응용통계연구 제 9권 2호
1996년 9월 pp. 135-143

반복시행 확률화 응답모형의 효율에 관한 연구

이 해 용[1], 강 현 철[2]

요 약

민감한 속성을 가진 그룹에 속하는 비율을 조사하고자 하는 경우에 응답자들의 응답기피나 거짓응답을 피하기 위한 조사방법들이 많은 연구자들에 의해서 개발되었다. 본 논문에서는 Warner(1965), Mangat & Singh(1990), Mangat(1994) 등이 제안한 방법들을 중심으로 확률화 응답모형을 반복시행함으로써 단일시행 모형에서의 단점을 보완하고 효율을 높일 수 있음을 보였다.

1. 소 개

낙태, 약물중독 등과 같은 민감한 속성을 가진 그룹에 속하는 비율을 조사하고자 하는 경우에 응답자들의 응답기피나 거짓응답은 편의를 발생시키는 중요한 원인이 된다. 따라서 응답기피나 거짓응답을 피하기 위한 조사방법들이 많은 연구자들에 의해서 개발되었다. Warner(1965)는 응답자의 비밀을 보장할 수 있는 확률화 응답장치(randomized response device)를 고안하였고, Mangat & Singh(1990)과 Mangat(1994)는 Warner(1965)의 장치에 비해 효율을 높일 수 있는 확률화 응답전략을 제안하였다. 이러한 확률화 응답기법들은 조사자에게 응답자의 비밀을 노출시키지 않는다는 장점을 가지고 있으나, 직접질문에 의한 조사에 비해 추정량의 분산이 상대적으로 크다는 단점이 있다. Liu & Chow(1976)는 Warner(1965)의 확률화 응답장치를 반복하여 사용함으로써 표본의 크기를 증가시키지 않고서도 추정량의 분산을 줄일 수 있음을 보였으며, 본 논문에서는 Mangat & Singh(1990)과 Mangat(1994)의 방법들을 반복하여 사용함으로써 효율을 높일 수 있음을 보이고 Liu & Chow(1976)의 반복시행모형과 비교하고자 한다.

2. 단일시행모형

2.1 Warner모형

Warner의 확률화 응답기법은 먼저 확률 p를 가진 부분(A)과 확률 $1-p$를 가진 부분(B)으로 구성된 확률장치(random device)로부터 응답자에게 A나 B가 선택되게 한다. 그 다음 A가

1) (136-742) 서울시 성북구 동선동 3가, 성신여자대학교 통계학과 교수
2) (136-701) 서울시 성북구 안암동 5가, 고려대학교 대학원 통계학과 박사과정 수료

선택된 응답자에게는 속성을 가지고 있으면 "예", 그렇지 않으면 "아니오"로 대답하게 하고, B가 선택된 응답자에게는 속성을 가지고 있으면 "아니오", 그렇지 않으면 "예"로 대답하게 한다. Warner의 확률화 응답기법으로부터 얻어지는 모비율 π의 최대우도 추정량은 n_1이 n명의 응답자들 중에서 "예"라고 대답한 사람의 수라고 할 때

$$\hat{\pi} = \frac{p-1}{2p-1} + \frac{n_1}{(2p-1)n} \quad , p \neq 1/2$$

이 된다. $\hat{\pi}$은 π의 불편추정량이며 분산은

$$Var(\hat{\pi}) = \frac{\lambda(1-\lambda)}{n(2p-1)^2}$$

이다.

2.2 Mangat-Singh모형

Mangat-Singh모형에서는 두 개의 확률장치가 이용된다. 장치 R_1은 선택될 확률이 각각 T와 $1-T$인 두 질문, (i)"나는 민감한 그룹에 속한다"와 (ii)"확률장치 R_2로 가시오", 으로 구성되어 있다. 장치 R_2는 Warner의 확률장치이며 선택될 확률이 p와 $1-p$인 두 질문, (i)"나는 민감한 그룹에 속한다"와 (ii)"나는 민감한 그룹에 속하지 않는다", 으로 구성되어 있다. n_1이 n명의 응답자들 중에서 "예"라고 대답한 사람의 수라고 할 때, 모비율 π의 최대우도 추정량은

$$\hat{\pi} = \frac{n_1/n-(1-T)(1-p)}{2p-1+2T(1-p)}$$

이다. $\hat{\pi}$은 π의 불편추정량이며 분산은

$$Var(\hat{\pi}) = \frac{\pi(1-\pi)}{n} + \frac{(1-T)(1-p)[\ 1-(1-T)(1-p)]}{n[\ 2p-1+2T(1-p)]^2}$$

이다.

