2000 Proceedings of the Autumn Conference, Korean Statistical Society

반복시행된 확률화 응답(RRD) 모형의 독립조건

이관제[1] · 국세정[2]

Independence Condition in the Repeated Randomized Response Models

Kwan J. Lee, Sejeong Kook
Department of Statistics, Dongguk University

- Abstract -

Krishnamoorphy and Raghavarao(1993) invented exact binomial and asymptotically normal test procedures for truthful answering in the repeated randomized response models under the assumption that two repeated response measures are independent.

Under the same assumption, Lakshmi and Raghavarao(1992) suggested asymptotic chi-square test for respondents' truthful answering in the same models.

In this article we detect the factors and the conditions with which two response variables might be independent, and find the condition for independence in the repeated randomized response models with considering untruthful answer. But, the condition of independence make the randomized model no meaning.

Under the assumption of conditional independence between two response variables, we can apply the same logical statements on deriving the tests for truthful answering in the repeated randomized response models as in Krishnamoorphy and Raghavarao(1993).

Keywords: Repeated ramdomized model, Independence, Conditional independence, Untruthful answering

I. 서 론

확률화응답모형은 어떠한 민감한 성질을 가지고 있는 모집단의 모비율에 관한 연구방법으로, 확률화 장치를 통하여 조사자는 응답자가 어느 질문에 대답을 하였는지를 모르게 함으로

1) 동국대학교 통계학과 부교수, 서울시 중구 필동 3가 26
2) 동국대학교 통계학과 대학원, 서울시 중구 필동 3가 26

서 응답자의 정보에 대한 비밀을 보장한다. 응답자들이 개인적으로 나타내기를 꺼리는 사항에 대하여 비밀이 보장되므로 진실된 응답을 한다. 확률화 장치를 이용하여 응답자들에게 솔직한 응답을 유도한다 하지만, 조사하고자하는 성질이 너무 민감한 경우, 예를 들면 법적으로 금하고 있는 마약의 사용여부, 또는 응답자들이 확률화 응답기법의 제대로 이해하지 못한 경우에는 응답자들로부터 솔직한 응답을 얻는다는 확신을 하기가 어렵다. 이러한 응답의 신뢰성의 문제는 Krishnamoorphy와 Raghavarao(1993)와 Lakshmi와 Raghavarao(1992)에 의하여 통계적 가설검정문제로 연구되었다. 이들 연구에서는 한 응답자에게 두 번의 확률화응답기법을 적용하여 과연 응답자들이 솔직한 응답 여부를 검정하고 이를 토대로 하여 민감한 집단에 속하는 모비율을 추정하였다.

확률화응답모형에 관한 내용은 Warner(1965)와 류제복(1993)등에 자세히 설명되었다.

Krishnamoorphy와 Raghavarao(1993)의 반복된 확률화응답의 확률모형을 유도하기 위한 독립 가정과 유도 과정을 살펴보자.

II. Krishnamoorphy와 Raghavarao(1993)의 반복된 확률화 응답 모형

모집단이 민감한 사항에 속하는 집단 C와 그렇지 않은 집단 $\overline{C}$로 나누어져 있고, 각 개인은 C, $\overline{C}$ 둘 중 하나의 집단에 속한다고 가정한다. 집단 C의 모비율을 π 라 하면, 집단 $\overline{C}$의 모비율은 $1-\pi$ 이다. 모집단으로 복원추출된 응답자에게는 아래와 같은 질문 1이나 질문 2 중 하나가 확률 장치에 의하여 주어진다.

질문 1 : 당신은 집단 C에 속합니까?
질문 2 : 당신은 집단 $\overline{C}$에 속합니까?

첫 번째 확률장치에서 질문 1이 선택될 확률을 p_1, 질문 2가 선택될 확률을 $1-p_1$이고, 두 번째 확률장치에서 질문 1이 선택될 확률을 p_2, 질문 2가 선택될 확률을 $1-p_2$이라 하자. i번째 응답자의 첫 번째 확률장치에 의한 응답확률변수를 X_i, 두 번째 확률장치에 의한 응답확률변수를 Y_i라 할 때, 각 응답변수는 "예"라고 응답하면 1이고, "아니오"라고 응답하면 0이라 정의한다.

