민감한 양적 정보을 얻기 위한 확률화응답시스템의 구현

박희창[1] · 남기성[2] · 이기성[3]

요약

본 논문에서는 민감한 양적 정보를 얻기 위한 조사에서 응답자들이 정직하게 응답하기를 꺼리는 질문들에 대하여 응답자의 비밀을 노출시키지 않고서 양적 정보에 대한 보다 정확한 정보를 얻을 수 있는 양적 확률화응답기법을 인터넷 상에서 사용할 수 있도록 구현하고자 한다. 본 시스템은 DB 환경에 바탕을 두어 기존의 온라인 설문조사 시스템 및 질적 확률화응답기법과 연계하여 자료를 공유할 수 있을 뿐만 아니라 독립된 스팟 서베이(spot survey)가 가능하도록 구현하고자 한다.

주제어 : 양적 확률화응답시스템, 질적 정보, 양적 정보

1. 서론

최근 들어 온라인 설문조사가 현장에서 많이 행해지고 있으나, 온라인 설문조사 역시 기존의 설문조사와 마찬가지로 조사자가 응답자들의 프라이버시나 사생활과 관련된 민감한 정보를 얻고자 할 경우에 정확한 정보를 얻기란 쉬운 일이 아니다. 한편, 응답자들은 기존의 설문조사와 마찬가지로 온라인 상에서 민감한 질문을 직접적으로 받게 되면 혹시 자신의 비밀이나 사생활이 노출될까 의심하여 정직한 응답을 꺼리게 된다. 따라서, 탈세량, 음주량, 흡연량 등과 같은 민감한 양적 정보를 얻고자 할 경우 간접질문기법인 Greenberg 등(1971)이 제안한 양적 모형의 확률화응답기법을 온라인 설문조사에 적용해 볼 수 있다. 그러면, 응답자들의 응답이 확률장치를 이용하여 간접적으로 이루어지게 되므로 응답자가 자신의 신상에 대한 불안이나 개인 정보의 유출을 이유로 정확하지 않은 응답을 할 가능성을 줄일 수 있게 된다. 그러므로 온라인 설문조사를 하는 데 이러한 양적 정보를 얻기 위한 확률화응답기법을 이용할 수 있는 시스템이 필요하다고 생각된다.

따라서, 본 연구에서는 민감한 양적 정보를 얻기 위한 양적 확률화응답시스템을 구현해 보고자 한다. 본 시스템은 기존의 온라인 설문조사 시스템과 더불어 사용할 수 있을 뿐만 아니라 독립된 스팟 서베이 등이 가능하도록 구현하고자 한다. 그리고 본 시스템은 동일 한 응답자가 여러 번 답하는 것을 막기 위해 로그인(log in)을 하는 사이트에서는 동일 아이디에 대하여 중복 응답을 하는 것을 막을 수 있고, 로그인을 하지 않는 사이트에서는 동일 IP에서 중복 응답하는 것을 막을 수 있도록 하며, 학교

1 창원대학교 통계학과 부교수, (641-773) 경남 창원시 사림동 9
2 창원대학교 통계학과 강사, (641-773) 경남 창원시 사림동 9
3 우석대학교 전산정보학부 부교수, (565-701) 전북 완주군 삼례읍 후정리 490

실습실 등과 같이 여러 명이 사용하는 경우에 대비하여 응답에 제한을 두지 않을 수 있도록 구현하고자 한다.

본 논문은 1장에서 서론으로 민감한 사항에 대한 인터넷 조사와 온라인 설문조사에서 확률화응답기법의 필요성을 설명하고, 2장에서는 확률화응답기법을 소개하며, 3장에서는 구현된 양적 확률화응답시스템에 대하여 살펴보고, 4장에서는 결론과 향후 연구과제를 다루고 있다.

