# A Comparison of Randomized Response Technique with Direct Response Method of having Nonresponse

홍기학[1], 이기성[2], 손창균[3]

요 약

민감한 모집단의 모수 추정 방법으로 확률화응답기법과 무응답을 고려한 직접조사 방법의 효율성을 비교 분석하였다.

주요용어 : 면대면조사, RRT, 무응답, 응답확률, PPSWR

## 1. 서론

일반적으로 직접조사방법이란 조사자와 응답자의 직접교류에 의해 정보가 얻어지는 방법을 총칭한다. 개인의 사생활 또는 사회적 이슈에 민감한 모수를 갖는 모집단을 대상으로 한 표본조사에서 정보수집방법으로 직접조사방법을 사용할 경우 응답자들로부터의 무응답으로 인해 그들 모수에 대한 타당성 있는 추정을 하기가 힘들다. 간접조사방법은 조사자와 응답자간의 연결매체를 통한 간접교류에 의해 정보가 수집되는 과정을 말한다.

대표적인 직접조사방법으로 면대면조사(face to face survey)를 들 수 있고, 간접조사방법으로는 확률화응답기법(randomized response technique : RRT)을 들 수 있다.

면대면조사는 조사시점에서 응답자들의 응답거부나 무응답으로 인한 무응답오차(nonresponse error)가 발생할 수 있고, RRT는 확률장치의 사용으로 인한 정보의 손실이 발생할 수 있다.

본 논문에서는 민감한 모집단에 대한 모수 추정에서 무응답을 고려한 직접조사방법과 정보의 손실을 고려한 양적 속성의 RRT를 제안하고 그 효율성을 비교 분석하고자 한다. 제 2장과 제 3장에서는 무응답을 고려한 직접조사방법과 정보의 손실을 고려한 양적 속성의 RRT에 대하여 살펴보고 제 4장에서는 이 들 두 방법의 추정량의 효율성을 분산 측면에서 비교 분석하고자 한다.

## 2. 무응답을 고려한 직접질문방법

크기 $N$의 유한모집단 $I=(1,\cdots,i,\cdots,N)$으로부터 크기 $n$의 표본 $s$를 확률비례복원추출(probability proportional to size with replacement : PPSWR)로 뽑는 경우를 가정한다.

$z_i(1,\cdots,i,\cdots,N)$를 $i$번째 응답자의 추출확률이라고 놓으면, 무응답을 고려한 모집단 평균 $\overline{Y}=\frac{1}{N}\sum_{i=1}^{N}y_i$에 대한 추정량은 다음과 같이 쓸 수 있다.

$$\hat{e}_{DR}=\frac{1}{N}\sum_{i=1}^{n}\frac{R_i y_i}{nz_i}. \tag{2.1}$$

1) 동신대학교 컴퓨터학과 교수, 전남 나주시 대호동 252
2) 우석대학교 e-정보공학과 교수, 전북 완주군 삼례읍 후정리 490
3) 협성대학교 교양학부 전임강사, 경기도 화성시 봉담읍 상리 14번지

위 식에서 $R_i$는 지시변수로서 표본으로 뽑힌 $i$번째 응답자가 응답을 하면 1, 그렇지 않으면 0값을 갖는다. 이 때 응답확률은 다음과 같이 표현할 수 있다.

$$P(R_i=1) = p_i.$$

$(E_1, E_2)$ 및 $(V_1, V_2)$를 각각 계획 및 응답변수에 대한 기대값과 분산이라고 놓으면, 식 (2.1)은 모집단 평균 $\overline{Y}$에 대한 편향추정량이다.

$$E(\hat{e}_{DR}) = E\left(\frac{1}{N}\sum_{i=1}^{n}\frac{R_i y_i}{nz_i}\right) = E_1E_2\left(\frac{1}{N}\sum_{i=1}^{n}\frac{R_i y_i}{nz_i}\right)$$
$$= \frac{1}{N}\sum_{i=1}^{N} p_i y_i. \tag{2.2}$$

또한, 추정량 $\hat{e}_{DR}$의 편향과 분산은 다음과 같다.

