The Korean Communications
in Statistics Vol. 8, No. 3, 2001
pp. 731-737

Implementation of Randomized Response Technique at Internet Survey

Hee Chang Park[1], Ki Seong Nam[2], Gi Sung Lee[3]

Abstract

In this paper, we suggest and implement an internet survey system cooperating the randomized response technique(RRT). We compare it with the direct survey. RRT is an indirect survey method to get true information from the respondent who is conceived to have sensitive character without revealing his/her status. We implement our system(method) based a data-base system, so it common all kind of data obtained through internet surveys. This system also can be used in spot survey independently.

Keywords : randomized response technique, internet survey system, spot survey

1. 서론

일반적으로 인터넷조사는 온라인조사의 하나로서 인터넷 망(www)을 통해서 행해지는 시장조사, 여론조사 및 사회조사를 의미한다. 전자조사(electronic research), 온라인조사(online survey), 인터넷조사(internet survey), 웹조사(web survey), 넷조사(net survey) 등 다양한 표현으로 명명되고 있는 온라인 조사법은 기존의 사회조사나 시장조사 등을 인터넷 망을 통해서 행하는 방법이다. 1980년 이후부터 꾸준히 조사업계 등에서는 컴퓨터를 이용한 전화조사(CATI)나 컴퓨터를 이용한 면접조사(CAPI), 팩스에 의한 자동입력 등 컴퓨터나 통신기기를 이용한 장치산업화가 추진되어 온 것은 사실이나 인터넷 망의 확산과 PC의 급속한 보급이 인터넷조사를 가능하게 하였다(Coomber, 1997). 1990년대 중반이후 인터넷을 이용한 사회조사는 비용과 속도에서 혁신을 가져와 확산의 과정을 밟고 있다. 인터넷조사의 확산을 가능케 한 기저에는 먼저 급속한 PC 보급의 결과

1) Associate Professor, Department of Statistics, Changwon National University, Changwon, Kyungnam, 641-773, Korea.
E-mail : hcpark@sarim.changwon.ac.kr
2) Lecturer, Department of Statistics, Changwon National University, Changwon, Kyungnam, 641-773, Korea.
E-mail : ksnam@sarim.changwon.ac.kr
3) Associate Professor, Division of Computer and Information Science, Woosuk University, Wanju-gun, Jeonbuk, 565-701, Korea.
E-mail : gisung@woosuk.ac.kr

어디에서도 접할 수 있는 단말기와 네트워크인프라가 광범위하게 구축됨으로써 접속비용의 저렴화, 그리고 응답자가 입력한 자료가 다른 프로세서를 거치지 않고 바로 디지털 자료로 축적되는 기술수준의 발전이 있었다.

최근 들어 인터넷조사가 현장에서 많이 행해지고 있으나, 인터넷조사 역시 기존의 설문조사와 마찬가지로 조사자가 응답자들의 프라이버시나 사생활과 관련된 민감한 정보를 얻고자 할 경우에 정확한 정보를 얻기란 쉬운 일이 아니다. 한편, 응답자들은 인터넷 상에서 민감한 질문을 직접적으로 받게 되면 면접조사나 전화조사와 같이 조사원에 의해 이루어지는 기존의 조사보다는 응답을 회피하는 정도가 덜 하겠지만 그래도 혹시 자신의 비밀이나 사생활이 노출될까 의심하여 정직한 응답을 꺼리게 된다. 또한, 패널로 구성된 응답자들을 대상으로 인터넷조사를 실시할 경우에는 면접조사나 전화조사 못지 않게 민감한 질문에 대한 진실된 응답을 얻는데 한계가 있다. 이러한 문제점을 해결하기 위하여 Warner(1965)가 제안한 확률장치를 이용한 간접질문방식인 확률화응답기법(randomized response technique ; RRT)을 인터넷 설문조사에 적용해 볼 수 있다. 이 때, 응답자들은 인터넷 상에서 구현된 확률장치를 통해 선택된 설문에 대하여 응답을 하게 되므로 응답자 자신의 비밀이나 개인 정보의 유출을 이유로 정확하지 않은 응답을 할 가능성을 줄일 수 있게 된다. 물론, 인터넷 상에서 응답자들에게 확률화응답기법을 이해시켜야 되므로 응답자들의 관심을 불러일으키지 못할 수 있다는 문제점이 있다. 하지만, 확률화응답기법을 보다 쉽게 이해할 수 있도록 충분히 설명하고, 인터넷 상에서 다양한 형태의 확률장치를 구현하여 응답자들의 흥미를 유발시킨다면 이러한 문제점은 해결될 수 있으리라 생각된다. 그러므로 인터넷 설문조사를 하는 데 이러한 확률화응답기법을 이용할 수 있는 시스템이 필요하다고 생각된다.

