한국통계학회논문집 제3권 2호
1996년 8월 pp. 139-150

2단계 확률화응답기법의 효율성 및 응답자보호수준[1)]

김 영 원[2)], 정 선 아[3)]

요 약

확률화응답기법은 민감한 사항에 대한 모비율의 효율적인 추정과 응답자에 대한 충분한 보호수준 제공이라는 두가지 측면을 동시에 추구하는 통계조사기법이다. 본 연구에서는 기법에서 적용된 질문선택확률이 이와 같은 두가지 측면에 미치는 영향을 정리하고, 이런 관점에서 최근에 제시된 2단계 확률화응답기법들은 기존의 기법에서 단순히 질문선택확률구조만을 수정한 것으로 해석될 수 있으며, 따라서 효율성 측면에서 본질적으로 기존의 방법과 동일한 기법임을 보였다.

1. 서 론

확률화응답기법(Randomized Response Technique)은 민감한 질문에 대한 응답자의 의도적인 거짓응답에 의한 편의와 무응답비율을 줄이기 위하여 고안된 통계 조사 방법으로 이분류 모집단(Dichotomous Population)에서 민감한 속성 (A)을 갖는 모집단 비율 (π)을 추정하는 경우 대칭이 되는 두개의 질문을 이용하는 Warner (1965)의 기법이 제시된 이후 보다 효율적인 기법을 구현하기 위하여 다양한 기법이 연구되었다. Greenberg 등 (1969)은 Simmons가 제시한 방법에 따라 무관질문기법을 체계화시켰다. 최근에는 2단계로 확장된 기법들이 Mangat와 Singh (1990), 김종호 등 (1992), Mangat (1994)에 의하여 제시되었고 그 기법들의 효율성이 연구되었다. 이 밖에도 다양한 각도에서 확률화응답기법에 관한 연구결과들이 발표되었으며 Chaudhuri와 Mukerjee (1988)와 류제복 등 (1993)은 확률화응답기법에 대한 기존의 연구 결과들을 체계적으로 정리하였다.

확률화응답기법의 사용 목적상 효율적인 확률화응답기법을 구현하는 경우, 모수 π를 효율적으로 추정한다는 측면과 응답자에게 충분한 보호 수준을 제공하여야 한다는 상반된 두개의 측면이 동시에 고려되어야 한다. Flinger 등 (1977), Anderson (1977) 등이 응답자 보호 수준에 관한 중요성을 지적하고 있음에도 불구하고 최근에 제안된 기법에 있어서 추정상의 효율성 측면만이 고려된 상태에서 기법들의 효율성이 부적절하게 비교 연구되어지고 있는 것이 현실이다.

본 연구에서는 이와 같은 두 가지 측면을 동시에 고려하여 기법들의 본질적인 차이점을 밝히고 아울러 기법의 질문선택확률과 두 가지 측면에서의 기법의 효율성간의 관계를 직관적인 측면에서

1) 이 연구는 94년도 한국과학재단 연구비지원에 의한 결과임. (과제번호 941-0100-032-1)
2) (140-742) 서울 용산구 청파동 2가 숙명여자대학교 통계학과 부교수.
3) (441-083) 경기도 수원시 권선구 매산로 3가 경기도청 통계전산담당관실.

정리하고자 한다. 2절에서는 Fisher의 정보를 통하여 확률화응답기법의 추정상의 효율성을 설명하고, 이분류 모집단의 경우 기본적인 Warner기법과 무관질문기법에서 추정상의 효율성과 질문선택확률간의 관계를 정리하였다. 3절에서는 기존에 제시된 응답자 보호 수준 척도들 간의 차이점을 정리하고 추정상의 효율성을 결정하는 요인이 응답자 보호 수준에 미치는 영향을 살펴보았다. 그리고 4절에서는 최근에 제시된 2단계 확률화응답기법들이 추정상의 효율성 측면과 응답자 보호 측면에서 기존의 기법들과 어떤 근본적인 차이점을 갖고 있는지 밝혔다.

2. 추정상의 효율성

확률변수 X의 밀도함수 $f_X(x;\pi)$가 정규적인 조건을 만족하는 경우, X를 관측하게 되는 통계적 실험을 통하여 미지의 모수 π를 추정하는데 사용되는 정보량은 일반적으로 Fisher의 정보,

$$I(X) = E\left\{ \left(\frac{\partial \log f_X(X;\pi)}{\partial \pi} \right)^2 \right\}$$

