Journal of the Korean Data & Information Science Society
2012, **23**(6), 1183–1194

http://dx.doi.org/10.7465/jkdi.2012.23.6.1183
한국데이터정보과학회지

설문조사에서의 무응답 처리

이화정[1] · 강석복[2]

[12]영남대학교 통계학과

접수 2012년 10월 4일, 수정 2012년 11월 6일, 게재확정 2012년 11월 22일

요약

설문조사를 실시할 경우 무응답이 발생하지 않는 경우는 매우 드문일이며 무응답이 발생할 경우 무응답처리에 대해서는 다양한 방법이 적용되고 있다. 본 논문에서는 무응답의 두 가지 유형인 단위무응답과 항목무응답의 발생비율을 기존에 조사된 실제 사례를 이용하여 파악하였으며, 무응답률에 따른 분석결과의 차이를 비교하였다. 또한, 대학생들을 대상으로 집단면접조사를 통해 직접 자료를 수집하였고 이 자료를 바탕으로 무응답률과 무응답이 발생하는 이유를 제시하였다.

주요용어: 가중값 조정, 결측값, 단위 무응답, 항목 무응답.

1. 서론

요즈음은 많은 분야에서 다양한 주제로 설문조사가 이루어지고 있으며 또한 다양한 분야에서 여러가지 실험자료들이 쏟아져 나오고 있다. 이처럼 수집된 자료를 바탕으로 분석을 실시할 경우 가장 먼저 직면하는 문제가 바로 결측값 또는 무응답의 처리이다. 계획된 실험에서 얻을 관측 값 중에서 어떤 사고로 관측되지 못한 측정값이 결측값 (missing value)이며 표본조사에서 표본으로 선택된 대상자들 중 일부로부터 원하는 정보를 얻지못하는 경우를 무응답 (nonresponse)이라 한다. 실험에서 발생되는 결측값보다 설문조사에서의 무응답이 좀 더 많이 발생되며 특히 설문조사에서의 무응답은 제어하기가 어렵다. 실험에서 결측값이 발생한 이유가 실험의 실패인지 아니면 연구자가 자리를 비운 사이에 반응이 일어나서 정확히 기록할 수 없는 경우인지 등의 어느 정도 그 원인을 파악할 수는 있으나, 조사의 경우 응답자가 어떤 이유에서 질문에 대한 응답을 하지 않은 것인지를 짐작하기는 쉽지 않다. 또한 실험 자료의 경우는 그 관측 수가 크지 않으며 실험에서 자료를 얻기까지 많은 시간과 비용이 소요되는 경우들이 많으므로 효율적인 측면에서도 실험에서의 결측값 발생 시 여러 가지 방법을 적용하여 결측값를 대체하는 방법이 좋을 것이다. 그러나 설문조사에서 무응답이 발생할 경우는 실험에 비해 표본의 크기도 크며 관련되는 측정문항이 많아 무응답을 대체하기가 쉽지 않다. 이처럼 실험에서의 결측값과 조사에서의 무응답은 발생 원인, 측정되는 자료의 특성 등이 매우 다르므로 그 처리 방법을 다르게 적용해야 한다. 연구자가 자료처리에 대한 경험에서 자주 직면하는 문제가 무응답의 처리에 대한 요구이다. 의뢰자가 무응답 처리에 대한 설명없이 모든 항목에서 무응답이 전혀 없는 결과를 이용한 기존 연구자들의 논문이나 보고서를 제시하며 무응답을 제거를 요구하거나 의뢰자의 설문 자료특성을 고려하지 않고 무응답의 대체방법을 지정하여 요구하는 경우가 있다. 이에 본 연구자는 무응답을 처리하는 방법 중 무응답 대체방법에 대해 무응답의 발생하는 상황이나

[1] (712-749) 경북 경산시 대동 214-1, 영남대학교 통계학과, 강사.
[2] 교신저자: (712-749) 경북 경산시 대동 214-1, 영남대학교 통계학과, 교수. E-mail: sbkang@yu.ac.kr

무응답 자료의 특성에 따라 살펴보고 실제 수집된 기존의 자료를 통해 설문조사에서 무응답률이 어느 정도인지에 조사하고자 한다. 패널자료를 이용하여 무응답의 발생정도에 따라 분석 결과의 차이를 비교하고자 하며 무응답의 처리에 있어 무응답의 원인이 무엇보다도 중요하다고 판단되므로 무응답이 발생되는 이유를 직접 설문조사를 통하여 파악하고자 한다.

2. 설문조사에서의 무응답 처리 방법

설문조사에서의 무응답은 크게 2가지로 나뉜다. 조사 대상자로부터 전혀 정보를 얻지 못한 경우의 단위 무응답 (unit nonresponse)과 질문 문항 중 일부 문항에 대해서 응답하지 않은 항목 무응답 (item nonresponse)이 있다.

2.1. 단위 무응답

일반적으로 단위 무응답의 처리 방법으로 가중값 조정 (weighting adjustment)과 무응답 자료를 분석에서 제외하는 방법이 있다. 가중값을 조정하는 방법에서 가중값은 전체 표본 집단을 여러 개의 무응답층으로 나눈 후 각 무응답층에서 응답률을 추정하여 그 역수를 무응답 조정 승수로 계산한 후 기본 가중값에 곱하여 최종 가중값를 산정한다.