$$B(\hat{e}_{DR}) = E(\hat{e}_{DR}) - \overline{Y} = -\frac{1}{N}\sum_{i=1}^{N}(1-p_i)y_i, \tag{2.3}$$

$$V(\hat{e}_{DR}) = V\left(\frac{1}{N}\sum_{i=1}^{n}\frac{R_i y_i}{nz_i}\right)$$
$$= \frac{1}{N^2}\left[V_1E_2\left(\sum_{i=1}^{n}\frac{R_i y_i}{nz_i}\right) + E_1V_2\left(\sum_{i=1}^{n}\frac{R_i y_i}{nz_i}\right)\right]$$
$$= \frac{1}{N^2}\left[\frac{1}{n}\sum_{i=1}^{N}\left(\frac{p_i y_i}{z_i} - \sum_{i=1}^{N} p_i y_i\right)^2 z_i + \frac{1}{n}\sum_{i=1}^{N}\frac{y_i^2 p_i(1-p_i)}{z_i}\right]. \tag{2.4}$$

따라서 식 (2.3)과 (2.4)로부터 추정량 $\hat{e}_{DR}$의 평균제곱오차는 다음과 같다.

$$MSE(\hat{e}_{DR}) = V(\hat{e}_{DR}) + \{B(\hat{e}_{DR})\}^2$$
$$= \frac{1}{N^2}\left[\frac{1}{n}\sum_{i=1}^{N}\left(\frac{p_i y_i}{z_i} - \sum_{i=1}^{N} p_i y_i\right)^2 z_i \right.$$
$$\left. + \frac{1}{n}\sum_{i=1}^{N}\frac{y_i^2 p_i(1-p_i)}{z_i} + \left(\sum_{i=1}^{N}(1-p_i)y_i\right)^2\right]. \tag{2.5}$$

## 3. 양적 속성의 RRT

크기 $N$의 유한모집단 $I=(1,\cdots,i,\cdots,N)$으로부터 크기 $n$의 표본 $s$를 확률비례복원추출로 뽑는 경우를 가정한다. 이 때 확률장치는 Eichhorn과 Hayre(1983)의 확률장치를 사용한다. 즉, 확률장치는 평균 $\overline{A}$와 분산 $\sigma_A^2$을 갖는 숫자 $A_1,\cdots,A_m$등을 갖는 표들로 구성된 상자로 이루어진다.

표본으로 뽑힌 $i$번째 응답자는 주어진 상자로부터 민감한 변수와는 독립적으로 $A_j(>0)$가 적힌 표를 뽑아서 $W_i = A_j y_i$로 응답한다. $E_2(W_j) = \overline{A}y_i$이고, 이로부터 확률화응답 $r_i = W_i/\overline{A}$를 얻는다.

이 때, $r_i$는 $y_i$에 대한 비편향추정량이 된다.

$$E_2(r_i) = y_i,$$
$$V_2(r_i) = \sigma_A^2 y_i^2/(\overline{A})^2 = V_i.$$

$z_i(1,\cdots,i,\cdots,N)$를 $i$번째 응답자의 추출확률, $r_i$를 $i$번째 응답자가 확률장치를 통해 $y_i$

값에 대하여 응답한 값이라고 놓으면, 모집단 평균 $\overline{Y}=\frac{1}{N}\sum_{i=1}^{N}y_i$에 대한 확률화응답추정량은 다음과 같이 쓸 수 있다.

$$\hat{e}_{RR}=\frac{1}{N}\sum_{i=1}^{n}\frac{r_i}{nz_i}. \tag{3.1}$$

$(E_1,E_2)$ 및 $(V_1,V_2)$를 각각 계획 및 확률응답변수에 대한 기대값과 분산이라고 놓으면, 식 (3.1)은 모집단 평균 $\overline{Y}$에 대한 비편향추정량이다.