로 설명된다. 따라서 통계적 모형(혹은 실험)을 통하여 얻게 되는 모수 추정상의 효율성은 $I(\cdot)$로 측정될 수 있고, Fisher의 정보는 확률화응답기법의 추정상의 효율성을 설명하여 주는 적절한 기준이 될 수 있을 것이다(Anderson (1977) 참조). 예를 들어 Warner기법과 무관질문기법의 추정상의 효율성 비교를 위하여 기존의 연구에서는 Warner (1965)와 Greenberg 등 (1969)이 제시한 추정량의 분산을 이용하고 있다. 이런 추정량들은 관심 모수공간을 벗어날 수가 있고 따라서 현실적으로는 절단된 형태의 추정량을 사용하고 있지만(Flinger 등 (1977) 참조), 기존의 연구에서는 이런 기법들의 추정상의 효율성을 파악하기 위하여 기존의 추정량의 분산을 그대로 사용하는 문제점을 갖고 있다. 이런 관점에서 Flinger 등 (1977)은 π의 최우추정량을 제시하고 추정상의 효율성비교를 위하여 추정량의 분산 대신 평균제곱오차를 사용하였다.

따라서 본 연구에서는 통계모형을 이용하여 모수를 추정하는 경우, 사용 추정량에 의존하지 않고 모형 자체가 지니고 있는 추정상의 정보량을 반영하는 Fisher의 정보, $I(\cdot)$를 확률화응답기법의 추정상 효율성의 척도로 사용하고자 한다. 물론 이런 접근방식이 기법의 추정상의 효율성에 관한 기존의 연구방식과 크게 상이한 결과를 도출하는 것은 아니지만 확률화응답기법의 본질을 이해하는데 도움이 될 것이다.

확률화응답기법은 민감한 사안에 대한 간접적인 답변을 유도하기 위하여 다음과 같은 질문에 대한 선택적인 응답을 하도록 구성된다.

'질문 1 : 당신은 그룹 A에 속합니까?'
'질문 2 : 당신은 그룹 A에 속하지 않습니까?'
'질문 3 : 당신은 그룹 B에 속합니까?'

여기서 속성 B는 응답자가 전혀 거부감없이 응답할 수 있는 것으로, '질문3'은 무관질문에 해당한다. 본 연구에서는 속성 B를 갖는 모집단 비율 (π_B)은 알고 있다고 가정한다. 이런 가정을 충족하는 무관질문의 실제적인 구현방법은 Horvitz 등 (1976)이 제시하였다.

이분류 모집단에서 민감한 속성을 갖는 모집단 비율 π를 추정하기 위하여, Warner (1965)는

'질문1'이 선택될 확률은 p_w이고 '질문2'가 선택될 확률은 $1-p_w$인 기법을 제안하였다. 랜덤하게 추출된 응답자의 답변을 확률변수 X로 표시하면 X는 성공확률 $\lambda = \pi p_w + (1-\pi)(1-p_w)$ 인 베르누이 분포에 따른다.

따라서 Warner기법의 추정상의 효율성을 나타내는 $I_W(\cdot)$ 는 다음과 같다.

$$I_W(X) = \frac{(2p_w-1)^2}{\{\pi p_w + (1-\pi)(1-p_w)\}\{\pi(1-p_w) + p_w(1-\pi)\}} , \tag{2.1}$$

여기서 $nI_W(X)$는 Warner (1965)가 제시한 추정량의 분산(Chaudhuri와 Mukerjee, 1988, p.4 참조)의 역수에 해당한다. 이제 Warner기법의 추정상의 효율성과 질문선택 확률구조를 결정하는 p_w와의 관계를 살펴보기 위하여 $I_W^{-1}(X)$를 고려하면 다음과 같다.

$$I_W^{-1}(X) = \frac{1}{4(2p_w-1)^2} - \left(\pi - \frac{1}{2}\right)^2 , \tag{2.2}$$

여기서 π는 미지의 상수이므로 $I_W(X)$는 $|2p_w-1|$에 대한 단조 증가 함수이므로 다음 결론을 얻는다.

정리 2.1

Warner기법의 추정상의 효율성은 $|2p_w-1|$에 비례한다.

따라서 랜덤하게 복원 추출된 n명의 응답자를 대상으로 Warner기법을 적용하는 경우, 추정상의 효율성은 $n|2p_w-1|$로 설명된다. 직관적인 측면에서 '질문1'과 '질문2'로부터 얻게 되는 π에 대한 정보는 질문의 대칭속성 때문에 서로 상쇄되는 효과를 가져다주고, 따라서 실제적으로 n명의 응답자를 고려할 때 얻게 되는 추정을 위한 정보량을 표본크기 개념으로 해석하면, 추정상의 효율성은 두 질문을 선택하게 되는 응답자의 수의 기대값의 차이, $|np_w - n(1-p_w)| = n|2p_w-1|$로 설명된다. 아울러 이때 발생하는 정보의 손실은 응답자 보호 측면에 기여하게 된다. 물론 알려진 바와 같이 $p=1/2$ 인 경우, Warner기법을 통하여 π의 추정을 위한 정보는 전혀 얻을 수 없다.

