한국통계학회논문집 제4권 2호
1997년 8월 pp. 411 - 419

확률화 응답모형의 한계에 대한 고찰

박 진 우1)

요 약

본 연구에서는 확률화응답모형이 가지는 두가지 한계에 대하여 고찰하였다. 첫째 민감한 속성을 갖는 모비율의 추정시 모비율 π_A가 매우 작은 값일 경우, 즉 희귀속성일 경우 확률화응답모형을 적용하게 되면 비밀보장의 효과를 감안한다고 해도 직접질문법에 비해 비효율적일 수 있음을 지적하였다. 둘째로 비밀보장에서 오는 이점과 그로 인한 효율의 손실이라는 서로 상충되는 면을 객관적으로 고려하는데 있어서 한계가 있음을 지적하였다.

1. 서 론

사람들을 대상으로 하는 표본조사에서 응답자들이 응답하기를 꺼리는 경향이 있는 민감한 문제에 대해 조사를 하는 경우 솔직하지 못한 응답이나 또는 무응답으로 인해 생기게 되는 비표본오차가 심각한 문제가 될 수 있다. 가령 마약의 복용여부 또는 낙태수술의 경험여부나 음주운전여부, 절도여부 등을 묻는 조사가 이에 해당된다.

Warner(1965)는 이러한 문제의 해결을 위해 응답자의 비밀을 어느 정도 보장해줌으로 보다 정직한 대답을 얻을 수 있도록 하는 확률화응답모형(randomized response model)을 제안하였다. 원래 Warner는 이지모형에서의 모비율 π_A를 추정하는 문제에 대해 이 방법을 적용하였다. 이후 수많은 연구자들에 의해 이 방법은 다양하게 연구, 발전되어져서 다지모형과 연속변량모형 등으로 확장되었다.. 그 밖에도 다양한 장치 및 추정방법들에 대한 연구가 활발하게 이루어져 왔다. 한편 Chaudhuri와 Mukerjee(1988), 류제복 외(1993)는 이 분야의 연구들을 종합, 정리하였다.

확률화응답모형은 기존의 직접질문법에 비해 응답자의 비밀을 보장해 줌으로써 보다 진실한 응답을 얻으려 하는 새로운 발상의 조사방법이다. 모비율의 추정에 관한 확률화응답모형의 기존연구들은 주로 여러가지 추정량들의 효율을 분산이나 평균제곱오차(mean square error : MSE) 등을 통해 비교하였다. 일반적으로 응답자들이 정직한 대답을 꺼려하는 경우에는 확률화응답모형을 사용하여 추정하는 것이 직접질문법에 비해 효과적임을 보여주고 있다. 그러나 이러한 확률화응답모형의 효용은 모비율 π_A 값과 상관없이 항상 동일한 것은 아니다. 그런데도 π_A 값이 달라짐에 따라 확률화응답모형의 효용도 달라진다는 사실에 대해서는 별다른 논의가 이루어지지 않았다. 또한 확률화응답모형은 응답자의 비밀을 어느 정도 보장해 준다는 장점이 있는 반면, 그 대가로 추정의

1) (440-600)경기 수원시 수원우체국 사서함 77 수원대학교 응용통계학과 조교수.

효율면에서의 손실을 야기시키게 된다. 따라서 비밀보장이라는 측면에서 얻게되는 유익이 효율의 손실에 비해 상대적으로 월등하다고 판단될 때에야 비로소 이 방법은 유용하다고 볼 수 있다. 확률화응답모형에서 비밀보장의 정도는 확률화장치에서 민감한 질문을 선택할 확률인 P 값에 의해 조절된다고 볼 수 있다. P 값이 달라짐에 따라 확률화응답모형에서 추정량의 효율에 변화가 생기게 되는데 이 변화가 어떤 양상을 띠는 지에 대해서도 별다른 논의가 없었다.

본 연구에서는 먼저 모비율이 달라짐에 따라 직접질문법과 확률화응답모형의 상대효율이 어떻게 변하는지를 고찰하였다. 아울러 확률화응답모형에서 생기는 비밀보장과 효율감소의 상충문제를 관찰해 보았다. 2절에서는 모비율이 달라짐에 따르는 확률화응답모형의 효율의 변화에 대해 다루었다. 3절에서는 응답자의 비밀보장과 효율의 상충문제를 살펴 보았다.

