연구논문

선거예측에서 편의의 감소: 거짓응답을 중심으로

Bias Reduction in Election Poll Analysis: Measurement Bias of Responses

박 용 치*
Yong-Chie Park

본 논문에서는 선거예측에서 당선자 예측이 빗나가는 요인과 선거예측의 편의를 감소시키는 방안을 검토하였다. 편의를 야기시키는 요인은 크게 무응답에 따른 편의, 대체 응답자 선정에서의 편의, 표본크기의 추정방법에 따른 편의, 추정분포의 적용에 따른 편의, 거짓응답에 따른 편의 등을 생각할 수 있다. 이들 가운데 특히 거짓응답에 따르는 편의를 감소시키기 위하여 거짓응답계수를 계산하고 이로부터 선거예측나무를 이용하여 편의를 수정하는 방법과 사후확률을 이용하여 편의를 수정하는 방법을 가설적 사례를 이용하여 설명하였다. 이에 의하면 당선자 예측에서 편의를 다소간 감소시킬 수 있을 것으로 기대된다.

Researchers in sample surveys have long recognized that what people say they are going to and what they do are not always the same, to say the least. Naturally this inconsistency distorts the survey results, especially in election poll surveys in Korea. How can one cope with this measurement bias? This paper suggests two ideas: election poll tree and using posterior probability. By using two ideas the inconsistency of the poll survey results could be improved much or less.

Ⅰ. 서론

2000년 4월 13일 실시된 제16대 총선에서 방송사와 여론조사 기관들이 공동으로 조사하여 발표한 총선 예측조사의 당선자 예측이 크게 빗나

* 서울시립대학교 법정대학 행정학과 교수

감으로써 선거조사에 대한 비난여론이 비등하였다. 이는 선거예측조사에서 오차가 크게 발생하고 이로 인하여 표본조사의 신뢰성에 대한 문제가 제기된 것이다.

방송사간에 다소 차이가 있으나 대체로 민주당이 한나라당에 7~17석 앞설 것이라고 예측하였으나 실제로는 한나라당이 16석 앞서는 것으로 나타나서 이러한 비난여론이 타당성을 갖게 되었으며 선거예측조사에 참가했던 여론조사기관으로서는 실로 감당하기 어려울 정도의 비난을 받았다. 그러나 여론조사기관을 위시한 표본 조사론자들은 이러한 비난은 표본조사가 반드시 오차를 갖고 있다는 사실을 일방적으로 무시하는 처사라고 반박하면서 그들의 조사는 상당한 정도의 신뢰성이 있으며, 제15대 총선에서 보다 오차의 범위도 줄어들었다고 항변하고 있다.

여기에서 생각할 것은 시민들은 표본 조사론의 이론을 알지 못하며 그들은 총선 예측에서 당선자와 낙선자의 숫자를 예측해 주기를 기대한다는 것이다. 그러나 여론조사기관과 표본 조사론자들은 당선자를 예측하는 것은 마치 점쟁이가 점을 치는 것과 같은 형태이며(류제복, 2000, p. 55), 언론에서도 예상했던 후보가 실제로 얼마나 당선 되었는가로 예측조사의 신뢰성을 평가할 것이 아니라, 좀더 과학적이고 통계적인 방법으로 이를 평가하고 아울러 국민들에게 선거예측 결과에 대한 올바른 판단을 할 수 있도록 유도해야 한다고 주장하고 있다.

시민의 요구는 당선자를 확실히 예측해 달라는 것이고 여론조사기관은 시민이 원하는 것을 해줄 수 없다는 것이다. 본 연구에서는 표본 조사론자의 주장이 틀렸다는 것은 아니다. 그러나 시민의 요구에 부응하지 못하기 때문에 좁은 의미의 표본 조사론에 머물지 말고 표본 조사론을 넘어서 시민의 요구에 부응하는 선거예측을 해야 한다 환언하면, 당선자 예측을 좀더 정확하게 해야 한다는 것이다. 본 연구에서의 주장은 시민들의 이러한 요구가 무리한 것이 아니라고 생각한다는 것이다. 따라서 본 연구에서는 시민의 요구에 좀더 부응하는 선거예측을 함으로써 선거예측에서의 편의를 감소시키는 방법을 모색해 보자는 것이다.

Ⅱ. 선거예측 편의의 설명: 표본 조사론

1. 표본조사론에 의한 설명

선거 예측조사의 이론적 배경은 표본 조사론으로 설명할 수 있다. 표본 조사론에 근거하여 선거예측에서 당선자 예측을 할 수 없다는 주장을 살펴보면 다음과 같다.

첫째, 추출된 표본의 통계량으로 전체 모집단의 모수를 추정하는 표본조사에서는 반드시 표본오차가 존재한다. 즉, 어떤 표본조사에서 표본의 평균값을 $\bar{X}$, 그 표준편차를 s라고 한다면 추정하려는 모집단의 모수 μ의 값은 95% 신뢰수준에서 $\bar{X} - 1.96\frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{s}{\sqrt{n}}$ 사이에 있으며 100번의 표본조사를 할 경우 95번은 이 범위 안에 있고, 5번은 이 범위를 벗어날 것이라는 것이다. 따라서 예측의 표본 통계량이 표본오차의 범위 내에 있다면 당선자 예측이 잘못되었다 하더라도 즉, 후보자의 득표순위 1 - 2위가 바뀌었다 하더라도 이는 잘못된 조사가 아니라 정확한 조사라는 것이며 (미디어 리서치, No.6 , 2000, 1) 거꾸로, 당선자 순위를 맞추었다 하더라도 표본오차를 벗어났다면 이는 부정확한 조사라고 결론지어야 한다는 것이다.