Mangat-Singh모형에서 $T=0$이면 Warner모형과 동일하게 된다.

2.3 Mangat모형

Mangat모형에서는 응답자가 속성을 가지고 있으면 "예"라고 대답하게 하고, 속성을 가지고 있지 않으면 Warner의 확률장치를 이용하여 "예" 또는 "아니오"로 대답하게 한다. n_1이 n명의 응답자들 중에서 "예"라고 대답한 사람의 수라고 할 때, 모비율 π의 최대우도 추정량은

$$\hat{\pi} = (n_1/n-1+p)/p$$

이다. $\hat{\pi}$은 π의 불편추정량이며 분산은

$$Var(\hat{\pi}) = \frac{\pi(1-\pi)}{n} + \frac{(1-\pi)(1-p)}{np}$$

이다.

3. 반복시행과 모비율의 추정

Liu & Chow(1976)는 Warner의 확률화 응답장치를 반복하여 사용함으로써 표본의 크기를 증가시키지 않고서도 추정량의 분산을 줄일 수 있음을 보였으며 Mangat-Singh모형과 Mangat 모형에서도 반복시행에 의해 추정량의 분산을 줄일 수 있다.

세 모형을 각 응답자들에게 m번 반복시행하였을 때, j번째 응답자가 i번 "예"라고 응답한 것을 $X_j=i$라고 표현하면 j번째 응답자가 i번 "예"라고 응답할 확률은

① Warner모형일 때

$$\begin{aligned} P(X_j=i|m) &= \binom{m}{i}[\ \pi p^i(1-p)^{m-i} + (1-\pi)p^{m-i}(1-p)^i] \\ &= w_i \quad , j = 1, 2, \cdots, n; \quad i = 0, 1, 2, \cdots, m \end{aligned}$$

② Mangat-Singh모형일 때

$$\begin{aligned} P(X_j=i|m) = & \binom{m}{i}[\ \pi\{T+(1-T)p\}^i\{(1-p)(1-T)\}^{m-i} \\ & +(1-\pi)\{(1-p)(1-T)\}^i\{T+(1-T)p\}^{m-i}] \\ = & \ w_i \quad , j = 1, 2, \cdots, n; \quad i = 0, 1, 2, \cdots, m \end{aligned}$$

③ Mangat모형일 때

$$\begin{aligned} P(X_j=i|m) &= \binom{m}{i}(1-\pi)(1-p)^i p^{m-i} \\ &= w_i \ , i<m \ , j=1, \cdots, n \end{aligned}$$

$$P(X_j=m|m) = \pi + (1-\pi)(1-p)^m = w_m \ , \ j=1, \cdots, n$$

이다.

세 모형에서의 우도함수와 로그우도함수는

$$L = \prod_{i=0}^{m} w_i^{n_i} \ , \quad \log L = \sum_{i=0}^{m} n_i \log w_i \ ; \ \sum_{i=0}^{m} w_i = 1$$

이 된다. π의 최대우도 추정량 $\hat{\pi}$을 구하기 위하여 방정식 $\partial \log L/\partial \pi = 0$을 계산하는 것은 이 방정식이 폐쇄형으로 표현되지 않기 때문에 쉽지 않다. 이러한 경우에 반복적인 스코어 방법

(score method)에 의해서 $\hat{\pi}$을 구할 수 있다(Rao(1965)). 스코어 방법에 대하여 간단히 설명하면 다음과 같다. 먼저 π의 임의의 초기 추정치 $\pi_0(0 \le \pi_0 \le 1)$를 선택하여 π의 첫 번째 근사치 π_1을 아래와 같이 구한다.

