응용통계연구 제 6권 2호
1993년 9월 pp.311-318.

대체 확률화응답기법[1)]

류 제 복[2)]

요 약

본 논문에서는 Warner의 확률응답모형 대신에 강요질문모형을 반복 사용하여 응답자들이 진실되게 응답했는 가에 대한 가설검정을 하고 Warner모형을 사용한 경우와 검정력을 비교하였다. 또한 강요질문모형을 사용하여 민감한 특성을 갖는 모집단비율의 추정량을 구하고 Warner모형을 사용하여 얻은 추정량들보다 효율적이 되는 조건을 찾았다.

1. 머 리 말

개인의 사생활이나 이해와 관계되는 민감한 질문에 대하여 응답자들이 응답을 회피하거나 고의적인 거짓응답으로 비표본오차가 발생하게 된다. 이는 응답자들이 자신의 신분이나 사생활이 보장되지 않는다고 생각하기 때문이다. 이에 Warner(1965)는 확률장치를 사용하여 응답자의 신분을 보장해 줌으로써 보다 진실된 응답을 얻을 수 있는 획기적인 확률화응답모형을 제시하였다. 그후 사용의 간편함이나 효율의 증대를 위한 많은 연구가 이루어지고 있다. 특히 Chaudhuri와 Mukerjee(1988)는 이분야에 대한 연구들을 정리하여 체계화 시켰다. 그러나 응답자들이 이러한 확률화응답모형에 대한 이해의 미흡함과 신분의 완전한 보장에 대한 두려움을 갖게 되므로 실제적으로 진실된 응답을 얻기가 힘들고 사용상에도 어려움이 따른다(Duffy와 Waterton, 1988).

진실한 응답이 곤란한 경우에 대하여 Warner(1965), Mangat와 Singh(1990) 등이 모집단비율을 추정하였고 Krishnamoorthy와 Raghavarao(1991) 그리고 Lakshmi와 Raghavarao(1992)는 Warner의 확률응답모형을 반복 사용하여 응답자들이 진실되게 응답했는 가에 대한 가설검정을 하였고 이로부터 거짓응답확률과 민감한 특성에 대한 모집단비율을 추정하였다. 그러나 Warner 모형에서는 두 질문이 모두 민감한 질문과 관련되어 있으므로 응답자들이 응답을 할때 선택된 질문이 무엇이든 "예"라고 응답하기가 거북스럽게 된다. 따라서 본 논문에서는 응답자들에게 이러한 부담을 덜어주기 위하여 하나의 질문에 "예"라는 응답을 강요하는 강요질문모형을 사용하고자 한다.

2장에서는 직접질문의 경우 거짓응답의 효과를 측정하였고 3장에서는 강요질문모형을 반복 사용하여 응답자들이 진실되게 응답했는가에 대한 가설검정과 Warner의 확률응답모형을 반복 사용한 경우와의 효율을 비교하였다.

---

1) 이 논문은 1991년도 교육부지원 한국학술진흥재단의 지방대학육성과제 학술연구조성비에 의하여 연구되었음

2) (360-764) 충북 청주시 내덕동 36 청주대학교 응용통계학과

2. 직접질문에서 거짓응답의 효과

단순임의복원추출된 n 명의 응답자들에게 직접 다음과 같은 민감한 질문을 하였을 때 거짓응답의 효과를 측정하고자 한다.

설문 : "나는 민감한 그룹(C그룹)에 속한다"

응답자들이 진실되게 응답한다고 할 때 n명의 응답자들 중에서 n' 명의 응답자가 위 설문에 "예" 라고 응답하였다면 민감한 그룹에 속하는 모집단비율 π의 추정량 $\hat{\pi}_t$ 는 다음과 같다.

$$\hat{\pi}_t = \frac{n'}{n} \tag{2-1}$$

이때 $\hat{\pi}_t$ 는 π의 최우추정량이고 추정량의 분산은 $\pi(1-\pi)/n$ 가 된다.

응답자들이 진실되게 응답하지 않을 경우, 응답확률변수 $R_i(i=1,2,\cdots,n)$를 i번째 응답자가 "예" 라고 응답하면 1 이고 "아니오" 라고 응답하면 0 이라 정의하고 모든 응답자에 대하여 다음과 같은 가정을 할 수 있다.

① $\delta = P(R = 0 \mid C)$

② $\beta = P(R = 1 \mid \overline{C})$

③ 응답은 각 표본단위에 대하여 독립이다.