Simmons는 무관질문(즉, '질문3')을 사용한 기법을 제안하였고 이는 Greenberg 등 (1969)에 의하여 이론적으로 체계화되었다. 이들은 '질문1'이 선택될 확률은 p_u이고 '질문3'이 선택될 확률이 $(1-p_u)$인 확률화응답기법을 고려하였다. 랜덤하게 추출된 응답자의 답변을 확률변수 X로 표시하면 X는 성공확률 $\lambda = p_u\pi + (1-p_u)\pi_B$인 베르누이 분포에 따른다.

따라서 무관질문기법의 추정상의 효율성을 나타내는 $I_U(X)$는 다음과 같다.

$$I_U(X) = \frac{{p_u}^2}{\{\pi p_u + \pi_B(1-p_u)\}\{(1-\pi)p_u + (1-p_u)(1-\pi_B)\}} , \tag{2.3}$$

여기서 $nI_U(X)$는 이 기법에서 제시된 추정량의 분산(Chaudhuri 와 Mukerjee, 1988, p.16 참조)의 역수에 해당한다. 무관질문의 추정상의 효율성과 질문선택확률구조를 결정하는 p_u와의 관계를 고려하면

$$I_U^{-1}(X) = \left\{\pi + \frac{(1-p_u)}{p_u}\pi_B\right\}\left\{(1-\pi) + \frac{1-p_u}{p_u}(1-\pi_B)\right\} \tag{2.4}$$

이고, 여기서 $0 < \pi < 1$, $0 \le \pi_B \le 1$ 그리고 $0 < p_u < 1$ 이므로,

$$\frac{\partial}{\partial p_u}\{I_U^{-1}(X)\} = -\frac{1}{p_u^2}\left\{\pi_B(1-\pi) + \pi(1-\pi_B) + 2\frac{1-p_u}{p_u}\pi_B(1-\pi_B)\right\} < 0 \tag{2.5}$$

즉, $I_U(X)$는 p_u에 대한 단조 증가 함수이므로 다음과 같은 결론을 얻는다.

정리 2.2

무관질문기법의 추정상의 효율성은 p_u에 비례한다.

따라서 랜덤하게 복원 추출된 n명의 응답자를 고려하면, 추정상의 효율성은 np_u로 설명된다. 직관적인 측면에서 두개의 질문 중 관심의 대상인 π에 대한 정보를 제공하는 것은 '질문1' 이고 '질문3'은 π에 대한 정보를 전혀 제공하고 있지 않다. 그러므로 통계적 관측을 통하여 얻게 되는 정보량을 표본크기 개념으로 해석하면, 무관질문기법의 경우 추정상의 효율성은 n명의 응답자 중 '질문1'을 선택할 것으로 기대되는 응답자의 수 np_u로 설명된다. 한편 이때 발생하는 정보의 손실, 즉 $n(1-p_u)$는 응답자 보호 측면에 기여하게 된다.

3. 응답자 보호 수준

민감한 속성에 관한 표본조사에 있어서 간접응답 방식인 확률화응답기법을 사용하는 이유는 응답자의 비밀을 보장함으로써 응답자로부터 진실된 답변을 얻을 수 있다는 것이다. 이런 확률화응답기법의 본래의 의도에도 불구하고 기존의 많은 연구에 있어서는 응답자 보호 수준에 관한 고려없이 추정상의 효율성에 관한 문제가 취급되고 있는 것이 현실이다. 통계적 관점에서 기법의 응답자 보호에 관한 척도는 Lanke (1976), Leysieffer와 Warner (1976), Flinger 등 (1977) 그리고 Anderson (1977) 등에 의하여 여러 형태로 제안되었다.

본 절에서는 기존의 다양한 응답자 보호 수준에 대한 척도들을 기준으로 대표적인 확률화응답기법인 Warner기법과 무관질문기법에서 질문선택확률 p_w와 p_u가 응답자 보호 수준에 미치는 영향에 대하여 살펴 보기로 한다.