Dufour 등 (2001)은 무응답층을 만드는 방법으로는 의사결정나무등을 이용한 세그맨테이션 모형 (segmentation model)이나 로지스틱 회귀모형 (logistic regression) 방법에 대해 소개하고 SLID (survey of labour and income dynamics)자료를 이용하여 세그맨테니션 모형으로 적용하는 것이 무응답 편의를 줄이는데 효과적이라고 하였다. Kim 등 (2004)은 무응답층을 이용한 가중값 조정방법으로 핫덱 대체법 (hot deck imputation)을 적용한 기존의 중복대체 방법과 무응답층을 고려하지 않은 단순 응답자 평균과 비교한 결과 단순 응답자 평균 방법이 편향이 가장 크며 변동계수 (CV)는 단순중복대체가 가장 크고 단순 응답자 평균방법과 무응답층을 이용한 가중값 조정이 비슷하게 나타나 이론적인 성질과 일치함을 보였다. 이 논문에서 무응답층을 만드는 방법으로는 의사결정나무 모형 중 CHAID (chisquare automatic interaction detection) 알고리즘을 사용하였으며 층내 분산을 줄이기 위해서는 ANOVA의 F-검정 통계량을 사용하였다.

이와 같은 단위 무응답에서의 가중값 조정은 표본에 따라 무응답 층을 다시 구분하는 것이 바람직하다고 이야기 하고 있다. 최근 국민건강영양조사와 같은 대규모 조사에서는 복합표본추출법을 이용하고 있으며 이를 이용하여 가중값 조정방법이 적용되고 있다. 그러나 가중값에 대한 자료는 무응답층에 대한 많은 논의와 연구의 결과로 제시되는 것으로 선거여론조사와 같이 조사에서 결과 발표까지의 소요기간이 매우 짧은 경우 무응답 층을 고려하여 가중값 조정을 한다는 것은 현실적으로 어려움이 많다. 현재 신문이나 방송에서 소개되고 있는 선거여론조사는 응답자들을 대상으로 분석한 결과를 사용하고 있으며 소규모 조사나 개별연구에서도 거의 단위무응답은 분석에서 제외되고 있는 실정이다. 다만 최근에 와서는 응답률을 제시함으로써 무응답의 정도를 알려주기 시작했다.

따라서 비전문가들도 어렵지 않게 적용할 수 있도록 모집단, 조사 주제, 표본추출방법, 응답률 등에 따라 단위 무응답에 사용할 수 있는 가중값의 기준이 제시될 수 있다면 무수히 많이 쏟아지고 있는 다양한 설문조사를 이용한 연구결과의 질을 높이는데 기여할 수 있을 것이라 생각된다.

2.2. 항목 무응답

항목 무응답의 처리방법으로는 완전제거법 (list-wise deletion), 단일대체방법 (single imputation)과 다중대체방법 (multiple imputation)으로 구분할 수 있다. 무응답이 포함되는 케이스를 분석에

서 제외시키는 완전제거법의 경우 편의가 발생되거나 전체표본에 비해 더 작은 자료를 이용하므로 정보의 손실을 가져오게 된다. 측정된 표본을 모두 사용하면서 무응답을 대체하는 방법 중 단일대체방법에는 무응답에 대체되는 값이 랜덤하게 적용되는 확률적 대체방법 (stochastic imputation)과 무응답에 하나의 대체값을 사용하는 결정적 대체방법 (determinaistic imputation)이 있다. 결정적 대체방법으로 연역적 대체 (deductive imputation), 평균 대체 (mean imputation), 비 대체 (ratio imputation), 회귀대체 (regression imputation), 최근방 이웃대체 (nearest neighbor imputation), 축차핫덱 대체 (sequential hot deck imputation) 등의 방법이 있고, 확률적 대체방법으로는 랜덤 핫덱 대체 (random hot dect imputation), 가중 핫덱 대체 (weighted hot dect imputation), 랜덤 비 대체 (random ratio imputation), 랜덤 회귀 대체 (random regression imputation) 등이 있다. 다중대체방법은 단순 대체를 반복하여 무응답을 여러개 대체하는 방법인 MI (multiple imputation)과 베이지안 (Bayesian) 방법 등이 있다.

항목무응답의 경우는 단위무응답에 비해 많은 처리방법이 있기 때문에 어떤 방법을 적용하는 것이 좋은지에 대해 판단하는 것이 쉽지 않다. 무응답 처리방법을 선택하는데 있어서는 우선적으로 자료의 특징을 고려해야 할 필요가 있다. 우선 실험자료에서의 결측값과 조사자료에서 무응답을 동일한 방법으로 처리하는 것은 적절하지 않다. 실험을 통해 자료를 수집한 경우는 표본의 수도 많지 않으며, 수치형 자료의 특성도 대부분 수치형이라 그 대체 방법이 다양하게 적용이 될 수 있다고 생각된다. 그러나 설문지를 통한 자료수집에서는 관측값이 거의 대부분 범주형 자료이므로 평균 대체, 최근방 이웃대체 등과 같은 방법은 적합하지 않다.

범주형 자료의 항목무응답 대체방법으로 적용이 가능한 방법으로는 다음과 같다. 관측된 응답 중 최대 빈도를 갖는 응답을 무응답에 대체하는 방법인 최빈범주법 (modal category model), 무응답이 있는 변수를 종속변수로 하고 무응답이 없는 변수를 설명변수로 하여 로지스틱 회귀분석을 적용한 로지스틱 회귀분석 (logistic regression), 둘 이상의 변수간의 연관성을 적용한 연관규칙에서 연관성이 높은 개체들의 반응 중에서 빈도가 높은 값을 무응답에 대응하는 연관규칙 (association rule), 여러 가지 대체방법 중 많이 선택된 값을 적용하는 투표융합 (voting fusion), 무응답이 없는 완전한 데이터를 학습용 자료와 검증용 자료로 나누어 개별 대체방법의 대체값을 역전파 (back-propagation) 신경망의 입력으로 사용하고 무응답이 발생하지 않는 실제 값을 신경망의 출력으로 하여 개별 대체 방법과 실제값의 관계를 학습한 후 무응답이 있는 데이터에 적용하는 방법인 신경망융합 (neural network fusion) 등이 있다 (Shin과 Sohn, 2002). 또한 랜덤 핫덱 대체를 응용한 방법으로 범주형의 자료에서는 대체군의 적용 시 항목값의 일치여부를 판단하여 점수를 부여한 후 모든 대체군의 항목들을 비교하여 가장 높은 개체들 중 임의로 선택하는 응용 핫덱 방법이 있다 (Choi, 2011).

