Journal of Korean
Data & Information Science Society
2002, Vol. 13, No.2 pp 341 354

Implementation of Qualitative Unrelated Question Model for Obtaining Sensitive Information at Internet Survey

Hee Chang Park[1], Ho Min Myung[2]

Abstract

This paper is planned to use randomized response technique which is an indirect response technique on internet as a way of obtaining much more precise information, not revealing secrets of responsors, considering that respondents are generally reluctant to answer in a survey to get sensitive information targeting employees, customers, etc.

Keywords : 확률화응답기법, 질적 무관질문기법, E-R diagram

1. 서 론

인터넷 설문조사는 대표성의 결여 등과 같은 단점이 있음에도 불구하고 오프 라인(off-line)의 조사에서 가지지 못한 많은 장점(비용의 절감, 신속한 응답 및 집계)이 있어서 급속도로 발전하고 있으며, 모집단이 제한적인 회원이나 기업체의 구성원 등을 대상으로 한 조사에서는 온라인 설문조사의 단점(대표성의 결여)을 보완할 수 있으므로 현실적으로 많이 사용하고 있다. 그러나 온라인 설문조사 역시 기존의 설문조사와 마찬가지로 조사자가 응답자들의 프라이버시나 사생활과 관련된 민감한 정보를 얻고자 할 경우에 정확한 정보를 얻기란 쉬운 일이 아니다. 따라서 기업 의사결정에 대한 찬·반 여부, 성적 질문 등과 같은 민감한 정보를 얻고자 할 경우 간접응답기법인 확률화응답기법을 온라인 설문조사에 적용할 수 있다.

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 간접응답을 하도록 하는 확률화응답기법(randomized response technique ; RRT)을 처음으로 제시하

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

였다. Warner는 응답자들에게 민감한 질문과, 민감한 질문과 배반되는 질문으로 구성된 확률장치를 사용하여 민감한 속성에 대한 질적 정보를 얻고자 하였다. 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 양적 속성기법으로 발전시켜 민감한 변수에 대한 양적 정보를 얻고자 하였다. Loynes(1976)는 Warner기법의 민감한 질문과 배반이 되는 질문 대신에 "예"라고 응답하도록 강요하는 강요질문기법(forced answer technique)을 제안하였다. 이후 미국, 캐나다, 영국, 호주 등 서구 여러 나라와 일본, 인도 등 몇몇 아시아 국가에서도 이 분야에 대한 연구가 활발히 진행되고 있다. 특히 Fox와 Tracy(1986), Chaudhuri와 Mukerjee(1988)는 확률화응답기법들을 정리, 요약하여 체계화하였다. 최근에는 이들 이론들의 실제적 활용에 많은 관심이 집중되고 있으며, 사회학, 경영학, 의학 등 여러 학문분야에서의 조사활동에도 이의 활용이 적극 모색되고 있다.

국내에서는 류제복 등(1993)이 확률화응답기법에 관한 책을 출간하여 그 중요성을 강조하였고, 류제복 등(1995)은 확률화응답기법이 적용된 사례들을 비교 분석하여 실용화를 위한 방안을 제시하였다. 박희창 등(2001)은 질적 자료의 관련질문기법에 대한 온라인 설문조사시스템을 구현한 바 있다.

본 연구에서는 확률화응답기법 중에서 질적 무관질문기법을 인터넷상에서 사용할 수 있는 시스템을 구현하고자 한다. 본 연구에서 개발된 시스템은 데이터베이스에 바탕을 두어 기존의 온라인 설문조사 시스템과 더불어 사용할 수 있을 뿐만 아니라 독립된 스팟 서베이(spot survey)가 가능하도록 개발하고자 한다. 그리고 본 시스템은 동일한 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 구현하고자 한다. 또한 방화벽이 구축된 사무실 및 공공 기관에서 IP가 동일시 될 경우 기계주소(Mac Address)로 구분을 하며, 학교 실습실 등과 같이 여러 명이 사용하는 경우에 대비하여 응답에 제한을 두지 않을 수 있도록 구현하고자 한다.

본 논문은 2절에서 확률화응답기법 중에서 질적 무관질문기법에 대해 전반적으로 살펴본 후, 3장에서 구현된 시스템의 개발환경과 구성에 대하여 기술하며, 4장에서는 예제를 보여주고, 5장에서는 결론과 향후 연구과제를 다루고자 한다.