2. 모비율에 따른 확률화응답모형의 효율

모집단내에서 어떤 속성 A를 지니는 부분의 비율 π_A를 추정하고자 한다. 이 속성은 사람들이 일반적으로 민감하게 느끼는 속성이라고 하자.

"당신은 속성 A를 지니고 있습니까?"

라는 질문을 표본으로 뽑힌 사람에게 직접 물어서 응답을 얻는 방법이 일반적인 직접질문법이다. 반면 Warner(1965)는 확률화장치를 사용하여 P의 확률로는 아래의 질문1에 응답하게 하고, 1-P의 확률로는 질문2에 응답하게 하는 확률화응답모형을 소개하였다.

질문1 : "당신은 속성 A를 지니고 있습니까?"

질문2 : "당신은 속성 A^c를 지니고 있습니까?"

한편 Greenberg 외(1969)는 Warner모형의 질문2 대신에 다음과 같은 질문을 사용하였다.

질문2 : "당신은 속성 Y를 지니고 있습니까?"

이 때 속성 Y는 속성 A와 전혀 무관하며 민감하지 않은 속성이어야 하는데 이와 같은 질문을 사용하는 방법을 무관질문형모형(unrelated question model)이라고 한다. 한편 이 때 질문1에 응답할 확률을 Q라고 하자.

모집단에서 n개의 표본을 단순임의추출법으로 추출하여 조사하는 경우를 생각하는데 모든 응답자들이 정직하게 대답한다고 가정하자. 응답자들의 대답을 각각 X_i라고 표시하고 '예'라고 대답하는 경우에는 1, '아니오'라고 대답하는 경우에는 0의 값을 부여하면 직접질문법에 의한 모비율의 추정량과 그 분산은 각각 식(2.1), (2.2)와 같다.

$$\widehat{\pi_{A,D}}=\frac{\sum X_i}{n} \tag{2.1}$$

$$Var(\widehat{\pi_{A,D}})=\frac{\pi_A(1-\pi_A)}{n} \tag{2.2}$$

그러나 이와 같은 문제에서는 민감한 속성을 가진 사람들이 정직하게 응답하지 않을 가능성이 많다. 민감한 속성을 지닌 사람이 정직하게 응답할 확률을 T_a라고 한다면 위의 추정식 (2.1)은 불편추정량이 되지 못한다. 이 때의 편의와 추정량의 분산은 아래의 식 (2.3), (2.4)와 같다.

$$bias(\widehat{\pi_{A,D}})=\pi_A(T_a-1), \tag{2.3}$$

$$Var(\widehat{\pi_{A,D}}) = \frac{\pi_A T_a (1-\pi_A T_a)}{n} . \tag{2.4}$$

Warner의 모형을 사용했을 때 응답자들이 모두 정직하게 응답한다고 가정한다면 그 추정량의 분산의 식은 (2.5)와 같다.

$$Var(\widehat{\pi_{A,W}}) = \frac{\lambda(1-\lambda)}{n(2P-1)^2} . \tag{2.5}$$

여기서 $\lambda = P\pi_A + (1-P)(1-\pi_A)$ 이다. Warner모형을 사용하는 경우에도 민감한 속성을 갖는 응답자가 부정직하게 응답할 수 있을 것을 고려하여 정직한 응답의 확률을 $T_{a'}$이라고 하자. 이 경우 편의와 분산의 식은 각각 (2.6), (2.7)식으로 바뀐다.

$$bias(\widehat{\pi_{A,W}}) = \pi_A(T_{a'} - 1) \tag{2.6}$$

$$Var(\widehat{\pi_{A,W}}) = \frac{\lambda'(1-\lambda')}{n(PT_{a'} + P - 1))^2} \tag{2.7}$$

여기서 $\lambda' = P\pi_A T_{a'} + (1-P)(1-\pi_A)$이다.