이러한 가정으로부터 응답자들이 민감한 그룹에 속한다고 응답할 확률은

$$\begin{aligned} \lambda &= P(R = 1) \\ &= \pi(1-\delta) + (1-\pi)\beta \end{aligned} \tag{2-2}$$

가 된다. 그런데 가정③으로부터 $\sum_{i=1}^{n} R_i$ 는 $b(n, \lambda)$가 되므로 거짓응답으로부터 π 의 최우추정량 $\hat{\pi}_r$ 을 얻는다.

$$\hat{\pi}_r = \frac{1}{n}\sum_{i=1}^{n} R_i \tag{2-3}$$

$\hat{\pi}_r$ 을 π의 추정량으로 사용하면 편의가 발생한다. 즉,

$$\lambda - \pi = (1 - \pi)\beta - \pi\delta \tag{2-4}$$

따라서 평균제곱오차 (MSE)는 다음과 같다.

$$MSE(\hat{\pi}_r) = \frac{\lambda(1-\lambda)}{n} + \{(1-\pi)\beta - \pi\delta\}^2 \tag{2-5}$$

(2-5)식은 Warner(1965, p67)의 경우와 같게 된다.

3. 확률화응답모형에서 거짓응답의 효과

2장에서는 직접질문에서의 거짓응답의 효과를 알아보았다. 그러나 질문이 민감한 질문일 수록 거짓응답의 가능성이 커지게 되므로 이를 방지하기 위하여 Warner(1965)가 확률장치를 이용한 확률화응답기법을 제시하였다. 그는 거짓응답의 가능성이 큰 민감한 질문에서는 확률응답모형이 바람직하다는 것을 입증하였다. 한편 Liu와 Chow

(1976)는 Warner모형에 대하여 확률장치를 $m(>1)$번 반복 시행한 후 그때까지 "예"라고 응답한 횟수를 가지고 모비율 π를 추정하였다. 이는 응답자로 하여금 시행을 여러번 반복하게 함으로써 모비율 π에 대한 정보를 더 많이 얻고, 또한 표본의 수를 증가시키지 아니하고 분산을 감소시켜줌으로써 효율을 높여 주었다. 그러나 반복 횟수를 증가시킴에 따라 거짓응답의 가능성은 커지게 된다.

여기서는 강요질문모형을 반복 사용하여 응답자들이 진실되게 응답했는 가에 대한 가설검정을 하였고 Warner의 확률응답모형을 반복 사용한 경우와 검정력을 비교하였다. 또한 강요질문모형을 사용하여 민감한 특성에 대한 모집단비율을 추정하였고 이 추정량이 Warner(1965)의 추정량이나 Krishnamoorthy와 Raghavarao (1991)가 Warner의 확률응답모형을 반복 사용하여 얻은 추정량보다 효율적이 되는 조건을 찾았다.

3-1. 가설검정

확률화응답모형의 사용여부에 대한 하나의 기준이 될 수 있는 거짓응답 확률에 관하여 Krishnamoorthy 와 Raghavarao(1991) 그리고 Lakshmi 와 Raghavarao(1992)는 다음과 같은 Warner의 확률응답모형을 반복 사용한 2×2분할표를 이용하여 응답자들이 진실되게 응답했는 지의 여부를 검정하였고 민감한 특성에 대한 모집단비율을 추정하였다.

(설문1) "나는 민감한 그룹에 속한다"
(설문2) "나는 민감한 그룹에 속하지 않는다"

본 논문에서는 Warner모형대신 다음과 같은 강요질문을 사용하였을 경우의 가설검정을 실시한다.

(설문1) "나는 민감한 그룹에 속한다"
(설문2) "예" 라고 응답

Liu와 Chow(1976)의 반복 시행에서와 같이 설문1과 설문2가 선택될 확률이 각각 p와 $1-p$인 확률장치를 두번 반복 시행하여 첫번째와 두번째의 응답확률변수 X_i 와 Y_i 를 i번째 응답자가 "예" 라고 응답하면 1 이고 "아니오" 라고 응답하면 0 이라 정의한다. 일반적으로 확률응답모형에서 p는 0.5보다 작다고 간주할 수 있으므로 여기서도 p가 0.5보다 작다고 가정한다. 그리고 민감한 그룹에 속하지 않는 사람이 거짓응답할 확률은 0으로 가정한다. 즉,

$$\beta = P(R = 1 \mid \overline{C}) = 0$$

이때, X_i 와 Y_i 는 독립이므로

$$\begin{aligned} P(X_i=1,\ Y_i=0) &= \pi P(X_i=1,\ Y_i=0 \mid C) + (1-\pi)P(X_i=1,\ Y_i=0 \mid \overline{C}) \\ &= \pi\delta p(1-\delta p) + (1-\pi)p(1-p) \end{aligned} \tag{3-1}$$

