The Korean Communications
in Statistics Vol. 5, No. 2, 1998
pp. 403-410

개선된 양적속성의 무관질문모형[1)]

이 기 성[2)]

요 약

Mangat(1994)는 Mangat-Singh(1990)이 제안한 2단계 관련질문모형의 사용 절차를 좀 더 단순화시킨 개선된 관련질문모형을 제안하여 민감한 질적 속성을 추정하였다. 본 논문에서는 Mangat의 개선된 관련질문모형을 양적속성의 무관질문모형으로 확장하고자 한다. 또한, 제안한 모형이 Greenberg et al.의 양적속성의 무관질문모형이나 최경호(1996)가 제안한 2단계 양적속성의 무관질문모형보다 효율적이 되는 조건을 제시하였다.

1. 서 론

Warner(1965)는 확률장치를 통한 간접응답으로 응답자의 신분이나 비밀을 노출시키지 않고서 민감한 질문에 대해 정보를 이끌어 낼 수 있는 확률화응답모형(randomized response model ; RRM)을 처음으로 제시하였다. 그 후 Greenberg et al.(1971)은 무관질문모형(unrelated question model)을 제안하여 양적속성(quantitative attribute)을 갖는 경우로 확장하였으며, 이러한 양적속성의 확률화응답모형은 그 이후 많은 학자들에 의해 연구, 발전되어 왔다.

최근에 Mangat-Singh(1990)은 2단계 관련질문모형을 제안하여 민감한 질적 속성을 추정하였으며, 최경호(1996)는 이 모형을 양적속성을 추정할 수 있는 2단계 무관질문모형으로 확장시켰다. 또한, Mangat(1994)는 Mangat-Singh의 2단계 관련질문모형에서 민감한 질적 속성을 추정하기 위해 사용한 2개의 확률장치를 1개로 줄여 그 사용 절차를 좀 더 단순화 한 개선된 관련질문모형을 제안하였다.

본 논문에서는 Mangat의 개선된 질적 속성의 관련질문모형을 양적속성의 무관질문모형으로 발전시켜 개선된 양적속성의 무관질문모형을 제안하고자 한다. 그리고, 제안한 모형이 Greenberg et al.의 양적속성의 무관질문모형이나 2단계 양적속성의 무관질문모형보다 효율적이 되는 조건을 제시하였다.

2. 양적속성의 무관질문모형

2.1 양적속성의 무관질문모형

1) 이 논문은 1998년도 우석대학교 학술연구조성비에 의하여 연구되었음.
2) (565-701) 전북 완주군 삼례읍 후정리 490 우석대학교 전산통계학과 조교수

Greenberg et al.은 질적 속성의 무관질문모형을 양적속성으로 확장하였는데 그 내용을 살펴보면 다음과 같다.

응답자들은 설문 1이 선택될 확률이 p이고 설문 2가 선택될 확률이 $1-p$인 다음과 같은 2개의 설문으로 구성된 확률장치를 통해 선택된 설문에 대해 응답하게 된다.

설문 1 : 당신의 민감한 변수 X에 대한 값은 얼마입니까?
설문 2 : 당신의 무관한 변수 Y에 대한 값은 얼마입니까?

민감한 질문에 대해 민감한 변수 X가 연속인 밀도함수 $g(\cdot)$를 갖는다고 가정하고, Y를 밀도함수 $h(\cdot)$를 갖는 무관속성이라 하자. 그리고 X의 모평균 μ_x를 추정하는데 있어서, 무관한 변수 Y의 모평균 μ_y를 알고 있다고 가정하자.

이 때, 응답자가 Z라고 응답하면 Z의 확률밀도함수는 다음과 같다.

$$f(z) = pg(z) + (1-p)h(z) . \tag{2.1}$$

따라서, 기대할 수 있는 응답의 평균 즉, Z의 모평균 μ_z와 모분산 σ_z^2은 다음과 같다.

$$\mu_z = p\mu_x + (1-p)\mu_y , \tag{2.2}$$

$$\sigma_z^2 = p\,\sigma_x^2 + (1-p)\,\sigma_y^2 + p(1-p)(\mu_x - \mu_y)^2 . \tag{2.3}$$

여기서, σ_x^2은 X의 모분산이며, σ_y^2은 Y의 모분산이다.

