Journal of Korean
Data & Information Science Society
2004, Vol. 15, No. 4, pp. 731~741

Implementation of Multi-Proportions Randomized Response Model for Sensitive Information at Internet Survey

Hee Chang Park[1] · Ho Min Myung[2]

Abstract

This paper is planned to use multi-proportions randomized response model for sensitive information on internet survey. This is an indirect response technique as a way of obtaining much more precise information. In this system we consider that respondents are generally reluctant to answer in a survey to get sensitive information targeting employees, customers, etc.

Keywords : 다지모집단, 확률화응답기법, E-R diagram

1. 서론

사회가 복잡하고 다양해짐에 따라 신속하고 보다 정확한 정보가 요구되고 이를 충족시키기 위하여 표본조사의 필요성이 점차 증대되고 있다. 사회 여러 분야의 표본조사에서 발생하는 오차에는 표본오차와 비표본오차가 있으며, 최근에 연구의 관심은 비표본오차를 줄이는 데 있다. 이러한 비표본오차는 응답자들이 민감하거나 개인적인 이해와 관계되는 질문을 받았을 경우 더욱 증가하게 된다. 예를 들어 음주운전, 낙태경험, 환각제사용, 동성연애 및 탈세여부 등과 같은 사회적으로나 개인적으로 매우 민감한 문제에 관한 조사에서 기존의 직접질문방식을 그대로 사용할 경우 응답자들이 응답을 회피하거나 거짓으로 응답하는 경향이 뚜렷이 나타나게 된다. 이는 응답자들이 민감한 질문에 응답함으로써 불이익을 받거나 사생활이 보장되지 않는다고 생각하기 때문이다. 민감한 질문에 대한 조사에서 발생하는 비표본오차를 줄이기 위하여 Warner(1965)는 응답자들에게 직접적인 응답을 요구하는 것이 아니라 확률장치를 통

1) First Author : Professor, Department of Statistics, Changwon National University, Changwon, Gyeongnam, 641-773, Korea
E-mail : hcpark@sarim.changwon.ac.kr
2) Graduate Student Department of Statistics, Changwon National University, Changwon, Gyeongnam, 641-773, Korea

한 간접적인 응답만을 요구함으로 응답자들의 신분을 보장해 주는 획기적인 확률화응답기법을 제시하였다.

Abul-Ela 등(1967)은 이지 모집단(dichotomous population)에 대한 Warner의 관련 질문기법을 다지 모집단(polychotomous population)의 경우로 확장하였다. Greenberg 등(1969)은 민감한 질문과 배반되는 질문 대신에 민감한 질문과 전혀 무관한 질문을 사용하는 무관질문기법을 제안하였으며, Greenberg 등(1971)은 이를 민감한 변수에 대한 양적 정보를 얻기 위해 양적속성기법으로 발전시켰다. Loynes(1976)는 Warner기법의 민감한 질문과 배반이 되는 질문 대신에 "예"라고 응답하도록 강요하는 강요질문기법(forced answer technique)을 제안하였다. 또한 Fox와 Tracy(1986), Chaudhuri와 Mukerjee(1988)는 확률화응답기법을 정리, 요약하여 체계화하였다.

국내에서는 류제복 등(1995)이 확률화응답기법이 적용된 사례들을 비교 분석하여 실용화를 위한 방안을 제시한 바 있으며, 이기성(1999)은 2단계 확률화응답모형에 관한 연구를 수행한 바 있다. 또한 박희창 등(2001a, b)은 질적 자료의 관련질문기법에 대한 온라인 설문조사시스템을 구현한 바 있으며, Park과 Myung(2002)은 질적 자료의 무관질문기법에 대한 인터넷시스템을 개발하였다.

본 연구에서는 Abul-Ela의 다지 모형 확률화 응답시스템을 인터넷상에서 구현하여 민감한 정보를 얻은 후 일반 질문 기법과 비교 민감한 정보에 있어서 Abul-Ela의 모형의 효율성을 보고자 한다. 2절에서 다지 모형에 대해 전반적으로 살펴본 후, 3절에서는 시스템의 개발환경과 구성에 대하여 기술하고, 4절에서는 예제를 통하여 구현된 시스템에 관해 토의하며, 5절에서 결론을 맺고자 한다.