2. 질적 무관질문기법

본 절에서는 확률화응답기법 중 질적 무관질문기법(qualitative unrelated question technique)에 관하여 기술하고자 한다. 관련질문기법에서는 민감한 질문과 서로 배반이 되는 2가지 질문을 작성하였다. 하지만 응답자들은 두 질문이 모두 민감한 속성을 내포하고 있으므로 반감을 가질 수 있다. 그래서 민감한 속성을 가진 질문과 민감한 질문과는 전혀 관련을 가지지 않는(민감한 그룹과 전혀 관련이 없는 그룹 Y)질문을 사용함으로써 응답자에 대한 신뢰를 더 얻을 수 있다. Greenberg 등(1969)은 이러한 질적 무관질문모형에 대한 전반적인 이론적 체계를 수립하였고, Moors(1971)와

Lanke(1975) 등도 질적 무관질문모형에 대한 추정문제를 다루었다.

질적 무관질문모형은 "민감한 그룹과 전혀 관련이 없는 그룹 Y"의 모집단비율 (π_y)를 알 경우와 (π_y)를 모를 경우 2개로 나누어진다. 먼저 "민감한 그룹과 전혀 관련이 없는 그룹 Y"의 모집단비율 (π_y)를 알고 있는 경우 Greenberg 등(1969)이 사용한 질적 무관질문기법의 확률장치는 다음과 같은 2개의 설문으로 구성되어 있다.

설문 1 : 당신은 민감한 속성 A를 가지고 있습니까?
설문 2 : 당신은 무관한 속성 Y를 가지고 있습니까?

이 때 단순임의복원으로 추출된 n명의 응답자들은 확률장치에 의해서 선택된 질문에 "예" 또는 "아니오"로만 응답한다. 설문 1이 선택될 확률을 p, 설문 2가 선택될 확률을 $1-p$ 라 하면 응답자들이 "예"라고 응답할 확률은 다음과 같다.

$$\lambda = p\pi + (1-p)\pi_y \tag{2.1}$$

여기서 π는 민감한 속성 A에 대한 모비율이며, π_y는 무관한 속성 Y에 대한 모비율이다.

n명의 응답자들 중에서 "예"라고 응답한 사람의 수를 n_1이라 하면 $\hat{\lambda} = \frac{n_1}{n}$ 이므로 민감한 속성 A에 대한 모비율 π의 추정량과 그 분산은 다음과 같다.

$$\hat{\pi} = \frac{\hat{\lambda} - (1-p)\pi_y}{p} \tag{2.2}$$

$$Var(\hat{\pi}) = \frac{\lambda(1-\lambda)}{np^2} \tag{2.3}$$

한편, $\hat{\pi}$의 분산추정량은 다음과 같다.

$$\widehat{Var}(\hat{\pi}) = \frac{\hat{\lambda}(1-\hat{\lambda})}{(n-1)p^2}, \ (n \neq 1, p \neq 0) \tag{2.4}$$

그러나 민감한 그룹과 전혀 관련이 없는 그룹 Y의 모집단비율 (π_y)을 알고 있다는 가정은 일반적으로 타당하지 않다. "민감한 그룹과 전혀 관련이 없는 그룹 Y"의 모집단비율 (π_y)를 모를 경우 π_y를 추정하여야 한다.

구하고자 하는 미지 모수가 π 하나일 때는 단지 하나의 표본이면 충분하나 두 개의 모수 π와 π_y를 모두 모를 때는 최소한 두 개의 표본이 필요하게 된다. 따라서 모

집단으로부터 단순임의복원으로 크기가 n_1과 n_2인 두 개의 독립표본을 추출하여 모비율 π의 추정량을 구한다. 두 개의 표본을 사용해야 되므로 i (i = 1,2) 번째 표본에서 민감한 질문이 선택될 확률이 p_i가 되는 두 개의 확률장치가 필요하게 된다. i 번째 표본에서 응답자들이 "예"라고 응답할 확률은 다음과 같다.

$$\lambda_i = p_i(\pi - \pi_y) + \pi_y \tag{2.5}$$

여기서 π는 민감한 속성 A에 대한 모비율이며 π_y는 무관한 속성 Y에 대한 모비율이다.

n_{i1}을 i번째 표본에서 "예"라고 응답한 사람들의 수라고 하면 $\hat{\lambda}_i = n_{i1}/n_i$이므로 민감한 속성 A에 대한 모비율 π의 추정량과 그 분산은 다음과 같다.