Lanke (1976)는 응답자가 "예(Y)" 또는 "아니오(N)"라는 응답을 하는 경우, 자신의 민감한 속성을 노출시킬 우려 때문에 이런 응답을 꺼린다는 점을 고려하여 $\max\{P(A|Y), P(A|N)\}$를

응답자 보호 척도로 사용하고 있으며 이 척도가 큰 값을 갖게 되면 응답자 보호수준은 낮아진다. 무관질문기법에서는 항상 $P(A|Y) > P(A|N)$ 의 관계가 성립되므로 이 경우 $\max\{P(A|Y), P(A|N)\} = P(A|Y)$이고, Warner기법에서는

$$P(A|Y)-P(A|N) = \frac{P(A)}{P(Y)P(N)}[\ P(Y|A)-P(Y)\ ]$$

이고, $P(Y|A) = p_w$, $P(Y) = p_w\pi + (1-p_w)(1-\pi)$ 이므로

$$\max\{\ P(A|Y)\ ,\ P(A|N)\ \} = \begin{cases} P(A|Y)\ , & p_w > \frac{1}{2} \\ P(A|N)\ , & p_w < \frac{1}{2} \end{cases}$$

이다. 무관질문기법에서

$$P(A|Y) = \frac{p_u\pi + (1-p_u)\pi\pi_B}{p_u\pi + (1-p_u)\pi_B} \tag{3.2}$$

이고

$$\frac{\partial}{\partial p_u}P(A|Y) = \frac{\pi\pi_B(1-\pi)}{\{\ p_u\pi + (1-p_u)\pi_B\}^2} > 0 \tag{3.3}$$

이므로 Lanke의 응답자 보호 척도를 기준으로 무관질문기법의 응답자 보호 수준은 p_u에 반비례한다. 한편 Warner기법에서는

$$P(A|Y) = \frac{p_w\pi}{p_w\pi + (1-p_w)(1-\pi)}$$
$$P(A|N) = \frac{(1-p_w)\pi}{(1-p_w)\pi + p_w(1-\pi)} \tag{3.4}$$

이다. 여기서 $P(A|N)$은 $P(A|Y)$에 p_w대신 $(1-p_w)$가 사용된 함수이므로 $\max\{P(A|Y), P(A|N)\}$은 $p_w=1/2$ 에 대칭인 볼록함수이고 $1/2<p_w<1$인 경우 관심대상인 $P(A|Y)$에 대하여

$$\frac{\partial}{\partial p_w}P(A|Y) = \frac{\pi(1-\pi)}{\{\ p_w\pi + (1-p_w)(1-\pi)\}^2} > 0 \tag{3.5}$$

이므로, Lanke의 척도를 기준으로 Warner기법의 응답자 보호 수준은 $|p_w - \frac{1}{2}| = |2p_w-1|$에 반비례한다.

Flinger 등(1977)은 다음의 응답자 보호 척도를 제시하였다.

$$J_1 = \frac{1 - \max\{\ P(A|Y)\ ,\ P(A|N)\ \}}{1-\pi}\ ,$$

여기서 $0 \le J_1 \le 1$ 이고, J_1이 0에 가까워지면 낮은 수준의 응답자 보호 수준을, 또한 J_1이 1에

가까워지면 높은 수준의 응답자 보호 수준을 의미한다. J_1은 결국 $\max\{P(A|Y), P(A|N)\}$에 반비례하게 되고, 따라서 이 척도는 기법간의 응답자 보호 수준의 비교에 있어서 Lanke (1976)의 척도와 동일한 결과를 나타내게 된다.

Leysieffer와 Warner (1976)는 응답자로부터 R이라는 응답을 얻었을 때 $P(A|R) > \pi$이면 R은 그룹 A에 대한 노출 위험을 의미하고 $P(A^c|R) > 1-\pi$ 이면 R은 그룹 A^c에 대한 노출 위험을 의미한다는 사실을 고려하여 다음 위험함수를 제시하였다.

$$g(R,A) = \frac{P(A|R)}{P(A^c|R)} \cdot \frac{1-\pi}{\pi} = \frac{P(R|A)}{P(R|A^c)} , \tag{3.6}$$

여기서 $P(A|R) > \pi$이면 $P(A^c|R) < 1-\pi$이고 $g(R,A) > 1$ 이 된다. 즉, $g(R,A)$를 응답자 보호 척도로 사용하면 $g(R,A)$ 이 1 보다 클수록 응답자가 R 이라는 응답을 했을 때 그룹 A에 속한다고 간주될 확률이 커지게 되므로 응답자 보호 수준은 낮아진다고 할 수 있다. Leysieffer와 Warner (1976)는 $g(R,A)$와 동시에 $g(R,A^c)$, 즉 응답자가 그룹 A^c에 속하게 될 위험함수도 고려하고 있으나, 실제적으로 확률화응답기법이 적용되는 대부분의 경우(예를 들어 그룹 A가 마약복용 경험을 갖는 집단을 의미하는 경우), 응답자가 그룹 A^c에 속하는 것으로 간주되는 것에 대해서는 개의치 않게 되므로 일반적으로 우리의 관심은 $g(R,A)$ 이다.