둘째, 선거 예측조사의 성패는 후보자들에 대한 실제 득표율을 얼마나 정확하게 예측하느냐 하는데 달려있다.[1] 따라서 예측조사의 정확도는 예측치와 실제치의 차이를 기준으로 평가하는 것이 올바른 방법이다(미디어 리서치, 2000, 2). 다시 말하면 당선자를 맞추지 못한 곳이 얼마나 되느냐가 문제가 아니라 후보자들의 실제 득표율에 대한 예측값들의 차가 어느 정도 허용오차의 범위를 벗어났는지, 허용오차의 범위를 벗어난 곳

1) 선거 예측조사의 성패는 후보자들에 대한 실제 득표율을 얼마나 정확하게 예측하느냐에 달려 있는데 각 방송사에서 예측 발표한 예상 득표율과 실제 득표율의 사이에는 허용오차의 범위(±5.0%p)를 넘는 것이 상당수에 이르러 이번 선거 예측결과는 통계학적으로 볼 때 신뢰성에 큰 문제가 있다는 평가도 있다 (류제복, 2000: p.46).

이 얼마나 되는지가 조사의 신뢰성을 판가름하는 기준이 된다는 것이다. 즉, 예측치와 실제치의 차이는 당선자에 대한 예측득표율과 실제득표율의 차이($|\hat{p}_1 - p_1| \times 100$)와 1-2위간 예측득표율의 차이와 실제 득표율의 차이 ($|(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)| \times 100$) 모두를 고려하는 것이 바람직하다.

셋째, 선거예측의 환경변화이다. 최근 들어 급속히 확산되어 가는 정치 불신층의 비율이 증가하고 이에 따른 부동층의 확산경향, 지역정서, 정권교체로 인한 기존 여야 지지 성향층의 혼란, 낮은 투표율, 전화 조사의 폭주로 인한 무응답과 거절층의 증가, 솔직하지 못한 조사 응답 등의 악조건 등이 조사 환경을 종합해서 고려해 볼 때 선거 예측만큼은 발전을 해왔다는 것이다.

2. 선거예측에 영향을 미치는 요인

선거예측의 정확성에 영향을 미치는 요인에는 i)무응답에 따른 편의 ii)무응답자 교체에 따른 편의 iii)거짓 응답에 따른 편의 iv)투표율(출구조사의 경우는 편의를 야기시키는 요인이 아니라고 생각됨) v) 표본의 크기 산정의 부적절성 vi) 추정분포의 수정 등으로 설명될 수 있다.

i) 무응답에 따른 편의 : 제 16대 총선에서 처음 실시된 출구조사에서도 무응답률이 약 30%에 이르렀다고 한다. 출구조사는 확률표본을 얻을 수 있고 이에 대한 통계적 방법의 적용도 다양하지만 제16대 총선의 출구조사에서도 상당수의 무응답이 발생하여 선거예측에 심한 편의가 생기지 않을 수 없었다.

ii) 무응답자 교체에 따른 편의 : 출구조사에서 응답자들이 응답을 거부할 때 다른 사람들로 교체하여 응답을 얻는 방법을 사용한다. 그러나 이러한 교체방법을 사용할 경우 편의가 발생할 수 있다.

iii) 표본의 크기산정의 부적절성: 선거예측조사에서 사용되는 표본의 크기는 특정후보의 득표율을 예측하기 위하여 비율추정의 분산공식을 사용한다. 그러나 사전조사에 의하여 경합지역으로 분류되는

지역에서는 두 후보간의 예상득표율의 분산공식을 사용하여 표본의 크기를 결정하는 것이 바람직하다.

iv) 투표율: 예측조사에서의 연령별 응답율이 실제 선거에서의 연령별 투표율과 다르기 때문에 예측의 편의가 발생한다.

v) 추정분포의 수정: 정규분포는 득표율 예측에는 강하지만 성공과 실패로 구분되는 당선자 예측에는 강하지 못하다.

vi) 거짓 응답에 따른 편의 : 예측 조사에서 오류발생의 큰 요인으로 지적되는 것이 거짓 응답이다. 즉, 출구조사에서 응답자들이 솔직한 응답을 하지 않았을 것이라는 점이다.

Ⅲ. 선거예측 편의의 감소방안

1. 선거예측 편의감소의 방안

(1) 무응답에 따른 편의: 선거예측에서 무응답에 따른 예측편의를 수정하기 위해서는 무응답을 줄여야 한다는데서 시작된다. 왜 사람들은 응답하지 않는가? 선거 예측조사는 전화조사에 의하여 이루어지는 것이 일반적이며 이러한 전화조사에서 무응답율은 30-40%에 이른다고 한다.

여론조사가 일상화된 미국에서도 처음 전화를 걸었을 때 응답 거부율은 60-70%나 된다고 한다(변상근, 중앙일보, 2000.3.28). 따라서 두 세번씩 전화를 하여 응답을 받아내는 끈기와 노하우가 여론조사의 생명이다. 응답 거부는 거부한 그들도 실제 투표에는 참여하므로 애국심이나 협조정신의 결여로 보기도 어렵다. 그렇다면 그들은 왜 응답을 거부하는가? 그 이유는 다음 몇 가지로 요약될 수 있겠다.

첫째, 시간을 빼앗기고 하던 일을 방해받기 때문이다. 둘째, 이런 저런 조사기관이라고 밝히면서 질문을 해오지만 실제로 조사기관을 지칭하는 사람들이 과연 누구인지 알 재간도 없다. 최소한 상대방은 나의 전화번호라도 알고 있는데 말이다. 셋째, 어떤 질문은 개인의 프라이버시를 노출시킬 위험도 있다. 일단 조사에 응하게 되면 중도에 답변을 그만두기

도 어려워 응답하고 싶지 않은 것에도 응답해야 한다. 넷째, 조사자들이 응답자에게 영향을 미치려 하는 경우도 적지 않다. 다섯째, 자신의 견해를 남에게 밝히려 하지 않는 사회적 태도도 있다. 자신의 견해를 밝혔다가 피해를 본다는 잠재의식도 무시하지 못하며, 사이비 조사기관들이 난립하고 심야에 상대후보의 이름을 거론하며 사칭하기도 한다.