$$\pi_1 = \pi_0 + S(\pi_0)/I(\pi_0)$$

여기에서 $S(\pi_0)$와 $I(\pi_0)$는 각각 π_0에서의 스코어(score)와 정보(information)이다. 다시 위 식의 우변에 있는 π_0대신에 π_1을 대입하여 두 번째 근사치 π_2를 얻는다. 변화량이 무시할 만큼 작아질 때까지 이러한 과정을 계속하여 최종적인 추정치 $\hat{\pi}$을 구한다. 이렇게 구한 $\hat{\pi}$은 π의 점근적(asymptotic) 불편추정량이 되며, $\hat{\pi}$의 분산은 점근적으로 정보의 역수가 된다. 즉,

$$var(\hat{\pi}) \simeq \frac{1}{I(\pi)}$$

세 반복시행 모형에서의 스코어와 정보를 유도하면 다음과 같다.

스코어 $S(\pi)$는

$$S(\pi) = \frac{\partial \log L}{\partial \pi} = \sum_{i=0}^{m} \frac{n_i}{w_i}\left(\frac{\partial w_i}{\partial \pi}\right)$$

이며, 세 모형에서의 스코어는 각각

① Warner모형일 때

$$S_1(\pi) = \sum_{i=0}^{m} \frac{n_i[\ p^i(1-p)^{m-i} - p^{m-i}(1-p)^i]}{\pi p^i(1-p)^{m-i} + (1-\pi)p^{m-i}(1-p)^i}$$

② Mangat-Singh모형일 때

$$S_2(\pi) = \sum_{i=0}^{m} \frac{n_i[\ \{T+(1-T)p\}^i\{(1-p)(1-T)\}^{m-i} - \{(1-p)(1-T)\}^i\{T+(1-T)p\}^{m-i}]}{\pi\{T+(1-T)p\}^i\{(1-p)(1-T)\}^{m-i} + (1-\pi)\{(1-p)(1-T)\}^i\{T+(1-T)p\}^{m-i}}$$

③ Mangat모형일 때

$$S_3(\pi) = \sum_{i=0}^{m-1} n_i\,(1-\pi) + n_m \frac{\pi + (1-\pi)(1-p)^m}{1-(1-p)^m}$$

이다.

$E(n_i) = n\ w_i$임을 이용하면 정보(information) $I(\pi)$는

$$I(\pi) = -E\left(\frac{\partial^2 \log L}{\partial \pi^2}\right) = n\sum_{i=1}^{m} \frac{1}{w_i}\left(\frac{\partial w_i}{\partial \pi}\right)^2$$

이며, 세 모형에서의 정보는

① Warner모형일 때

$$I_1(\pi) = n\sum_{i=0}^{m}\frac{\binom{m}{i}[\ p^i(1-p)^{m-i}-p^{m-i}(1-p)^i]^2}{\pi p^i(1-p)^{m-i}+(1-\pi)p^{m-i}(1-p)^i}$$

② Mangat-Singh모형일 때

$$I_2(\pi)=n\sum_{i=0}^{m}\frac{\binom{m}{i}[\ \{T+(1-T)p\}^i\{(1-p)(1-T)\}^{m-i}-\{(1-p)(1-T)\}^i\{T+(1-T)p\}^{m-i}]}{\pi\{T+(1-T)p\}^i\{(1-p)(1-T)\}^{m-i}+(1-\pi)\{(1-p)(1-T)\}^i\{T+(1-T)p\}^{m-i}}$$

③ Mangat모형일 때

$$I_3(\pi) = n\left(\sum_{i=0}^{m-1}\frac{\binom{m}{i}(1-p)^i\ p^{m-i}}{1-\pi}+\frac{(1-(1-p)^m)^2}{\pi+(1-\pi)(1-p)^m}\right)$$

이다.