따라서, 본 연구에서는 확률화응답기법을 이용한 인터넷조사 시스템을 구현해 보고자 한다. 본 시스템은 기존의 인터넷조사 시스템과 더불어 사용할 있을 뿐만 아니라 독립된 스팟 서베이(spot survey) 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 실습실 등과 같이 여러 명이 사용하는 경우에 대비하여 응답에 제한을 두지 않을 수 있도록 구현하고자 한다. 또한, 본 시스템은 개인의 신상보호를 위해 응답자가 확률장치를 이용한 응답에서 민감한 속성이 있는지에 대한 설문에 응답을 하였는지, 민감한 속성이 없는지에 대한 설문에 응답을 하였는지를 조사자가 모르도록 구현하였다.

본 논문은 1장에서 서론으로 민감한 사항에 대한 인터넷조사에서 확률화응답기법의 필요성을 설명하고, 2장에서는 확률화응답기법을 소개하며, 3장에서는 구현된 인터넷조사 시스템에 대하여 자세히 살펴보고, 4장에서는 결론과 향후 연구과제를 다룬다.

2. 확률화응답기법

사회 여러 분야의 조사에는 응답자들이 응답을 회피하거나 정직하게 응답하지 않는 질문들이 종종 포함된다. 예를 들면, 불로소득, 탈세여부, 전과경험, 컨닝경험, 음주운전경험, 알코올중독, 환각제사용, 성경험, 낙태경험, 동성연애 등과 같은 사항들은 응답자들이 응답을 꺼려하는 질문들이다. 이러한 민감한 질문을 공개적으로 하게 되면 무응답이나 거짓응답 또는 응답을 회피함으로써 응답자들로부터 정확한 정보를 얻을 수 없게 된다.

따라서, 민감한 질문에 대해 보다 신뢰할 수 있는 정보를 얻기 위해서는 직접질문보다는 간접적인 대체 질문방식이 필요하게 된다. 이에 Warner(1965)는 응답자들에게 민감한 질문과, 민감한 질문과 배반되는 즉, 부의 관계를 갖는 질문 중에서 확률장치를 통해 선택된 질문에 응답하게 함으로써 응답자의 신분이나 비밀을 노출시키지 않고서 민감한 질문에 대한 정보를 이끌어 낼 수 있는 확률화응답기법을 처음으로 제시하였다.

응답자들은 확률장치에 의해 선택된 질문에 응답하면 되며, 이 때 조사자는 응답자가 어떠한 질문에 응답을 했는지를 알 수 없게 되므로 응답자는 솔직하게 응답할 수 있다. Warner가 사용한 관련질문기법의 확률장치는 다음과 같은 2개의 설문으로 구성되어 있다.

설문 1 : 당신은 민감한 속성 A를 가지고 있습니까?
설문 2 : 당신은 민감한 속성 A를 가지고 있지 않습니까?

단순임의복원으로 추출된 n명의 응답자들은 확률장치에 의해서 선택된 설문에 "예" 또는 "아니오"로만 응답한다. 설문 1이 선택될 확률을 p, 설문 2가 선택될 확률을 $1-p$라 하자. 이 때, 설문 1이 선택될 확률 p는 조사자가 확률장치에서 사전에 조정할 수 있다.