따라서 이러한 무응답을 줄이기 위해서는 첫째, 조사가 응답자의 생활을 방해하지 않도록 배려하여야 한다. 이를 위해서는 짧은 시간에 끝내는 것도 한가지 방법이 되겠고 직업 등을 고려하여 잠재적 응답자가 바쁠 것으로 추정되는 시간을 피하여 조사를 진행하는 것도 다른 한가지 방법이 되겠다. 둘째, 조사기관의 신분을 명백히 밝혀야 한다. 일방적으로 자신의 신분을 밝히기 보다 응답자가 확신할 수 있도록 예를 들면 응답자가 전화하게 하는 방법도 있겠다. 셋째, 응답자의 프라이버시를 확실히 보장해야 한다. 즉, 응답결과에 대하여 응답자의 신분을 철저히 보장해주어야 하며 그것을 믿게 해야 한다. 넷째, 응답자의 생각에 영향을 미쳐 선거 결과를 일정한 방향으로 유도하려는 행위를 해서는 안되며 이를 차단하는 법적 방법도 모색되어야 한다. 다섯째, 어떠한 응답을 하더라도 응답자에게 어떠한 해악이 미치지 않는다는 생각을 응답자가 할 수 있도록 사회적인 인식을 바꾸어 가도록 노력하여야 한다.

(2) 대체 응답자 선정: 무응답자를 교체하는 경우 편의를 줄이기 위해서는 앞에서 언급했듯이 무응답을 줄이는 것과 대체 응답자 선정에서 무응답자와 성향이 비슷한 표본으로 교체하여 대표성을 확보하는 것이 중요하다. 이를 위해서는 무응답자 대체규칙을 상세히 만들어 현장에서 기계적으로 적용할 수 있게 해야 한다. 현장의 조사요원이 반자의적으로 대체응답자를 고르지 않도록 해야 한다.

(3) 표본크기의 추정방법: 경합지역과 비경합지역의 경우 표본크기는 주어진 여건을 감안하여 표본의 크기를 결정하여야 한다는 것이다. 일반으로 선거예측조사에서 사용되는 표본의 크기는 특정 후보의 득표율을 예측하기 위한 것이므로 비율추정의 분산을 이용하여 예측한다. 그러나 선거예측조사에서 관심의 초점은 당선이 확실한 지역보다 1, 2위 후보들 간의 당선 가능성이 백중하게 경합을 벌리는 곳이므로 이러한 지역에 대

한 선거예측에서는 두 후보에 대한 득표율의 차이에 대한 분산을 이용하여 예측하는 것이 더 타당할 것이다(류제복, 2000, 49). 즉, 표본크기의 추정방법을 다르게 해야 하며 득표율의 차이에 대한 추정을 할 때는 득표율을 추정할 때보다는 95% 신뢰수준에서 약 4배의 표본을 조사하여야 한다.[2)]

(4) 추정분포의 수정: 추정분포의 수정이란 지금까지 흔히 사용되어오던 추정분포를 정규분포에서 이항분포 또는 beta 분포를 사용하는 방법을 생각할 수 있다. 정규분포는 득표율 추정에는 강점이 있지만 성공과 실패와 같은 상황인 당선자 예측에는 약점이 있기 때문이다.

첫째, 이항분포의 경우를 보면 이항분포는 다음과 같은 특징을 갖는 상황에서 적용될 수 있다.

i) 2분법적 결과 : 불확실한 사건이 계속적으로 발생하고 각각의 사건은 두 가지 결과 중의 하나를 갖는 경우이다. 성공/실패, 예/아니오, 당선/낙선 등

ii) 일정한 확률 : 각각의 사건은 동일한 성공의 확률(p)를 갖는다.

iii) 독립성 : 각각의 사건의 결과는 다른 사건의 결과와 독립적인 경우이다. 즉, 성공 확률이 선행하는 결과에 의존하지 않는다.

주요 방송사나 여론 조사기관은 선거의 결과를 예측하는데 출구조사의 결과를 어떻게 외삽(extrapolate)시키는가? 이항분포는 결과표본에 기초한 선거에서 누가 당선될 것인가에 대하여 확률진술을 하는 기초를 형성한다. 이제 n명의 유권자 가운데 r명이 D당 후보에 찍을 확률은 $P_B(R = r \mid n, p) = \frac{n!}{r!(n-r)!} p^r (1-p)^{n-r}$ 이 된다. D당 후보의 기대치는 E(R)=np, 그 분산은 Var(R)=np(1-p) 이다.

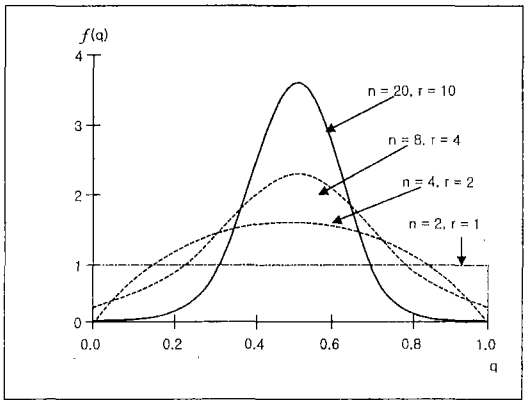
---

2) 허용오차를 B라고 할 때, 표본의 크기 n은 i) 득표율 추정의 경우 $z_{0.025} \fallingdotseq 2$ 이므로

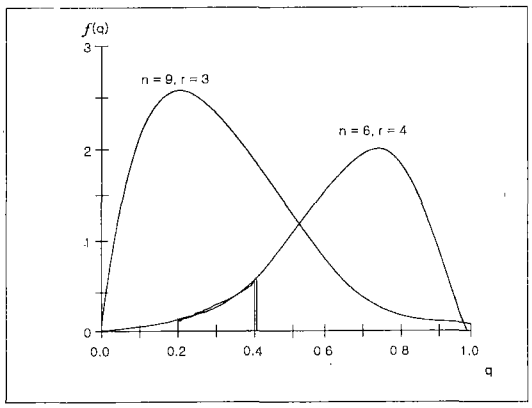
$$n = \frac{z^2_{0.025}\, p(1-p)}{B^2} \leq \frac{2^2}{4B^2}$$ 이 되고, ii) 득표율 차이의 추정의 경우

$$n = \frac{z^2_{0.025}\{p_1(1-p_1) + p_2(1-p_2) + 2p_1p_2\}}{B^2} \leq \frac{2^2}{B^2}$$ 이 된다.

