한국통계학회논문집 제4권 1호
1997년 4월 pp. 161-165

양적속성 추정을 위한 2단계 확률화응답기법[1)]

최 경 호[2)]

요 약

양적속성의 추정을 위한 확률화응답기법은 Greenberg et al.(1971)로부터 시작된다. 이후 새로운 방법에 대한 제시(Dalenius와 Vitale(1974)) 및 분포함수에 대한 추정(Duffy와 Waterton(1984))등이 연구되어 오고 있다. 본 연구에서는 2단계 확률화응답기법을 이용하여 양적속성을 추정하는 방법을 제시하고 이의 효율성 비교를 행해 보고자 한다.

1. 서 론

본질적으로 확률화응답기법은 무응답이나 거짓응답으로 인하여 유발되는 비표본오차를 줄이기 위하여 개발된 간접조사방식이다. Warner(1965)에 의하여 이 방법이 소개된 이후, 이 기법은 양적(quantitative)으로 민감한 속성이나 변수를 다루는 경우로까지 발전하게 되었다. 특히 Greenberg et al.(1971)은 무관질문기법을 양적속성으로 확장하여 이 부분의 기초를 마련하였다. 예컨데 양적으로 민감한 속성이란 낙태횟수나, 절도횟수, 그리고 주어진 기간동안의 술 소비량 등을 의미한다.

양적속성을 추정하기 위한 연구로는 Dalenius와 Vitale(1974), Warner(1971), Parzen(1962), Scheult(1970)등을 들 수 있다.

본 논문에서는 질적속성의 추정을 위한 Mangat와 Singh(1990)의 2단계 확률화응답기법을 양적속성의 경우로 응용하는 방법을 제시하고 이의 효율을 Greenberg et al.의 경우와 비교해 보고자 한다.

2. 양적속성기법

민감한 변수 X의 미지의 모평균 μ_x를 추정하는데 있어 무관질문기법을 양적속성으로 확장한 Greenberg et al.의 기법은 다음과 같다.

민감한 변수 X가 연속인 밀도함수 $g(\cdot)$를 갖고, Y를 밀도함수 $h(\cdot)$를 갖는 무관속성이라 하며, Y의 모평균 μ_y는 알고 있다고 하자.

단순임의복원으로 추출된 n명의 응답자에 대해서 확률 p로 X값에 응답하고 확률 q=1-p로 Y값에 응답하게 한다. 이 때 임의의 응답자가 Z라고 응답하면 Z의 밀도함수는 다음과 같고

1) 본 연구는 1996년도 제2차 전주대학교 학술연구비 지원에의하여 수행 되었음.
2) (560-759) 전주시 완산구 효자동 1200 전주대학교 응용통계학과 조교수

$$f(z)=pq(z)+qh(z) \tag{2.1}$$

이 때, 기대할 수 있는 평균응답 즉 Z의 모평균 μ_z는 다음과 같다.

$$\mu_z = p\mu_x + q\mu_y \tag{2.2}$$

한편, Z의 모분산 σ_z^2는 다음과 같다.

$$\begin{aligned}\sigma_z^2 &= E(Z^2)-[E(Z)]^2 \\ &= p\sigma_x^2+q\sigma_y^2+pq(\mu_x-\mu_y)^2\end{aligned} \tag{2.3}$$

단, σ_x^2과 σ_y^2은 각각 X와 Y의 모분산.

이제, $Z_1, Z_2, \cdots, Z_n$을 n명의 응답자로부터의 응답이라고 하면 표본평균과 표본분산은 다음과 같다.

$$\bar{Z}=\sum_{i=1}^{n} Z_n/n,$$

$$s_z^2 = \sum_{i=1}^{n}(Z_i-\bar{Z})^2/(n-1)$$

나아가,

$$E(\bar{Z}) = \mu_z \tag{2.4}$$

$$E(s_z^2) = \sigma_z^2 \tag{2.5}$$

이므로 μ_x의 불편추정량과 $\hat{\mu}_x$의 분산은 다음과 같다.

$$\hat{\mu}_x = \frac{\bar{Z}-q\mu_y}{p} \tag{2.6}$$

$$Var_1(\hat{\mu}_x) = \frac{\sigma_z^2}{np^2} \tag{2.7}$$

3. 2단계기법

n명의 단순임의복원 추출된 응답자에 대하여, 2단계 확률화응답기법을 이용한 민감변수 X의 모평균 μ_x를 추정하는 방법을 살펴보자.