이처럼 항목 무응답의 처리방법에는 그 종류와 방법이 매우 다양하며 많은 학자들에 의해 논의된 결과가 많으나 일부만 소개하고자 한다.

표본자료에서 발생한 무응답을 무시하고 분석하였을 때의 문제점으로 편의가 발생하는 것과 분산이 과소추정 되는 것이다. 무응답자와 응답자의 평균이 동일한 경우는 편의가 발생되지 않으나 이런 가정이 현실적이지 못하다 (Kim과 Cho, 1996; Kalton, 1983). 무응답을 대체한 경우는 대체한 후의 결과가 무응답이 있는 자료에 기초했을 때보다 편의가 항상 더 작다고 보장할 수 없으며 대체한 자료를 완전한 자료로 간주하면 상관관계와 같은 이변량 또는 다변량 모수에 대한 추정에서 편의가 발생되기 쉽다 (Santos, 1981). 또한 실제 분산을 과소 추정하는 단점이 있다.

Hedderley와 Wakeling (1995)는 무응답을 대체하는데 데이터의 크기가 매우 중요한 인자이며 변수의 크기에 비하여 관측치 개수와 변수의 개수의 관계에 따라 무응답대체 방법들의 성능이 달라진다고 하였다. Kim과 Cho (1996)는 무응답률의 증가하는 비율이 완만하거나 작은 경우에는 대체효과가 크지 않다고 하였으며 무응답률이 급격히 증가하거나 무응답률이 큰 경우는 일반적으로 평균 대체

방법보다는 랜덤 대체방법이 실제값 분포에 더 가까운 대체값을 얻으며 회귀대체보다는 더 안전한 장점이 있다고 하였다.

Kwon (1997)은 선거여론조사에서의 무응답자 처리방법을 다음과 같이 5가지로 소개하고 있다.

첫째, 재계산법으로 지지후보 미결정자와 응답을 거부한 무응답자들을 제외한 상태에서 주요 후보의 지지율을 재계산한 다음 후보자들의 지지율에 비례하여 무응답자들을 할당하는 방법이다. 이는 무응답자의 인구 통계학적 특성과 정치적 성향에 대한 아무가정도 없으며 미결정자들이 투표할 것이라는 가정이 필요하다.

두 번째 방법으로는 무응답자들을 잠재적 투표자에서 제외하고 응답자들만을 대상으로 각 후보에 대한 지지도를 계산하는 방법이며 이것은 무응답자들은 투표일에 투표를 하지 않을 것이라는 가정을 하며 이는 접전을 벌이는 선거에는 맞지 않으나 그 외 미국 여러 선거에서 정확히 들어 맞는 경우가 있었다.

세 번째는 무응답자들을 주요 후보들에게 똑같이 할당하는 균등할당법, 이는 첫 번째 재계산법과 동일한 가정이 적용된다.

네 번째는 무응답자는 도전자에게 지지할 가능성이 높다는 가정하에 도전자에게 모든 무응답을 할당하는 방법이며, 현직자의 재선상황이나 주요후보가 2명인 상황에서만 적용될 수 있는 한계를 가진다.

마지막으로 판별분석을 이용하는 방법으로 미결정자들이 모두 투표일에 투표할 것이라는 가정과 미결정자들이 성향이 비슷하지 않을 것이란 가정하에서 많은 인구 통계학적 특성을 고려하여 적용하는 방법이다.

무응답 대체를 통해 완전자료로 분석하게 될 경우 고려해야 되는 문제에 대해서 Kim (2000)은 첫 번째로 무응답 대체 후 추정량이 구조적 편향 (systematic bias)를 가질 수 있다는 것이고 두 번째로는 추정량의 분산이 과소추정 (underestimation)될 수 있다는 것이다. 특히 분산의 과소 추정문제를 보완한 방법으로는 대체모형을 이용한 방법, 수정된 잭나이프 방법, 다중 대체방법, 붓스트랩을 이용한 방법, 균형이분표본을 이용한 방법에 대해 소개하였으나 실제 문제에서는 개별적 특성에 따라 선택하도록 제안하고 있다.

Shin과 Sohn (2002)은 범주형 자료의 대체방법 중 데이터의 크기가 상대적으로 작으며 무응답 비율이 크면 로지스틱회귀를 피하는 것이 바람직하며 데이터의 크기가 작으면서 한 변수의 무응답이 무응답이 없는 변수와 높은 의존관계가 있으면 최빈범주법이나 연관 규칙의 사용을 피하는 것이 좋으며 이 경우 신경망 융합이 가장 우수한 것으로 소개하고 있다.

Choi (2011)는 경제총조사 항목 무응답의 대체방법으로 2010년 경제 총조사 시범예행조사자료를 이용하여 수치형 자료들을 평균 대체, 회귀 대체, 비 대체, 최빈범주를 이용한 대체, 응용 핫덱 대체 방법을 이용하여 비교한 결과 경제 총조사 무응답 자료로는 응용핫덱대체방법이 가장 적절하다고 이야기하고 있다.

Yoon과 Choi (2012)는 다차원 분할표로 정리된 범주형 자료가 무시할 수 없는 무응답이 있을 경우 최대우도추정값를 구하면 분할표상에서 추정된 무응답 빈도에 대한 확률이 특정 칸에서 0을 가지는 변방값 문제가 발생된다. 이 문제를 해결하면서 선거 자료를 가장 잘 설명하고 예측의 정확도를 높일 수 있는 모형의 추정방법을 제시하였다. 즉, 무응답의 대체와 모수의 추정을 함께 수행하는 계층적 베이지안 방법을 제시하였다.