무관질문형모형을 사용했을 때 응답자들이 모두 정직하게 응답한다고 가정한다면 그 추정량의 분산의 식은 (2.8)과 같다.

$$Var(\widehat{\pi_{A,U}}) = \frac{\lambda(1-\lambda)}{n\ Q^2} \tag{2.8}$$

여기서 $\lambda = Q\pi_A + (1-Q)\pi_Y$ 이다. 무관질문에서 속성 Y의 비율 π_Y는 알려져 있다고 한다.또한 무관질문형모형을 사용하는 경우에도 민감한 속성을 갖는 응답자가 부정직하게 응답할 수 있을 것을 고려하여 정직한 응답의 확률을 $T_{a'}$이라고 하자. 이 경우 편의와 분산의 식은 각각 (2.9), (2.10)식으로 바뀐다.

$$bias(\widehat{\pi_{A,W}}) = \pi_A(T_{a'} - 1) \tag{2.9}$$

$$Var(\widehat{\pi_{A,U}}) = \frac{\lambda'(1-\lambda')}{n(QT_{a'})^2} \tag{2.10}$$

여기서 $\lambda' = Q\pi_A T_{a'} + (1-Q)\pi_Y$ 이다.

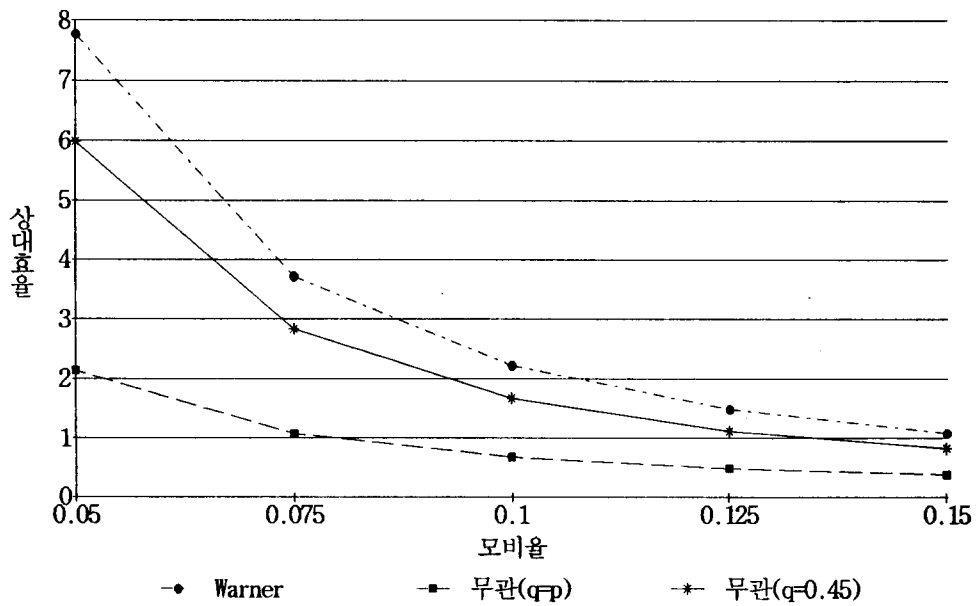
Warner(1965)나 Greenberg 외(1969)는 직접질문법에 대해서는 응답자들이 거짓응답을 할 가능성이 있지만, 확률화응답모형을 사용하면 모두 정직하게 응답한다고 가정하고 거짓응답이 있는 직접질문법과 확률화응답모형의 효율을 MSE를 써서 비교한 결과 확률화응답모형이 유용한 방법이라는 것을 보였다. 하지만 그들은 π_A의 변화에 따라 이러한 효율이 어떻게 변하는 가에 대해서는 특별히 고려하지 않았다. 그러나 가령 민감한 속성 A의 비율이 매우 낮은 경우를 생각한다면, 이 때에는 속성 A를 지닌 사람들만 거짓응답을 할 우려가 있고 속성 A를 지니지 않은 대부분의 사람들은 정직하게 대답하리라고 기대할 수 있다. 따라서 이 경우 모든 사람에게 확률화응답모형을 적용한다면 결과적으로 불필요한 효율의 손실을 가져오게 되어 확률화응답모형에서 얻는 비밀보장의 효과에도 불구하고 전체적인 효율은 떨어져서 그 효용성이 의문시된다고 하겠다. 뿐만 아니라 π_A가 매우 낮을 때에는 확률화응답모형을 사용한다고 해도 속성 A를 지닌 사람이 정직하게 응답하지 않을 수도 있으므로 이러한 점도 고려되어야 할 것이다.