4. 세 모형의 비교와 반복시행에서의 효율

<표 1>은 Warner모형, Mangat-Singh모형, Mangat모형의 정보, 즉, $I_1(\pi)$, $I_2(\pi)$, $I_3(\pi)$을 제시한 것이다. 모비율 π는 0.3으로 고정하였고 p는 0.2, 0.4, 0.6, 0.8, T는 0.1, 0.3, 0.5, 0.7, 0.9, 반복시행의 수 m은 1, 2, 3, 4, 5, 10 의 경우에 대하여 결과를 제시하였다. 정보는 점근분산의 역수이므로 정보가 클수록 효율이 좋다고 말할 수 있다. Warner모형은 p가 0또는 1에 가까울수록 효율이 좋고 0.5에 가까울수록 효율이 좋지 않다. Mangat모형은 p가 1에 가까울수록 효율이 좋고 0에 가까울수록 효율이 좋지 않다. Mangat-Singh모형은 T와 p가 1에 가까울수록 효율이 좋다. 동일한 p에 대해서 T가 0.9일 때의 Mangat-Singh모형의 효율이 가장 좋다. 일반적으로 Mangat-Singh모형의 추정량은

$$T>(1-2p)/(1-p)$$

일 때 Warner모형의 추정량보다 효율적이며, Mangat모형의 추정량은

$$\pi>1\ -\ \frac{p(1-T)\{1-(1-T)(1-p)\}}{\{2p-1+2T(1-p)\}^2}$$

일 때 Mangat-Singh모형의 추정량보다 효율적이다.

<표 2>는 세 모형에서 각 반복시행의 수에 대하여 시행의 수가 1인 경우에 대한 정보의 비 ($I_i(\pi|m)/I_i(\pi|m=1)$, m=2, 3, 4, 5, 10)를 나타낸 것이다. <표 1>과 <표 2>에서 모든 경우에 대하여 반복시행의 수가 커질수록 효율이 좋아짐을 알 수 있다. 또한 <표 2>는 효율이 좋지

<표 1> 세 모형의 정보($I_1(\pi)$, $I_2(\pi)$, $I_3(\pi)$)

π	p	T / m	Mangat	Warner	0.1	0.3	0.5	0.7	0.9
					Mangat-Singh				
0.3	0.2	1	45	152	79	5	16	113	318
		2	87	241	142	11	31	190	391
		3	129	303	191	17	46	247	440
		4	169	347	232	22	60	292	456
		5	208	378	265	27	73	326	467
		10	359	450	371	53	132	419	475
	0.4	1	131	16	2	10	65	175	353
		2	227	31	5	20	119	267	414
		3	305	46	7	30	163	329	454
		4	364	60	10	39	201	371	465
		5	404	73	12	48	233	400	472
		10	470	132	24	90	340	460	476
	0.6	1	256	16	31	79	152	254	391
		2	374	31	60	142	241	343	436
		3	432	46	87	191	303	402	466
		4	458	60	111	232	347	430	471
		5	468	73	133	265	378	449	474
		10	476	132	221	371	450	473	476
	0.8	1	395	152	175	226	285	353	432
		2	459	241	267	319	368	414	456
		3	472	303	329	379	422	454	473
		4	475	347	371	413	444	465	475
		5	476	378	400	436	459	472	476
		10	476	450	460	471	475	476	476

않은 경우(Warner모형에서는 p가 0.5에 가까울 때, Mangat모형에서는 p가 0에 가까울 때, Mangat-Singh모형에서는 T와 p가 모두 0에 가까울 때)에 반복시행의 효과가 더욱 크다는 것을 보여주고 있다.