그러나 이와 같은 방법은 분석한 결과마다 매번 추정하여 결과를 해석하는 것으로 학문적으로 접근하는 것이 아닌 대부분의 통계분석을 소비하는 사람에게는 쉽게 적용하기 어렵다. 그러므로 빠르고 쉽게 사용가능한 일반적인 무응답의 처리방법을 제시하는 것이 좀 더 현실적이라고 생각된다. 무응답 처리는 무응답률과 무응답 발생이유에 따라 우선적으로 그 기준을 다르게 설정해야 할 것이다.

3. 무응답 발생 및 처리

대부분의 설문조사에서는 무응답 발생이 당연시 되고 있으나 설문조사를 활용한 논문이나 보고서에서 무응답률이나 무응답을 처리한 방법을 언급한 경우는 많지 않다. 1993년부터 1997년까지 5년간 출판된 American Political Science Review, American Journal of Political Science, British Journal of Political Sscience에 실린 논문들 중 응답자의 50% 이상이 적어도 1~2개 질문에 응답하지 않았으나 무응답의 처리에 대한 연구자들의 언급이 19%만 언급된 것으로 보고하고 있다 (King 등, 2001).

2000년부터 2004년까지 5년간 행정학 분야 주요 학회지 (한국행정학보, 한국정책학회보, 정책분석평가학회보)에 발표된 일반기고논문을 대상으로 조사한 결과 전체 논문 660편 중 219개의 논문이 설문조사방법을 사용하였으며 이 중 73.1% (160편)의 논문은 회수율에 대해 언급하였으나 회귀분석 논문 중 따로 무응답을 언급한 논문은 174편 중 4.6% (8편)로 미비한 수준인 것으로 나타났다 (Kang과 Kim, 2006).

Groves (2006)는 미국 내 주부들 대상의 여론조사에서 설문조사의 거절률이 매우 높아지고 있다고 언급하며 무응답 편의는 무응답률에 간접적으로 영향을 받고 있다고 하였다. 특히 무응답률의 기준에 대해 Babbie (2007, p. 262)와 Singleton과 Straits (2005, p. 145)의 말을 인용하고 있다.

Babbie는 조사연구에서 중요한 것은 무응답률을 낮추는 것이며 사회 연구에서의 분석 및 보고서에서는 적어도 50%의 응답률이 적절하며 60%는 좋으며 70%의 응답률은 아주 좋은 것이라고 하였다. Singleton과 Straits는 면접 조사의 경우, 85%의 응답 비율은 최소한 적절하다고 하였으며 70% 아래는 편견의 심각한 가능성이 있다고 하였다.

응답률은 조사방법과 조사기관, 조사 내용 등에 따라 차이가 많으므로 연구자는 응답률이나 무응답률에 대해 언급함으로 조사 결과를 활용하는데 있어 판단할 수 있는 정보를 제공하는 것이 바람직할 것으로 보인다. 실제 사례를 통해 무응답의 발생정도, 무응답률에 따른 분석결과의 차이와 무응답 발생원인을 파악하기 위해 한국청소년정책연구원에서 제공하는 "한국 청소년 패널조사" 자료, "제 19대 국회의원선거에 관한 유권자 의식조사" 보고서, 193명의 대학생들을 대상으로 설문조사한 자료를 활용하였으며 각 자료는 minitab 16을 이용하여 분석하였다.

3.1. 청소년 패널조사

청소년 패널조사는 한국청소년정책연구원에서 청소년들의 잠재적 직업선택, 향후 진로설정 및 준비, 일탈행위, 여가참여 등의 생활실태를 파악하기 위해 전국 중학교 2학년과 초등학교 4학년 청소년들을 대상으로 표본추출하여 조사한 것이다. 중학교 2학년은 2003년부터 2008년 즉, 고등학교 졸업 후 1년 차까지 6년 동안 조사하였으며 초등학교 4학년은 2004년부터 2008년까지 5년간 조사한 것이다. 패널조사의 이탈률은 단위 무응답률의 의미와 동일하지 않으므로 단위무응답의 사례로 패널조사는 적절한 자료는 아니다. 그러나 여기서는 1차 표본을 완전한 자료로 사용하고 2차에서 5차까지 자료를 무응답이 발생했을 때의 자료로 이용하면, 무응답률에 따른 분석결과의 차이를 비교할 수 있을 것으로 판단하여 이 자료를 활용하였다.

먼저, 패널조사의 경우 이탈자의 비율이 어느 정도인지 파악하기 위해 초등학교 4학년 2844명을 대상으로 동일 표본을 5차례 패널 조사한 결과와 중학교 2학년을 대상으로 동일 표본을 6차례 조사한 결과의 1차 이후 대상자들의 무응답률을 요약한 결과는 Table 3.1과 같다. 초등학교 4학년 중 2844명을 추출하여 매년 추적 조사한 자료의 경우 2차년도에서는 학생 4.8%와 학부모 6.1%가 이탈된 것으로 나타났으며, 초등학생 4학년이 중학교 2학년이 되는 기간동안 추적되지 않아 조사가 이루어지지 않은 비율이 10%를 초과하였다. 중학교 2학년은 3449명을 추출하여 6차례 조사한 것으로 2차년도에서는 학생 7.6%와 학부모 9.9%가 이탈된 것으로 나타났으며 6차년도는 학생 17.9%와 학부모 17.6%가 이탈된 것으로 나타났다. 이 패널자료에서는 초등학생이 중학교 진학 후 측정된 4차년도 (2007년)에서 이탈률의 증가가 가장 크게 나타났다.