먼저 다음과 같이 구성된 두개의 확률장치를 준비한다. 즉, 확률장치 R_1은 (i)"당신의 민감변수 X에 대한 값은 얼마입니까 ?"와 (ii)"확률장치 R_2로 가시오"가 각각 추출확률 T와 (1-T)로 구성되어 있고, 확률장치 R_2는 (i) "당신의 민감변수 X에 대한 값은 얼마입니까?"와 (ii)"당신의 무관속성 Y에 대한 값은 얼마입니까?"가 각각 추출확률 p와 (1-p)로 구성된 두개의 확률장치를 준비한다.

응답자로부터 응답 $Z_1, Z_2, \cdots, Z_n$을 얻은 과정을 나타내 보면 다음의 그림과 같다.

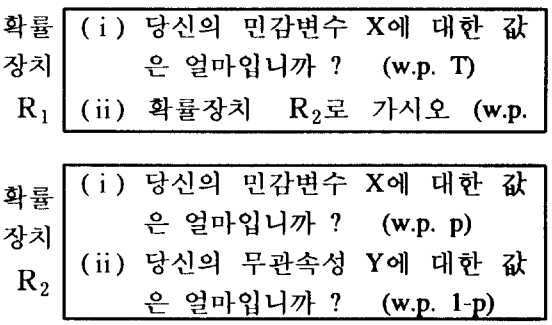

그림 2.1 양적속성 추정을 위한 2단계 확률장치

무관속성 Y에 대한 모평균 μ_y가 알려져 있고 모든 응답자가 선택된 질문에 정직하게 응답하였다면, 응답에 대한 밀도함수는 다음과 같다.

$$f(z) = Tg(z) + (1-T)[pq(z) + qh(z)] \tag{3.1}$$

따라서 이로부터 μ_z와 σ_z^2는 다음과 같다.

$$\mu_z = T\mu_x + (1-T)[p\mu_x + q\mu_y] \tag{3.2}$$

$$\begin{aligned}\sigma_z^2 &= E(z^2) - [E(z)]^2 \\ &= [T+(1-T)p]\sigma_x^2 + q(1-T)\sigma_y^2 + [T+(1-T)p][q(1-T)](\mu_x-\mu_y)^2\end{aligned} \tag{3.3}$$

식 2.3에서 $p=[T+(1-T)p]$, $q=[q(1-T)]$이면 이는 식 3.3과 같게 됨을 알 수 있다.

[정리 1] 양적속성을 추정하기 위한 2단계 확률화응답기법에서 모평균의 추정량 $\hat{\mu}_x$은

$$\hat{\mu}_x = \frac{\bar{Z} - (1-T)q\mu_x}{T + (1-T)p} \tag{3.4}$$

이고 이는 모평균 μ_x의 불편추정량이다.

(증명) $Z_1, Z_2, \cdots, Z_n$을 n명의 응답자로 부터의 응답이고 $\bar{Z} = \sum_{i=1}^{n} Z_i/n$이라면

$E(\bar{Z}) = \mu_z$이므로 식 3.2로부터 식 3.4와 이의 불편성이 증명된다. ∎

한편 식 3.4로부터 $\hat{\mu}_x$의 분산은 다음과 같다.

$$Var_2(\hat{\mu}_x) = \frac{\sigma_z^2}{n[T + (1-T)p]^2} \tag{3.5}$$

[정리 2] 양적속성에 대한 추정시, 2단계 확률화응답기법을 이용한 모평균의 추정량 $\hat{\mu}_x$의 분산 $Var_2(\hat{\mu}_x)$은 Greenberg et al.기법에서의 분산 $Var_1(\hat{\mu}_x)$보다 항상 작다.