$$\hat{\pi}_{u2} = \frac{\hat{\lambda}_1(1-p_2) - \hat{\lambda}_2(1-p_1)}{p_1 - p_2} \quad , p_1 \neq p_2 \tag{2.6}$$

$$Var(\hat{\pi}_{u2}) = \frac{\dfrac{(1-p_2)^2\lambda_1(1-\lambda_1)}{n_1} + \dfrac{(1-p_1)^2\lambda_2(1-\lambda_2)}{n_2}}{(p_1 - p_2)^2} \quad , \quad p_1 \neq p_2 \tag{2.7}$$

한편, $\hat{\pi}_{u2}$의 분산추정량은 다음과 같다.

$$\widehat{Var}(\hat{\pi}_{u2}) = \frac{\dfrac{(1-p_2)^2\hat{\lambda}_1(1-\hat{\lambda}_1)}{n_1 - 1} + \dfrac{(1-p_1)^2\hat{\lambda}_2(1-\hat{\lambda}_2)}{n_2 - 1}}{(p_1 - p_2)^2} \quad , p_1 \neq p_2 \tag{2.8}$$

3. 시스템의 구현

구현된 시스템의 개발 환경에서 개발 언어는 수식의 계산을 위하여 GNU C compiler를 사용하였으며, 화면 구성을 위하여 JAVA, HTML을 사용하였고, 데이터를 읽고, 쓰고, 검색하기 위하여 SQL구문을 사용하였다. 운영 체제는 Linux를 사용하였으며, 데이타베이스는 MySQL-Ver 3.23.39를 사용하였다. 시스템은 관리자(조사자) 모드와 응답자 모드의 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 때는 접속 순서에 따라 표본 1과 표본 2로 나누어 응답에 참여하도록 하였다. 응답의 결과는 데이타베이스에 저장되며, 민감한 속성(변수)에 대한 모비율의 추정값과 분산추정값을 계산한 후에 응답자에게는 민감한 속성(변수)에 대한 모비율의 추정값만을 보여주고, 관리자에게는 모비율의 추정값과 분산추정값에 대한 결과 모두를 보여 주도록 구성하였다.

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

본 시스템의 구현을 위해 설계한 테이블은 다음의 <표 3.1>부터 <표 3.4>와 같다. <표 3.1>의 메인 테이블은 고유한 테이블이다. 여기에 모든 설문 문항들에 대한 정보

<표 3.1> 메인 테이블

메 인			
Logical Name	Physical Name	Data Type	비 고
인덱스 키	idx	integer	pk, auto_increment
설문 작성일	day	date	
설문 작성시간	time	time	
설문의 주제	subject	varchar(79)	
설문 문항수	number	tinyint	
설문 작성 종료 검사	check	char(2)	
설문의 소개 및 인사말	title	blob	

를 보관하고 있으며 아래에 나오는 설문지라는 테이블과 연계하여 분석을 수행한다. 여기에는 설문 작성일시, 주제, 문항수, 소개 등이 기록되며 고유한 인텍스 키로 설문지 테이블을 참조한다.

<표 3.2>에서 제시하는 설문 테이블은 실제 설문 문항의 정보를 보유하는 테이블이다. 설문 문항의 내용, 확률장치, 확률화응답기법 종류, 설문 1이 선택될 확률 등 상위 테이블인 메인에서 인덱스 키를 참조하여 설문 생성시 추가로 생성된다.

<표 3.3>과 <표 3.4>의 테이블은 응답자 집계용 테이블이다. 설문지 테이블에서 넘어온 문항을 응답한 수치가 기록되는 테이블이다. 표본 2테이블은 질적 이표본 무관 질문기법 - 무관한 속성의 모비율을 모르는 경우에는 최소한 두 개의 표본이 필요하게 된다. 따라서 모집단으로부터 단순임의복원추출법으로 크기가 n_1, n_2인 두 개의 독립 표본을 추출할 경우 사용된다.

<표 3.2> 설문 테이블

설 문			
Logical Name	Physical Name	Data Type	비 고
인덱스 키	idx	integer	pk, auto_increment
설문 1	q1	varchar(79)	
설문 2	q2	varchar(79)	
확률화응답기법종류	mode	varchar(2)	
확률장치종류	type	varchar(2)	
표본 1에서 설문1이 선택될 확률	p1	float	
표본 2에서 설문1이 선택될 확률	p2	float	
알고있는 모비율	p3	float	
표본 1참여자	n1	smallint	
표본 2참여자	n2	smallint	