단순임의복원추출된 n명의 응답자들이 확률장치를 통해 $z_i (i = 1, 2, \cdots, n)$라고 응답했을 때, 모평균 μ_x의 추정량 $\hat{\mu}_x$는 다음과 같다.

$$\hat{\mu}_x = \frac{\bar{z} - (1-p)\mu_y}{p} , \tag{2.4}$$

여기서, $\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$이고, $E(\bar{z}) = \mu_z$이므로 $\hat{\mu}_x$는 μ_x의 불편추정량이다.

또한, 추정량 $\hat{\mu}_x$의 분산은 다음과 같다.

$$V_1(\hat{\mu}_x) = \frac{\sigma_z^2}{np^2} . \tag{2.5}$$

2.2 2단계 양적속성의 무관질문모형

최경호는 Mangat-Singh(1990)의 2단계 질적 속성의 관련질문모형을 2단계 양적속성의 무관질문모형으로 확장하였는데 그 내용을 살펴보면 다음과 같다.

응답자들은 1단계로 설문 1이 선택될 확률이 T이고 설문 2가 선택될 확률이 $1-T$인 다음과 같은 2개의 설문으로 구성된 확률장치 R_1에서 선택된 설문에 대해 응답하게 된다.

설문 1 : 당신의 민감한 변수 X에 대한 값은 얼마입니까?
설문 2 : 확률장치 R_2로 가시오.

한편, 응답자들은 확률장치 R_1에서 설문 2가 선택된 경우 2단계로 설문 1이 선택될 확률이 p이고 설문 2가 선택될 확률이 $1-p$인 다음과 같은 2개의 설문으로 구성된 확률장치 R_2에서 선택된 설문에 대해 응답하게 된다.

설문 1 : 당신의 민감한 변수 X에 대한 값은 얼마입니까?
설문 2 : 당신의 무관한 변수 Y에 대한 값은 얼마입니까?

민감한 변수 X의 모평균 μ_x를 추정하는데 있어서, 무관한 변수 Y의 모평균 μ_y를 알고 있다고 가정하자.

이 때, 응답자가 Z라고 응답하면 Z의 확률밀도함수는 다음과 같다.

$$f(z) = Tg(z) + (1-T)[pg(z) + (1-p)h(z)]. \tag{2.6}$$

따라서, Z의 모평균 μ_z는

$$\mu_z = T\mu_x + (1-T)[p\mu_x + (1-p)\mu_y] \tag{2.7}$$

이고, 모분산 σ_z^2은 다음과 같다.

$$\sigma_z^2 = \{p + T(1-p)\}\,\sigma_x^2 + (1-T)(1-p)\,\sigma_y^2 + \{p + T(1-p)\}(1-T)(1-p)(\mu_x-\mu_y)^2, \tag{2.8}$$

여기서, σ_x^2은 X의 모분산이며, σ_y^2은 Y의 모분산이다.

단순임의복원추출된 n명의 응답자들이 확률장치를 통해 $z_i (i = 1, 2, \cdots, n)$라고 응답했을 때, 모평균 μ_x의 추정량 $\hat{\mu}_x$는 다음과 같다.

$$\hat{\mu}_x = \frac{\bar{z} - (1-T)(1-p)\mu_y}{p + T(1-p)} \ , \tag{2.9}$$

여기서, $\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$이고, $E(\bar{z}) = \mu_z$이므로 $\hat{\mu}_x$는 μ_x의 불편추정량이다.

또한, 추정량 $\hat{\mu}_x$의 분산은 다음과 같다.

$$V_2(\hat{\mu}_x) = \frac{{\sigma_z}^2}{n\{p + T(1-p)\}^2} \ . \tag{2.10}$$

3. 개선된 양적속성의 무관질문모형

단순임의복원추출된 n명의 응답자들은 민감한 속성 X를 가지고 있으면 "당신의 민감한 변수 X에 대한 값은 얼마입니까?"에 직접 응답하게 되고, 민감한 속성을 가지고 있지 않고 무관한 속성 Y를 가지고 있으면 다음과 같은 2개의 설문으로 구성된 양적속성의 무관질문모형의 확률장치를 이용하게 된다.

설문 1 : 당신의 민감한 변수 X에 대한 값은 얼마입니까?
설문 2 : 당신의 무관한 변수 Y에 대한 값은 얼마입니까?