π_A가 0.025, 0.050, 0.075, 0.100, 0.125, 0.150인 각각의 상황에 대해서 직접질문법과 확률화응답

모형의 MSE를 이용한 상대효율의 비교결과가 아래의 <표 1>에 나와 있다. MSE는 각각 위의 식 (2.3)-(2.10)을 이용하여 계산하였다. 표본의 크기 n은 1000인 경우를 고려했으며 Warner모형인 경우 확률화장치에서 민감한 질문을 선택할 확률 P가 0.7, 0.8인 경우를 생각하였다. 한편 무관질문형 모형인 경우 민감한 질문을 선택할 확률 Q를 한 번은 Warner모형에서의 P값과 같게 주었고, 또 한 번은 Fligner외(1977)가 제안한 바대로 Warner모형과 무관질문형모형의 일차적인 보호(primary protection)정도를 같게 해주는 값으로서 $1-Q=(1-P)\ /\ [(1-P)+(2P-1)\pi_y]$를 만족시키는 Q의 값을 주었다. 무관질문모형의 경우 무관질문속성의 비율 π_Y는 알고 있다고 가정했으며 π_Y값이 클수록 효과적이라는 Lanke(1975)의 제안에 따라 그 값이 0.9 인 경우를 생각하였다. 또한 민감한 속성 A를 지니지 않은 응답자들은 어떤 방법을 사용하든 정직하게 대답한다는 것을 가정하였다. 민감한 속성을 지닌 응답자들이 직접질문법에서 정직하게 대답할 확률 T_a는 0.5, 0.7, 0.9로 변화시켜가며 고려하였고, 확률화응답모형을 사용할 때 정직하게 대답할 확률인 $T_{a'}$과 직접질문법의 T_a의 차이가 각각 0.1, 0.2, 0.3, 0.4, 0.5 일 때를 고려하였다.

<표 1>을 보면 π_A와 민감한 질문을 선택할 확률인 P와 Q값, 정직하게 응답할 확률인 T_a와 $T_{a'}$ 값등이 상대효율에 영향을 미치는 요소임을 알 수 있다. 이 표를 통해서 관찰되는 바를 요약하면 다음과 같다.

1) 표에서 값이 1보다 크면 직접질문법보다 확률화응답모형의 효율이 떨어지는 것을 나타내는데, π_A값이 작을수록 직접질문법에 대한 확률화응답모형의 상대효율은 나빠진다는 사실을 알 수 있다. 아래의 [그림 1]은 P=0.7, T_a=0.7, $T_a-T_{a'}$=0.2 인 경우 π_A값의 변화에 따른 직접질문법과 확률화응답모형의 상대효율을 나타낸 것이다. Warner형이든 무관질문형이든 모두 π_A값이 작아질수록 직접질문법에 비해 상대효율이 나빠진다.



