 확률은 다음과 같다.

$$\lambda_2 = \frac{p\pi_2}{\pi_1} + (1-p) \ . \tag{2.2}$$

여기서, $\pi_2 = \frac{N_A}{N}$ 는 민감한 속성 A에 대한 모비율이다.

n_1명의 응답자들 중에서 "예"라고 응답한 사람들의 수를 n_2라 하였으므로 $\hat{\lambda}_2 = \frac{n_2}{n_1}$ 가 되어 민감한 속성 A에 대한 모비율 π_2의 추정량 $\hat{\pi}_2$는 다음과 같다.

$$\hat{\pi}_2 = \hat{\pi}_1[\ \hat{\lambda}_2 - (1-p)]/p$$

$$= \frac{1}{np}[n_2-(1-p)n_1]\,. \tag{2.3}$$

<정리 1> 추정량 $\hat{\pi}_2$는 모비율 π_2의 비편향추정량이다.

(증명)

$n_1 \sim b(n, \pi_1)$이고, $n_2 \sim b(n, \pi_1\lambda_2)$이므로, $\hat{\pi}_2$의 기대값을 구해 보면 다음과 같다.

$$\begin{aligned} E(\hat{\pi}_2) &= \frac{1}{np}[E(n_2)-(1-p)E(n_1)] \\ &= \frac{1}{np}[n\pi_1\lambda_2-(1-p)n\pi_1] \\ &= \pi_2\,. \end{aligned}$$

■

<정리 2> 추정량 $\hat{\pi}_2$의 분산은 다음과 같다.

$$Var(\hat{\pi}_2) = \frac{\pi_1(1-p)-\pi_2(1-2p+p\pi_2)}{np}\,. \tag{2.4}$$

(증명)

$$\begin{aligned} Var(\hat{\pi}_2) &= Var\left[\frac{n_2-(1-p)n_1}{np}\right] \\ &= \frac{Var(n_2)+(1-p)^2\,Var(n_1)-2(1-p)Cov(n_1, n_2)}{(np)^2} \end{aligned} \tag{2.5}$$

에서 $n_1 \sim b(n, \pi_1)$이고 $n_2 \sim b(n, \pi_1\lambda_2)$이며 $n_2 \mid n_1 \sim b(n_1, \lambda_2)$이므로, 다음을 얻을 수 있다.

$$\begin{aligned} Var(n_1) &= n\pi_1(1-\pi_1)\,, \\ Var(n_2) &= E[\,Var(n_2 \mid n_1)] + Var[E(n_2 \mid n_1)] \\ &= E[n_1\lambda_2(1-\lambda_2)] + Var(n_1\lambda_2) \\ &= n\pi_1\lambda_2(1-\lambda_2) + n\pi_1(1-\pi_1)\,\lambda_2^{\,2} \\ &= n[p\pi_2+(1-p)\pi_1][1-p\pi_2-(1-p)\pi_1]\,, \end{aligned}$$

$$\begin{aligned} Cov(n_1, n_2) &= E(n_1 n_2) - E(n_1)E(n_2) \\ &= E[n_1 E(n_2 \mid n_1)] - (n\pi_1)(n\pi_1\lambda_2) \\ &= \lambda_2 E(n_1^2) - n^2 \pi_1^2 \lambda_2 \\ &= n\pi_1(1-\pi_1)\lambda_2 \\ &= n(1-\pi_1)[p\pi_2 + (1-p)\pi_1]. \end{aligned}$$

따라서, 위의 세 식을 식(2.5)에 대입하여 $Var(\hat{\pi}_2)$를 구해보면 식(2.4)를 얻을 수 있다.

■

2.2 응답자가 직접질문에서 진실된 응답을 하지 않았을 경우

첫 번째 단계에서 덜 민감한 속성 B를 가지고 있는 응답자에게 직접질문을 함으로써 응답자가 완전히 진실된 응답을 하지 않은 경우에 대하여 살펴보기로 하자. 이 때, 민감한 속성 A를 가지고 있는 응답자들은 확률장치를 이용하여 응답하므로 거짓으로 응답할 이유는 없다고 가정한다. 따라서, 조건부 확률화응답모형에서 제안된 절차에 대하여 응답자들이 "예"라고 응답할 확률은 다음과 같다.

$$\lambda_2' = \frac{p\pi_2}{\pi_1\theta} + (1-p). \tag{2.6}$$

여기서, θ는 덜 민감한 속성 B를 가지고 있는 응답자가 진실로 응답할 확률이다.