2. 다지모형

본 절에서는 다지모집단에 대한 확률화응답모형에 관하여 기술하고자 한다. 모집단은 $t(\geq 2)$개의 상호 배반인 그룹으로 이루어져 있으며, 그 중에서 적어도 하나, 또는 최대로 $t-1$개의 그룹을 민감한 그룹으로 분류할 수 있고, 추정하고자 하는 모집단의 비율을 $\pi_1, \pi_2, \cdots, \pi_t$라고 하자.

여기서 $0 < \pi_j < 1$, $j = 1,2,3,\ldots,t$, 그리고 $\sum_{j=1}^{t} \pi_j = 1$이다.

모집단으로부터 단순임의복원으로 크기가 $n_1, n_2, n_3, \ldots, n_s$인 s개의 독립표본을 추출한다. 여기서 $s = t-1$이다. 확률장치는 s개의 확률장치로 구성되어 있으며, 확률장치 i $(i = 1,2,3,\ldots,s)$는 i번째의 표본을 의미하며, 각각의 확률장치는 t개의 다른 종류의 항목을 가진다. $j(j = 1,2,3,\ldots,t)$번째 종류의 카드는 j번째 그룹을 나타내며, 응답자는 지정된 확률장치를 이용하여 다음과 같이 선택된 설문에 대하여 응답하게 된다.

설문 i : "당신은 j번째 속성을 가지고 있습니까?"

i번째 표본에서 "당신은 j번째 속성을 가지고 있습니까?"라는 설문이 적혀 있는

확률장치항목의 비율을 p_{ij}라고 하자.

여기서 $i=1,2,3,\ldots,s$, $j=1,2,3,\ldots,t$이고, $\sum_{j=1}^{t} p_{ij}=1$이다. 그러면 i번째 표본에 속하는 응답자가 "예"라고 응답할 확률 λ_i는 다음과 같다.

$$\lambda_i = \sum_{j=1}^{t} p_{ij}\,\pi_j, \tag{2.1}$$

여기서 $i=1, 2, 3, \ldots, s$이다. 그런데 $s=t-1$이고, $\sum_{j=1}^{t} \pi_j=1$이므로, 다음의 식 (2.2)가 성립한다.

$$\sum_{j=1}^{s} (p_{ij}-p_{it})\pi_j=\lambda_i-p_{it} \tag{2.2}$$

이 식을 행렬로 표현하면 식 (2.3)와 같다.

$$\mathrm{p}\pi=\xi \tag{2.3}$$

여기는 p, π, ξ는 다음과 같다.

$$\mathrm{p} = \begin{bmatrix} p_{11}-p_{1t} & p_{12}-p_{1t} & \cdots & p_{1s}-p_{1t} \\ p_{21}-p_{2t} & p_{22}-p_{2t} & \cdots & p_{2s}-p_{2t} \\ \vdots & \vdots & & \vdots \\ p_{s1}-p_{st} & p_{s2}-p_{st} & \cdots & p_{ss}-p_{st} \end{bmatrix},$$

$$\pi=(\pi_1, \pi_2, \pi_3, \ldots, \pi_s)',$$

$$\xi=(\lambda_1-p_{1t}, \ldots, \lambda_s-p_{st})'.$$

i번째 표본에서 "예"라고 응답한 수를 n_{i1}이라 하면, λ_i의 불편 추정량 $\hat{\lambda}_i$는 식 (2.4)과 같다.

$$\hat{\lambda}_i = \frac{n_{i1}}{n_i} \tag{2.4}$$

만약 p가 정칙행렬이면 π의 불편 추정량은 식 (2.5)과 같이 얻어진다.

$$\hat{\pi} = (\hat{\pi}_1, \hat{\pi}_2, \hat{\pi}_3, \ldots, \hat{\pi}_s)' = \mathrm{p}^{-1}\mathcal{L}, \tag{2.5}$$

여기서 $\mathcal{L} = (\hat{\lambda}_1 - p_{1t}, \ldots, \hat{\lambda}_s - p_{st})'$이며, π_t의 추정량 $\hat{\pi}_t$는 다음과 같다.