$$\begin{aligned}E(\hat{e}_{RR})&=E\left(\frac{1}{N}\sum_{i=1}^{n}\frac{r_i}{nz_i}\right)=\frac{1}{N}E_1E_2\left(\sum_{i=1}^{n}\frac{r_i}{nz_i}\right)\\&=\frac{1}{N}\sum_{i=1}^{N}y_i\\&=\overline{Y}.\end{aligned} \tag{3.2}$$

이 때, 추정량 $\hat{e}_{RR}$의 분산은 다음과 같다.

$$\begin{aligned}V(\hat{e}_{RR})&=V\left(\frac{1}{N}\sum_{i=1}^{n}\frac{r_i}{nz_i}\right)\\&=\frac{1}{N^2}\left[V_1E_2\left(\sum_{i=1}^{n}\frac{r_i}{nz_i}\right)+E_1V_2\left(\sum_{i=1}^{n}\frac{r_i}{nz_i}\right)\right]\\&=\frac{1}{N^2}\left[\frac{1}{n}\sum_{i=1}^{N}\left(\frac{y_i}{z_i}-Y\right)^2z_i+\frac{1}{n}\sum_{i=1}^{N}\frac{V_i}{z_i}\right].\end{aligned} \tag{3.3}$$

## 4. 효율성 비교

식 (2.5)와 (3.3)을 이용해서 두 추정방법의 효율성을 비교해 보면

$$\begin{aligned}N^2\{MSE(\hat{e}_{DR})-V(\hat{e}_{RR})\}&=\frac{1}{n}\sum_{i=1}^{N}\left(\frac{p_iy_i}{z_i}-\sum_{i=1}^{N}p_iy_i\right)^2z_i+\frac{1}{n}\sum_{i=1}^{N}\frac{y_i^2p_i(1-p_i)}{z_i}\\&\quad+\left(\sum_{i=1}^{N}(1-p_i)y_i\right)^2-\frac{1}{n}\sum_{i=1}^{N}\left(\frac{y_i}{z_i}-Y\right)^2z_i-\frac{1}{n}\sum_{i=1}^{N}\frac{V_i}{z_i}\\&=\frac{1}{n}\left[\sum_{i=1}^{N}\left(\frac{p_iy_i}{z_i}-\sum_{i=1}^{N}p_iy_i\right)^2z_i-\sum_{i=1}^{N}\left(\frac{y_i}{z_i}-Y\right)^2z_i\right]\\&\quad+\frac{1}{n}\sum_{i=1}^{N}\frac{1}{z_i}(y_i^2p_i(1-p_i)-V_i)+\{\sum_{i=1}^{N}(1-p_i)y_i\}^2\end{aligned} \tag{4.1}$$

이 되고, 식 (4.1)에 $V_i=\sigma_A^2y_i^2/(\overline{A})^2$를 대입해서 정리하면,

$$\begin{aligned}N^2\{MSE(\hat{e}_{DR})-V(\hat{e}_{RR})\}&=\frac{1}{n}\left[\left\{\sum_{i=1}^{N}y_i(1-p_i)\right\}\left\{\sum_{i=1}^{N}y_i(1+p_i)\right\}-\sum_{i=1}^{N}\frac{y_i^2}{z_i}(1-p_i^2)\right]\\&\quad+\frac{1}{n}\sum_{i=1}^{N}\frac{1}{z_i}y_i^2\left(p_i(1-p_i)-\frac{\sigma_A^2}{(\overline{A})^2}\right)+\left\{\sum_{i=1}^{N}(1-p_i)y_i\right\}^2.\end{aligned} \tag{4.2}$$

이 된다.