응답자들이 진실되게 응답했다는 가정 하에 응답자들이 "예"라고 응답할 확률은 다음과 같다.

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

여기서, π는 민감한 속성 A에 대한 모비율이다.

n명의 응답자들 중에서 "예"라고 응답한 사람들의 수를 n'라고 하면, $\hat{\lambda} = \frac{n'}{n}$ 이다. 따라서, 민감한 속성 A에 대한 모비율 π의 추정량과 그 분산은 다음과 같다.

$$\hat{\pi}_w = \frac{\hat{\lambda}-(1-p)}{2p-1}\,,\quad p \neq \frac{1}{2}\,, \tag{2.2}$$

$$Var(\hat{\pi}_w) = \frac{\pi(1-\pi)}{n} + \frac{p(1-p)}{n(2p-1)^2}\,. \tag{2.3}$$

확률화응답기법에 의해서 얻어진 추정량의 분산 식(2.3)은 직접질문으로부터의 분산과 확률장치를 사용함으로써 생기는 분산의 합으로 되어있다.

한편, $\hat{\pi}_w$의 분산추정량은 다음과 같다.

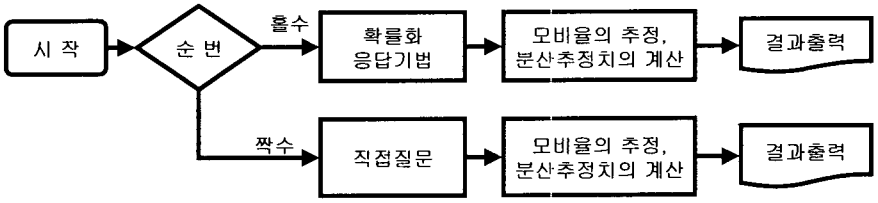
$$\widehat{Var}(\hat{\pi}_w) = \frac{\hat{\lambda}(1-\hat{\lambda})}{(n-1)(2p-1)^2}\,,\quad p \neq \frac{1}{2}\,. \tag{2.4}$$

3. 인터넷조사에서 확률화응답기법의 구현

3.1 시스템 개발 환경 및 시스템 흐름도

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

확률화응답시스템의 전체적인 흐름은 홈페이지에서 설문에 참여하는 응답자의 접속 순번에 따라 홀수 번째는 확률화응답기법을 이용한 설문에 응답하도록 하였고, 짝수 번째는 직접질문에 참여하도록 하였다. 응답의 결과는 DB로 저장되며, 모비율의 추정치와 분산추정치를 계산한 후에 응답자에게는 모비율의 추정치만을 보여주고, 조사자(관리자)에게는 모비율의 추정치와 분산추정치에 대한 결과를 모두 보여 주도록 구성하였다. 이러한 절차를 흐름도로 표현해 보면 <그림 3.1>과 같다.



<그림 3.1> 흐름도

본 시스템은 자료의 입력에서 처리, 결과를 모두 DB를 바탕으로 이루어져 있다. 이로 인하여 동일 응답자 등의 반복 측정에서도 DB를 사용함으로 인하여 기존의 설문응답시스템과 쉽게 합쳐서 사용할 수 있다. 또한 처리과정에서도 DB를 사용함으로 인하여 쿼리(query)를 사용하여 수행속도에서 파일 시스템 보다 빠르게 진행할 수 있다. 또한 기본적인 결과를 DB를 연동함으로 계속적인 조사 등에서 추세 분석이 가능하다.

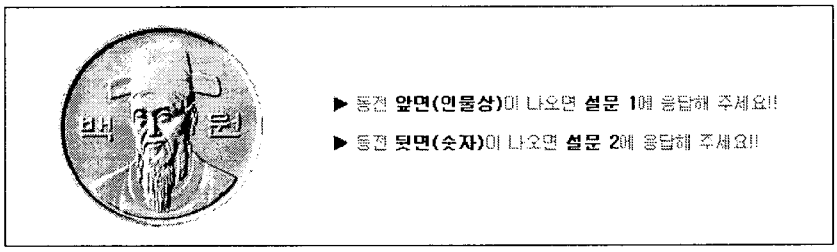
본 시스템의 구현을 위해 설계한 테이블의 항목은 <표 3.1>과 같다.