이된다. 그리고 $P(X_i=0,\ Y_i=1)$도 $P(X_i=1,\ Y_i=0)$와 같은 값을 가지므로 두번의 응답 $(X_i,\ Y_i)$이 일치하지 않을 확률은

$$\begin{aligned} \theta_f &= P(\text{불일치}) \\ &= P(X_i=1,\ Y_i=0) + P(X_i=0,\ Y_i=1) \\ &= 2[\ \pi\delta p(1-\delta p) + (1-\pi)p(1-p)\ ] \end{aligned} \tag{3-2}$$

이 된다. 한편, Warner모형을 두번 사용한 경우 두번의 응답이 일치하지 않을 확률은 (Krishnamoorthy와 Raghavarao, 1991) 다음과 같다.

$$\theta_{2w} = 2[\ p(1-p) + \delta(1-\delta)\pi(1-2p)^2\ ] \qquad (3\text{-}3)$$

단순임의추출한 n명의 응답자들에게 강요질문모형을 두번 반복하였을 경우의 응답이 다음과 같을 때

Y \ X	1	0	
	n_{11}	n_{01}	
	n_{10}	n_{00}	
			n

이로부터 거짓응답확률 $\delta = 0$ 에 대한 검정을 실시한다. 즉, 귀무가설($H_0 : \delta = 0$) 하에서 $(n_{01} + n_{10})$ 는 $b(n, \theta_{f_0})$에 따르므로 (3-2)식으로 부터

$$\begin{aligned} \theta_{f_0} &= P(\text{ 불일치 } | \ \delta=0\ ) \\ &= 2(1-\pi)p(1-p) \end{aligned} \qquad (3\text{-}4)$$

가 된다. 이는 π와 p값에 의존하게 되므로 대립가설 $H_1 : \delta > 0$ 에 대한 귀무가설 $H_0 : \delta = 0$ 을 검정하기 위한 기각역은 $n_{01} + n_{10} > b_\alpha$ 이고 b_α 는 모수 n 과 θ_{f_0} 를 갖는 이항분포의 $100(1-\alpha)$분위수가 된다. 한편 n이 큰 경우의 기각역은

$$\{\ \frac{(n_{01}+n_{10})}{n} - \theta_{f_0}\} \ / \ \{\ \frac{\theta_{f_0}(1-\theta_{f_0})}{n}\ \}^{1/2} \geq z_\alpha$$

가 된다. 이때 z_α 는 표준화정규분포에서의 $100(1-\alpha)$분위수이다.

Warner모형과 본 논문에서 다룬 강요질문모형을 반복 사용한 경우 각각의 검정력을 n=30과 n=1,000인 경우에 계산한 결과가 표1과 표2에 있다.

표1과 표2는 각각 이항분포와 표준화정규분포를 이용하였다. 두 확률화응답모형을 사용하였을 경우의 검정력은 π와 δ가 증가함에 따라 모두 증가한다. 그러나 Warner 모형을 사용한 경우는 p가 작을 때는 검정력이 좋으나 p가 증가함에 따라 검정력은 급격히 감소하는 반면에 강요질문모형을 사용한 경우에는 p가 증가함에 따라 검정력이 증가함을 알 수 있다. 따라서 개인의 사생활을 보장해 주기 위하여 p를 어느 정도 크게 한 경우 Warner모형을 사용한 경우보다 강요질문모형의 사용이 바람직하다고 본다.

귀무가설이 기각된 경우에 거짓응답확률 δ의 추정식은 (3-2)식으로 부터 얻을 수 있다. 즉,

$$\begin{aligned} E[(n_{10} + n_{01}) / n] &= 2[\pi\delta p(1-\delta p) + (1-\pi)p(1-p)\ ] \\ &= \theta_{f_0} + 2\pi\delta p(1-\delta p) \end{aligned} \qquad (3\text{-}5)$$

(3-5)식으로부터 거짓응답확률 δ의 추정량은 다음과 같다.

$$\hat{\delta} = \frac{1}{2p} - \left[\frac{1}{4p^2} - \frac{\{ (n_{10} + n_{01})/n - \theta_{f_0} \}}{2\pi p^2} \right]^{1/2} \quad (3\text{-}6)$$