여기서, 설문 1이 선택될 확률은 p이고, 설문 2가 선택될 확률은 $1-p$이다. 이 때, 응답자들은 확률장치에 의해서 선택된 설문에 대해 정직하게 응답한다. 응답자에 의해 이루어지는 이러한 모든 절차는 조사자가 관찰할 수 없다. 따라서, 응답자들의 응답 결과는 민감한 변수 X에 대한 응답인지 무관한 속성 Y에 대한 응답인지를 조사자가 알 수 없으므로 응답자 자신의 신분이나 프라이버시를 보호받게 된다.

민감한 변수 X의 모평균 μ_x를 추정하는데 있어서, 무관한 변수 Y의 모평균 μ_y를 알고 있다고 가정하자.

이 때, 응답자가 Z라고 응답하면 Z의 확률밀도함수는 다음과 같다.

$$f(z) = g(z) + (1-p)h(z) \ . \tag{3.1}$$

따라서, 기대할 수 있는 응답의 평균 즉, Z의 모평균 μ_z는

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

이며, 모분산 σ_z^2을 구해보면 다음과 같다.

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

여기서, σ_x^2은 X의 모분산이며, σ_y^2은 Y의 모분산이다.

단순임의복원추출된 n명의 응답자들이 $z_i (i = 1, 2, \cdots, n)$라고 응답했을 때, 개선된 양적속성의 무관질문모형에서 민감한 속성에 대한 모평균 μ_x의 추정량 $\hat{\mu}_x$는 다음과 같다.

$$\hat{\mu}_x = \bar{z} - (1-p)\mu_y . \tag{3.4}$$

여기서, $\bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i$이다.

한편, 추정량 $\hat{\mu}_x$의 기대값은

$$\begin{aligned}E(\hat{\mu}_x) &= E[\bar{z} - (1-p)\mu_y] \\ &= \mu_z - (1-p)\mu_y \\ &= \mu_x\end{aligned}$$

가 되므로 $\hat{\mu}_x$는 μ_x의 불편추정량이다.

<정리 1> 개선된 양적속성의 무관질문모형에서 민감한 속성에 대한 모평균 μ_x의 추정량 $\hat{\mu}_x$의 분산은 다음과 같다.

$$V_3(\hat{\mu}_x) = \frac{\sigma_x^2 + (1-p)\sigma_y^2 + (1-p)(p\mu_y - 2\mu_x)\mu_y}{n} . \tag{3.5}$$

(증명)

$$V_3(\hat{\mu}_x) = V_3[\bar{z} - (1-p)\pi_y]$$
$$= \frac{\sigma_z^2}{n}$$
$$= \frac{\sigma_x^2 + (1-p)\sigma_y^2 + (1-p)(p\mu_y - 2\mu_x)\mu_y}{n}.$$

■

4. 효율성 비교

본 논문에서 제안한 개선된 양적속성의 무관질문모형의 분산 식 (3.5)와 Greenberg et al.의 양적속성의 무관질문모형의 분산 식 (2.5)를 이용하여 다음과 같은 정리를 얻을 수 있다.

<정리 2> 제안한 추정량 $\hat{\mu}_x$의 분산 $V_3(\hat{\mu}_x)$는 Greenberg et al.의 양적속성의 무관질문모형의 분산 $V_1(\hat{\mu}_x)$보다 항상 작다.

(증명)

$V_1(\hat{\mu}_x) - V_3(\hat{\mu}_x) \geq 0$을 이용하여 $V_1(\hat{\mu}_x) \geq V_3(\hat{\mu}_x)$임을 보이기로 하자. 식 (2.3)과 식 (2.5) 및 식 (3.5)로부터

$$V_1(\hat{\mu}_x) - V_3(\hat{\mu}_x)$$
$$= \frac{p\sigma_x^2 + (1-p)\sigma_y^2 + p(1-p)(\mu_x - \mu_y)^2}{np^2} - \frac{\sigma_x^2 + (1-p)\sigma_y^2 + (1-p)(p\mu_y - 2\mu_x)\mu_y}{n}$$
$$= \frac{1}{np^2}[p\sigma_x^2 - p^2\sigma_x^2 + (1-p)\sigma_y^2 - p^2(1-p)\sigma_y^2$$
$$+ p(1-p)(\mu_x - \mu_y)^2 - p^2(1-p)(p\mu_y - 2\mu_x)\mu_y]$$
$$= \frac{1}{np^2}[p(1-p)\sigma_x^2 + (1-p)(1-p^2)\sigma_y^2 + p(1-p)\{\mu_x - (1-p)\mu_y\}^2 + 2p(1-p)\mu_y^2\}] \geq 0$$

이므로, $V_1(\hat{\mu}_x) \geq V_3(\hat{\mu}_x)$이 성립된다.