2. 확률화응답기법

사회적으로 민감한 조사에서 응답자들이 응답을 회피하거나 정직하게 응답하지 않는 질문들에 대하여 응답자의 신분이나 비밀을 노출시키지 않고서 민감한 질문에 대한 정보를 이끌어 내기 위하여 Warner(1965)는 확률장치를 이용하여 간접 응답을 하도록 하는 확률화응답기법(randomized response technique ; RRT)을 처음으로 제시하였다. Warner는 응답자들에게 민감한 질문과, 민감한 질문과 배반되는 질문으로 구성된 확률장치를 사용하여 민감한 속성에 대한 질적 정보를 얻고자 하였다. 그 후, Greenberg 등(1969)은 민감한 질문과 배반되는 질문 대신에 민감한 질문과 전혀 무관한 질문을 사용하는 무관질문기법(unrelated question technique)을 제안하였으며, Greenberg 등(1971)은 이를 양적속성 기법으로 발전시켜 민감한 변수에 대한 양적 정보를 얻고자 하였다.

이 장에서 민감한 양적 정보를 얻기 위한 Greenberg 등(1971)의 무관질문기법에 대하여 간략히 소개하고자 한다.

Greenberg 등이 사용한 무관질문기법의 확률장치 R은 다음과 같이 두 가지 설문으로 구성되어 있다.

<확률장치 R>

	설문내용	선택확률
설문 1	당신의 민감한 변수 X에 대한 값은 얼마입니까? (당신은 하루에 몇 개피의 담배를 피우십니까?) ( )개	p
설문 2	당신의 무관한 변수 Y에 대한 값은 얼마입니까? (당신은 하루에 전화를 몇 통화나 하십니까?) ( )통화	$1-p$

응답자들이 선택된 설문에 대하여 Z라고 응답할 확률은 다음과 같다.

$$\mu_z = p\mu_x + (1-p)\mu_y \tag{2.1}$$

여기서, μ_x는 민감한 변수 X에 대한 모평균이고, μ_y는 무관한 변수 Y에 대한 모평균으로 알고 있다고 가정한다.

단순임의복원으로 추출된 n명의 응답자들이 확률장치에 의해서 선택된 설문에 대하여 $z_i(i=1,2,\cdots,n)$라고 응답했을 때, 민감한 변수 X에 대한 모평균 μ_x의 추정량 $\hat{\mu}_x$는 다음과 같다.

$$\hat{\mu}_x = \frac{\bar{z}-(1-p)\mu_y}{p}\ ,\ \bar{z} = \frac{1}{n}\sum_{i=1}^{n} z_i\ . \tag{2.2}$$

이 때, 추정량 $\hat{\mu}_x$의 분산은

$$Var(\hat{\mu}_x) = \frac{{\sigma_z}^2}{np^2} \tag{2.3}$$

이며, $\hat{\mu}_x$의 분산추정량은 다음과 같다.

$$\widehat{Var}(\hat{\mu}_x) = \frac{{s_z}^2}{np^2}\ ,\quad {s_z}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(z_i-\bar{z})^2\ . \tag{2.4}$$

다음으로 무관한 변수 Y에 대한 모평균 μ_y을 모르고 있을 때, 양적 정보를 얻을 수 있는 이표본 무관질문기법에 대하여 살펴보기로 하자. 구하고자 하는 미지 모수가 μ_x와 μ_y이므로 최소한 두개의 표본이 필요하며, 모집단으로부터 단순임의복원으로 크기가 n_1과 n_2인 두개의 독립 표본을 추출한다. 두 개의 표본을 사용해야 하므로 $i(i=1,2)$번째 표본에서 민감한 설문이 선택될 확률이 p_i가 되는 다음과 같은 두 개의 확률장치가 필요하게 된다.