확률화응답기법이 무응답을 수반한 직접질문 방법보다 더 효율적이기 위해서는 다음의 조건을 만족해야 한다.

$$MSE(\hat{e}_{DR})-V(\hat{e}_{RR})\geq 0$$

즉, 식 (4.2)로부터 우변의 세 번째 항은 항상 0보다 크거나 같으므로, 첫 번째 항과 두 번째 항이 0보다 크거나 같으면 된다. 그런데, 첫 번째 항의 괄호 식을 살펴보면

$$\left\{\sum_{i=1}^{N} y_i(1-p_i)\right\}\left\{\sum_{i=1}^{N} y_i(1+p_i)\right\} - \sum_{i=1}^{N}\frac{y_i^2}{z_i}(1-p_i^2)$$
$$= \left\{\sum_{i=1}^{N} y_i(1-p_i)\right\}\left\{\sum_{i=1}^{N} y_i(1+p_i)\right\} - \sum_{i=1}^{N}\frac{1}{z_i}\{y_i(1-p_i)\}\{y_i(1+p_i)\}$$

로 표현할 수 있고, 이는 Cauchy-Schwarz 부등식의 형태로서, $1/z_i$로 인해 음의 값을 가질 가능성이 크다. 두 번째 항의 경우 직접질문으로부터 얻을 수 있는 응답률과 확률장치를 어떻게 구성하느냐에 따라 양수 또는 음수의 값을 가질 수 있다.

따라서 민감한 모집단의 모수 추정을 위한 표본조사에서 무응답을 수반한 직접질문 방법에 대한 확률화응답기법의 효율성은 응답률을 고려한 확률장치의 구조에 많은 영향을 받는다고 볼 수 있다. 그런데 확률장치의 구조는 조사자가 사전에 결정할 수 있으므로 무응답률에 대한 사전 정보에 따라 직접질문 보다는 확률화응답을 통한 간접질문에 의해 추정의 효율성을 얻을 수 있을 것으로 기대된다.

예제) 간단한 수치적인 비교를 위해 $N=10$, $n=5$라 하고, $y_i=\{1,2,3,4,5,6,7,8,9,10\}$, $A_i^1=\{5,6,7,8,9,5,6,7,8,9\}$, $A_i^2=\{1,2,3,4,5,6,7,8,9,10\}$, $A_i^3=\{10,9,8,7,6,5,4,3,2,1\}$ 라고 가정하면, 응답률 $p_i$와 추출확률 $z_i$에 따른 직접적인 비교 결과는 다음과 같다.

<표 4.1> 인위적인 모집단 가정 하에서 효율성 비교

| $z_i$ | $1/N=1/10$ | | | $A_i/\sum A_i$ | | |
|---|---|---|---|---|---|---|
| | $MSE(\hat{e}_{DR}) - V(\hat{e}_{RR})$ | | | | | |
| $p_i$ | $A_i^1$ | $A_i^2$ | $A_i^3$ | $A_i^1$ | $A_i^2$ | $A_i^3$ |
| 0.1 | 28.15 | 21.46 | 21.46 | 28.46 | 23.40 | 6.05 |
| 0.3 | 17.59 | 12.84 | 12.84 | 17.83 | 14.44 | 0.06 |
| 0.5 | 9.45 | 6.15 | 6.15 | 9.62 | 7.42 | -4.00 |
| 0.7 | 3.73 | 1.40 | 1.40 | 3.84 | 2.34 | -6.13 |
| 0.9 | 0.43 | -1.42 | -1.42 | 0.47 | -0.80 | -6.32 |

<표 4.1>로부터 응답률 $p_i$가 작아질수록, 확률장치의 분산이 작을수록 직접질문에 비해 확률화응답기법의 효율성이 증대됨을 알 수 있다.

## 참고문헌

Bansal, M.L., Singh, S., and Singh, R. (1994), Multi-character survey using randomized response technique, *Communication Statistics -Theory & Methods-*, 23(6), 1705-1715.

Chaudhuri, A., and Adhikary, A.K. (1990), Variance estimation with random response, *Communication Statistics -Theory & Methods-*, 19(3), 1119-1125.

Eichorn, B.H., and Hayre, L.S. (1983), Scrambled randomized response methods for obtaining sensitive quantitative data, *Journal of Statistical Planning and*

*Inference*, 7, 307-316.

Lesser, J.T., and Kalsbeeek, W.D.(1992), *Nonsampling error in surveys*, Wiley, U.S.A.

Shaul, K.B., Elizabeta, B., and Benzion, B. (2004), A note on randomized response models for quantitative data, *Metrika*, 60, 255-260.