[그림 1] π_A값의 변화에 따른 상대효율의 비교

<표 1> π_A와 T_a및 $T_{a'}-T_a$의 변화에 따른 직접질문법과 확률화응답모형과의 상대효율

π_A	T_a	$T_{a'}-T_a$	P = 0.7 Warner	P = 0.7 무관(Q=0.7)	P = 0.7 무관(Q=0.45)	P = 0.8 Warner	P = 0.8 무관(Q=0.8)	P = 0.8 무관(Q=0.27)
0.025	0.5	0.1	87.58	7.38	20.93	13.01	4.59	51.23
		0.2	35.15	5.34	15.28	7.90	3.29	37.50
		0.3	18.80	3.99	11.59	5.23	2.43	28.57
		0.4	11.65	3.09	9.08	3.71	1.85	22.47
		0.5	7.93	2.48	7.32	2.78	1.48	18.16
	0.7	0.1	43.15	9.17	26.60	12.05	5.58	65.60
		0.2	26.75	7.09	20.84	8.53	4.26	51.60
		0.3	18.20	5.69	16.81	6.38	3.40	41.69
	0.9	0.1	47.33	14.80	43.70	16.60	8.85	108.40
0.050	0.5	0.1	23.33	2.42	5.90	3.92	1.70	13.69
		0.2	9.46	1.68	4.22	2.37	1.16	9.93
		0.3	5.05	1.18	3.12	1.53	0.78	7.48
		0.4	3.10	0.86	2.38	1.03	0.54	5.81
		0.5	2.09	0.67	1.90	0.76	0.41	4.67
	0.7	0.1	12.68	2.96	7.84	3.83	1.96	18.76
		0.2	7.77	2.14	5.98	2.60	1.36	14.59
		0.3	5.26	1.67	4.77	1.90	1.04	11.71
	0.9	0.1	20.01	6.36	18.14	7.24	3.94	44.58
0.075	0.5	0.1	10.93	1.45	3.00	2.15	1.13	6.47
		0.2	4.49	0.97	2.10	1.29	0.73	4.63
		0.3	2.39	0.63	1.49	0.80	0.45	3.43
		0.4	1.44	0.42	1.09	0.51	0.28	2.61
		0.5	0.96	0.31	0.85	0.36	0.20	2.08
	0.7	0.1	6.19	1.64	3.87	2.06	1.18	8.89
		0.2	3.72	1.08	2.84	1.31	0.72	6.78
		0.3	2.49	0.80	2.21	0.92	0.51	5.39
	0.9	0.1	11.59	3.74	10.32	4.31	2.39	25.16
0.100	0.5	0.1	6.48	1.11	1.97	1.51	0.93	3.92
		0.2	2.72	0.71	1.34	0.90	0.58	2.76
		0.3	1.43	0.43	0.91	0.53	0.33	1.99
		0.4	0.84	0.26	0.63	0.31	0.18	1.48
		0.5	0.55	0.18	0.48	0.21	0.12	1.16
	0.7	0.1	3.78	1.14	2.41	1.40	0.88	5.25
		0.2	2.22	0.68	1.67	0.83	0.48	3.91
		0.3	1.45	0.47	1.27	0.55	0.31	3.07
	0.9	0.1	7.71	2.52	6.74	2.94	1.65	16.29
0.125	0.5	0.1	4.41	0.94	1.49	1.21	0.83	2.73
		0.2	1.88	0.59	0.99	0.71	0.51	1.89
		0.3	0.99	0.34	0.64	0.41	0.28	1.33
		0.4	0.56	0.18	0.42	0.22	0.14	0.96
		0.5	0.36	0.12	0.31	0.14	0.08	0.74
	0.7	0.1	2.63	0.90	1.71	1.08	0.74	3.54
		0.2	1.49	0.49	1.12	0.59	0.36	2.55
		0.3	0.96	0.32	0.82	0.37	0.21	1.97
	0.9	0.1	5.55	1.83	4.77	2.16	1.23	11.44
0.150	0.5	0.1	3.28	0.85	1.23	1.04	0.78	2.09
		0.2	1.43	0.52	0.79	0.61	0.46	1.42
		0.3	0.74	0.29	0.49	0.33	0.24	0.96
		0.4	0.41	0.14	0.30	0.17	0.11	0.67
		0.5	0.25	0.08	0.21	0.10	0.06	0.51
	0.7	0.1	1.99	0.77	1.33	0.90	0.65	2.59
		0.2	1.09	0.38	0.82	0.46	0.29	1.81
		0.3	0.68	0.23	0.57	0.27	0.15	1.37
	0.9	0.1	4.21	1.40	3.56	1.67	0.96	8.47

※ 상대효율 = MSE (확률화응답) / MSE (직접질문)

2) 직접질문법이나 확률화응답모형에서 정직하게 응답할 확률인 T_a , $T_{a'}$값에 따라서도 상대효율은 변하게 된다. $T_{a'}-T_a$값을 일정하게 고정시키면 T_a 값이 커질수록 확률화응답모형의 직접질문법에 대한 상대효율은 낮아지는 경향을 보인다. 또한 $T_{a'}-T_a$ 값의 차이가 줄어들수록 확률화응답모형의 상대효율은 낮아지게 된다. 이 사실은 π_A값이 작은 경우, 직접질문법을 사용하게 될 때 정직한 응답을 기대하기가 매우 어려운 반면 확률화응답모형을 사용할 때 정직한 응답의 가능성이 확연히 높아지는 경우에 한해서 확률화응답모형을 사용하는 것이 바람직하다는 것을 시사한다.