무관질문기법에서는 $P(Y|A) > P(Y|A^c)$이고 $P(N|A) < P(N|A^c)$이므로 $R=Y$인 경우에 $g(R,A)$ 는 1 보다 크므로 응답자 보호측면에서 $g(Y,A)$ 가 고려대상이 된다. $g(Y,A)$ 와 $P(A|Y)$ 의 관계는

$$g(Y,A) = \frac{P(A|Y)}{1-P(A|Y)} \cdot \frac{1-\pi}{\pi} \tag{3.7}$$

이므로 $g(Y,A)$ 는 결국 $P(A|Y)$ 의 증가함수이다. 또한 Warner기법에서는 $p_w > 1/2$ 이면 $P(Y|A) > P(Y|A^c)$ 이고 반면에 $p_w < 1/2$ 이면 $P(N|A) > P(N|A^c)$ 이므로 응답자 보호측면에서 $p_w > 1/2$ 인 경우 $g(Y,A)$ 가, 그리고 $p_w < 1/2$ 인 경우 $g(N,A)$ 가 고려대상이 된다. 여기서 $g(N,A)$은 (3.7)식에서 $P(A|Y)$대신 $P(A|N)$ 를 사용한 형태이므로 $g(Y,A)$ 와 $g(N,A)$는 각각 $P(A|Y)$ 와 $P(A|N)$ 의 증가함수이다. 따라서 Leysieffer와 Warner (1976)의 응답자 보호 척도는 Lanke (1976)가 제시한 척도와 동일한 결과를 갖게 된다.

Anderson (1977)은 민감한 질문에 대한 진실한 응답을 확률변수 W라고 하고 확률화응답기법에서 응답이 $X=x$ 일 때, 응답자 보호 수준을 $Var(W|X=x)$로 설명하고 특정기법의 응답자 보호 수준 척도로 $E\{Var(W|X)\}$를 제안하였다. 또한 Anderson (1977)은 Warner기법과 무관질문기법에서(즉, 베르누이 모형에서) $E\{Var(W|X)\}$와 X가 주어진 W의 조건부 분포의 Fisher의 정보에 대하여 다음 관계식을 보였다.

$$I(W|X) = \frac{E\{Var(W|X)\}}{\pi^2(1-\pi)^2} .$$

한편, Fisher의 조건부 정보, $I(W|X) = I(W) - I(X)$ 이고 $I(W) = 1/\pi(1-\pi)$ 이므로 $I(W|X)$는 확률화응답기법의 Fisher의 정보 $I(X)$에 반비례하게 된다. 따라서 Warner기법과 무관질문기법의 질문선택확률 p_w와 p_u가 Anderson의 척도를 기준으로 응답자 보호 수준에 미치는 영향을 고려하면 응답자 보호 수준은 Warner기법의 경우 (2.2)식에 의하여 $|2p_w-1|$에 반비례하고 무관질문기법의 경우 (2.5)식에 의하여 p_u에 반비례한다.

이와 같은 내용들을 정리하면 기존에 여러 형태로 Lanke (1976), Flinger 등 (1977), Leysieffer와 Warner (1976) 그리고 Anderson (1977) 등이 제시한 응답자 보호 척도를 기준으로 다음과 같은 결론을 얻는다.

정리 3.1

Warner기법의 경우 응답자 보호 수준은 $|2p_w-1|$에 반비례하고 무관질문기법의 경우 응답자 보호 수준은 p_u에 반비례한다.

따라서 정리 2.1, 2.2 및 3.1을 고려하면 두 기법에 있어서 추정상의 효율성과 응답자 보호 수준은 질문선택확률에 의존하게 되고 서로 상반된 결과를 보여준다는 사실을 알 수 있다.

4. 기법간의 동질성

최근에 보다 효율적인 확률화응답기법을 구현하기 위하여 2단계에 걸친 질문선택구조를 갖는 확률화응답기법들이 Mangat와 Singh (1990), 김종호 등 (1992) 그리고 Mangat (1994)에 의하여 제시되었다.

4.1 Mangat와 Singh의 기법 (M-S기법)

Mangat와 Singh (1990)은 Warner기법을 수정하여 1단계에서는 다음 2개의 질문 중 '질문1'을 선택할 확률이 t, '질문2'를 선택할 확률이 $(1-t)$가 되도록 고안된 확률장치 R_1을 이용하여 응답자는 선택된 질문에 응답하도록 한다.

'질문1 : 당신은 그룹 A에 속합니까?', '질문2 : 확률장치 R_2로 가시오.'

만약 응답자가 '질문2'를 선택하는 경우 또 다른 확률장치 R_2를 이용하여 다음 2개의 질문 중 하나의 질문에 답하게 된다. R_2에서는 '질문1'이 선택될 확률이 s, '질문2'가 선택될 확률이 $(1-s)$이다.

'질문1 : 당신은 그룹 A에 속합니까?', '질문2 : 당신은 그룹 A에 속하지 않습니까?'