$$\hat{\pi}_t = 1 - \sum_{j=1}^{s} \hat{\pi}_j \tag{2.6}$$

s개의 표본들이 독립적으로 추출되었기 때문에 $n_{11}, n_{21}, n_{31}, \ldots, n_{s1}$과 $\hat{\lambda}_1, \hat{\lambda}_2, \hat{\lambda}_3, \ldots, \hat{\lambda}_s$들은 독립이므로, 각각의 i에 대하여 $n_{i1} \sim b(n_i, \lambda_i)$를 따르게 된다. 그러므로 $\mathcal{L}$의 산포행렬(dispersion matrix) $disp(\mathcal{L})$는 식 (2.7)과 같다.

$$disp(\mathcal{L}) = diag(V_{11}, V_{22}, V_{33}, \ldots, V_{ss}), \tag{2.7}$$

여기서 $V_{ii} = \lambda_i(1-\lambda_i)/n_i$이다. 따라서 $\hat{\pi}$의 산포행렬 $disp(\hat{\pi})$은 식 (2.8)과 같이 나타난다.

$$disp(\hat{\pi}) = \mathrm{p}^{-1}\mathrm{diag}(\mathrm{V}_{11}, \mathrm{V}_{22}, \mathrm{V}_{33}, \ldots, \mathrm{V}_{ss})(\mathrm{p}^{-1})' \tag{2.8}$$

각각의 i에 대해 V_{ii}의 불편 추정량은 식 (2.9)와 같다.

$$\hat{V}_{ii} = \frac{\hat{\lambda}_i(1-\hat{\lambda}_i)}{n_i - 1} \tag{2.9}$$

또한 $disp(\hat{\pi})$의 불편 추정량은 식 (2.10)과 같다.

$$\widehat{disp}(\hat{\pi}) = \mathrm{p}^{-1}\mathrm{diag}(\hat{\mathrm{V}}_{11}, \hat{\mathrm{V}}_{22}, \hat{\mathrm{V}}_{33}, \ldots, \hat{\mathrm{V}}_{ss})(\mathrm{p}^{-1}) \tag{2.10}$$

3. 다지모형의 구현

3.1. 시스템 개발 환경 및 시스템 흐름

구현된 시스템의 개발 환경에서 개발 언어는 gnu c compiler, java, html등이며, 운

영 체제는 Linux를 사용하였다. 또한 데이터베이스는 MySQL-Ver 3.23.39를 이용하였다.

확률화응답시스템은 관리자(조사자) 모드와 응답자 모드 두 가지 부분으로 구성되어 있다. 관리자 모드에서는 설문을 작성하는 에디터와 확률장치의 선택 및 민감한 설문이 선택될 확률을 입력하는 부분으로 이루어져 있고, 기타 계산에 필요한 정보를 입력하도록 되어 있다. 응답자 모드는 실제 응답자가 응답할 수 있도록 이루어져 있으며, 무관질문기법에서 무관한 속성에 대한 모비율을 모를 때와 무관한 변수에 대한 모평균을 모를 때는 접속 순서에 따라 표본 1과 표본 2로 나누어 응답에 참여하도록 하였다. 응답의 결과는 데이터베이스에 저장되며, 민감한 속성(변수)에 대한 모비율(모평균)의 추정값과 분산추정값을 계산한 후에 응답자에게는 민감한 속성(변수)에 대한 모비율(모평균)의 추정값만을 보여주고, 관리자에게는 모비율(모평균)의 추정값과 분산추정값에 대한 결과 모두를 보여 주도록 구성하였다.

본 시스템은 자료의 입력에서 처리, 결과를 모두 데이터베이스를 바탕으로 이루어져 있다. 이로 인하여 동일 응답자 등의 반복 측정에서도 기존의 설문응답시스템과 쉽게 합쳐서 사용할 수 있다. 또한 처리과정에서도 데이터베이스를 사용함으로 인하여 쿼리(query)를 사용하여 수행 속도면에서 파일시스템보다 빠르게 진행할 수 있으며, 기본적인 결과를 데이터베이스에 저장함으로 계속적인 조사에서 추세 분석이 가능하다.

본 시스템의 구현을 위해 설계한 테이블은 <표 3.1>부터 <표 3.4>와 같다.