<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	

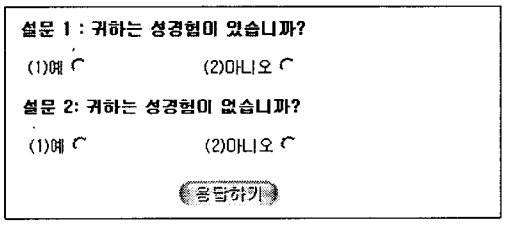
3.2 시스템 예

본 시스템은 확률장치로서 <그림 3.2>와 같은 동전을 이용하였다. 동전의 앞면이 나올 확률은 조사자가 조정할 수 있도록 하여 각각의 확률에 대하여 효율성을 비교할 수 있도록 하였다.

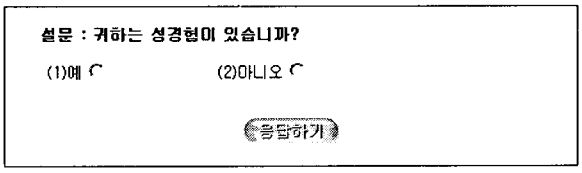


<그림 3.2> 동전을 이용한 확률장치

본 시스템에서 예제로 사용된 질문은 학생들에게 민감한 질문인 성경험의 유무에 대하여 질문을 하였다. 질문은 <그림 3.3>과 <그림 3.4>와 같다.



<그림 3.3> 확률화응답기법의 질문



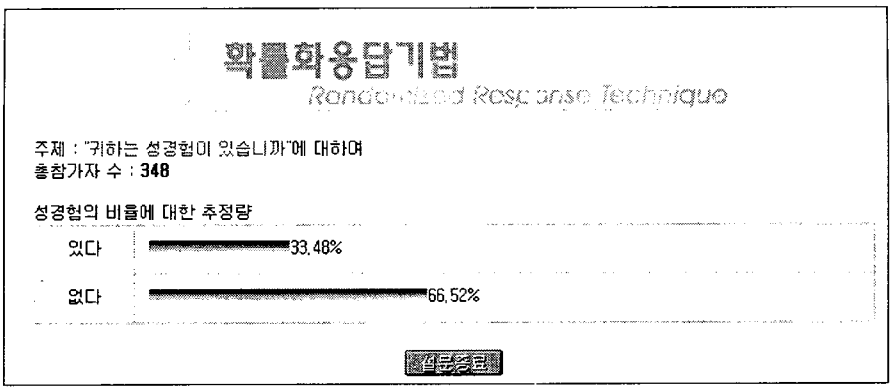
<그림 3.4> 직접질문

본 시스템은 창원대학교 홈페이지에 일정기간 동안 올려놓았으며, 따라서 학교 실습실 등과 같이 여러 명이 사용하는 경우에 대비하여 동일 IP에 대하여 중복 응답을 하는 것을 허용하였다.

3.3 응답결과

응답결과는 응답자용과 조사자용으로 구분되어 있으며, 응답자용은 모비율의 추정치만을 <그림 3.5>와 같이 보여주고 있다.

조사자용 결과표는 <표 3.2>와 같이 확률화응답기법을 이용한 결과와 직접질문을 비교 가능하도록 표의 형식으로 만들었다. 이 표에서는 확률장치에서 민감한 설문 1이 선택될 확률 p, 표본의 크기(총참여자 수), "예"라고 응답한 응답자 수, 모비율 π의 추정량 $\hat{\pi}_w$와 $\hat{\pi}_w$의 분산추정치를 나타내고 있다.