이 경우 응답자가 "예"라고 응답할 확률은 다음과 같다.

$$\lambda_{MS} = t\pi + (1-t)\{s\pi + (1-s)(1-\pi)\} \tag{4.1}$$

이 기법을 랜덤하게 추출된 응답자에게 적용하여 얻은 확률변수를 X라 하면 X는 성공확률 λ_{MS}인 베르누이 분포에 따른다. 이 기법의 추정상의 효율성을 나타내는 Fisher의 정보는 다음과 같다.

$$I_{MS}(X) = \frac{\{t+(1-t)(2s-1)\}^2}{[\ \pi\{t+(1-t)s\}+(1-\pi)(1-t)(1-s)]} \times \frac{1}{[\ \pi(1-t)(1-s)+(1-\pi)\{t+(1-t)s\}]} \tag{4.2}$$

또한, Lanke(1975, 1976), Flinger 등(1977), Leysieffer(1976)이 제시한 응답자 보호 척도의 기준이 되는 $P(A|Y)$와 $P(A|N)$은 다음과 같다.

$$P(A|Y) = \frac{\{t+(1-t)s\}\pi}{\{t+(1-t)s\}\pi+(1-t)(1-s)(1-\pi)}$$
$$P(A|N) = \frac{(1-t)(1-s)\pi}{\{t+(1-t)s\}(1-\pi)+(1-t)(1-s)\pi} \tag{4.3}$$

만약, $p_w = t+(1-t)s$라 하면 $1-p_w = (1-t)(1-s)$이고 따라서 (4.2)식은 (2.1)식과 동일하고, (4.3)식은 (3.4)식과 동일하다. 즉, 추정상의 효율성 측면이나 응답자 보호 수준 측면에서 M-S 기법은 $p_w = t+(1-t)s$인 Warner기법과 동일하다는 사실을 알 수 있다.

본질적으로 Mangat와 Singh (1990)이 제안한 기법은 비록 2단계에 걸친 확률장치를 사용하고 있으나 질문선택 확률구조를 정리하면 응답자가 Warner기법의 '질문1'을 선택할 확률이 $t+(1-t)s$이고 '질문2'를 선택할 확률이 $(1-t)(1-s)$인 확률화응답기법으로 정리될 수 있다. 따라서 다음 결론을 얻을 수 있다.

정리 4.1

Mangat와 Singh (1990)의 기법은 $p_w = t+(1-t)s$인 Warner기법과 동일하다

Mangat와 Singh (1990)은 Warner기법의 추정량의 분산과 M-S기법의 추정량의 분산을 직접 비교하여 M-S기법에서 $s = p_w$라는 가정 하에서

$$t > (1-2p_w)/(1-p_w) \tag{4.4}$$

의 조건을 만족하면 M-S기법이 Warner기법보다 효율적이라는 것을 밝히고 있으나 여기서 $s = p_w$라는 가정은 응답자 보호 수준이 전혀 고려되지 않은 부적절한 가정이므로 이런 가정 하에서의 두 기법의 효율성 비교는 의미가 없다고 판단된다. 한편, 정리 2.1을 고려하면, 추정상의 효율성만을 고려할 때,

$$|2p_w-1| < |2\{t+(1-t)s\}-1| \tag{4.5}$$

이면, M-S기법이 보다 효율적이라고 할 수 있고 (4.4)식은 $s = p_w$를 가정하는 경우 (4.5)식으로

부터 간단하게 유도될 수 있는 결과이다. 또한, Warner기법의 응답자 보호 수준은 $|2p_w-1|$에 반비례하므로 (4.4)식이 성립되는 경우 응답자 보호 측면에서는 오히려 Warner기법이 M-S기법보다 바람직하다는 사실에 유의하여야 한다.

적절한 추정상의 효율성 비교를 위하여 동일한 조건, 즉 동일한 응답자 보호 수준 하에서 두 기법이 비교되어야 하므로 정리 4.1에 의하여 Warner 기법과 M-S 기법의 효율성 비교는 무의미하다는 사실을 알 수 있다.