Table 3.1 Nonresponse rate of youth panel data (unit: person (%))

		Year	2003	2004	2005	2006	2007	2008
4th grade of elementary school	Youth	Response	.	2844	2707	2672	2511	2448
					(95.18)	(93.95)	(88.29)	(86.08)
		Nonresponse	.	0	137	172	333	396
					(4.82)	(6.05)	(11.71)	(13.92)
	Parents	Response	.	2844	2670	2630	2499	2451
					(93.88)	(92.48)	(87.87)	(86.18)
		Nonresponse	.	0	174	214	345	393
					(6.12)	(7.52)	(12.13)	(13.82)
2nd grade of middle school	Youth	Response	3449	3188	3125	3121	2967	2833
				(92.43)	(90.61)	(90.49)	(86.02)	(82.14)
		Nonresponse	0	261	324	328	482	616
				(7.57)	(9.39)	(9.51)	(13.98)	(17.86)
	Parents	Response	3449	3106	3081	3093	2950	2841
				(90.06)	(89.33)	(89.68)	(85.53)	(82.37)
		Nonresponse	0	343	368	356	499	608
				(9.94)	(10.67)	(10.32)	(14.47)	(17.63)

위의 청소년 패널조사 (KYPS)에서의 무응답 조정은 항목 무응답을 말하며 항목 무응답에 대한 보정 작업은 실시하지 않고 개별 연구자에게 일임하고 있다. 위 사례에서와 같이 5차례 또는 6차례에 걸쳐 이루어지는 조사에서 최초 표본 이후 무응답 (이탈) 대상들에 대한 대체방법에 대해서는 언급하고 있지 않다. 청소년 패널자료의 초등학교 4학년 대상 패널자료에서 1차 자료와 2차, 3차, 4차, 5차 자료 중 항목무응답이 없었던 질문 중 범주형 분석이 가능한 문항을 선별하여 교차분석을 실시한 결과가 Table 3.2이다. 이 분석은 성별과 학생들의 거주지역 (동, 읍, 면)의 연관성을 분석한 것으로 각각의 결과는 통계적으로 연관성이 없는 것으로 나타났다. 그러나 각 분석에서의 χ^2 통계량과 유의확률 (p-value)의 차이로 무응답률에 따른 분석결과를 비교하고자 한다. 1차 자료 즉 무응답이 없는 자료에서는 χ^2 통계량이 0.626, 유의확률이 0.731이었으나 2차자료는 χ^2 통계량이 2.211,유의확률이 0.331 (무응답률 4.82%), 3차 자료에서는 χ^2 통계량이 1.574, 유의확률이 0.455 (무응답률 6.05%), 4차는 χ^2 통계량이 3.048, 유의확률이 0.218 (무응답률 11.71%), 5차는 χ^2 통계량이 2.022, 유의확률이 0.364 (무응답률 13.92)로 나타났다. 이 결과에서는 무응답이 가장 많은 5차 자료가 1차 자료와 가장 차이가 크거나 무응답이 가장 적은 2차 자료가 차이가 가장 작게 나타난 것은 아니었다. 즉, 무응답률에 낮을수록 완전한 자료의 분석결과와 비슷하다고 말할 수는 없다. 실제 조사자료로 이와 같은 무응답률의 발생에 따라 분석결과의 차이를 비교할 수 없다. 따라서 본 연구자는 이 같은 패널자료들을 적용하면 무응답률에 따른 분석결과 비교나 무응답 대체방법 적용 시 완전자료와의 비교가 가능할 것으로 생각된다.

3.2. 19대 국회의원 선거

실제 여론조사에서 항목 무응답이 어느 정도 발생하는지에 대한 사례로 제 19대 국회의원 선거에 관한 유권자 의식조사보고서 (National Election Commission, 2012)의 분석자료를 이용하였다. 선거관련된 공개된 자료 중 객관적인 자료로 사용할 수 있을 것이라 판단하였으며 항목무응답의 경우는 보고서의 분석결과를 이용하여 항목무응답의 비율을 파악할 수 있다. 단, 실제 조사된 문항이나 보고서에 언급되지 않은 내용이 있어, 설문문항의 수와 분석에 사용된 문항의 수가 차이가 있으며, 다중응답으로 측정된 문항은 항목무응답률에서 제외하였다.

Table 3.2 Analysis results according to the degree of occurrence of nonresponse (unit: person (%))

		A	B	C	Total	Missing	χ^2 (p-value)
1st	Male	1314 (86.22)	167 (10.96)	43 (2.82)	1524 (100)	0	0.626 (0.731)
	Female	1143 (86.59)	146 (11.06)	31 (2.35)	1320 (100)	0	
	Total	2457 (86.39)	313 (11.01)	74 (2.60)	2844 (100)		
2nd	Male	1261 (86.97)	138 (9.52)	51 (3.52)	1450 (100)	0	2.21 (0.331)
	Female	1108 (88.15)	117 (9.31)	32 (2.55)	1257 (100)	0	
	Total	2369 (87.51)	255 (9.42)	83 (3.07)	2707 (100)		
	Missing	0	0	0		137	
3rd	Male	1244 (87.73)	130 (9.17)	44 (3.10)	1418 (100)	0	1.574 (0.455)
	Female	1118 (89.15)	105 (8.37)	31 (2.47)	1254 (100)	0	
	Total	2362 (88.40)	235 (8.79)	75 (2.81)	2672 (100)		
	Missing	16	1	0		155	
4th	Male	1156 (86.98)	126 (9.48)	47 (3.54)	1329 (100)	0	3.048 (0.218)
	Female	1052 (89.00)	100 (8.46)	30 (2.54)	1182 (100)	0	
	Total	2208 (87.93)	226 (9.00)	77 (3.07)	2511 (100)		
	Missing	0	0	0		333	
5th	Male	1140 (87.49)	121 (9.29)	42 (3.22)	1303 (100)	0	2.022 (0.364)
	Female	1022 (89.26)	94 (8.21)	29 (2.53)	1145 (100)	0	
	Total	2162 (88.32)	215 (8.78)	71 (2.90)	2448 (100)		
	Missing	5	0	0		391	