둘째, β-분포의 경우를 보자. 다음번 선거에서 D당 후보에 투표할 유권자의 비율에 관하여 관심을 갖는다고 하자. 이 비율이 불확실한 것이라면 이 불확실성을 연속확률분포로 생각할 수 있다. 그리고 이 비율은 0과 1사이의 값을 가질 수 있으므로 지수분포나 정규분포도 이 불확실성을 정확하게 반영할 수는 있다. 그러나 β 분포가 적절할 수도 있다. 0과 1사이의 값을 가질 수 있는 불확실한 량을 Q라고 하면 β 분포의 밀도함수는 $f_\beta(q \mid r, n) = \frac{(n\text{-}1)!}{(r\text{-}1)!(n\text{-}r\text{-}1)!} q^{r\text{-}1} \cdot (1\text{-}q)^{n\text{-}r\text{-}1}$ 이다. 여기에서 n이 크지면 분포는



〈그림-1〉 대칭적 beta 분포



〈그림-2〉 비대칭적 beta 분포

타이트해지고, n이 작으면 분포는 더 흩어진다. $r < \frac{n}{2}$이면 오른쪽으로 skew되고, $r > \frac{n}{2}$ 이면 왼쪽으로 skew된다. β 확률변수의 기대치는 $E(Q) = \frac{r}{n}$ 이고, 그 분산은 $Var(Q) = \frac{r(n-r)}{n^2(n+1)}$이다. 대체적으로 말하면 이것은 n명의 투표자 가운데 r명이 D당 후보에 투표할 확률로 해석될 수 있다.

2. 거짓응답에 따른 편의

표본조사론에 근거하고 있는 여론조사는 응답자들의 진실한 응답을 전제하고 있다. 만일 응답자들의 응답이 진실에 토대를 두지 않는다면 그러한 여론조사는 의미가 없어진다. 그러나 최근의 선거예측조사에서 응답자들이 과연 진실한 응답을 하고 있는지에 대하여 의심을 제기하지 않을 수 없는 상황이 전개되었다. 왜 응답자들은 진실한 응답을 하지 않는가? 이에 대한 대답은 그리 간단하지가 않은 것 같다.[3] 그런데 응답자의 응답이 확신할 수 없는 경우가 자주 발생하고 있다는 생각이 든다. 한국의 상황에서 투표 등 정치적 의견을 조사할 때 그 응답을 진실을 말한다고 생각할 수 없는 경우가 그것이다.[4]

표본 서베이를 연구하는 사람들은 그들이 하려고 하는 것과 그들이 행하는 것은 반드시 동일한 것이 아님을 오래 전부터 인정해 왔다. 당연히 이러한 비일관성은 서베이 결과를 왜곡시킨다. 우리는 이러한 측정편의(measurement bias)에 어떻게 대처해야 하는가?

투표 날의 출구조사에서 응답자들은 우리의 정치상황에서 자신이 투표한 것을 사실대로 응답하지 않는 경향이 있다. 따라서 출구조사의 응답을 그대로 선거예측에 적용하는 것은 잘못을 범할수 있다.

거짓응답에 따른 편의를 수정하는 방법은 응답과 관련된 추가적인 정보를 찾아내어 응답결과를 수정하는 방법이다. 이러한 정보의 출처는 경험적인 자료와 주관적인 견해로 분류될 수 있다.

---

3) 이에 대한 이유는 크게 문화적 요인과 정치상황적 요인으로 구분될 수 있다(박용치, 1976).

4) 다른 예를 들면 스무고개(twenty questions)에서 어린애들의 응답이 그렇다.

i) 경험적 자료: 경험적 자료(empirical data)는 확률분포를 직접 부과하기 위해 사용될 수 있다. 이 경우에 사전의 믿음(prior belief)은 상대도수가 옳은 확률을 나타낸다는 것을 검증하기 위해 사용되거나, 또는 상대도수가 선거예측자에 의해 부과된 확률을 잘 나타내도록 하기 위해 그들을 매끄럽게 하는 데에 사용된다. 또한 경험적 자료는 사후분포를 획득하기 위하여 사전분포와 우도분포(likelihood distribution)와 함께 사용될 수도 있다.

ii) 전문가의 주관적 견해 : 선거예측에서는 경험적 자료를 이용할 수 없는 경우도 많다. 이러한 경우에 유일한 정보의 출처는 주관적 견해 (subjective opinion)이다. 문제를 분석할 때 도움을 얻기 위해 전문가를 활용하는 것은 적절한 방법 중의 하나라 할 수 있다. 그리고 분석과정에서 전문가는 여러 역할을 하고 있다. 첫째, 분석과정을 창안하고 지도하는 것이고 둘째, 선거예측자의 입장에서 선거예측을 대행하는 것이며 셋째, 선거예측을 위한 정보를 공급하는 것이다. 이 정보는 어떤 경우에는 전문가가 가지고 있는 경험적 자료일 수 있으나, 어떤 경우에는 어떤 불확실 상황(미지의 결과 또는 확률변수)에 대한 전문가의 주관적 견해(즉, 믿음)이다. 전문가의 견해를 다룰 때 선거예측자는 전문가의 활용 방안과 획득한 정보의 처리방법을 결정해야 한다.

(1) 전문가의 활용방안: 전문가를 활용하는 방안에는 전문가를 교사로서 활용하는 방안과 전문가의 판단을 활용하는 방안이 있다.

i) 전문가를 교사로서 활용 : 선거예측자 자신이 전문가가 되겠다는 시도 아래 전문가와 그 문제에 대해 토론하는 것이다. 즉, 근본적으로 전문가와 같은 전반적인 지식을 획득하는 것이다. 예를 들어, 선거예측의 편의 감소와 관련된 선거예측 문제에서 선거예측자는 전문가를 교사로서 활용할 수 있다.

ii) 전문가의 판단을 활용: 전문가를 활용하는 다른 방법은 문제가 되고 있는 특정의 불확실 상황에 대해 전문가의 판단을 요구하는 것이다. 즉, 전문가와 같은 정도의 깊이로 문제를 이해하려고 시도하

지는 않는 것이다. 그보다는 오히려 선거예측자가 전문가에게 판단을 요구하는 것이다. 전문가의 판단을 끌어내는 방법에는 여러 가지가 있다.