<표 3.1> 메인테이블

메 인			
Logical Name	Physical Name	Data Type	비 고
인덱스 키	idx	integer	pk, auto_increment
설문 작성일	day	date	
설문의 주제	subject	time	
하부그룹 인덱스	sub_group	tinyint	

<표 3.1>의 메인테이블은 고유한 테이블이다. 여기에 모든 설문 문항들에 대한 정보를 보관하고 있으며 아래에 나오는 설문지라는 테이블과 연계하여 분석을 수행한다. 여기에는 설문 작성일시, 주제, 등이 기록되며 하부그룹 인덱스를 이용하여 각각의 그룹의 내용을 가져온다.

<표 3.2> 표본 그룹 테이블

표본 그룹			
Logical Name	Physical Name	Data Type	비 고
인덱스 키	idx	integer	pk, auto_increment
확률	probility	float	
독립표본의 수	number	mediumint	
독립표본의 번호	first	tinyint	
그룹의 번호	second	tinyint	

<표 3.2>에서 제시하는 표본 그룹 테이블은 독립표본의 수와 독립표본이 선택될 확률을 저장한다.

<표 3.3> 확률 그룹 테이블

확률 그룹			
Logical Name	Physical Name	Data Type	비 고
인덱스 키	idx	integer	pk, auto_increment
확률	probility	float	
독립표본의 번호	first	tinyint	
그룹의 번호	second	tinyint	

<표 3.3>의 확률 그룹 테이블은 각각의 독립표본 그룹이 선택될 확률을 저장한다.

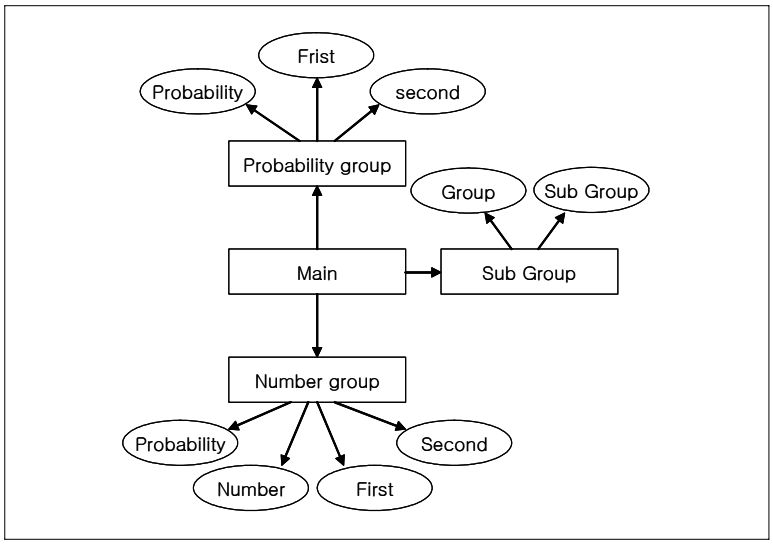
<표 3.4> 하부 그룹 테이블

하부 그룹			
Logical Name	Physical Name	Data Type	비 고
인덱스 키	idx	integer	pk, auto_increment
확률	group_subject	archar(255)	
메인키 인덱스	main_idx	integer	

<표 4.4>의 하부 그룹 테이블에서는 각각의 그룹에 대응하는 문구가 기록된다.

본 시스템에 사용된 데이터베이스 구성은 <그림 3.1>과 같은 구조를 가지고 있다. 메인테이블은 설문이 선택될 확률 및 주제, 작성일자 등의 자료를 가지고 있으며, 표본 테이블은 설문 테이블에서 응답자들이 응답한 결과를 집계하여 결과를 보여주는