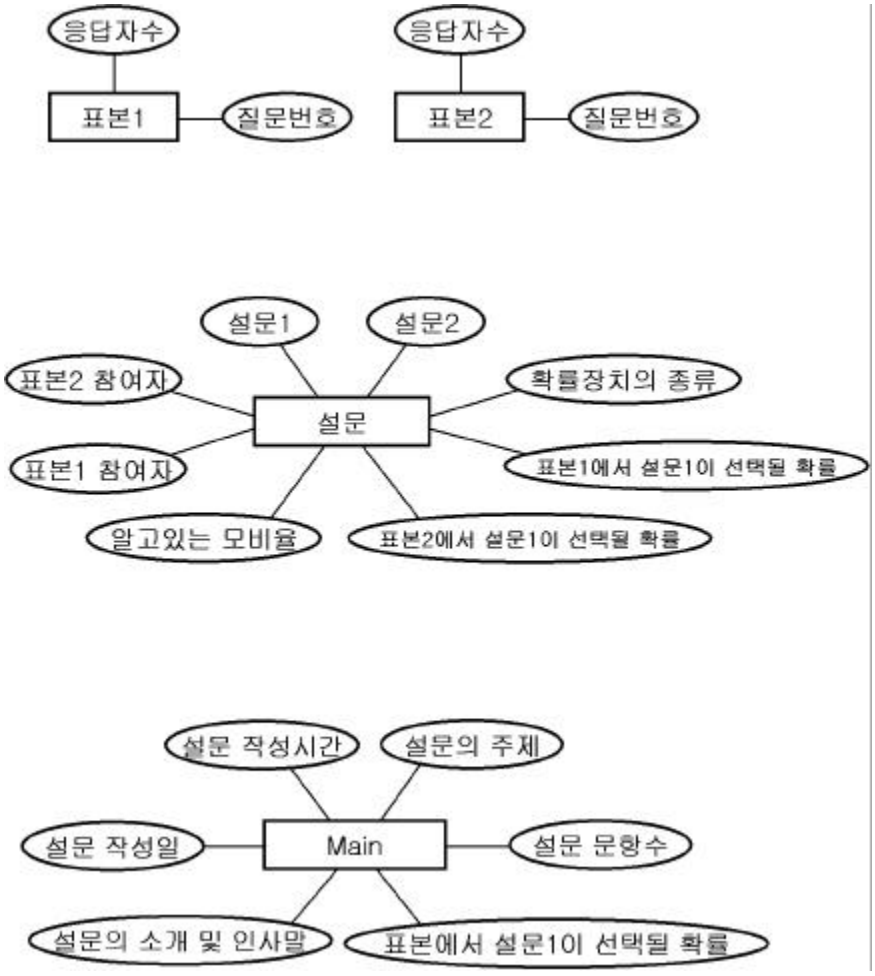
<표 3.3> 표본 1 테이블

표본 1			
Logical Name	Physical Name	Data Type	비 고
인덱스 키	idx	integer	pk, auto_increment
“예”라고 응답한 응답자 수	s1	mediumint	

<표 3.4> 표본 2 테이블

표본 1			
Logical Name	Physical Name	Data Type	비 고
인덱스 키	idx	integer	pk, auto_increment
"예"라고 응답한 응답자 수	s1	mediumint	

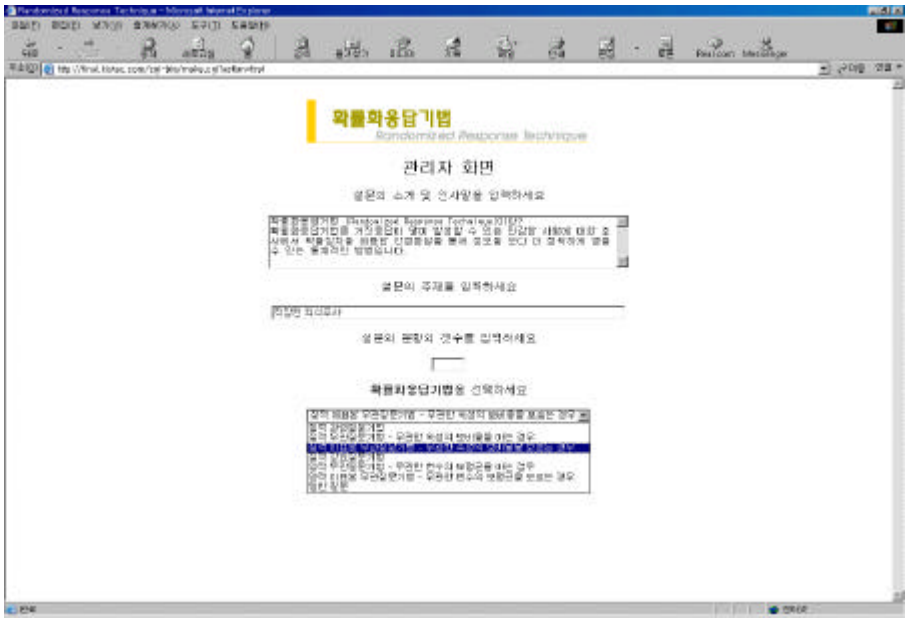
본 시스템에 사용된 데이터베이스 구성은 <그림 3.1>과 같은 구조를 가지고 있다. 메인 테이블은 설문의 주제, 문항의 수, 작성일자 등을 저장하고 있으며 설문 테이블과 연계하여 구성된다. 설문 테이블은 메인 테이블의 인덱스 키를 이용하여 설문의 주제에 해당되는 설문문항, 확률장치종류, 확률화응답기법 종류 등을 기록하며, 표본 1, 표본 2 테이블은 설문 테이블의 인덱스 키 값을 이용하여 순차적으로 응답 결과가 기록된다.