제 19대 국회의원선거에 관한 유권자 의식조사는 선거기간 전 (2012. 3. 19. ∼ 3. 20.), 선거기간 중 (2012. 4. 1.∼ 4. 3.) 그리고 선기기간 후 (2012. 4. 12.∼ 4. 21.) 3차례에 걸려 이루어졌다. 3차례의 조사는 전국 16개 시도에 거주 중인 만 19세 이상 유권자를 대상으로 유효표본 총 1,500명을 조사한 것이며 1차, 2차 조사는 전화면접조사 (computer aided tehephone interview)와 유선전화 RDD (random digit dialing)조사를 하였으며 3차조사에서는 면접원을 이용한 1대 1 개별면접 방법이었다. 표본설정은 1차 2차는 지역/성/연령별 비례할당추출 (quota sampling)이었으며 3차는 그리드 방법 (grid method)을 이용한 조사 대상자를 선택한 것이다. 유권자 의식조사의 경우 1차와 2차의 조사에서는 전화조사의 응답율이 각각 14.6%와 21.6%였다. 따라서 이와 같은 경우는 단위무응답률이 각각 85.4%, 78.4%로 볼 수 있다.

Table 3.3 Item nonresponse rate in the voter survey for the 19th election of members of the National Assembly (unit: number (%))

	0%	Less then 10%	Between 10% and 20%	More then 20%	Total
1st	4 (16.67)	14 (58.33)	4 (16.67)	2 (8.33)	24 (100)
2nd	4 (16.67)	15 (62.5)	4 (16.67)	1 (4.17)	24 (100)
3rd	48 (73.85)	17 (26.15)	0 (0.00)	0 (0.00)	65 (100)

Table 3.3는 설문문항에서 항목 무응답이 발생한 비율을 요약한 것이다. 응답자를 대상으로 무응답을 분석한 것이 아니라 설문문항의 수나 질문유형에 따라 무응답의 발생정도를 파악하기 위한 것이다. 1차, 2차 조사에서는 24개 질문에서 항목 무응답이 발생하지 않은 비율은 모두 16.67%였으며 선거기간 후 조사에서는 65개 문항에서 73.85%로 높게 나타났다. 이는 선거 전에 자신의 생각을 결정하지 못한 부분이 선거 후에 자신의 결정을 응답하는 것으로 조사시기에 따른 차이라고 보여진다. 항목 무응답이 전혀 발생하지 않은 질문들을 살펴보면 1차의 경우 선택형 질문 1개와 3문항이 모두 예/아니오 (yes/no) 형식의 질문이었고, 2차에서는 3개 문항이 3차에서는 29개가 예/아니오 형식의 문항이었으며, 12개가 선택형 질문으로 48개 문항에서 41개 문항이 두 가지 선택형이거나 응답 후 세부질문에 대한 문항으로 대부분 여러 가지 보기 중 선택하는 문항에 있어서 무응답이 발생되는 것으로 나타났다. 이 유권자 조사에서는 여러가지 보기 중 선택하는 문항보다 예/아니오 형식이 무응답이 적게 발생하는 것으로 설문조사에서 질문에 대한 응답의 선택항목 수를 줄이는 것이 무응답률을 낮출 수 있는 방법 중 하나라고 보여진다. 이처럼 무응답률은 조사방법, 조사시기, 질문문항 특성에 따라 크게 차이가 있을 것으로 보여진다. 다만, 이 유권자 의식 조사의 설문지의 경우로 모든 경우에 적용이 가능하다고 볼 수는 없으므로 설문지에 선택 항목수에 따른 무응답률을 비교도 필요할 것이라 생각된다.

3.3. 실제 설문조사에서 무응답율 및 이유

많은 조사에서 응답자가 질문에 대답하지 않는 문항이 왜 발생하는지 그 원인은 무엇인가? 이 질문에 대해 Lee와 Kim (1997)과 Ku (2008)의 논문에는 다음과 같이 소개하고 있다. Lee와 Kim은 자료수집단계에서 무응답원인으로는 연구모집단의 특성, 조사에 대한 부담감, 사생활 침해에 대한 염려, 연구주제에 대한 무관심 또는 거부감, 자료의 수집방법 및 조사기간, 강제적 또는 자발적인 참여에 의한 것이라고 언급하고 있다.

무응답에 대한 처리 및 이해에 관해서 Ku는 선거여론조사에서 무응답을 해석하는데 있어 문제를 제기하였다. 지역 언론인을 대상으로 심층인터뷰와 설문조사를 실시한 결과 대다수의 언론인들은 무응답을 단순히 무관심으로 인해 답하지 않는다고 인식하는 것으로 나타났다. 이 조사에 의하면 전체 조사자 74명 중 무응답이 의미하는 바를 설명하도록 하는 질문에서 58명이 응답하였으며 무응답의 의미가 부동층이라고 응답한 언론인이 8명, 무관심 44명, 두 가지 측면을 모두 지적한 언론인은 6명으로 나타났다. 이 조사에 참여한 대다수의 언론인들은 무응답을 제한적인 하나의 의미로 이해하는 것은 지양되어야 한다고 말하고 있다.

본 연구에서도 무응답의 발생이유를 파악하기 위해 먼저 설문조사를 한 후 응답하지 않은 문항수와 응답하지 않은 주된 이유에 대해 조사하였다. 조사목적으로 인한 응답 거부나 사생활 침해와 같은 이유가 발생하지 않도록 교육용 자료 수집을 목적으로 하여 대학생들을 대상으로 자신들의 음주실태, 여행, 미의 기준에 관련한 질문으로 충분이 응답할 수 있다고 판단되는 문항으로 구성하여 설문조사를 실시하였다.

이 조사는 연구자가 2012년 3월 26일~4월 3일 동안 대구, 경북에 소재한 2개 대학의 총 193명 대학생들을 대상으로 집단면접조사를 실시하였다. 그들의 취향 및 생각에 관한 내용으로 총 25개 문항-즉 범주형 질문 19개 (선별형 질문 1개, 다중응답 2개 포함)와 수치형 질문 6개 문항-으로 구성된 질문지로 설문조사 후 자신이 응답한 설문지에서 응답을 하지 않은 문항의 수와 무응답 이유에 대해서 설문조사하였다.