<확률장치 R_i>

	설문내용	선택확률
설문 1	당신의 민감한 변수 X에 대한 값은 얼마입니까? (당신은 하루에 몇 개피의 담배를 피우십니까?) (　　)개	p_i
설문 2	당신의 무관한 변수 Y에 대한 값은 얼마입니까? (당신은 하루에 전화를 몇 통화나 하십니까?)　(　　)통화	$1-p_i$

$i\,(i=1,2)$번째 표본의 응답자들이 Z_i라고 응답할 확률은 다음과 같다.

$$\mu_{zi} = p_i\mu_x + (1-p_i)\mu_y\ . \tag{2.5}$$

n_i명의 응답자들이 확률장치에 의해서 선택된 설문에 대하여 $z_{ij}(i=1,2,j=1,$

$2, \cdots, n_i$)라고 응답했을 때, 민감한 변수 X에 대한 모평균 μ_x의 추정량 $\hat{\mu}_x$는 다음과 같다.

$$\hat{\mu}_x = \frac{(1-p_2)\,\bar{z}_1 - (1-p_1)\,\bar{z}_2}{p_1 - p_2}\ ,\ p_1 \neq p_2\ . \tag{2.6}$$

여기서, $\bar{z}_1 = \frac{1}{n_1}\sum_{j=1}^{n_1} z_{1j}$, $\bar{z}_2 = \frac{1}{n_2}\sum_{j=1}^{n_2} z_{2j}$이다.

이 때, 추정량 $\hat{\mu}_x$의 분산은

$$Var(\hat{\mu}_x) = \frac{\frac{(1-p_2)^2\,\sigma_1^{\,2}}{n_1} + \frac{(1-p_1)^2\,\sigma_2^{\,2}}{n_2}}{(p_1-p_2)^2}\ ,\ p_1 \neq p_2 \tag{2.7}$$

이며, $\hat{\mu}_x$의 분산추정량은 다음과 같다.

$$\widehat{Var}(\hat{\mu}_x) = \frac{\frac{(1-p_2)^2\,s_1^{\,2}}{n_1} + \frac{(1-p_1)^2\,s_2^{\,2}}{n_2}}{(p_1-p_2)^2}\ ,\ p_1 \neq p_2\ . \tag{2.8}$$

여기서, $s_1^{\,2} = \frac{1}{n_1-1}\sum_{j=1}^{n_1}(z_{1j} - \bar{z}_1)^2$, $s_2^{\,2} = \frac{1}{n_2-1}\sum_{j=1}^{n_2}(z_{2j} - \bar{z}_2)^2$이다.

3. 양적 확률화응답시스템의 구현

3.1 시스템 개발 환경 및 시스템 흐름

구현된 시스템의 개발 환경에서 개발 언어는 GCC, Java, HTML 등이며, 운용 환경은 Linux용으로 개발하였다. 또한 DB는 MySQL을 이용하였다.

양적 확률화응답시스템은 관리자 모드와 응답자 모드의 두 가지 모드로 구성되어 있다. 관리자 모드에서는 설문을 작성하는 에디터와 확률장치의 선택 및 민감한 설문이 선택될 확률을 입력하는 부분으로 이루어져 있고, 무관한 변수 Y의 모평균을 알 때와 모를 때를 선택할 수 있다. 응답자 모드는 실제 응답자가 응답을 할 수 있도록 이루어져 있으며, 무관한 변수 Y의 모평균을 모를 때는 접속 순서에 따라 표본 1과 표본 2로 나누어져 응답에 참여하도록 하였다. 응답의 결과는 DB로 저장되며, 민감한 변수 X의 모평균의 추정값과 분산추정값을 계산한 후에 응답자에게는 민감한 변수 X의 모평균의 추정값만을 보여주고, 조사자(관리자)에게는 모평균의 추정값과 분산추정값에 대한 결과를 모두 보여 주도록 구성하였다.

본 시스템은 자료의 입력에서 처리, 결과를 모두 DB를 바탕으로 이루어져 있다. 이로 인하여 동일 응답자 등의 반복 측정에서도 DB를 사용함으로 인하여 기존의 설문

응답시스템과 쉽게 합쳐서 사용할 수 있다. 또한 처리과정에서도 DB를 사용함으로 인하여 쿼리(query)를 사용하여 수행 속도면에서 파일시스템보다 속도에서 보다 빠르게 진행할 수 있다. 또한 기본적인 결과를 DB를 연동함으로 계속적인 조사 등에서 추세 분석이 가능하다.