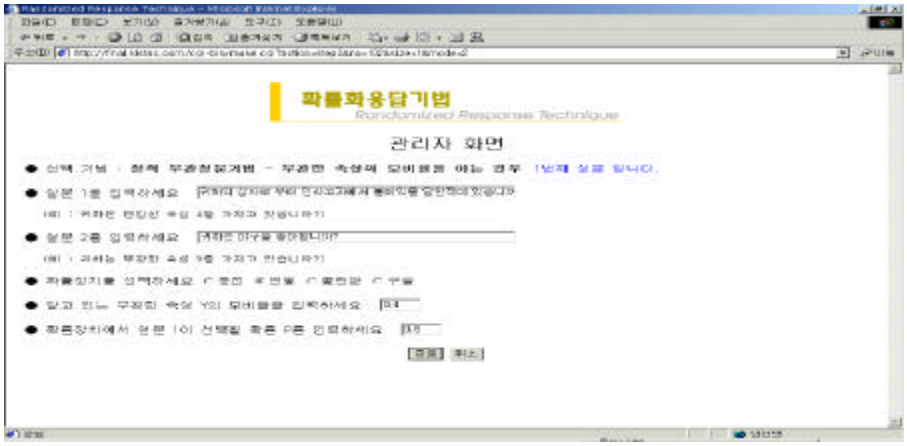
<그림 3.1> E-R Diagram

본 시스템은 관리자 모드와 응답자 모드의 두 부분으로 구성되어 있다. 관리자 모드는 설문의 생성 및 삭제를 담당하는 부분이며, 응답자 모드는 실제응답자들이 설문에 응답을 하는 부분이다.

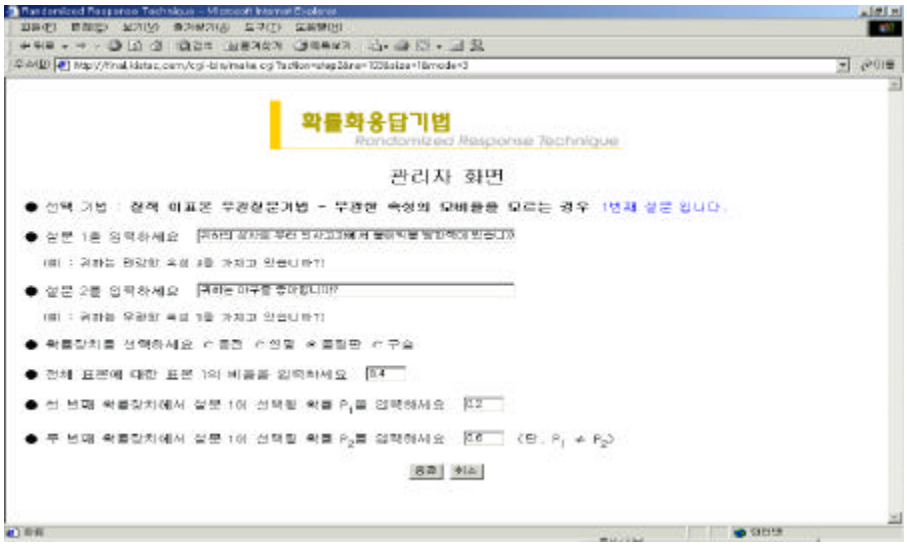
설문을 생성하기 위해서 <그림 3.2>와 같이 관리자 초기화면에 접속하여 설문의 소개, 설문의 주제, 확률화응답기법의 종류, 문항의 수 등을 입력한다. 다음 단계에서는 <그림 3.3> 및 <그림 3.4>와 같이 설문 문항을 입력할 수 있는 에디터가 나타나며, 설문 내용과 확률장치 등의 내부 사항을 모두 입력하면 설문이 만들어진다.