Table 3.4를 살펴보면 25개 설문문항에 대해 응답자의 69.43%가 모든 질문에 응답하였으며, 1개 질문에 응답하지 않은 경우가 24.87%, 2개 질문은 3.63%로 나타났다. 3개 질문 이상에서 응답하지

않은 경우는 2.08%이며 거의 대부분의 문항에 응답하지 않은 경우는 193명 중 1명 즉 0.52%인 것으로 나타났다. 질문에 따라 살펴보면 무응답이 전혀 없는 질문이 32% (8개)로 나타났으며 전체 문항의 88% (22개)가 질문당 무응답이 4명 이하로 나타났다 . 193명 중 30.57% (59명)가 무응답한 질문이 있는 것으로 나타났다. Table 3.5에서 응답하지 않은 이유를 답한 52명 중 69.23% (36명)가 무엇을 답해야 할지 결정할 수 없어서였으며, 귀찮아서가 11.54% (6명), 기타가 19.23% (10명)로 응답자의 본인도 결정하지 못한 대답에 대해 무응답 대체방법을 적용하는 것은 심사숙고할 필요가 있다고 생각된다. 본 조사에서는 대부분 4개 이하의 문항에서만 응답하지 않았으며 22개를 응답하지 않은 경우는 무응답 이유도 응답하지 않았다. 항목 무응답의 비율이 높은 응답자의 응답은 분석에서 제외하는 것이 바람직할 것으로 보인다. 다만 그 기준에 대해서는 많은 논의가 필요하다. Table 3.5에서 무응답 발생 이유 중 귀찮아서를 기타에 포함시킨 후 무응답의 비율에 따라 무응답 발생이유의 연관성을 Fisher의 정확 검정으로 분석한 결과 유의확률이 0.701로으로 나타나 무응답의 비율에 따라 응답하지 않은 이유와는 상관이 없는 것으로 보여진다. 그러나, 본 연구에서 사용된 질문지는 질문문항이 25개로 질문의 수가 많지 않으며 응답자가 응답하지 않은 문항이 대부분 4개 이하 (15.4%)로 나타나 무응답률과 그 이유에 관해서는 일반화 하기가 어렵다. 추후 다양한 주제의 설문조사에서 무응답의 발생원인에 대한 별도조사가 이루어진다면 무응답의 원인을 구체적으로 파악하여 무응답 처리에 대한 개략적 기준을 제시하는데 도움이 될 것이라 생각된다.

Table 3.4 Nonresponse rate in the survey

Respondents (unit:person (%))			Questions (unit: number (%))		
No. of nonresponse	Persons	%	No. of nonresponse	Questions	%
0	134	69.43	0	8	32.00
1	48	24.87	1	3	12.00
2	7	3.63	2	5	20.00
3	1	0.52	3	3	12.00
4	2	1.04	4	3	12.00
22	1	0.52	10	2	8.00
			15	1	4.00
Total	193	100.00	all	25	100

Table 3.5 Nonresponse and nonresponse reasons (unit: person (%))

	Undecided	Etc.		Total	Missing	Fisher' exact p-value
		Annoying	Etc.			
One	28 (66.67)	5 (11.9)	9 (21.43)	42 (100)	6	0.701
		14 (33.33)				
Two or more	7 (77.78)	1 (11.11)	1 (11.11)	9 (100)	2	
		2 (22.22)				
Total	35 (68.63)	16 (31.37)		51 (100)		
Missing	1	0			8	

4. 결론

많은 연구에서 무응답처리에 대한 논의가 이루어지고 있으며, 최근에는 많은 부분이 무응답 대체 방법에 관심을 두고 있다. 또한 최근 국가통계에 사용되는 자료들의 경우는 단위무응답 및 모집단의 특성을 반영할 수 있는 가중값를 제공하고 있거나 무응답의 처리를 연구자에게 일임하는 방식으로 무응답 처리부분에 대해 상기시키고 있다. 이러한 분위기로 인해 소규모 설문조사나 조사설계가 이루어지지 않은 경우에서조차 무분별하게 무응답 대체을 적용해야한다고 인식되고 있다. 단위 무응답의 경우는 무응답 대체를 하기위해서는 무응답층의 특성 연구가 무엇보다 중요하며, 항목 무응답의 경우는 수집된 자료의 특성을 우선적으로 고려하고 모집단의 특성과 조사목적, 표본크기, 무응답 비율 등을 고려해서 무응답 처리방법을 결정하는 것이 필요하다. Table 3.2의 청소년 패널 자료를 통해 비교한 결과 무응답률이 높다고 해서 반드시 무응답이 발생하지 않은 경우 차이가 크다고 단정할 수는 없다. 이와 마찬가지로 모든 자료에 무응답의 대체를 무조건 적용하는 것 또한 바람직하다고 볼 수는 없다. 분석에 앞서 조사 목적과 대상, 무응답의 자료 특성 등에 따른 무응답의 발생원인을 먼저 고려한 후 무응답의 처리방법을 결정해야 할 것이다. 무응답의 처리에 있어서 무응답의 발생원인에 대해 파악이 불가능할 때 무응답의 대체방법을 성급히 결정하거나 무응답이 발생한 자료를 모두 제거하는 방법보다는 무응답을 하나의 반응으로 처리하거나 개별적 항목에서만 무응답을 분석에서 제외하는 것을 제안한다. 무응답의 대체로 인해 잘못된 결론이 도출되거나 항목 무응답이 발생한 자료를 모든 분석에서 완전히 제거할 경우 의외로 제거되는 수가 많아 자료 손실이 큰게 발생하는 일이 빈번하다. 따라서 항목 무응답이 한 두개 정도로 무응답의 발생비율이 높지 않은 경우 그 항목에서만 무응답을 분석에서 제외하고 무응답의 발생비율이 높은경우 무응답을 하나의 반응으로 하여 처리하는 것이 좋을 것이라 생각된다. 무응답의 처리방법의 선택도 중요하지만 실제 대학생들을 대상으로 개인적인 생각, 의견, 습관 등 민감하지 않고 쉽게 응답이 가능하리라고 판단되는 질문으로 구성된 조사에서조차 무응답이 발생한 이유로 무응답자들의 69.23%가 무엇을 답해야할지 결정할 수 없어서라는 것으로 보아 무응답의 발생을 줄이는 방법으로 질문에 응답할수 있는 반응으로 “결정하지 못함”이라는 부분을 추가하거나, 유권자 자료를 이용한 분석에서 yes/no 형식의 질문이 무응답이 적게 나타난 것처럼 선택항목수를 줄여서 설문문항을 구성할 것을 제안한다. 또한 연구자가 다뤄 본 여러 설문조사에서 1개의 응답을 요구한 경우에도 다중응답하는 경우가 흔히 나타나는 것으로 보아 설문에서 한 가지 응답만을 너무 강요해서 무응답률을 높이는 이유가 아닌가 하는 생각이 든다. 본 논문에서 사용된 사례와 대학생들을 대상으로 한 설문조사로 무응답의 원인을 일반화 하기는 어려우므로 좀더 다양한 주제의 사례와 설문조사를 통한 무응답의 원인에 대한 연구가 필요하다.