또한, 민감한 집단에 속하는 응답자가 거짓응답을 할 확률을 L_1이라 하고, 민감하지 않는 집단에 속한 응답자가 거짓 응답할 확률을 L_2라 한다. 즉,

$$L_1 = P(\text{거짓응답}|C), \quad L_2 = P(\text{거짓응답}|\overline{C}).$$

위와 같이 정의된 반복된 확률화응답모형에서 Krishnamoorphy와 Raghavarao(1993)는 응답확률변수 X_i와 Y_i가 독립이라는 가정하에, i번째 응답자의 연속된 두 번의 응답 (X_i, Y_i)에 대한 결합 확률을 다음과 같이 제시하였다.

$$P(X_i=1, Y_i=1) = \pi[p_1(1-L_1)+(1-p_1)L_1][p_2(1-L_1)+(1-p_2)L_1] + (1-\pi)[p_1L_2+(1-p_1)(1-L_2)][p_2L_2+(1-p_2)(1-L_2)]. \quad (1)$$

$$P(X_i=1, Y_i=0) = \pi[p_1(1-L_1)+(1-p_1)L_1][p_2L_1+(1-p_2)(1-L_1)] + (1-\pi)[p_1L_2+(1-p_1)(1-L_2)][p_2(1-L_2)+(1-p_2)L_2]. \quad (2)$$

$$P(X_i=0, Y_i=1) = \pi[p_1L_1+(1-p_1)(1-L_1)][p_2(1-L_1)+(1-p_2)L_1] + (1-\pi)[p_1(1-L_2)+(1-p_1)L_2][p_2L_2+(1-p_2)(1-L_2)]. \quad (3)$$

$$P(X_i=0, Y_i=0) = \pi[p_1L_1+(1-p_1)(1-L_1)][p_2L_1+(1-p_2)(1-L_1)] + (1-\pi)[p_1(1-L_2)+(1-p_1)L_2][p_2(1-L_2)+(1-p_2)L_2]. \quad (4)$$

식 (2)과 (3)을 이용하여, 서로 일치하지 않는(discordant) 응답이 나올 확률을 구하면

$$\begin{aligned} P(\text{불일치 응답}) &\equiv P(X_i=1,\ Y_i=0) + P(X_i=0,\ Y_i=1) \\ &= \theta_0 + 2(1-2p_1)(1-2p_2)[\pi L_1(1-L_1)+(1-\pi)L_2(1-L_2)] \\ &= \theta(L_1, L_2) \end{aligned} \quad (5)$$

이고, 여기서 $\theta_0 = p_1+p_2-2p_1p_2$ 이다.

거짓응답확률 L_1과 L_2에 대한 귀무가설과 대립가설을 다음과 같이 설정할 수 있다.

$$H_0 : L=0, \qquad H_1 : 0<L_1<1,\ 0<L_2<1. \quad (6)$$

만일 $L_1=L_2=L$을 가정하면, $\theta(L)$는

$$\theta(L) = \theta_0 + 2(1-2p_1)(1-2p_2)L(1-L)$$

과 같고, 식 (5)을 이용하여, 식 (6)의 귀무가설과 대립가설은 일반성을 잃지 않고 $p_1,\ p_2<0.5$ 일 때, 다음과 같이 다시 나타낼 수 있다.

$$H_0 : \theta(L)=\theta_0, \qquad H_1 : \theta_0<\theta(L)\leq 0.5.$$

표본의 크기가 n일 때, $N_{xy}(x, y=0,1)$는 반응확률변수가 $(X=x,\ Y=y)$을 갖는 응답자의 빈도확률변수이다. 거짓 응답이 없다는 귀무가설이 참이면, 불일치응답을 보인 반응확률변수 $N_{10}+N_{01}$는 이항 분포 $b(n, \theta_0)$을 따른다. 이항분포를 이용한 유의수준 α를 갖는 기각역은 $(N_{10}+N_{01}) > b_\alpha$이다. 여기서, b_α는 모수 n과 θ_0를 갖는 이항분포의 $(1-\alpha)$ 분위수이다.