첫째, 가장 발생가능성이 높은 결과를 나타내는 하나의 값, 즉 점추정치(point estimate)를 끌어내는 것이다. 예를 들어 출구조사의 예에서, 전문가에게 "거짓응답의 여부에 대해 당신의 판단을 얘기해 주십시오"라고 질문할 수 있다. 또는 특정 후보의 당선가능성에 대한 불확실성 상황에 대해 전문가에게 "이 후보의 당선가능성에 대해 당신의 판단을 얘기해 주십시오"라고 질문할 수 있을 것이다.

둘째, 더 많은 정보가 필요하거나 또는 결과치의 범위에 관한 정보가 필요한 문제라면, 전문가에게 최악의 결과치와 최선의 결과치(즉, 확률변수의 최대치와 최소치)에 대한 판단을 요구할 수 있을 것이다. 결과치의 범위에 관해 전문가의 판단을 듣고 선거예측자는 전문가가 자신의 진술에 얼마나 확신을 갖고 있는지를 알고 싶어할 것이다. 만약 "매우 확신한다" 또는 "확신하지 못한다"라고 한다면 전문가로 하여금 확률을 부과하도록 해야 한다. 확률분포는 전문가가 선거예측자에게 자신의 판단을 전달하는 궁극적인 수단이다. 확률분포는 전문가가 그의 판단에 대해 줄 수 있는 모든 정보를 제공한다.

(2) 전문가의 판단의 처리: 선거예측자는 전문가가 제공한 정보를 처리해야 한다. 수정된 믿음은 P(관심을 두고 있는 결과 | 전문가로부터의 정보)이다. 이것은 Bayes의 정리를 사용하면 다음과 같이 표현된다.

$$P(\text{관심을 두고 있는 결과} \mid \text{전문가로부터의 정보}) = \frac{P(\text{전문가로부터의 정보} \mid \text{관심을 두고 있는 결과}) \times P(\text{관심을 두고 있는 결과})}{P(\text{전문가로부터의 정보})}$$

비록 이 방법이 믿음을 수정하기 위해 실제 사용되지는 않지만 전문가를 활용할 때 고려해야 할 중요한 사항을 강조하고 있다. 우도확률 P(전문가로부터의 정보 | 관심을 두고 있는 결과)는 본질적으로 전문가에게

대한 신뢰성의 척도이다. 만약 전문가를 완전하게 신뢰한다면, 우도확률은 이를 반영하게 될 것이다. 예를 들어, 선거예측의 경우에서 당신은 원래 이 응답자가 거짓응답을 할 가능성이 0.3 이라고 믿었고, 전문가는 이 응답자가 확실히 거짓응답할 것이라고 진술하였다고 가정하자, 그러면

$$P(\text{거짓응답함} \mid \text{거짓응답할 것이라고 전문가가 말함}) = \frac{P(\text{거짓응답할 것이라고 전문가가 말함} \mid \text{거짓응답함}) \times P(\text{거짓응답함})}{P(\text{거짓응답할 것이라고 전문가가 말함})}$$

만약 전문가를 완전하게 신뢰할 수 있다고 당신이 믿는다면 P(거짓응답할 것이라고 전문가가 말함 | 거짓응답함) = 1.0 이다.

$$\begin{aligned} &P(\text{거짓응답할 것이라고 전문가가 말함}) \\ &= P(\text{거짓응답할 것이라고 전문가가 말함} \mid \text{거짓응답함}) \times P(\text{거짓응답함}) \\ &\quad + P(\text{거짓응답할 것이라고 전문가가 말함} \mid \text{거짓응답하지 않음}) \times P(\text{거짓응답하지 않음}) \\ &= 1.0 \times 0.3 + 0 \times 0.7 \\ &= 0.3 \end{aligned}$$

따라서, 다음과 같이 기대했던 결과를 얻게 된다.

$$P(\text{거짓응답함} \mid \text{거짓응답할 것이라고 전문가가 말함}) = \frac{1.0 \times 0.3}{0.3} = 1.0$$

만약 전문가를 완전하게 신뢰할 수는 없다고 당신이 믿는다면 우도확률(likelihood probability)로 1.0보다 작은 확률을 부과하게 될 것이다. 여기에서 제공된 정보가 선거예측자에 의해 처리되는 과정에서 전문가에 대한 신뢰성이 매우 중요하다. 전문가가 그의 견해를 왜곡시킬 어떤 동기를 가지고 있을 때에는 그의 신뢰성은 당연히 낮아지며, 만약 수정절

차를 사용하면 이 사실은 정보에 대해 부과한 우도함수에 반영될 것이다. 물론 어려운 점은 전문가의 믿음을 직접 관측할 수 없다는 점이다. 오직 확률이 부과되었던 불확실 상황의 결과만을 관측할 수 있을 뿐이다.[5] 전문가가 그의 견해를 왜곡시킬 동기를 가질 수 있다는 잠재성을 선거예측자는 인식해야 하며 이 왜곡은 확률수정과정에서 고려되어야 한다.

여러 명의 전문가를 활용할 경우에는 이들의 견해를 통합하여야 한다. 평균치나 가중평균치와 같이 간단하다는 매력이 있는 여러 방법이 있다. 그러나 면밀히 살펴보면 이 방법들은 적절하지 않다. 전문가들로 하여금 서로 토론하도록 하여 그들의 견해차를 파악하도록 하는 것이 대개의 경우 가장 효과적인 방법이다. 기본적으로, 전문가들 스스로 그들의 견해를 통합하도록 하는 것이다. 이것은 그들의 견해차가 각 전문가가 가지고 있는 정보가 서로 다르기 때문에 발생한다는 가정과 일치한다. 만약 견해차가 파악될 수 있으면 전문가들은 합의에 도달할 수 있을 것이다. 이와 같은 방법에서는 합의에 영향을 미칠 집단역학(group dynamics)에 의한 작용이 있을 수 있으므로, 이를 최소화하기 위해 전문가들이 토론할 환경을 주의 깊게 설계하여야 한다.[6]

Ⅳ. 분석결과: 가설적 사례

1. 선거예측나무의 이용

여기에 출구조사의 응답과 관련된 몇 가지 통계적 자료가 있다. 응답자는 D당과 H당간의 선택을 해야 하며 그는 D당에 투표한 경우 80%만이 정확하게(바르게) 응답하고, H당에 투표한 경우 70%만이 정확하게(바

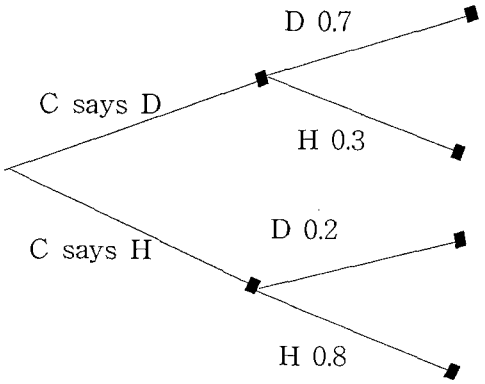
5) 이러한 경우 도덕적 위해(moral hazard : 신뢰성이 불확실한데서 발생하는 위험)의 문제가 있다고 말한다.