참고문헌

Babbie, E. (2007). *The practice of social research*, 11th ed., Wadsworth, Belmont, CA.

Choi, P. G. (2011). *The study of imputation method for item nonresponse in the economic census,* Research report, the first half of 2011, Statistical Center, Statistical Research Institute, Daejeon.

Dufour, J., Gagnon, F., Morin, Y., Renaud, M. and Sarndal, C.-E. (2001). A better understanding of weight transformation through a measure of change. *Survey Methodology*, **27** 97-108.

Hedderley, D. and Wakeling, I. (1995). A comparison of imputation techniques for internal prefeerence mapping, using monte carlo simulation. *Food Quality and Preference*, **6**, 281-297.

Kalton, G. (1983). *Compensating for missing survey data,* Survey Research Center, Institute for Social Research, the University of Michigan, Ann Arbor, MI.

Kang, M. A. and Kim, K. A. (2006). Handling of missing data in public policy studies. *Public Administration Review*, **40**, 31-52.

Kim, J. K., Han, G. S. and Yoon, Y. (2004). Nonresponse weighting adjustment inn Korea household income and expenditure survey. *Journal of the Korean Official Statistics*, **9**, 79-102.

Kim, K. S. (2000). Imputation methods for nonresponse and their effect. *The Korean Association for Survey Research*, **1**, 1-12.

Kim, Y. W. and Cho, S. K. (1996). Imputation method of item nonresponse in sample survey. *Communications of the Korean Statistical Society*, **3**, 145-159.

King, G., Honaker, J., Joseh, A. and Scheve, K. (2001). Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political Science Review*, **95**, 49-69

Ku, T. G. (2008). A study of Korean journalists' perception on poll reporting. *Journal of Political Communication*, **7**, 47-82.

Kwon, H. N. (1997). The 15th presidential election and the polls. *Proceedings of The Korean Association of Broadcasting and Telecommunication*, 159-166.

Kwon, H. N. (2006). *Theory and practice of media election*, Comunicationbooks, Seoul.

National Election Commission. (2012). *Voter opinion survey for the 19th assembly elections*, National Election Commission, Seoul.

Lee, S. W. and Kim, E. G. (1997). Statistical techniques for treatment of nonresponses in public health categorical data. *Journal of the Korean Society of Health Statistics*, **22**, 114-132.

Groves, R. M. (2006). Nonresponse rates and nonresponse bias in household surveys. *Public Opinion Quarterly*, **70**, 646-675.

Santos, R. L. (1981). *Effects of imputation on complex statistics, income survey development program, survey development research center in nonresponse and imputation report on additional task 2,* Survey Research Center, Institute for Social Research, University of Michigan, Ann Arbor, MI.

Shin, H. W. and Sohn, S. Y. (2002). Comparing accuracy of imputation methods for categorical incomplete data. *The Korean Journal of Applied Statistics*, **15**, 33-43.

Singleton, R. and Bruce S. (2005). *Approaches to social research,* 4th ed., Oxford University Press, New York.

Yoon, Y. H. and Choi, B. (2012). Model selection method for categorical data with non-response. *Journal of Korean Data & Information Science Society,* **23**, 627-641.

Journal of the Korean Data & Information Science Society
2012, **23**(6), 1183–1194

http://dx.doi.org/10.7465/jkdi.2012.23.6.1183
한국데이터정보과학회지

Handling the nonresponse in sample survey

Hwa-Jung Lee[1] · Suk-Bok Kang[2]

[12]Department of Statistics, Yeungnam University

Abstract

When it comes to a survey, no answer would occur frequently. Therefore various methods for handling nonresponse have been applied to analyse the survey. In this paper, the ratio of occurrence of two type of nonresponse cases - unit nonresponse and item nonresponse - is presented using previous real survey data, and we compared complete data and data with nonresponse. We suggest the reason of happening of nonresponse and the ratio of nonresponse using data collected through group interviews.

Keywords: Item nonresponse, missing value, unit nonresponse, weighting adjustment.

[1] Instructor, Department of Statistics, Yeungnam University, Gyeongsan, Gyeongbuk 712-749, Korea.
[2] Corresponding author: Professor, Department of Statistics, Yeungnam University, Gyeongsan, Gyeongbuk 712-749, Korea. E-mail: sbkang@yu.ac.kr