■

또한, 본 논문에서 제안한 개선된 양적속성의 무관질문모형의 분산 식 (3.5)와 2단계 양적속성의 무관질문모형의 분산 식 (2.10)을 이용하여 다음과 같은 정리를 얻을 수 있다.

<정리 3> 제안한 추정량 $\hat{\mu}_x$는 다음과 같은 조건에서 2단계 양적속성의 무관질문모형의 추정량보다 더 효율적이다.

$$\mu_x \approx \mu_y \text{이면서 } T < \frac{1}{2} \text{ 이고 } 1-p > T \text{이다.}$$

(증명)

$V_2(\hat{\mu}_x) - V_3(\hat{\mu}_x)$을 이용하여 $V_2(\hat{\mu}_x) \geq V_3(\hat{\mu}_x)$을 만족하는 조건을 구해보기로 하자. 식 (2.8)과 식 (2.10) 및 식 (3.5)로부터

$$V_2(\hat{\mu}_x) - V_3(\hat{\mu}_x)$$

$$= \frac{\{p+T(1-p)\}\,\sigma_x^2 + (1-T)(1-p)\,\sigma_y^2 + \{p+T(1-p)\}(1-T)(1-p)(\mu_x-\mu_y)^2}{n\{p+T(1-p)\}^2}$$

$$- \frac{\sigma_x^2 + (1-p)\,\sigma_y^2 + (1-p)(p\mu_y - 2\mu_x)\mu_y}{n}$$

이므로, 수식의 표현을 간소화하기 위하여 $p+T(1-p) = A(>0)$로 놓고 정리하면 다음과 같다.

$$V_2(\hat{\mu}_x) - V_3(\hat{\mu}_x) = \frac{1}{nA^2}\big[A(1-A)\,\sigma_x^2 + (1-p)(1-T-A^2)\,\sigma_y^2 + A(1-T)(1-p)(\mu_x-\mu_y)^2 - A^2(1-p)(p\mu_y-2\mu_x)\mu_y\big].$$

첫 번째 항에서 $1-A = (1-p)(1-T) > 0$이므로 $A(1-A)\,\sigma_x^2 \geq 0$이 성립된다.

두 번째 항에서 $1-T-A^2$이 양수가 되기 위한 조건은 $1-A > T$이고 $1+A > T$이다. 따라서, $1+A > T$에서 $T < \frac{1+p}{p}$ 의 관계를 얻을 수 있는 데 이는 항상 성립됨을 알 수 있다. 또한, $1-A > T$에서 $\frac{T}{1-T} < 1-p$의 관계로부터 $1-T > T$이고 $1-p > T$임을 알 수 있다. 따라서, $(1-p)(1-T-A^2)\,\sigma_y^2 \geq 0$가 성립되기 위한 조건은 $T < \frac{1}{2}$ 이고 $1-p > T$이다.

세 번째 항인 $A(1-T)(1-p)(\mu_x-\mu_y)^2$는 항상 양수이다.

네 번째 항에서 만약 $\mu_x \approx \mu_y$라 하면 $A^2(1-p)(p\mu_y-2\mu_x)\mu_y$는 음수가 되며, 결과적으로 $-A^2(1-p)(p\mu_y-2\mu_x)\mu_y$는 양수가 된다.

그러므로, $V_2(\hat{\mu}_x) \geq V_3(\hat{\mu}_x)$를 만족하는 조건은 $\mu_x \approx \mu_y$이면서 $T < \frac{1}{2}$ 이고 $1-p > T$이 됨을 알 수 있다.

■

참고문헌

[1] 류 제복, 홍 기학, 이 기성 (1993). 「확률화응답모형」, 자유아카데미, 서울.

[2] 최 경호 (1996). 양적속성 추정을 위한 2단계 확률화응답기법, 「한국통계학회논문집」, 제 4권 1호, 161-165.

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

[4] Greenberg, B. G., Kubler, R. R., Abernally, J. R., and Horvitz, D. G. (1971). Applications of the Randomized Response Technique in Obtaining Quantitative Data, *Journal of the American Statistical Association*, 66, 243-250.

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

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

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