6) 이를 위하여 delphi방법을 생각할 수 있을 것이다.

르게) 응답한다고 하자 이 정보를 표와 그림으로 나타내면 〈표-1〉 및 〈그림-1〉과 같다.

〈표-1〉 응답자의 응답의 조건확률

유권자의 응답	실제투표		합 계
	D당에 투표	H당에 투표	
D당에 투표	0.70	0.30	1.00
H당에 투표	0.20	0.80	1.00



〈그림-1〉 예측득표율 모형

우선 이러한 정보를 갖고 표본을 추출할 수 있겠는가? 그렇다면 어떤 실험을 할 수 있겠는가? 본 논문에서는 거짓 응답계수를 계산하고 이를 토대로 하여 선거예측에서의 순위가 뒤바뀌는지를 검토해 보기로 한다.

〈표-2〉 거짓응답계수

D당 지지율이 높은 지역		H 당 지지율이 높은 지역	
D 당	H 당	D 당	H 당
0.77771	0.477433	0.645028	0.844034

(1) 거짓응답계수의 계산: 사전조사(출구조사)에서의 결과와 실제 득표율을 토대로 거짓응답계수를 구할 수 있다. 거짓응답계수는 사전조사(출구조사)에서의 결과와 실제득표율간의 관계로부터 회귀분석계수로 계산되었다. 이것은 D당을 지지하는 비율이 높은 지역과 H 당을 지지하는 비율이 높은 지역에서 차이가 날 것으로 생각되어 이를 분리하여 계산하였다. 그 결과는 〈표-2〉와 같다.

(2) 예측득표율의 계산: 거짓응답을 고려하는 예측득표율을 계산하면 〈표-3, 4〉와 같으며 실제득표율과 비교할 때 1, 2위의 순위는 실제득표율과 일치한다.

〈표-3〉 D당 지지의 경우

단위: %

지역구	사전조사 D당	사전조사 H당	실제득표율 D당	실제득표율 H당	예측득표율 D당	예측득표율 H당
강 서 을	42.6	47.5	46.1	41.4	58.0	32.1
강 동 을	43.9	47.6	48.4	42.3	59.0	32.5
인천/남동을	38.9	44.9	43.3	40.0	53.7	30.1
인천/계양	41.8	51.9	48.4	44.4	59.6	34.1
안양/만안	30.7	34.0	38.8	32.0	41.6	23.1
부천/원미을	45.5	46.5	50.9	41.3	59.7	32.3
평 택 을	25.7	34.2	35.9	33.7	37.9	22.0
고양/덕양갑	44.3	47.5	50.0	41.0	59.3	32.5
시 흥 시	27.5	35.4	40.1	37.9	39.9	23.0
안 성 시	40.5	50.6	49.6	43.3	57.9	33.2
강원/원주	26.7	48.7	35.1	33.8	46.2	29.2
제 주	43.9	46.5	55.2	39.8	58.4	32.0

거짓응답계수	
D당	H당
0.77771	0.47743

	A	B	C	D	E	F	G
1	<표-4> H당 지지의 경우						
2							단위: %
3		사전조사		실제득표율		예측득표율	
4	지역구	D당	H당	D당	H당	D당	H당
5	광 진 갑	49.1	44.2	41.4	50.8	44.2	49.1
6	은 평 갑	35.5	33.7	37.7	43.5	33.7	35.5
7	수원/권선	42.3	36.3	34.3	40.5	36.3	42.3
8	안양/동안	49.0	47.0	48.4	49.1	47.0	49.0
9	광 명 시	46.2	45.0	46.0	47.4	45.0	46.2
10	남양주시	49.3	31.2	35.8	36.8	31.2	49.3
11	군 포 시	48.6	41.5	45.3	45.6	41.5	48.6
12	하 남 시	43.1	38.4	34.9	40.1	38.4	43.1
13	광 주 군	39.9	36.4	34.1	34.2	36.4	39.9
14							
15	거짓응답계수						
16	D당	H당					
17	0.645028	0.84403					
18							

2. 사후확률의 이용

(1) 사후확률의 계산:

① 1단계: 사전확률은 새로운 정보를 얻기 전에 알고 있던 확률로써 이전의 선거에서의 지지율로써 가늠한다. 이 확률은 P(D투표), P(H투표), P(J투표) 표현된다.

② 2단계: 새로운 정보는 출구조사에 의하여 얻어진 정보이며 이전의 선거에서 D투표이면서 D출구라고 대답하는 확률로 추정한다. 이 확률은 실제 투표에서의 출구조사에서 (D투표수)에 대한 (D출구수)로 계산하고 그 비율은 전문가의 판단에 의존할 수 있다. Bayes의 정리에 따르면

$$P(D\text{출구} \mid D\text{투표}) = \frac{P(D\text{출구 and } D\text{투표})}{P(\text{투표})} \text{ 이므로}$$

$P(D\text{출구 and } D\text{투표}) = P(D\text{출구} \mid D\text{투표}) \cdot P(\text{투표})$ 가 된다. 이로부터 P(D출구 | D투표)를 구할 수 있고 다음과 같은 조건확률을 구할 수 있다.