조건부 확률화응답모형의 절차에 따라 첫 번째 단계에서 n명의 응답자들 중 "예"라고 응답한 사람들의 수를 n_1'이라 하고, 두 번째 단계에서 n_1'명의 응답자들 중 "예"라고 응답한 사람들의 수를 n_2'이라 하면 $\hat{\lambda}_2' = \frac{n_2'}{n_1'}$이 되므로 민감한 속성 A에 대한 모비율 π_2의 추정량 $\hat{\pi}_2'$는 다음과 같다.

$$\hat{\pi}_2' = \hat{\pi}_1\theta[\hat{\lambda}_2' - (1-p)]/p. \tag{2.7}$$

이 때, $n_1' \sim b(n, \pi_1\theta)$이고, $n_2' \sim b(n, \pi_1\theta\lambda_2)$이므로 $\hat{\pi}_2'$는 모비율 π_2의 편향추정량이며, 그 편향은 다음과 같다.

$$B(\hat{\pi}_2') = \pi_2(\theta - 1). \tag{2.8}$$

따라서, 제안한 조건부 확률화응답모형의 절차에 따라 덜 민감한 속성 B를 가지고 있는 응답자가 진실로 응답하지 않을 경우에 민감한 속성 A에 대한 모비율의 추정량 $\hat{\pi}_2'$의 평균제곱오차는 다음과 같다.

$$MSE(\hat{\pi}_2') = \frac{\pi_1\theta(1-p)-\pi_2(1-2p+p\pi_2)}{np} + \{\pi_2(\theta-1)\}^2 . \tag{2.9}$$

2.3 비밀보장의 정도

응답자가 민감한 속성에 대하여 응답을 함으로써 입게 되는 비밀보장의 정도를 측정하기 위하여 베이즈 확률 측면에서 제안한 조건부 확률화응답모형에서 민감한 속성을 가지고 있는 응답자가 "예(Y)"라고 응답했을 때 자신의 민감한 속성 A를 드러낼 확률을 구해보면 다음과 같다.

$$P_2(A \mid Y) = \frac{\pi_1\pi_2}{p\pi_2+(1-p)\pi_1} . \tag{2.10}$$

이 때, 위의 값이 작을수록 비밀보장의 정도가 잘되는 확률화응답모형이라 할 수 있다.

3. 효율성 비교

3.1 분산 측면

이 절에서는 제안한 조건부 확률화응답모형이 분산 측면에서 Loynes의 모형과 Carr 등의 모형보다 효율적이 되는 조건을 제시하여 그 효율성을 비교해 보고자 한다.

우선, Loynes의 강요질문모형에서 민감한 속성 A에 대한 모비율 π_2의 추정량 $\hat{\pi}_l$의 분산은 다음과 같다.

$$Var(\hat{\pi}_l) = \frac{(1-\pi_2)(1-p+p\pi_2)}{np} . \tag{3.1}$$

제안한 조건부 확률화응답모형의 분산 식(2.4)와 Loynes의 모형의 분산 식(3.1)을 이용하여 $V(\hat{\pi}_2) < V(\hat{\pi}_l)$를 만족하는 조건을 구해보도록 하자.

$$\begin{aligned} Var(\hat{\pi}_l) - Var(\hat{\pi}_2) &= \frac{(1-\pi_2)(1-p+p\pi_2)}{np} - \frac{\pi_1(1-p)-\pi_2(1-2p+p\pi_2)}{np} \\ &= \frac{(1-p)(1-\pi_1)}{np} > 0 . \end{aligned} \tag{3.2}$$

식(3.2)를 만족하는 조건은 $p \neq 1$, $\pi_1 \neq 1$이며, 이는 일반적으로 성립되므로 제안한 조건부 확률화응답모형이 Loynes의 모형보다 항상 효율적이라는 사실을 알 수 있다.

다음으로, Carr 등의 모형에서 민감한 속성 A에 대한 모비율 π_2의 추정량 $\hat{\pi}_c$의 분산은 다음과 같다.

$$Var(\hat{\pi}_c) = \frac{(1-p+p\pi_1)(1-p)-\pi_2(1-2p+p\pi_2)}{np} . \tag{3.3}$$

마찬가지로, 제안한 조건부 확률화응답모형의 분산 식(2.4)와 Carr 등의 모형의 분산 식(3.3)을 이용하여 $V(\hat{\pi}_2) < V(\hat{\pi}_c)$를 만족하는 조건을 구해보도록 하자.

$$\begin{aligned} Var(\hat{\pi}_c) - Var(\hat{\pi}_2) &= \frac{(1-p+p\pi_1)(1-p)-\pi_2(1-2p+p\pi_2)}{np} \\ &\quad - \frac{\pi_1(1-p)-\pi_2(1-2p+p\pi_2)}{np} \\ &= \frac{(1-p)^2(1-\pi_1)}{np} > 0 . \end{aligned} \tag{3.4}$$

식(3.4)를 만족하는 조건은 $p \neq 1$, $\pi_1 \neq 1$이며, 이는 일반적으로 성립되므로 제안한 조건부 확률화응답모형이 Carr 등의 모형보다 항상 효율적이라는 사실을 알 수 있다.