표본의 크기 n이 충분히 클 때 $E(N_{10}+N_{01}) = n\theta_0$, $Var(N_{10}+N_{01}) = n\theta_0(1-\theta_0)$이므로, 중심극한 정리를 이용해 얻어진 근사적 기각역은 다음과 같이 나타낼 수 있다.

$$\frac{\frac{(n_{10}+n_{01})}{n}-\theta_0}{\sqrt{\frac{\theta_0(1-\theta_0)}{n}}} \geq z_\alpha .$$

여기서, z_α는 표준정규분포에서의 $(1-\alpha)$ 분위수이다.

$L_1=L_2=L$하에서 L에 대한 점추정과 점추정량의 분산은 다음과 같다.

$$\hat{L}=0.5\pm\sqrt{0.25-\frac{(n_{01}+n_{10})/n-\theta_0}{2(1-2p_1)(1-2p_2)}}$$

$$Var(\hat{L})=\frac{\theta(1-\theta)}{4n\{(1-2p_1)(1-2p_2)\}^2\{1-4L(1-L)\}} .$$

그리고, 귀무가설 $L=0$ 이 참이고, $p_1=p_2=p$ 에서 모비율 π에 대한 최우추정량은

$$\hat{\pi}=\frac{\{(1-p)^2+p^2\}n_{00}}{(1-2p)(n_{00}+n_{11})}-\frac{p^2}{1-2p}$$

이다. 귀무가설 $L=0$을 만족하지 않고, $L_1=L_2$ 일 때 모비율 π의 추정량은

$$\hat{\pi}=\frac{n_{11}/n-(p_1\hat{L}+\overline{p_1}(1-\hat{L})(p_2\hat{L}+\overline{p_2}(1-\hat{L}))}{(1-2\hat{L})(p_1+p_2-1)}$$

이다.

Lakshmi와 Raghavarao(1992)는 Krishnamoorphy와 Raghavarao(1993)에서 전개된 결합확률을 이용하여 응답의 신뢰성에 대한 χ^2-검정을 제안하였다. Krishnamoorphy와 Raghavarao(1993) 그리고 Lakshmi와 Raghavarao(1992)의 논문에서는 확률변수 X와 Y가 독립이라는 동일한 가정에서 결과들이 유도되었다. 보다 자세한 과정은 Krishnamoorphy와 Raghavarao(1993)와 Lakshmi와 Raghavarao(1992)를 참조한다.

Krishnamoorphy와 Raghavarao(1993) 그리고 Lakshmi와 Raghavarao(1992) 논문이 발표된 이후에 확률화 응답모형에서의 거짓응답의 검정에 대한 연구는 확률변수 X와 Y의 독립성을 가정하고 전개되었으며 유도된 결과가 발표되었다.

III. 반복된 확률화 응답모형에서의 독립성

두 확률변수 X와 Y은 독립은 독립의 정의에 따라 $x\in\{0,1\}$와 $y\in\{0,1\}$에 대하여

$P(X=x, Y=y) = P(X=x)P(Y=y)$을 만족한다. X와 Y의 주변확률은 식 (1) - (4)을 이용하여

$$
\begin{aligned}
P(X=1) &= 1-\theta+L_1\theta+L_2\{-1-2p_1(-1+\theta)+\theta\}+p_1(-1+2\theta-2L_1\theta),\\
P(X=0) &= p_1+L_2(-1+2p_1)(-1+\theta)+\theta-L_1\theta-2p_1\theta+2L_1p_1\theta,\\
P(Y=1) &= 1-\theta+L_1\theta+L_2\{-1-2p_1(-1+\theta)+\theta\}+p_2(-1+2\theta-2L_1\theta),\\
P(Y=0) &= p_2+L_2(-1+2p_2)(-1+\theta)+\theta-L_1\theta-2p_2\theta+2L_1p_2\theta
\end{aligned}
$$

이 된다.