<그림 3.5> 응답자용 결과

<표 3.2> 조사자용 결과

확률화응답기법과 직접질문기법의 결과 비교표		
	확률화응답기법	직접질문기법
민감한 설문의 선택확률	0.30	None
표본의 크기	348	347
"예"라고 응답한 응답자 수	197	100
모비율 π의 추정량 $\hat{\pi}_w$	0.33477	0.28818
$\hat{\pi}_w$의 분산추정치	0.00442	0.00059

확률화응답기법을 이용한 인터넷조사에 참여한 사람의 수는 348명이었고, "예"라고 응답한 사람의 수는 197명이었으며, 이로부터 성경험이 있는 사람들의 모비율을 추정해 본 결과 33.4%로 나타났으며, 분산추정치는 0.00442이었다. 그리고, 직접질문기법을 이용한 인터넷조사에 참여한 사람의 수는 347명이었고, "예"라고 응답한 사람의 수 100명이었으며, 이로부터 성경험이 있는 사람들의 모비율을 추정해 본 결과 28.8%로 나타났으며, 분산추정치는 0.00059이었다. 이와 같은 결과에서 알 수 있듯이 확률화응답기법을 이용한 인터넷 설문조사에서 구한 모비율의 추정치가 직접질문을 이용한 인터넷 설문조사에서 구한 모비율의 추정치보다 4.6%정도 높게 나타났다.

그러나, 두 기법들 사이에 통계적으로 유의한 차이가 나타나지는 않았다. 그 이유로는 대학생들에게 더 이상 성경험이 민감한 사항이 아닐 수도 있지만 그것보다는 동일한 IP에 대하여 중복응답을 허용하였기 때문에 응답자들이 조사자에게 직접응답을 하는 것보다는 덜 부담스러울 수 있으므로 직접질문에서도 진실된 응답을 했을 것이라고 생각된다. 마찬가지로 인터넷조사의 경우 확률화응답기법이나 직접질문 모두 모니터 상에서 이루어지는 간접조사방법으로 생각해 볼 수도 있을 것이다. 한편, 학생이 아닌 일반 교직원들도 인터넷조사에 참여를 함으로써 결혼한 사람들에게 성경험이 있느냐는 질문을 하게 된 데에도 그 원인을 찾을 수 있다.

4. 결론 및 향후과제

본 연구에서는 인터넷조사에서 확률화응답기법을 이용할 수 있는 시스템을 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. 이 시스템은 인터넷 상에서 응답자들에게 확률화응답기법을 이해시켜야 되므로 응답자들의 관심을 불러일으키지 못할 수 있다는 문제점이 있기는 하지만, 확률화응답기법을 보다 쉽게 이해할 수 있도록 충분히 설명하고, 인터넷 상에서 다양한 형태의 확률장치를 구현하여 응답자들의 흥미를 유발시킨다면 이러한 문제점은 해결될 수 있으리라 생각된다.

한편, 본 연구에서 구축한 확률화응답기법을 이용한 인터넷조사시스템은 기존의 각종 인터넷조사에 민감한 사항에 대한 설문을 추가하여 확률장치를 통해 조사함으로써 좀 더 타당한 정보를 얻는데 활용될 수 있을 것으로 기대된다. 또한, 이 시스템은 기업체나 단체에서 개인의 민감한 사항에 대한 정확한 정보를 얻는데 사용 가능하다고 생각된다.

향후의 과제로는 질적인 민감한 질문뿐만 아니라 양적인 민감한 질문에 대한 시스템의 개발로 다양한 질문에 대한 대처 방법과 응답자에게 흥미를 유발할 수 있는 디자인을 도입한 확률장치의 고안이 필요하다.

감사의 글

본 시스템의 구축을 위해 도움을 준 창원대학교 대학원생 명호민과 유지현에게 감사드립니다.

참고문헌

[1] 김정기, 김희재, 남기성, 박희창, 이성철, 정정현(1999), 사회조사분석론, 창원대학교 출판부.

[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*, Vol. 2, No. 2, <http://www.socresonline.org.uk/socresonline/2/2/2.html>

[5] 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*, Vol. 5, No. 2, <http://www.amstat.org/publications/jse>

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