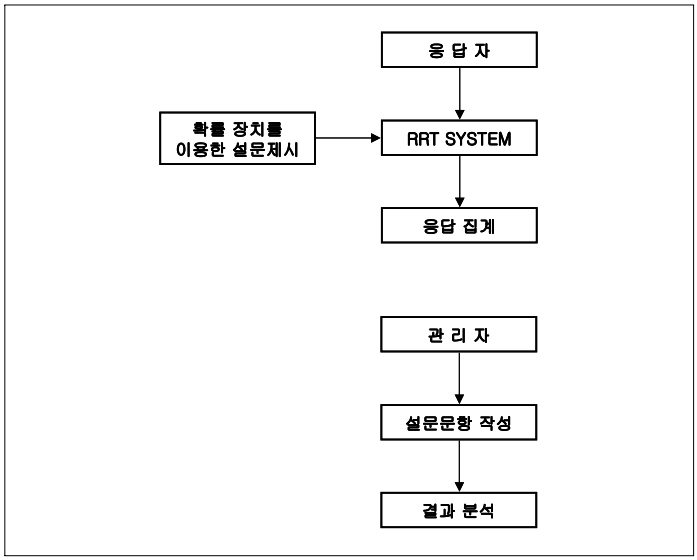
구조를 하고 있다.



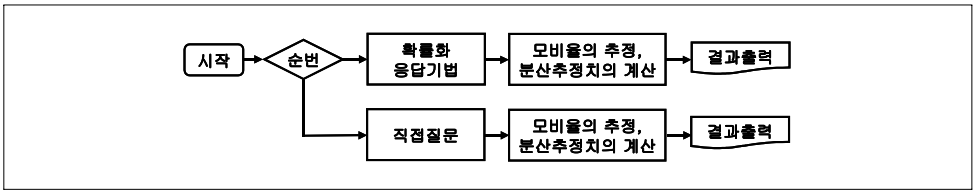
<그림 3.1> E-R Diagram

3.2 시스템의 구성

본 시스템의 구성은 응답자 모드와 관리자 모드 2개로 구분되며, 시스템 구성도는 <그림 3.2>와 같고 시스템의 흐름도는 <그림 3.3>과 같다.



<그림 3.2> 시스템 구성도



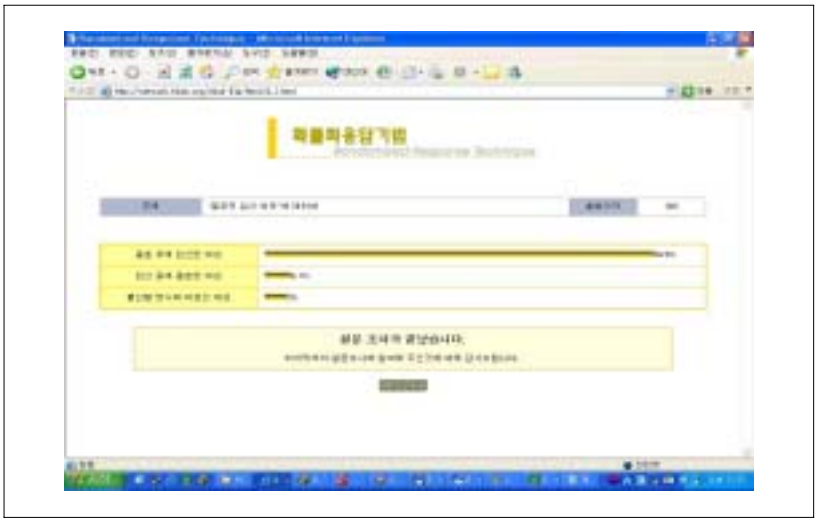
<그림 3.3> 시스템 흐름도

4. 예 제

이 장에서는 다지모형 확률화응답기법을 적용하여 실시한 설문 조사로부터 얻어진 결과에 대해 논의하고자 한다. 기간은 2002년 6월 7일부터 6.30까지 30대의 여성 380명을 대상으로 하여 결혼 전 임신 여부를 설문의 주제로 하였다. 그룹은 다음과 같이 세 그룹으로 나누었다.

그룹 1 : 결혼 후에 임신한 여성
그룹 2 : 임신 중에 결혼한 여성
그룹 3 : 출산할 당시에 미혼인 여성

응답결과는 응답자용과 관리자용으로 구분되어 있으며, 응답자용은 <그림 4.1>에서 보는 바와 같이 민감한 속성을 지닌 그룹의 모비율 추정값만을 보여준다.



<그림 4.1> 응답자용 결과

<그림 4.1>에서와 같이 “결혼전 임신여부”라는 민감한 질문에 대하여 380명이 참가

하였고 결혼 후에 임신한 여성의 추정치는 88.9%, 임신 중에 결혼한 여성의 추정치는 6.5%, 출산한 당시에 미혼인 여성의 추정치는 5%로 추정되었다.