표1. n=30 일때 π, p 및 δ 의 변화에 따른 검정력
(괄호안은 Warner모형일 경우의 검정력)

π / p / δ	0.1			0.3			0.5		
	0.1	0.3	0.4	0.1	0.3	0.4	0.1	0.3	0.4
0	.1001	.0604	.0966	.0751	.0735	.0951	.1277	.0806	.0833
	(.0763)	(.0754)	(.0668)	(.0763)	(.0754)	(.0668)	(.0763)	(.0754)	(.0668)
.1	.1057	.0689	.1121	.0913	.1085	.1497	.175	.1591	.1888
	(.1044)	(.0802)	(.0678)	(.1773)	(.0903)	(.0699)	(.2701)	(.1012)	(.0721)
.2	.1114	.0776	.1278	.1092	.1502	.2139	.2273	.2605	.3268
	(.1301)	(.084)	(.0686)	(.2872)	(.1032)	(.0724)	(.4835)	(.1252)	(.0764)
.3	.1172	.0865	.1433	.1284	.1967	.283	.2828	.3732	.4719
	(.1505)	(.0868)	(.0692)	(.377)	(.1131)	(.0743)	(.6345)	(.1445)	(.0796)
.4	.123	.0953	.1583	.1489	.2461	.3525	.3399	.4852	.6028
	(.1636)	(.0885)	(.0695)	(.4332)	(.1194)	(.0754)	(.715)	(.157)	(.0816)
.5	.1289	.1041	.1724	.1705	.2965	.4182	.3971	.5873	.7085
	(.1681)	(.0891)	(.0697)	(.4521)	(.1216)	(.0758)	(.7395)	(.1613)	(.0822)

표2. n=1,000 일때 π, p 및 δ 의 변화에 따른 검정력
(괄호안은 Warner모형일 경우의 검정력)

π / p / δ	0.1			0.3			0.5		
	0.1	0.3	0.4	0.1	0.3	0.4	0.1	0.3	0.4
0	.0495	.0495	.0495	.0495	.0495	.0495	.0495	.0495	.0495
	(.0495)	(.0495)	(.0495)	(.0495)	(.0495)	(.0495)	(.0495)	(.0495)	(.0495)
.1	.0704	.1026	.1236	.1439	.3332	.4579	.2978	.7197	.8722
	(.2466)	(.0716)	(.0543)	(.8682)	(.1371)	(.0651)	(.9974)	(.2345)	(.0775)
.2	.0966	.1816	.2396	.3051	.7512	.8982	.6817	.994	.9989
	(.5137)	(.0934)	(.0583)	(.9989)	(.2537)	(.0798)	(1.000)	(.4961)	(.1067)
.3	.1283	.2819	.382	.5052	.9545	.9938	.9167	1.000	1.000
	(.7032)	(.1118)	(.0614)	(1.000)	(.3608)	(.0917)	(1.000)	(.6909)	(.1318)
.4	.1654	.3936	.5261	.6944	.9955	.9998	.9867	1.000	1.000
	(.7961)	(.124)	(.0632)	(1.000)	(.4314)	(.0994)	(1.000)	(.7889)	(.1488)
.5	.2037	.5051	.6518	.8361	.9997	1.000	.9986	1.000	1.000
	(.8225)	(.1283)	(.0639)	(1.000)	(.4556)	(.1021)	(1.000)	(.817)	(.1547)

3-2. π의 추정

민감한 특성의 모집단비율 π를 추정하기 위하여 거짓응답확률을 $\delta = 0$로 가정한다. 즉, 귀무가설 $H_0 : \delta = 0$.

본 논문에서 제시한 강요질문모형을 두번 사용한 경우, 응답이 일치하지 않는 사람수 $(n_{01} + n_{10})$ 가 $b(n, \theta_{f_0})$ 을 하므로 θ_{f_0}에 대한 최우추정량은 다음과 같다.

$$\hat{\theta}_{f_0} = \frac{(n_{01} + n_{10})}{n} \tag{3-7}$$

최우추정량의 불변성에 의하여 (3-4)식과 (3-7)식으로부터 다음과 같은 π의 최우추정량과 추정량의 분산을 얻는다.

$$\hat{\pi}_f = 1 - \frac{(n_{10} + n_{01}) / 2n}{p(1-p)} \tag{3-8}$$

$$\begin{aligned} V(\hat{\pi}_f) &= \frac{1}{4n^2p^2(1-p)^2} n2(1-\pi)p(1-p)[1 - 2(1-\pi)p(1-p)] \\ &= \frac{\pi(1-\pi)}{n} + \frac{(1-\pi)\{0.5 - p(1-p)\}}{np(1-p)} \end{aligned} \tag{3-9}$$