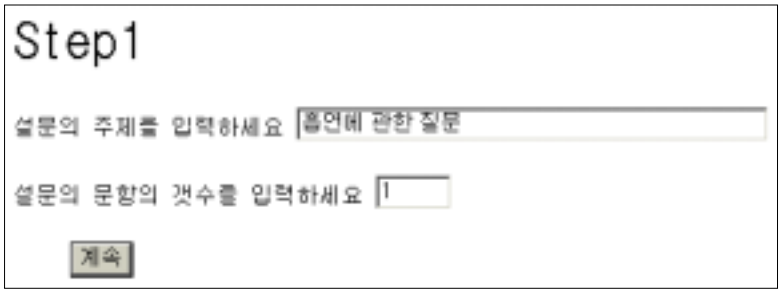
본 시스템의 구현을 위해 설계한 테이블의 항목은 <표 3.1>과 같다.

<표 3.1> 시스템 DB 테이블

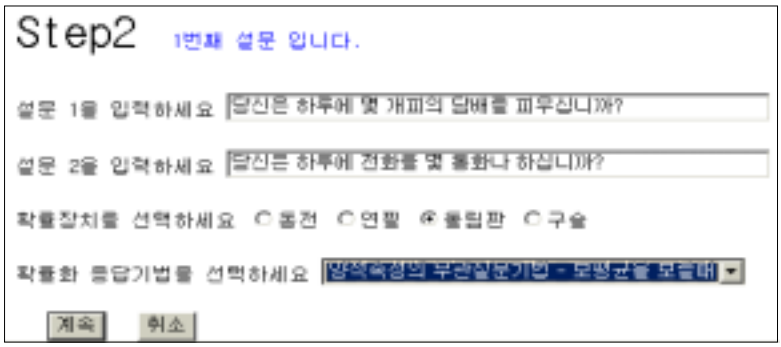
테이블 : Survey_Item (설문지별 구성 문항정보)			
Logical Name	Physical Name	Data Type	비 고
회원 아이디	Id	Varchar	PK, Index
독립 문항 Index	Spot_number	Integer	PK, Index
확률	Probability	Float	
확률장치 Type	Probability_type	Integer	
설문일자	Day	Date	
RRT type	Rrt_type	Integer	
응답	Response	Integer	

3.2 시스템 예

본 시스템은 관리자 모드와 응답자 모드의 두 부분으로 이루어져 있다. 관리자 모드로 접속하여 <그림 3.1>과 같이 설문의 주제, 문항의 수를 입력한 후 계속을 누르게 되면 다음 단계로 <그림 3.2>와 같이 설문 문항을 입력할 수 있는 에디터가 나타나며, 설문 내용과 확률장치, 확률화응답기법의 종류 등 입력사항을 입력하면 설문이 만들어진다.

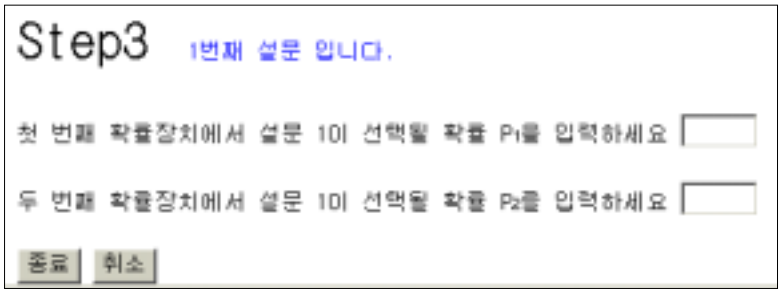


<그림 3.1> 관리자 모드 Step 1



<그림 3.2> 관리자 모드 Step 2

본 시스템은 <그림 3.2>에서 알 수 있듯이 4가지의 확률장치 중 하나를 선택하도록 되어 있으며, <그림 3.2>는 돌림판이 선택된 경우를 나타내고 있다. 그리고, 확률화응답기법은 무관한 변수 Y의 모평균을 알고 있는 경우와 모르고 있을 경우를 선택할 수 있는 데 모르고 있는 경우를 선택하여 계속을 누르게 되면 <그림 3.3>과 같은 화면이 나타나게 된다.