3) Warner모형에서 P의 값이 0.7일 때에는 π_A값이 0.075 이하, P의 값이 0.8일 때에는 대략 π_A값이 0.050 이하일 때 확률화응답모형의 효율이 항상 직접질문법의 효율에 비해 낮음을 알 수 있다. 그 밖의 경우에는 T_a 값이 매우 작고 반면 $T_{a'}-T_a$값이 0.3이상 차이가 나는 경우에 한해서만 확률화응답모형의 효율이 더 우수하나 그렇지 않은 경우에는 오히려 확률화응답모형의 효율이 더 떨어지는 것을 관찰할 수가 있다.

4) 무관질문형 모형에서는 Q의 값에 따라서 양상이 달라진다. Q의 값을 P의 값과 같게 할 때에는 Warner모형에 비해 무관질문형 모형이 훨씬 더 효율적이어서 π_A의 값이 0.05 아래일 경우에만 직접질문법보다 효율이 떨어지게 된다. 반면 Fligner외(1977)가 말한 바대로 두 모형간의 일차적인 보호를 같게 해주는 수준에서 Q를 정해주게 되면 Warner 모형의 P가 0.7일 때 무관질문형 모형의 Q는 0.45가 되고 Warner 모형의 P가 0.8일 때 Q는 0.27이 되는데 이 때에는 위의 3)에서 언급한 Warner모형의 경우와 비슷한 양상을 띠게 된다.

이상의 결과들을 통해 확률화응답모형과 직접질문법의 상대효율은 모비율 π_A의 값과 $T_{a'}-T_a$ 값에 따라 양상이 달라진다는 것을 알 수 있다. π_A값이 0.05 이하일 때에는 MSE의 측면에서 볼 때 확률화응답모형을 사용하는 것이 직접질문법을 사용하는 것에 비해 효율이 떨어진다. 또한 π_A 값이 대략 0.15 이하의 작은 값일 때에는, T_a값이 작고 동시에 $T_{a'}-T_a$ 값이 0.3 이상 현저히 차이가 나지 않는 한, 확률화응답모형은 직접질문법에 비해 그 효율이 떨어진다는 사실을 알 수 있다. 그 이유는 앞에서도 언급한 바와 같이 π_A값이 작은 모집단에 대한 표본조사에서 확률화응답모형을 사용하게 되면 민감한 속성을 지니지 않은 대부분의 응답자들에 대해서도 불필요하게 확률화장치를 사용함으로 인해 생겨나는 효율의 감소 때문이라고 할 수 있다.

3. 비밀보장과 효율의 문제

확률화응답모형의 기본적인 생각은 응답자가 민감하게 느낄 수 있는 문제에 대해 비밀을 보장해 줌으로 진실된 응답을 얻고 그 결과 보다 신뢰할 만한 추정값을 얻고자 하는 것이다. 직접조사법에서는 응답자가 주어진 문제에 대해 직접 대답할 것을 요구하는데 반해 확률화응답모형에서는 확률화장치(randomizing device)를 사용하여 몇 개의 문항중 어느 하나에 대답하도록 해 줌으로써 응답자에게 어느 정도 비밀을 보장해 주게 된다. 가령 단순한 이지모형에서 Warner의 모형을 생각해 보자. 이 때 π_A 의 추정량과 분산은 아래의 식 (3.1), (3.2)와 같다.

$$\widehat{\pi_A} = \frac{P-1}{2P-1} + \frac{n_1}{(2P-1)n} \tag{3.1}$$

$$Var(\widehat{\pi_A}) = \frac{\pi_A(1-\pi_A)}{n} + \frac{1}{n}\left[\frac{1}{16(P-0.5)^2} - \frac{1}{4}\right] \tag{3.2}$$