4.2 김종호 등의 기법

김종호 등 (1992)은 Mangat와 Singh (1990)의 2단계 확률화응답기법에서 2번째 단계의 '질문2', 즉 '당신은 그룹 A에 속하지 않습니까?'를 '"예"라고 대답하시오' 라는 강요질문으로 대체한 새로운 2단계 확률화응답기법을 제안하였다. 이 경우 응답자가 "예"라고 응답할 확률은 다음과 같다.

$$\lambda_K = t\pi + (1-t)\{s\pi + (1-s)\}$$

이 때, 응답자의 답을 나타내는 확률변수 X는 성공확률 λ_K인 베르누이 분포에 따른다. 이 기법의 추정상의 효율성을 나타내는 Fisher의 정보는 다음과 같다.

$$I_K(X) = \frac{\{t+(1-t)s\}^2}{[\ \{t+(1-t)s\}\pi+(1-t)(1-s)]\ \{t+(1-t)s\}(1-\pi)} \cdot \tag{4.6}$$

이 기법의 경우 $P(A|Y) > P(A|N)=0$ 이므로 Lanke (1976), Flinger 등 (1977) 그리고 Leysieffer (1976)이 제시한 응답자 보호 척도를 고려할 때, 관심대상인 $P(A|Y)$는 다음과 같다.

$$P(A|Y) = \frac{\pi}{\{t+(1-t)s\}\pi + (1-t)(1-s)} \cdot \tag{4.7}$$

만약, $p_u = t+(1-t)s$이고 $\pi_B=1$이라고 하면 (4.6)식은 (2.3)식과 동일하고, (4.7)식은 (3.2)식과 동일하다. 즉 추정상의 효율성 측면이나 응답자 보호 수준 측면에서 김종호 등의 기법은 $p_u = t+(1-t)s$이고 $\pi_B=1$인 무관질문기법과 동일하다는 사실을 알 수 있다.

본질적으로 김종호 등이 제안한 기법은 2단계에 걸친 확률장치를 사용하고 있으나 질문선택 확률구조를 정리하면 응답자가 '질문1 : 당신은 그룹 A에 속합니까?'를 선택할 확률인 $t+(1-t)s$이고 '질문2 : "예"라고 대답하시오'를 선택할 확률이 $(1-t)(1-s)$인 기법이다. 여기서 '질문2'는 $\pi_B=1$ 인 무관질문에 해당하므로 다음 결론을 얻을 수 있다.

정리 4.2

김종호 등 (1992)의 기법은 $p_u = t+(1-t)s$ 이고 $\pi_B=1$인 무관질문기법과 동일하다.

따라서, 이 기법은 $t+(1-t)s$ 가 커지면 정리 2.2에 의하여 추정량의 효율성이 증가하고 반면에 정리 3.1에 의하여 응답자 보호 수준은 감소한다.

4.3 Mangat 기법

Mangat (1994)는 Mangat와 Singh (1990)이 제시한 기법이 2개의 확률장치를 사용하는데 따르는 면접상의 복잡성을 극복하기 위하여 다음과 같은 새로운 확률화응답기법을 제안하였다. 응답자 보호 측면에서 전체 응답과정(특히, 확률장치의 사용여부)은 다른 사람이 관측할 수 없도록 비밀을 보장하는 상황하에서 응답자는 첫 번째 단계에서 그룹 A에 속하는 경우 "예" 라고 응답하고 그룹 A에 속하지 않는 응답자의 경우에는 확률장치를 이용하여 Warner기법에서 사용된 '질문1'과 '질문2'를 각각 확률 p_m과 $1-p_m$으로 선택하여 응답하게 하는 기법을 제시하였다. 이 경우 응답자가 "예" 라고 응답할 확률은

$$\lambda_m = \pi + (1-\pi)(1-p_m) = p_m\pi + (1-p_m)$$

이고 응답결과는 성공확률이 λ_m인 베르누이 분포로 설명될 수 있다.

이 기법의 추정상의 효율성을 나타내는 Fisher의 정보는 다음과 같다.

$$I_m(X) = \frac{p_m^2}{\{p_m\pi + (1-p_m)\}\, p_m(1-\pi)} \ . \tag{4.8}$$

이 경우 $P(A|N)=0$이고 응답자 보호 수준을 고려할 때, 관심대상인 $P(A|Y)$는 다음과 같다.

$$P(A|Y) = \frac{\pi}{p_m\pi + (1-p_m)} \ . \tag{4.9}$$

만약, $p_m = p_u$, $\pi_B=1$이라고 하면 (4.8)식은 (2.3)식과 동일하고 (4.9)식은 (3.2)식과 동일하다. 즉, 추정상의 효율성 측면이나 응답자 보호 수준 측면에서 Mangat (1994)가 제시한 기법은 $p_u = p_m$이고 $\pi_B=1$인 무관질문기법과 동일하다는 사실을 알 수 있다. 따라서 다음 결론을 얻을 수 있다.

정리 4.3

Mangat (1994)의 기법은 $p_u = p_m$이고 $\pi_B=1$인 무관질문기법과 동일하다.

결국 Mangat기법은 김종호 등 (1992)의 기법과 $p_m = t + (1-t)s$인 경우 동일한 기법이고, 또한 $p_u = p_m$, $\pi_B=1$인 무관질문기법에 해당한다.