<그림 3.3> 관리자 모드 Step 3

무관한 변수 Y의 모평균을 모르고 있을 경우에는 두 개의 확률장치가 필요하게 되며, 따라서 <그림 3.3>에서와 같이 첫 번째 확률장치에서 민감한 설문 1에 선택될 확률 p_1과 두 번째 확률장치에서 민감한 설문 1에 선택될 확률 p_2를 입력하도록 되어 있다.

만약 무관한 변수 Y의 모평균을 알고 있을 경우에는 그 모평균을 직접 입력하면 되며, 이 때 확률장치에서 민감한 설문 1이 선택될 확률 p을 입력하도록 되어 있다.

<그림 3.4>는 응답자가 설문 사이트에 접속하였을 때 나타나는 화면이다.

<그림 3.4> 응답자 화면

응답자가 20이라고 응답을 하였으나, 조사자는 응답자들이 설문 1에 응답을 하였는지, 설문 2에 응답을 하였는지 알 수 없도록 설계되어 응답자가 진실되게 응답할 수 있도록 하였다.

3.3 응답결과

응답결과는 응답자용과 조사자용으로 구분되어 있으며, 응답자용은 민감한 변수 X의 모평균의 추정값만을 보여주고 있다. 관리자는 <그림 3.5>와 같이 확률화응답기법을 이용한 응답자 수, 민감한 변수 X의 모평균의 추정값, 모분산의 추정값 등을 볼 수 있도록 구현하였다. 또한 다른 설문과 같이 복합적으로 이용할 때는 문항에 따른 즉, 남·녀별 모평균 등을 비교할 수 있도록 구현되었다.

양적 속성의 무관 질문에 대한 결과표

n_1	15
n_2	15
모평균의 추정량	23.027
모분산의 추정량	6.539
P_1	0.7
P_2	0.3

<그림 3.5> 관리자용 결과표

4. 결론 및 향후과제

본 연구에서는 민감한 양적 정보를 얻기 위한 양적 확률화응답시스템을 구현하여 실제조사에서 사용할 수 있도록 하였다. 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. 그리고, 이 시스템은 무관한 변수 Y에 대한 모평균을 알고 있을 때 뿐만 아니라 모르고 있을 때에도 민감한 변수 X에 대한 모평균을 추정할 수 있으므로 더욱 실용성이 높다고 할 수 있겠다.

향후의 과제로는 양적인 응답뿐만 아니라 이지 응답을 확장한 다지 응답에 대한 민감한 질문에 대한 확률화응답시스템의 개발이 필요하다고 생각된다. 또한 다양한 질문에 대한 대처 방법과 응답자에게 흥미를 유발할 수 있는 디자인을 도입한 확률장치의 고안이 필요하다.

참고문헌

1. 김정기, 김희재, 남기성, 박희창, 이성철, 정정현 (1999). 사회조사분석론, 창원대학교출판부.
2. 류제복, 홍기학, 이기성 (1993). 「확률화응답모형」, 자유아카데미, 서울
3. Chaudhuri, A. and Mukerjee, R. (1988). *Randomized Response : Theory and Techniques*, Marcel Dekker, Inc., New York.
4. Coomber, R. (1997). Using the Internet for Survey Research, *Sociological Research Online*, 2, No. 2, <http://www.socresonline.org.uk/socresonline/2/2 / 2.html>
5. Greenberg, B. G., 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.
6. Greenberg, B. G., Kubler, R. R., Abernathy, J. R., and Horvitz, D. G. (1971). Applications of the RR Technique in Obtaining Quantitative Data, *Journal of the American Statistical Association*, 66, 243-250.
7. Schwarz, C. J. (1997). StatVillage : An On-line, WWW-Accessible, Hypothetical City Based on Real Data for Use in an Introductory Class in Survey Sampling, *Journal of Statistics Education*, 5, No. 2, <http://www.amstat.org/ publications/jse>
8. Warner, S. L. (1965). Randomized Response ; A Survey Technique for Eliminating Evasive Answer Bias, *Journal of the American Statistical Association*, 60, 63-69.