여기서 P는 확률화장치에서 민감한 속성의 질문을 선택할 확률이며, n은 표본의 크기, n_1은 표본중 '예'라고 대답한 사람의 수를 나타낸다. 위의 분산식 (3.2)에서 처음의 항은 직접조사법을 직용했을 때의 추정량의 분산에 해당되며 나중의 항은 확률화응답모형에서 응답자의 비밀을 보장해 주기 위해 초래하게 되는 분산의 증가분에 해당된다. 이 분산의 증가분은 P의 값이 0이나 1에 가까울수록 작아지고 0.5에 가까워 질수록 커진다. P의 값이 0이나 1인 경우는 직접조사법이 되며, P의 값이 0.5에 가까와지면 질수록 응답자의 비밀이 더 잘 보장된다. 위의 식을 통해 볼 때 응답자의 비밀을 보장해 주면 줄수록 추정량의 효율은 더 떨어지게 된다는 점을 알 수 있다. 따라서 확률화응답모형에서 응답자의 비밀을 보장해야 한다는 점과 추정량의 효율을 보다 높이고자 하는 서로 상충되는 요구를 어떻게 적절히 만족시켜야 할 지를 결정하는 문제는 중요한 논점이 된다.

Leysieffer와 Warner(1976)는 비밀보장의 정도를 알려주는 척도로서 위험함수(jeopardy function)라는 개념을 도입하여 다음과 같이 정의하였다. 여기서 R은 응답을 의미하며 $g(R,A)$는 응답 R의 속성 A에 대한 위험척도, $g(R,A^c)$는 응답 R의 속성 A^c에 대한 위험척도로 쓰여지고 있다.

$$g(R,A) = \frac{P(A \mid R)}{P(A^c \mid R)} \frac{1-\pi}{\pi} = \frac{P(R \mid A)}{P(R \mid A^c)} ,$$

$$g(R,A^c) = \frac{P(A^c \mid R)}{P(A \mid R)} \frac{\pi}{1-\pi} = \frac{P(R \mid A^c)}{P(R \mid A)} .$$

Leysieffer와 Warner는 속성 A를 지닌 사람이 '예' 라고 대답할 때의 위험척도를 $g(Y,A)$라고 하고, 속성 A^c를 지닌 사람이 '아니오' 라고 대답할 때의 위험척도를 $g(N,A^c)$라고 한다면

$$1 < g(Y,A) < K_1 ,$$

$$1 < g(N,A^c) < K_2$$

의 조건하에서 $VAR(\widehat{\pi_A})$를 최소로 하는 계획이 확률화응답모형에서 최소분산을 얻는 계획임을 보였다.

가장 단순한 경우로 속성 A가 민감한 속성일 때 A^c 은 전혀 민감하지 않다고 한다면, $K_2 = \infty$로 둘 수 있으므로 이 때 최소분산을 갖게 하는 P의 값은 $P = \frac{K_1}{K_1+1}$가 된다. 따라서 위험척도의 한계 K_1을 효과적으로 정할 수만 있다면 주어진 위험척도내에서 효율을 극대화할 수 있는 계획을 세울 수 있게 된다. 그러나 문제는 K_1을 어떻게 정할 수 있느냐 하는 것이다. 상황에 따라서 K_1의 값은 달라지게 될 것이며 그 값을 미리 정할 수 있는 어떠한 객관적인 기준도 없다. 그 결과 실제 확률화응답모형을 사용할 때에는 K_1을 결정한 후 거기에 따라서 P 값을 정하는 것이 아니라 처음부터 바로 P 값을 정하게 된다.

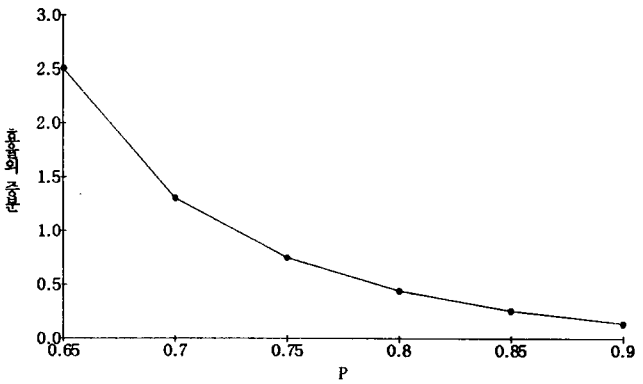
그러면 P의 값이 달라짐에 따라 효율에 어떤 변화가 생기는 지를 살펴 보자. 위의 식 (3.2)를