한편, Mangat (1994)는 만약

$$\pi > 1 - \frac{s(1-t)\{1-(1-t)(1-s)\}}{\{2s-1+2t(1-s)\}^2} \tag{4.10}$$

이 성립하면, Mangat기법이 M-S기법보다 효율적이라는 사실을 밝히고 있으나 이 결과는 $p_w = t+(1-t)s$, $p_u = s$, $\pi_B=1$이라는 가정하에서 (2.1)식과 (2.3)식의 단순비교에 의하여 얻을 수 있는 결과이고, 이 때 사용된 가정은 두 기법의 응답자 보호수준이 고려되지 않은, 추정상의 효율성 비교를 위하여 부적절한 가정이라는 사실에 유의하여야 한다.

5. 결 론

본 연구에서는 확률화응답기법의 효율성 분석에 있어서 고려되어야 할 추정상의 효율성과 응답자 보호 수준이라는 두 가지 상반된 측면과 기법의 질문선택확률간의 관계를 정리하였다. 확률화응답기법이란 일반적인 직접 조사에서 얻을 수 있는 정보 중 일부분은 응답자의 진실된 응답을 얻기 위하여 희생하고 나머지 표본 정보를 이용하여 조사 목적에 해당하는 모비율을 추정하는 표본조사 기법으로 설명될 수 있다.

Warner기법의 경우 추정 목적으로 사용되는 정보량은 $n|2p_w-1|$로 설명되고 나머지 표본정보는 응답자 보호를 위하여 희생된다고 해석될 수 있으며, 따라서 이 기법의 응답자 보호 수준은 $n|2p_w-1|$에 반비례하게 된다. 이런 관점에서 Warner기법들 간의 추정상의 효율성 비교는 그 의미가 없으며, 4장에서 밝혀진 바와 같이 Mangat와 Singh (1990)이 제시한 기법은 본질적으로 Warner기법으로 해석될 수 있으므로 이들의 연구 결과에서 제시한 효율성 비교는 적절하지 못한 연구 결과라는 것을 알 수 있다.

무관질문기법의 경우 마찬가지로 추정 목적에 사용되는 정보량은 np_u로 설명될 수 있고 나머지 표본정보는 응답자 보호를 위하여 희생된다고 해석될 수 있다. 한편 김종호 등 (1992)이 제시한 기법과 Mangat (1994)가 제시한 기법은 모두 $\pi_B=1$인 무관질문기법으로 해석될 수 있으므로, 이들 기법과 무관질문기법 간의 효율성 비교는 적절하지 못하고 이들 기법과 Warner기법과의 비교는 기존의 Warner기법과 무관질문기법간의 효율성 비교에 관한 연구 결과들(Chaudhuri와 Mukerjee, 1988, 참조)이 그대로 적용될 수 있다.

한편 기존의 연구에서 사용된 확률화응답기법의 응답자 보호 수준에 관한 기준들은 일반적인 통계적 개념을 활용한다는 관점에서 제안된 것들이므로 실제적으로 응답자가 심리적으로 느끼게 되는 보호 수준을 정확히 반영하고 있다고 볼 수는 없다. 따라서 심리적인 측면에서 각 기법들의 응답자 보호 수준에 관한 경험적인 고찰을 통하여, 실제 표본조사에 있어서 보다 효과적인 확률화응답기법의 구현에 관한 연구가 필요하다고 판단된다.

참 고 문 헌

[1] 김종호, 류제복, 이기성 (1992). 새로운 2단계 확률화응답모형, 「응용통계연구」, 제5권, 제2호, 157-167.

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

[3] Anderson, H. (1977). Efficiency versus Protection in a General Randomized Response Model, *Scandinavian Journal of Statistics*, Vol. 4, 11-19.

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

[5] Flinger, M. A., Policello, G. E. and Singh, J. (1977). A Comparison of Two RR Survey Methods with Consideration for the Level of Respondent Protection, *Communications in Statistics - Theory and Methods*, Vol. 6, 1511-1526.

[6] Greenberg, B. G., Abul-Ela, Abdel-Latif, A., Simmons, W. R. and Horvitz, D. G. (1969). The Unrelated Question RR Model: Theoretical Framework, *Journal of the American Statistical Association*, Vol. 64, 520-539.

[7] Horvitz, D. G., Greenberg, B. G., and Abernathy, J. R. (1976). Randomized Response : A Data Gathering Device for Sensitive Questions, *International Statistical Review*, Vol. 44, 181-196.

[8] Lanke, J. (1976). On the Degree of Protection in Randomized Interviews, *International Statistical Review*, Vol. 44, 197-203.

[9] Leysieffer, R. W. and Warner, S. L. (1976). Respondent Jeopardy and Optimal Design in RR Models, *Journal of the American Statistical Association*, Vol. 71, 649-656.

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

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

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