<그림 3.2> 관리자 초기화면

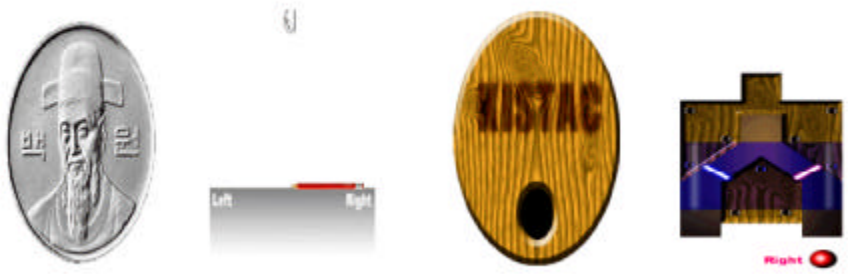


<그림 3.3> 질적 무관질문기법 - 무관한 속성의 모비율을 아는 경우



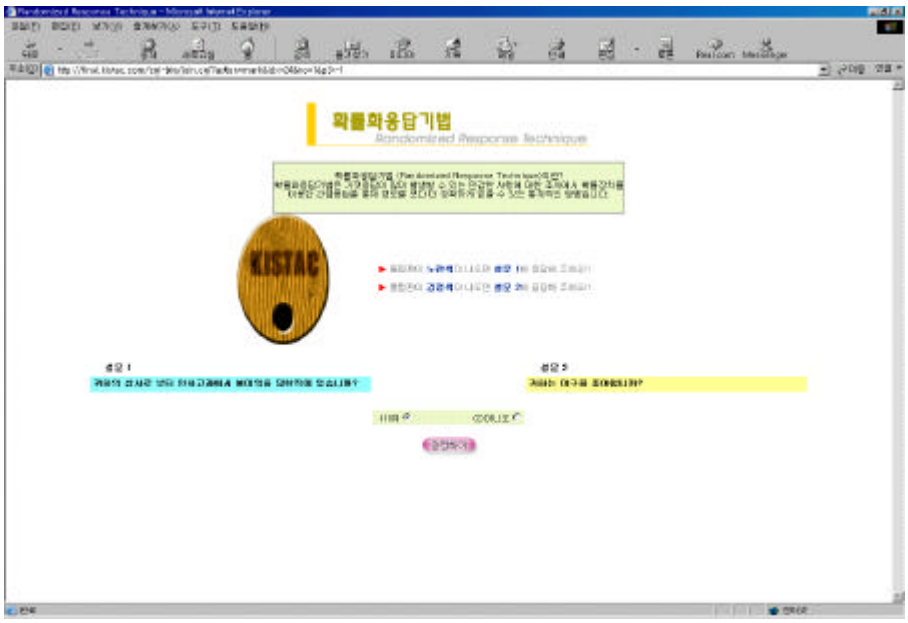
<그림 3.4> 질적 이표본 무관질문기법 - 무관한 속성의 모비율을 모르는 경우

본 시스템에서는 4종류의 확률장치 중 하나를 선택하도록 되어 있으며, <그림 3.5>는 이들 4가지를 나타낸 것이다. 이들 확률장치는 관리자 화면에서 선택할 수 있으며, 확률장치에서 설문 1이 선택될 확률을 입력하면 그 확률에 따라 동전의 앞면 혹은 뒷면, 연필의 오른쪽 혹은 왼쪽과 같이 두 사건 중 하나의 형태로 나타난다.



<그림 3.5> 4종류의 확률장치

응답자 화면은 <그림 3.6>과 같으며, 여기서 응답자들은 설문 문항에 대하여 응답할 수 있다.



<그림 3.6> 응답자 화면

관리자는 응답자가 설문 1에 대하여 "예" 라고 응답을 하였는지 설문 2에 대하여 "예"라고 응답을 하였는지 알 수 없도록 설계되어 응답자가 솔직하게 응답할 수 있도록 하였다.

4. 예 제

이 장에서는 질적 무관응답기법 시스템을 적용하여 실시한 설문 조사로부터 얻어진 결과에 대해 논의하고자 한다.