한편 확률장치 R_1에 기초한 Warner의 추정량과 추정량의 분산은 다음과 같다.

$$\hat{\pi}_w = \frac{(n_{11} + n_{10}) / n - (1-p)}{2p-1} \tag{3-10}$$

$$V(\hat{\pi}_w) = \frac{\pi(1-\pi)}{n} + \frac{p(1-p)}{n(2p-1)^2} \tag{3-11}$$

그리고 Warner모형을 두번 사용한 경우(Krishnamoorthy와 Raghavarao,1991) 적률법에 의한 π의 추정량과 추정량의 분산을 구하면 다음과 같다.

$$\hat{\pi}_{2w} = \frac{n_{11} / n - (1-p)^2}{(2p-1)} \tag{3-12}$$

$$\begin{aligned} V(\hat{\pi}_{2w}) = &\frac{\pi(1-\pi)}{n} + \\ &\frac{1}{n(2p-1)^2}[2\pi p(1-p)(2p-1) + (1-p)^2 + (1-p)^4] \end{aligned} \tag{3-13}$$

(3-9)식과 (3-11)식으로부터 $V(\hat{\pi}_f) < V(\hat{\pi}_w)$ 이 되는 조건은

$$1 - \frac{p^2(1-p)^2}{(2p-1)^2\{0.5 - p(1-p)\}} < \pi \tag{3-14}$$

이고 (3-9)식과 (3-13)식으로부터 $V(\hat{\pi}_f) < V(\hat{\pi}_{2w})$ 이 되는 조건은 다음과 같다.

$$\frac{A + B^2 - 2B/p}{A + 2B} < \pi \tag{3-15}$$

여기서, $A = 1 / 2p(1-p) - 1$ 이고 $B = p(1-p) / (2p-1)$ 이다.

4. 맺 음 말

개인의 사생활이나 이해와 관련되는 민감한 질문에 대하여 응답자들이 자신의 신분이 보장되지 않는다고 생각함으로써 무응답이나 거짓응답을 하게 된다.

본 논문에서는 응답자들이 진실되게 응답했는 가에 대한 가설검정과 거짓응답확률및 민감한 모집단비율을 추정하기 위하여 Warner의 확률응답모형대신 강요질문모형을 반복 사용하였다.

Warner모형을 사용한 경우의 검정력은 민감한 질문이 선택될 확률 p가 작을수록 좋은 반면에 강요질문을 사용한 경우는 p가 클수록 검정력이 좋아짐을 알 수 있다. 그리고 강요질문을 사용하여 얻어진 모집단비율 π에 대한 추정량이 Warner모형을 사용하여 얻은 추정량들보다 효율적이되는 조건을 찾았다.

참고 문헌

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

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

[3] 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.

[4] Duffy, J. C. and Waterton, J. J. (1988), "Randomized Response vs. Direct Questioning : Estimating the Prevalence of Alcohol Related Problems in a Field Survey," *The Australian Journal of Statistics*, **30**(1), 1-14.

[5] Krishnamoorthy, K. and Raghavarao, D. (1991), "Unthruthful Answering in Repeated Randomized Response Procedures," submitted to *Canadian Journal of Statistics*.

[6] Lakshmi, D. V. and Raghavarao, D. (1992), "A Test for Detecting Untruthful Answering in Randomized Response Procedure," *Journal of Statistical Planning and Inference*, **31**, 387-390.

[7] Liu, P. T. and Chow, L. P. (1976), "The Efficiency of the Multiple Trial Randomized Response Technique," *Biometrics*, 32, 607-618.

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

[9] Tenebein, A. (1970), "A Double Sampling Scheme for Estimating from Binomial Data with Misclassifications," *Journal of the American Statistical Association*, **65**, 1350-1361.

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

An Alternative Randomized Response Technique[1)]

Jea-Bok Ryu[2)]

Abstract

In this paper, we consider the test based on using Forced question model instead of Warner model and compare the power of two randomized respose models. The estimator for the proportion of the individuals belonging to the sensitive group is obtained by using Forced question model and the conditions that the estimator by Forced question model will be more efficient than the estimators by Warner model are found when the respondents are truthful in their answers.

1) This research was supported by Korea Research Foundation, 1991

2) Department of Applied Statistics, Chongju University, 36 Naedok-dong, Chongju-si, Chungbuk, 360-764, Korea