즉,

P(D출구 | D투표), P(H출구 | D투표), P(J출구 | D투표), P(NRes | D투표)
P(D출구 | H투표), P(H출구 | H투표), P(J출구 | H투표), P(NRes | H투표)
P(D출구 | J투표), P(H출구 | J투표), P(J출구 | J투표), P(NRes | J투표)을 구할 수 있다.

③ 3단계: 위의 결과를 토대로 하여 우리가 구하고자 하는 사후확률은[7]

$$P(D\text{투표} \mid D\text{출구}) = \frac{P(D\text{투표 and } D\text{출구})}{P(D\text{출구})} \text{ 의 유형들이다.}$$

이번 선거의 출구조사에서 P(D출구)를 구할 수 있고 P(D투표 and D출구)는 이전선거에서 구할 수 있다. 그러므로 P(D투표 | D출구)를 계산할 수 있다. 즉, 우리가 구하는 확률은 이와 같은 사후확률인데 그것들은 다음과 같다. 즉,

P(D투표 | D출구), P(H투표 | D출구), P(J투표 | D출구)
P(D투표 | H출구), P(H투표 | H출구), P(J투표 | H출구)
P(D투표 | J출구), P(H투표 | J출구), P(J투표 | J출구)
P(D투표 | NRes), P(H투표 | NRes), P(J투표 | NRes)을 구한다.

이 논리를 적용하여 사후확률을 계산한 것이 〈표 -5, 6, 7〉와 같다. 편의상 무응답이 10%, 20%, 30% 로 구분하여 사후확률을 계산하였다.

(2) 득표수의 계산: 선거예측에서는 사전조사(출구조사)에서 무응답이 다수 있고 이를 여러 가지 방식에 의하여 적절하게 분류하여 최종 조사결과에 조정 사용할 것으로 생각된다. 본 연구에서는 사전조사(출구조사)에서의 응답을 사후확률을 사용하여 예측결과를 수정하는 방법을 검토

7) 우리가 구하는 사후확률은 다음과 같이 구할 수 있다.

$$P(D\text{투표} \mid D\text{출구}) = \frac{P(D\text{투표 and } D\text{출구})}{P(D\text{출구})} = \frac{P(D\text{출구} \mid D\text{투표}) \cdot P(D\text{투표})}{P(D\text{출구})}$$

해 본 것이다. 사후확률을 이용하여 득표수를 계산하는 것은 간단하며 출구조사에서의 숫자에 해당하는 사후확률을 곱하여 더하기만 하면 된다. 가설적 사례[8]에 대한 그 결과를 제시하면 〈표-5, 6, 7〉의 아래 부분과 같다.

본 가설적 사례에서는 투표자수를 96,000명으로, 무효투표는 없는 것으로 가정하였다. 또한 출구조사에서 무응답이 많은 것을 고려하여 무응답 비율이 10%, 20%, 30%를 가정하여 계산해 보았다. 이 계산 결과에 의하면 출구조사에서는 H당 후보〉 D당 후보〉 J당 후보의 순서이었지만 사후확률을 사용하여 이를 수정하면 D당 후보〉 H당 후보〉 J당 후보로 순위가 바뀌고 D당 후보가 당선되게 된다. 이를 무응답 비율에 따라 따로 살펴보면 다음과 같다.

i) 무응답 비율을 10%로 가정한 경우: 이 경우는 출구조사에서 D당에 투표했다고 응답한 자는 37,000명, H당에 투표했다고 응답한 자는 38,000명, J당에 투표했다고 응답한 자는 11,400명, 무응답자는 9,600명이다. 그러나 사후확률을 사용하여 계산한 결과는 D당 후보가 43,771표, H당 후보가 43,509표, J당 후보가 8,720표로 되어 출구조사에서는 H당 후보가 당선될 것으로 예측되었으나 실제로는 D당 후보가 당선되는 결과를 가져오게 된다.

ii) 무응답 비율을 20%로 가정한 경우: 이 경우에도 앞에서와 같은 방식이 된다. 출구조사에서 D당에 투표했다고 응답한 자는 33,500명, H당에 투표했다고 응답한 자는 34,500명, J당에 투표했다고 응답한 자는 8,800명, 무응답자는 19,200명이다. 그러나 사후확률을 사용하여 계산한 결과는 D당 후보가 44,589표, H당 후보가 43,679표, J당 후보가 7,732표로 되어 출구조사에서는 H당 후보가 당선될 것으로 예측되었으나 실제로는 D당 후보가 당선되는 결과를 가져오게 된다.

---

8) 본 연구에서 사용된 가설적 사례는 투표자 수가 96,000명이고 출구조사 결과 37,000명이 D당에, 38,000명이 H당에, 11,400명이 J당에, 9,600명이 무응답한 경우를 상정하였다(무응답 비율이 10% 인 경우).

	A	B	C	D	E
1	<표-5> 사후확률의 추정 (무응답: 10%)				
2					
3	사전확률				
4		D당에 투표	H당에 투표	J당에 투표	
5		0.48	0.43	0.09	
6					
7	조건부확률				
8		D당에 투표	H당에 투표	J당에 투표	
9	CsaysD	0.700	0.200	0.050	
10	CsaysH	0.150	0.650	0.050	
11	CsaysJ	0.050	0.050	0.800	
12	NoResp	0.100	0.100	0.100	
13		1.000	1.000	1.000	
14	결합확률				
15		D당에 투표	H당에 투표	J당에 투표	주변확률
16	CsaysD	0.336	0.086	0.005	0.427
17	CsaysH	0.072	0.280	0.005	0.356
18	CsaysJ	0.024	0.022	0.072	0.118
19	NoResp	0.048	0.043	0.009	0.100
20					1.000
21	사후확률				
22		D당에 투표	H당에 투표	J당에 투표	
23	CsaysD	0.787808	0.201641	0.010551	1.000000
24	CsaysH	0.202247	0.785112	0.012640	1.000000
25	CsaysJ	0.204255	0.182979	0.612766	1.000000
26	NoResp	0.480000	0.430000	0.090000	1.000000
27					
28	출구조사응답자수				
29	CsaysD	37000			
30	CsaysH	38000			
31	CsaysJ	11400			
32	NoResp	9600			
33					
34	예상득표수				
35		D당에 투표	H당에 투표	J당에 투표	
36	CsaysD	29149	7461	390	37000
37	CsaysH	7685	29834	480	38000
38	CsaysJ	2329	2086	6986	11400
39	NoResp	4608	4128	864	9600
40	총득표수	43771	43509	8720	96000
41					