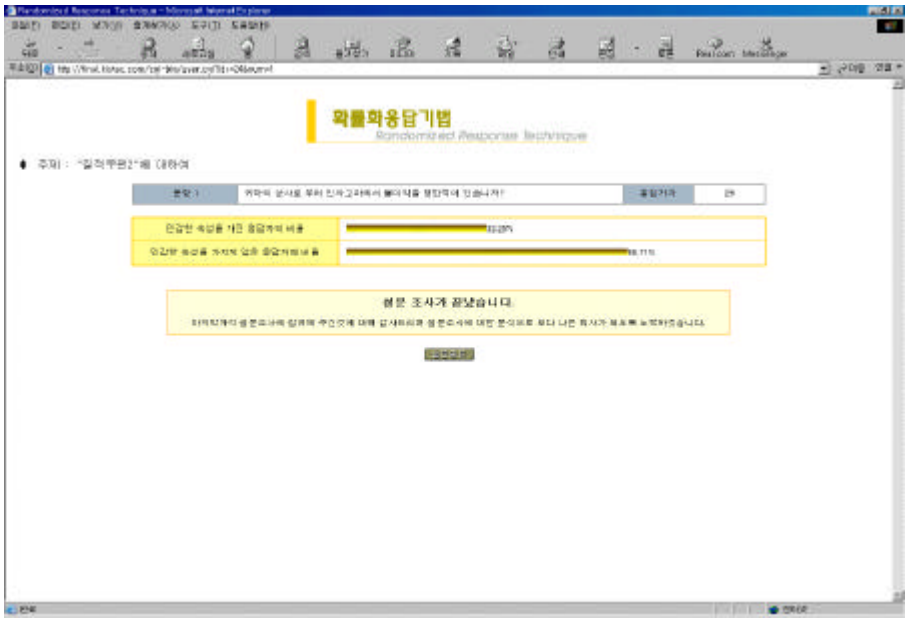
기간 : 2002. 1.4 ~ 1.31

대상 : (주)Esab 일반 사무직 사원(150명)

문항 : 귀하의 상사로부터 인사고과에서 불이익을 당한 적이 있습니까?

응답결과는 응답자용과 관리자용으로 구분되어 있으며, 응답자용은 민감한 속성 A의 모비율의 추정값만을 <그림 4.1>과 같이 보여준다.

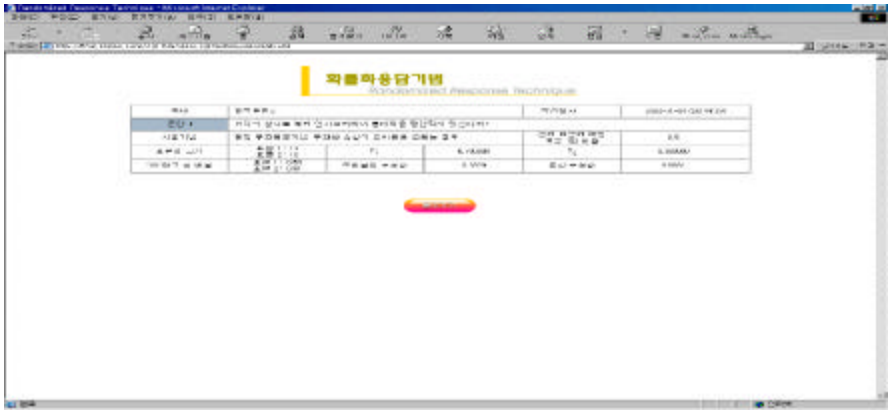
<그림 4.1>은 "귀하의 상사로부터 인사고과에서 불이익을 당한 적이 있습니까"라는 민감한 질문에 대하여 29명이 참여하였고 모비율 π의 추정량 $\hat{\pi}$은 0.3329임을 보여주고 있다.



<그림 4.1> 응답자용 결과 화면

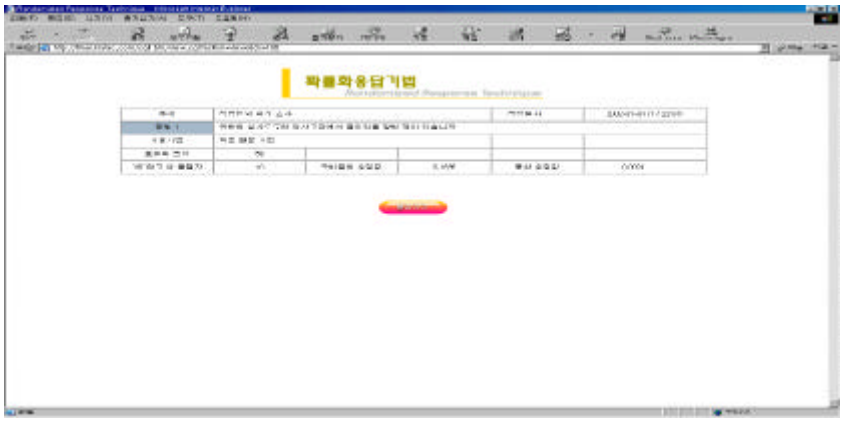
관리자용 결과는 <그림 4.2>와 같다. 여기서는 응답자 화면에는 볼 수 없는 설문 1이 선택될 확률, 표본 1의 크기, 표본 2의 크기, 표본 1이 선택될 확률 p_2, 표본 1에 "예"라고 응답한 응답자수, 표본 2에 "예"라고 응답한 응답자수, 모비율 π의 추정량