(증명) 식 2.3과 식 2.7로 부터

$$Var_1(\hat{\mu}_x)=\frac{p\sigma_x^2+q\sigma_y^2+pq(\mu_x-\mu_y)^2}{np^2}$$ 이다.

또한 식 3.3과 식 3.5로부터

$$Var_2(\hat{\mu}_x)=\frac{[T+(1-T)p]\sigma_x^2+(1-T)q\sigma_y^2+[T+(1-T)p](1-T)q(\mu_x-\mu_y)^2}{n[T+(1-T)p]^2}$$

이다.

이제 각 항별 비교로 통하여 $Var_1(\hat{\mu}_x) \geq Var_2(\hat{\mu}_x)$을 보이자.

첫째항에서, $\frac{1}{p} \geq \frac{1}{T+(1-T)p}$ 이고

세째항에서, $(1-T)p \leq T+(1-T)p$의 관계가 성립하므로 $\frac{q}{p} \geq \frac{(1-T)q}{T+(1-T)p}$ 이다.

마지막으로 둘째항의 관계에 대하여 다음을 고려해 보자

$$\frac{q}{p^2} \geq \frac{(1-T)q}{[T+(1-T)p]^2} = \frac{q}{\left[\frac{T}{1-T}+\frac{1-t}{\sqrt{1-T}}\cdot p\right]^2} \cdots (*)$$

위의 (*)식이 성립한다면 다음의 관계가 만족된다.

$$p \leq \frac{T}{\sqrt{1-T}} + \frac{(1-T)}{\sqrt{1-T}}p$$

$$\sqrt{1-T}\cdot p \leq T+(1-T)p$$

이제, $k(p) = T+(1-T)p-\sqrt{1-T}p$

$= T+\sqrt{1-T}(\sqrt{1-T}-1)p$

라면 이는 p에 대한 1차함수로 기울기가 음이므로 $k(1)>0$이면 $k(p)\geq 0$이다. 그런데

$$k(1)=T+(1-T)-\sqrt{1-T}$$

$$= 1-\sqrt{1-T} \geq 0$$

이므로 (*)식 역시 성립한다. 따라서 $Var_1(\hat{\mu}_x) \geq Var_2(\hat{\mu}_x)$이다. ∎

4. 결 론

본 연구에서 우리는 양적 확률화응답기법의 한 일환으로 2단계 확률화응답기법을 제시하고 이의 효율성을 검토해 보았다. 결과적으로 제시된 기법이, 분산의 측면에서 기존의 방법보다 효율적임을 알 수 있었다. 그런데 확률화응답기법을 이용한 조사시 상존하는 문제는 효율성 뿐만 아니라 신분보호(privacy protection)문제이다. 모든 반복조사기법이 그렇듯이 본 연구에서 고려된 2단계 확률화응답기법 역시 신분보호측면에서는 타당도가 떨어질 것으로 생각된다. 이러한 부분에 대해서는 향후 연구를 통하여 보완할 생각이며 끝으로 확률화응답을 이용한 조사기법을 연구하는데 있어 많은 도움을 주신 원광대학교 김혁주교수님께 감사를 드린다.

참고문헌

[1] Dalenius, T., and Vitale, R.A. (1974). A new RR design for estimating the mean of a distribution. *Tech. Rep.*, 78, Brown University, Providence, R.J.

[2] Duffy, J.C., and Waterton, J.J. (1984). RR model for estimating the distribution function of a quantitative character. *Internat. Statist. Rev.*, 52, 165-171

[3] Greenberg, B.G., Kubler, R.R., Abernally, J.R., and Horvitz, D.G. (1971). Applications of the RR technique in obtaining quantitative data. *Journal of American Statistical Association,* 66, 243-250.

[4] Mangat, N.S., and Singh, R. (1990). An alternative randomized response procedure, *Biometrika,* 77, 439-442.

[5] Parzen, E. (1962). On estimation of a probability density and mode. *Annals of Mathmatical Statistics,* 33, 1065-1076.

[6] Scheult, A.H. (1970). On unbiased estimation of density function. Ph.D. thesis, North-Carolina State University.

[7] Warner, S.L. (1971). The linear RR model. *Journal of American Statistical Association,* 66, 884-888.