<그림 4.2> 관리자용 결과

관리자는 <그림 4.2>와 같이 확률장치에서 각각의 그룹이 선택될 확률, 그룹에서 "예"라고 답변한 응답자의 수, 모비율 π의 추정량, 모비율 π의 분산 추정치 등을 볼 수 있다. "결혼전 임신여부"에 대한 직접질문에 대한 결과는 <그림 4.3>과 같다.

<그림 4.3> 직접 질문 기법

이와 같은 결과에서 알 수 있듯이 확률화응답기법을 이용한 온라인 설문조사에서 얻어진 모비율의 추정치가 직접질문을 이용한 설문조사에서 얻어진 모비율의 추정치보다 높게 나타났다는 것을 알 수 있다.

5. 결론

본 논문에서는 민감한 정보를 얻기 위한 조사에서 응답자들이 정직하게 응답하기를 꺼리는 질문들에 대하여 직접응답 대신에 간접응답을 통해 응답자의 비밀을 노출시키지 않고서 보다 정확한 정보를 얻을 수 있는 간접응답기법인 확률화응답기법 중 Abul-Ela의 다지 모형 확률화응답시스템을 인터넷상에서 사용할 수 있도록 구현하였다. 그 결과 확률화응답기법을 이용한 온라인 설문조사에서 얻어진 모비율의 추정치가 직접질문을 이용한 설문조사에서 얻어진 모비율의 추정치보다 높게 나타나는 결과를 얻었다.

본 시스템은 기존의 설문조사 시스템과 연계하여 민감한 질문에만 확률장치를 이용할 수 있도록 하여 다른 속성에 따라 민감한 질문에 대한 차이도 볼 수 있을 뿐만 아니라 독립된 단일문항 질문으로도 사용이 가능하도록 하였다. 기업내의 인트라넷이나 공공기관의 경우 민감한 질문에 대한 응답자들의 좀 더 정확한 응답을 기대할 수 있을 것으로 생각된다.

참고문헌

1. 류제복, 이계오, 이기성 (1995). 확률화응답기법의 실용화 방안, 응용통계연구 8(1), 9-26.
2. 박희창, 이기성, 김희재, 남기성 (2001a), 인터넷조사와 설문조사시스템, 자유아카데미, 서울.
3. 박희창, 남기성, 이기성 (2001b). 인터넷조사에서의 확률화응답기술의 구현, 한국통계학회논문집, 제8권, 제3호, 731-737.
4. 이기성(1999). 2단계 집락추출법에 의한 양적속성의 무관질문모형, 한국자료분석학회, 제1권, 제1호, 115-221.
5. Abul-Ela, A., Greenberg, B., and Horvitz, A.(1967), "A multi-proportions randomized response model." *Journal of the American Statistical Association*, 62. 990-1008.
6. Chaudhuri, A. and Mukerjee, R. (1988). *Randomized Response : Theory and Techniques*, Marcel Dekker, Inc., New York.
7. Fox, J. A. and Tracy, P. E. (1986). *Randomized Response : A Method for Sensitive Survey,* Sage Publications.
8. Greenberg, B.E., Abul-Ela, Abdel-Latif A., Simmons, W. R., and Horvitz, D. G. (1969). The Unrelated Question Randomized Response Model : Theoretical Framework, *Journal of the American statistical Association*,

64, 520-539.
9. Greenberg, B. G., Kubler, R. R., Abernathy, 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.
10. Loynes, R. M. (1976). "Asymptotically Optimal Randomized Response Procedures," *Journal of the American Statistical Association*, 71, 924-928.
11. Park, H. C. and Myung, H. M. (2002), "Implementation of Qualitative Unrelated Question Model for Obtaining Sensitive Information," *Journal of the Korean Data & Information Science Society*, 13(2), (Unpublished).
12. Warner, S. L. (1965). "Randomized Response ; A Survey Technique for Eliminating Evasive Answer Bias." *Journal of the American Statistical Association,* 60, 63-69.

[2004년 8월 접수, 2004년 10월 채택]