보면 Warner모형의 경우 확률화장치를 사용함으로 인해 잃게 되는 효율의 크기가 $\frac{1}{n}\left(\frac{1}{16(P-1/2)^2}-\frac{1}{4}\right)$ 임을 알 수 있다. 이 값은 P 값이 0.5와 아주 가깝지만 않다면,직접질문법에서 $T_a=1$일 때의 분산인 $\frac{\pi_A(1-\pi_A)}{n}$ 에 비해 상대적으로 매우 크므로 추정량의 효율은 주로 이 값에 의해 영향을 받게 된다. 아래의 <표 2>와 [그림 2]는 P 값이 0.65에서 0.90사이의 값일 경우 각각에 대한 효율의 증가분의 추이를 보여주고 있는데 P 값의 변화와 효율의 변화의 정도가 비례적이지 않음을 알 수 있다. P가 작은 값일수록 P값의 변화에 따른 효율의 변화의 정도가 심해지는 반면, P가 큰 값일 때에는 P값의 변화가 효율의 변화에 미치는 영향이 작아진다. 가령 확률화응답모형에서 P를 결정할 때 0.7과 0.75 중 어느 하나를 택하는 경우와 0.75와 0.8 중 어느 하나를 택하는 경우를 생각한다면 두 가지 경우 모두 P의 값의 차이는 0.05이지만 그로 인한 효율의 변화의 정도는 서로 다르다는 사실을 명심하여야 할 것이다.

<표 2> P 값과 효율의 증분

P	효율의 증가분 ($\times 10^{-3}$)
0.65	2.51
0.7	1.31
0.75	0.75
0.8	0.44
0.85	0.26
0.9	0.14

[그림 2]



한편 P가 달라짐에 따라 비밀보장의 정도의 변화가 어떤 양상을 띠게 되는지를 안다면 위에서 언급한 효율의 변화양상을 고려하여 최적의 절충점을 찾을 수 있을 것이다. 하지만 P의 변화에 따른 비밀보장의 정도를 계량적으로 제시하여 줄 수 있는 객관적인 장치가 없는 실정이므로 실제 확률화응답모형을 적용할 때에는 연구자 나름의 주관적인 판단에 의해 P를 결정해야 한다는 한계를 지니게 된다.

4. 결 론

본 연구에서는 확률화응답모형이 갖는 한계에 대해 두가지 사항을 생각해 보았는데 그것은 다음과 같이 요약될 수 있다.

1) 민감한 속성 A의 모비율 π_A를 추정하는 조사에서 직접질문법에 대한 확률화응답모형의 효율은 모비율 π_A값과 응답자들이 정직하게 대답할 확률에 따라서 달라진다. 민감한 속성의 비율이 매우 낮은 경우 확률화응답모형을 사용하면 민감한 속성을 지닌 응답자에 대한 비밀보장에서 오는 효과보다 속성을 지니지 않은 대다수 응답자들에게 확률화장치를 적용함으로 야기되는 효율의

저하의 영향이 더욱 커져서 그냥 직접질문법을 쓰는 것보다 효율이 떨어질 수도 있다.

2) 근본적으로 확률화응답모형은 응답자의 비밀보장을 위해 추정량의 효율을 희생하는 방법이다. 추정량의 효율은 비밀보장의 정도의 차이에 따라 민감하게 달라지므로 항상 비밀보장이라는 측면과 효율의 저하라는 측면의 균형을 고려하여 최적의 계획을 마련해야 한다. 그러나 현실적으로 최적의 기준을 마련할 수 있는 객관적인 장치를 갖고 있지 못하므로 상황에 따라 연구자가 주관적인 판단을 내려야 한다는 한계를 지니고 있다.

확률화응답모형이 본질적으로 가지는 위와 같은 한계들을 올바로 인식할 때 이 방법의 올바른 적용이 가능할 것이다.

참고 문헌

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

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

[3] Fligner,M.A., Policillo,G.W. and Singh,J. (1977). A Comparison of Two Randomized Response Survey Methods with Considerations for the Level of Respondent Protection, *Communications in Statistics-Theory and Methods,* A6(15), 1511-1524.

[4] 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.

[5] Lanke,J. (1975). On the Choice of Unrelated Question in Simmons' Version of Randomized Response, *Journal of the American Statistical Association,* 70,80-83.

[6] Leyseiffer,R.W. and Warner,S.L. (1976). Respondent Jeopardy and Optimal Designs in RR models, *Journal of the American Statistical Association,* 71,649-656.

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