3.2 비밀보장 측면

이 절에서는 제안한 조건부 확률화응답모형과 Loynes의 모형 그리고 Carr 등의 모형과의 효율성을 비밀보장 측면에서 비교해 보고자 한다.

베이즈 확률 측면에서 Loynes의 모형에 대하여 민감한 속성을 가지고 있는 응답자가 "예"라고 응답했을 때 자신의 민감한 속성 A를 드러낼 확률을 구해보면 다음과 같다.

$$P_l(A \mid Y) = \frac{\pi_2}{p\pi_2 + (1-p)} . \tag{3.5}$$

따라서, 제안한 조건부 확률화응답모형의 식(2.10)과 식(3.5)로부터 다음과 같은 관계를 얻어낼 수 있다.

$$P_2(A \mid Y) \leq P_l(A \mid Y) . \tag{3.6}$$

그러므로, 제안한 조건부 확률화응답모형이 Loynes의 모형에 비하여 비밀보장 측면에서 더욱 효율적임을 알 수 있다.

또한, 마찬가지 방법으로 베이즈 확률 측면에서 Carr 등의 모형에 대하여 민감한 속성을 가지고 있는 응답자가 "예"라고 응답했을 때 자신의 민감한 속성 A를 드러낼 확률을 구해보면 다음과 같다.

$$P_c(A \mid Y) = \frac{\pi_2\{p\pi_1 + (1-p)\}}{p\pi_2 + (1-p)\{p\pi_1 + (1-p)\}} \ . \tag{3.7}$$

따라서, 제안한 조건부 확률화응답모형의 식(2.10)과 식(3.7)로부터 다음과 같은 관계를 얻어낼 수 있다.

$$P_2(A \mid Y) \leq P_c(A \mid Y) \ . \tag{3.8}$$

그러므로, 제안한 조건부 확률화응답모형이 Carr 등의 모형에 비하여 비밀보장 측면에서 더욱 효율적임을 알 수 있다.

4. 결론

본 논문에서는 덜 민감한 속성 B에 대하여 응답자들에게 직접질문을 하여 "예"라고 응답한 사람들에 한하여 Loynes의 강요질문모형을 사용하여 민감한 속성 A에 대한 모비율을 추정해 내는 조건부 확률화응답모형을 제안하였다. 이 모형은 응답자들 중에서 민감한 속성 A와 별로 관계가 없을 것이라고 생각되는 일부를 덜 민감한 속성 B에 대한 직접질문을 통해 "아니오"라고 응답한 사람들을 사전에 제외할 수 있는 방법이다. 따라서, 민감한 속성 A와 관계가 있을 만한 응답자들을 대상으로 함으로써 좀 더 심층적인 조사가 될 수 있는 장점이 있다.

하지만 덜 민감한 속성 B에 대하여 직접질문을 하게 되면 응답자에 따라 완전히 진실된 응답을 하지 않는 경우가 있을 수 있기 때문에 그런 상황도 고려하여, 민감한 속성 A에 대한 모비율의 추정과 그 평균제곱오차를 구하였다. 또한, 제안된 조건부 확률화응답모형과 Loynes의 모형 그리고 Carr 등의 모형과 효율성을 분산측면에서 비교한 결과, $p \neq 1$, $\pi_1 \neq 1$을 만족하면 늘 제안된 조건부 확률화응답모형이 효율적이었다. 이 조건은 일반적으로 만족된다고 생각할 수 있으므로 제안된 조건부 확률화응답모형이 분산 비교 측면에서 효율이 뛰어남을 알 수 있었다. 한편, 비밀보장 정도 측면에서 효율성을 비교해 본 결과 제안된 조건부 확률화응답모형이 Loynes의 모형이나 Carr 등의 모형보다 더 효율적임을 알 수 있었다.

마지막으로 제안된 조건부 확률화응답모형은 표본으로 추출된 응답자들 모두에 대하여 확률장치를 사용하지 않아도 되므로 비용이나 시간을 절약할 수 있을 뿐만 아니라 좀 더 심층적인 정보를 얻을 수 있기 때문에 민감한 사항에 관한 실제조사에서 많이 활용될 수 있을 것으로 기대된다.

참고문헌

[1] 류제복, 홍기학, 이기성 (1993). 「확률화응답모형」, 자유아카데미, 서울.

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

[4] Chaudhuri, A. and Mukerjee, R. (1988). *Randomized Response : Theory and Techniques*, Marcel Dekker, Inc., New York.

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

[6] 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.