$\hat{\pi}$와 $\hat{\pi}$의 분산 추정치 등을 볼 수 있도록 구현하였다.



<그림 4.2> 관리자용 결과 화면

"귀하의 상사로부터 인사고과에서 불이익을 당한 적이 있습니까?"에 대한 직접질문에 대한 결과는 <그림 4.3>과 같다.



<그림 4.3> 직접 질문 기법

확률화응답기법을 이용한 온라인 설문조사에 참여한 사람의 수는 29명이었고, "예"라고 응답한 사람의 수는 20명이었으며, 이로부터 인과고과에서 불이익을 당한 적이 있는 사람들의 모비율을 추정해 본 결과 33.29%로 나타났으며, 분산추정치는 0.0691이었다. 그리고 직접질문을 이용한 온라인 설문 조사에 참여한 사람의 수는 59명이었고, "예"라고 응답한 사람의 수는 10명이었으며, 이로부터 인과고과에서 불이익을 당한 적이 있는 사람들의 모비율을 추정해 본 결과 16.95%로 나타났으며, 분산추정치는 0.0024이었다. 이와 같은 결과에서 알 수 있듯이 질적 무관기법을 이용한 온라인 설문

조사에서 구한 모비율의 추정치가 직접질문을 이용한 온라인 설문조사에서 구한 모비율의 추정치보다 높게 나타났다.

5. 결론 및 향후과제

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 질적 무관기법을 인터넷 상에서 사용할 수 있도록 구현하였다. 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.

향후의 과제로는 다양한 표본추출방법을 적용하여 회원이나 구성원들을 대상으로 목적에 맞는 응답자를 추출하는 기능을 부과하여 기업 등에서 사원이나 고객을 대상으로 한 조사에서 실용성을 보다 높이는 방법이 필요하다고 생각된다. 또한 다양한 질문에 대한 대처 방법과 응답자에게 흥미를 유발할 수 있는 멀티미디어적인 요소를 도입한 확률장치의 고안이 필요하다.

참고문헌

1. 김정기, 김희재, 남기성, 박희창, 이성철, 정정현 (1999). 사회조사분석론, 창원대학교출판부.
2. 류제복, 홍기학, 이기성 (1993). 확률화응답모형, 자유아카데미.
3. 류제복, 이계오, 이기성 (1994). 확률화응답기법의 실용화 방안, 응용통계연구 제8권, 제1호, 9-26.
4. 박희창, 이기성, 김희재, 남기성 (2001). 인터넷조사와 설문조사시스템, 자유아카데미.
5. 박희창, 남기성, 이기성 (2001). 인터넷조사에서의 확률화응답기술의 구현, 한국통계학회논문집, 제8권, 제3호, 731-737.
6. 이기성 (1992). 2단계 확률화응답모형에 관한 연구, 박사학위논문, 동국대학교 통계학과.
7. Chaudhuri, A. and Mukerjee, R. (1988). *Randomized Response : Theory and Techniques*(1st ed.). New York : Marcel Dekker, Inc..
8. Coomber, R. (1997). "Using the Internet for Survey Research." *Sociological Research Online*, 2(2).
   <http://www.socresonline.org.uk/socresonline/2/2/2.html>
9. Fox, J. A. and Tracy, P. E. (1986). *Randomized Response : A Method for Sensitive Survey*, Sage Publications.
10. Greenberg, et. al. (1969). The Unrelated Question Randomized Response

Model : Theoretical Framework, Journal of the American Statistical Association, 64, 520-539.
11. 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(2). <http://www.amstat.org/publications/jse>
12. Warner, S. L. (1965). "Randomized Response ; A Survey Technique for Eliminating Evasive Answer Bias." *Journal of the American Statistical Association* 60 : 63-69.

[2002년 8월 접수, 2002년 10월 채택]