	A	B	C	D	E
1	<표-6> 사후확률의 추정 (무응답: 20%)				
2					
3	사전확률				
4		D당에 투표	H당에 투표	J당에 투표	
5		0.48	0.43	0.09	
6					
7	조건부확률				
8		D당에 투표	H당에 투표	J당에 투표	
9	CsaysD	0.600	0.200	0.050	
10	CsaysH	0.150	0.550	0.050	
11	CsaysJ	0.050	0.050	0.700	
12	NoResp	0.200	0.200	0.200	
13		1.000	1.000	1.000	
14	결합확률				
15		D당에 투표	H당에 투표	J당에 투표	주변확률
16	CsaysD	0.288	0.086	0.005	0.379
17	CsaysH	0.072	0.237	0.005	0.313
18	CsaysJ	0.024	0.022	0.063	0.109
19	NoResp	0.096	0.086	0.018	0.200
20		0.480	0.430	0.090	1.000
21	사후확률				
22		D당에 투표	H당에 투표	J당에 투표	
23	CsaysD	0.760898	0.227213	0.011889	1.000000
24	CsaysH	0.230032	0.755591	0.014377	1.000000
25	CsaysJ	0.221198	0.198157	0.580645	1.000000
26	NoResp	0.480000	0.430000	0.090000	1.000000
27					
28	출구조사응답자수				
29	CsaysD	33500			
30	CsaysH	34500			
31	CsaysJ	8800			
32	NoResp	19200			
33					
34	예측득표수				
35		D당에 투표	H당에 투표	J당에 투표	
36	CsaysD	25490	7612	398	33500
37	CsaysH	7936	26068	496	34500
38	CsaysJ	1947	1744	5110	8800
39	NoResp	9216	8256	1728	19200
40	총득표수	44589	43679	7732	96000
41					

	A	B	C	D	E
1		<표-7> 사후확률의 추정 (무응답: 30%)			
2					
3	사전확률				
4		D당에 투표	H당에 투표	J당에 투표	
5		0.48	0.43	0.09	
6					
7	조건부확률				
8		D당에 투표	H당에 투표	J당에 투표	
9	CsaysD	0.500	0.200	0.050	
10	CsaysH	0.150	0.450	0.050	
11	CsaysJ	0.050	0.050	0.600	
12	NoResp	0.300	0.300	0.300	
13		1.000	1.000	1.000	
14	결합확률				
15		D당에 투표	H당에 투표	J당에 투표	주변확률
16	CsaysD	0.240	0.086	0.005	0.331
17	CsaysH	0.072	0.194	0.005	0.270
18	CsaysJ	0.024	0.022	0.054	0.100
19	NoResp	0.144	0.129	0.027	0.300
20					1.000
21	사후확률				
22		D당에 투표	H당에 투표	J당에 투표	
23	CsaysD	0.726172	0.260212	0.013616	1.000000
24	CsaysH	0.266667	0.716667	0.016667	1.000000
25	CsaysJ	0.241206	0.216080	0.542714	1.000000
26	NoResp	0.480000	0.430000	0.090000	1.000000
27					
28	출구조사응답자수				
29	CsaysD	30000			
30	CsaysH	31000			
31	CsaysJ	6200			
32	NoResp	28800			
33					
34	예측득표수				
35		D당에 투표	H당에 투표	J당에 투표	
36	CsaysD	21785	7806	408	30000
37	CsaysH	8267	22217	517	31000
38	CsaysJ	1495	1340	3365	6200
39	NoResp	13824	12384	2592	28800
40	총득표수	45371	43747	6882	96000
41					

iii) 무응답 비율을 30%로 가정한 경우: 이 경우에도 앞에서와 같은 방식이 된다. 출구조사에서 D당에 투표했다고 응답한 자는 30,000명, H당에 투표했다고 응답한 자는 31,000명, J당에 투표했다고 응답한 자는 6,200명, 무응답자는 28,800명이다. 그러나 사후확률을 사용하여 계산한 결과는 D당 후보가 45,371표, H당 후보가 43,747표, J당 후보가 6,882표로 되어 출구조사에서는 H당 후보가 당선될 것으로 예측되었으나 실제로는 D당 후보가 당선되는 결과를 가져오게 됨은 앞에서의 경우와 일치한다.

V. 결어

본 논문에서는 선거예측에서 당선자 예측이 빗나가는 요인과 선거예측의 편의를 감소시키는 방안을 검토하였다. 편의를 야기시키는 요인 가운데 특히 거짓응답에 따르는 편의를 감소시키기 위하여 거짓응답계수를 계산하고 이로부터 선거예측나무를 이용하여 편의를 수정하는 방법과 사후확률을 이용하여 편의를 수정하는 방법을 가설적 사례를 이용하여 설명하였다. 이에 의하면 당선자 예측에서 편의를 다소간 감소시킬 수 있을 것으로 기대된다.

참고문헌

류제복 (2000), "선거예측조사의 신뢰성 증진방안: 출구조사를 중심으로." 《조사연구의 방법론적 쟁점》. (서울: 한국조사연구학회, 2000 춘계학술대회).

미디어 리서치 사회조사실(2000), "16대 총선 예측조사 평가 및 신뢰성 증진 방안." *MRI Newletter*, no. 6. (서울: 미디어 리서치), pp.1-4.

박용치(1976), 태도조사에서의 응답투 《논문집》 3-2, 대전: 충남대학교 사회과학연구소.

변상근 (2000), 나는 왜 무응답자인가, 중앙일보 2000.3.28.

Raiffa, H. (1970), *Decision Analysis: Introductory Lectures on Choices under Uncertainty*. (Reading, Mass.: Addison-Wesley).

Clemen, R.T.(1996), *Making Hard Decision: An Introduction to Decision Analysis*, 2nd ed. (Pacific Grove: Duxbury Press).

von Winterfeldt, D. and W. Edwards(1986), *Decision Analysis and Behavioral Research.* (Cambridge, Mass.: Cambridge University Press).