확률변수 X와 Y의 $P(X=x, Y=y) - P(X=x)P(Y=y)$를 구하면

$$
\begin{aligned}
&P(X=1, Y=1) - P(X=1)P(Y=1)\\
&= \pi[p_1(1-L_1)+(1-p_1)L_1][p_2(1-L_1)+(1-p_2)L_1]\\
&+ (1-\pi)[p_1L_2+(1-p_1)(1-L_2)][p_2L_2+(1-p_2)(1-L_2)]\\
&- [1-\theta+L_1\theta+L_2\{-1-2p_1(-1+\theta)+\theta\}+p_1(-1+2\theta-2L_1\theta)]\\
&\quad [1-\theta+L_1\theta+L_2\{-1-2p_1(-1+\theta)+\theta\}+p_2(-1+2\theta-2L_1\theta)]\\
&=(L_1+L_2-1)^2(1-2p_1)(1-2p_2)(1-\theta)\theta \qquad (7)
\end{aligned}
$$

이고, 이와 유사하게 다음 식을 유도할 수 있다.

$$
\begin{aligned}
&P(X=1, Y=0) - P(X=1)P(Y=0)\\
&\quad =(L_1+L_2-1)^2(1-2p_1)(1-2p_2)(1-\theta)\theta
\end{aligned}
$$

$$
\begin{aligned}
&P(X=0, Y=1) - P(X=0)P(Y=1)\\
&\quad =(L_1+L_2-1)^2(1-2p_1)(1-2p_2)(1-\theta)\theta
\end{aligned}
$$

그리고

$$
\begin{aligned}
&P(X=1, Y=0) - P(X=1)P(Y=0)\\
&\quad =(L_1+L_2-1)^2(1-2p_1)(1-2p_2)(1-\theta)\theta.
\end{aligned}
$$

확률변수 X와 Y가 독립이기 위한 조건인 $P(X=x, Y=y) - P(X=x)P(Y=y) = 0$을 만족하기 위해서는 $L_1=1-L_2$ 이거나 $p_1=0.5$, 또는 $p_2=0.5$, 그리고 $\theta=0$ 또는 $\theta=1$ 중에서 하나만 만족되면 된다.

독립을 만족하는 조건 중에서 집단의 거짓응답확률 L_1 과 L_2 그리고 민감집단의 비율 θ는 모집단의 특징으로 임의적인 조정이 불가능하지만, 민감 질문의 비율 p_1과 p_2는 확률화 응답 모형을 고려할 때 조정 할 수 있다.

두 확률변수 X와 Y가 독립이기 위해 민감질문 확률을 $p_1=0.5$ 또는 $p_2=0.5$ 로 모형화 한다면, (1) - (4)에서 제시된 결합확률 $P(X=x, Y=y)$는 모두 0을 갖게 된다. 그리고 $p_1=0.5$ 또는 $p_2=0.5$ 이면 π 의 최우추정량이 ∞가 되며, L의 분산 또한 ∞가 되어 통계적으로 의미가 없게 된다.

Krishnamoorphy와 Raghavarao(1993)에 의하여 유도된 결합확률을 살펴보면, 민감 집단 또는 민감하지 않은 집단의 조건부 결합확률을 이용하였다는 것을 쉽게 알 수 있다. 즉,

$$P(X=x, Y=y|C) = P(X=x|C)P(Y=y|C)$$
$$P(X=x, Y=y|\overline{C}) = P(X=x|\overline{C})P(Y=y|\overline{C})$$

을 이용하여 모든 식 (1) - (4)의 결과를 얻을 수 있다.

참고 문헌

[1] Abul-Ela, A. A., Greenberg, B. G., and Horvits, D. G. "A Multi-Proportions Randomized Response Model", Journal of the American Statistical Association, 62, 990-1000, 1967.

[2] Agresti, A.: An Introduction to Categorical Data Analysis, Wiley, 1996.

[3] Krishnamoorthy, K. and Raghavarao, D. "Untruthful answering in repeated randomized response procedures", Canadian Journal of Statistics, V21, 233-236, 1993.

[4] Lakshmi, D. and Raghavarao, D. "A test for detecting untruthful answering in randomized response procedures", Journal of Statistical Planning and Inference 31, 387-390, 1992.

[5] Lee, K. J. "A test for detecting untruthful answering in repeated randomized response models", The Korean Journal of Applied Statistics (submitted for publication). vol. 12, 1999.

[6] Mangat, N. S. "An Improved Randomized Response Strategy", Journal of Royal Statistical Soc. B, 56, 93-95, 1994.

[7] Warner, S. L. "Randomized response: a survey technique for eliminating evasive answer bias", Journal of American Statistics Association., 60, 63-69, 1965.