5. 결론

각 모형을 실제 조사에 사용하는 경우에 p또는 T를 선택하는 것이 문제가 될 수 있다. 예를 들어 Warner모형을 사용하는 경우에 p를 0또는 1에 가까웁게 정하면 추정량의 분산을 작

<표 2> m(반복 횟수)=1일 때를 기준으로 한 정보비

π	T		0	0.1	0.3	0.5	0.7	0.9
	p m	Mangat	Warner	Mangat-Singh				
0.3	0.2 2	1.909	1.583	1.778	1.985	1.957	1.687	1.231
	3	2.813	1.985	2.394	2.954	2.871	2.192	1.383
	4	3.698	2.271	2.904	3.909	3.744	2.585	1.435
	5	4.546	2.477	3.325	4.848	4.578	2.888	1.468
	10	7.834	2.948	4.650	9.322	8.246	3.714	1.496
	0.4 2	1.733	1.957	1.993	1.973	1.818	1.527	1.172
	3	2.329	2.871	2.980	2.918	2.491	1.882	1.287
	4	2.774	3.744	3.960	3.836	3.063	2.119	1.316
	5	3.083	4.578	4.933	4.728	3.549	2.285	1.335
	10	3.579	8.246	9.697	8.825	5.181	2.624	1.347
	0.6 2	1.456	1.957	1.914	1.778	1.583	1.351	1.113
	3	1.682	2.871	2.745	2.394	1.985	1.579	1.191
	4	1.783	3.744	3.506	2.904	2.271	1.691	1.204
	5	1.825	4.578	4.204	3.325	2.477	1.765	1.213
	10	1.853	8.246	6.984	4.650	2.948	1.861	1.216
	0.8 2	1.162	1.583	1.527	1.411	1.291	1.172	1.056
	3	1.196	1.985	1.882	1.679	1.481	1.287	1.095
	4	1.203	2.271	2.119	1.828	1.560	1.316	1.099
	5	1.204	2.477	2.285	1.929	1.612	1.335	1.101
	10	1.205	2.948	2.624	2.083	1.666	1.347	1.101

게 할 수 있지만, 확률장치의 한 부분이 나올 확률이 지나치게 크기 때문에 조심스러운 응답자라면 조사에 거부감을 느낄 수 있으며 또다른 형태의 거짓응답이나 응답기피를 유발시킬 수 있다. 따라서 추정량의 효율만을 고려하여 p와 T를 선택하는 데에는 어려움이 있으며 응답자가 거부감을 느끼지 않게 하는 적당한 p와 T를 선택하여야 한다. 효율이 좋지 않은 p와 T가 선택된 경우에 각 모형을 반복하여 시행함으로써 추정량의 효율을 높일 수 있을 것이다. 본 논문에서는 모비율 π가 0.3인 경우의 결과만을 제시하였지만 다른 경우에도 유사한 결과를 얻을 수 있었다.

본 논문에서는 거짓응답이 존재하지 않는 경우로 연구가 제한되었지만 거짓응답이 존재하는 경우에는 세 모형에서의 최대우도 추정량은 일반적으로 불편추정량이 아니다. Mangat & Singh(1990)과 Mangat(1994)는 거짓응답이 존재하는 경우에서의 평균제곱오차를 구하고 비교

하였으며 이를 반복시행의 경우로 확장하여 비교하는 것도 좋을 것이라고 생각된다.

참고문헌

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

[2] Mangat, N, S.(1994) An Improved Randomized Response Strategy. *Journal of Royal Statistical Society series B*, 56, 93-95.

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

[4] Rao. D. R.(1965) *Limear Statistical Inference and its Applications*. Wiley, New York.

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

A Study on the Efficiency of the Multiple Trial Randomized Response Technique

Hae-Yong Lee[3], Hyun-Cheol Kang[4]

Abstract

In surveys on certain social problems which are sensitive in nature, many techniques have been introduced in order to protect evasive or untruthful answers. We suggest a multiple trial randomized response technique(MRRT) and it turns out that MRRT is feasible and more efficient by reducing the variance of the estimate than single trial RRT's investigated by Waner (1965), Mangat & singh (1990), and Mnagat (1994).

3) Professor, Department of Statistics, Sungshin Women's University, Dongsun-Dong 3-ka, Sungbuk-Ku, Seoul, 136-742, Korea.

4) Ph. D. Candidate, Department of Statistics, Korea University, Ahnam-Dong, Seongbuk-Ku, Seoul, 136-701, Korea.