선거예측조사의 신뢰성 증진방안

-출구조사를 중심으로-

A plan of improving the reliability of the election forecast survey
- A case of exit poll -

류 제 복*

요 약

지난 4월 13일에 실시된 16대 총선에서 방송사와 조사기관들이 공동으로 조사하여 발표한 선거예측조사에서 많은 오류가 발생하고 이로 인해 표본조사에 대한 신뢰성에 큰 문제가 제기되었다. 이에 향후 선거예측조사의 신뢰성을 회복하고 보다 정확한 예측을 위해 기 발표된 예측조사내용을 다각도로 심층분석하고 조사의 오류가 발생한 원인과 이를 해결하는 방안을 제시하였다. 아울러 이번에 처음으로 실시된 출구조사에 대한 문제점과 개선 방안도 함께 살펴보았다.

I. 서 론

2000년 4월 13일 제16대 국회의원선거 마감시간인 오후 6시가 지나자마자 KBS와 SBS, 그리고 MBC가 동시에 선거예측조사의 결과를 발표하기 시작하였다. KBS와 SBS는 미디어리서치, 코리아리서치, TN소프레스 글로벌리서치, 그리고 한국리서치 등 4개 조사전문회사와 콘소시엄 형태로, MBC는 한국갤럽과 선거예측조사를 실시하여 그 결과를 일반시청자들에게 자신있게 공표하였다. 그러나 그 결과는 몇 시간이 지나지 않아 엄청나 오류의 파고 속에서 공용방송과 조사기관의 실수를 만천하에 공개하는 상황이 되어버렸다.

4년 전인 1996년 4월 11일의 제15대 국회의원선거에서도 TV 3사와 CBS가 5개 여론조사기관과 합동으로 투표자전화조사를 실시하여 투표가 끝나는 시간에 조사결과를

*청주대학교 통계학과

발표하였으나, 실제 개표결과와 39곳이나 차이가 났다하여 공영방송의 공신력과 여론조사의 신뢰성에 커다란 상처를 받았다. 당시에는 지금보다도 더 열악한 조사환경과 적절치 못한 조사방법으로 인하여 정확치 못한 조사의 가능성이 높았다. 96년의 15대 총선에서 조사의 오류에 대한 많은 지적과 함께 개선 방향으로 한결같이 제안한 방법이 바로 출구조사로 이것만이 오류를 줄일 수 있는 최선의 방법이라 하였다. 당시 문제점으로 지적되었던 방송사의 경쟁적 방송태도, 윤리성의 미비에 대한 질책, 그리고 조사기관의 안일한 조사방법과 얼버무리기 식 대처가 4년간의 세월 속에 고스란히 묻혀 있다 그대로 발굴되어 다시 사용되었다는 것이다.

이미 많은 선진국에서 실시하여 그 신뢰성을 인정받고 있는 출구조사만이 최선의 방법이고 이 방법을 국내에 도입한다면 정확한 조사결과를 얻을 수 있어 국민들의 알 권리를 충실하게 충족시켜주어 선거방송의 효시가 될 것이라는 기대를 모았다. 그러나 국내에서 실질적으로 처음 실시된 출구조사(경합 또는 혼선지역인 80개 정도의 지역에 대해서만 출구조사를 실시하였다)를 병행한 선거예측조사의 결과가 지난 96년의 조사에서의 오류와 마찬가지로 개선된 점이 보이지 않자 이에 대한 국민들의 기대와 조사관련 집단의 희망이 무너지고 조사 자체의 무용론과 출구조사를 아예 허용하지 말자는 주장들이 한편에서 거세게 일고 있다. 이는 단순한 출구조사의 문제가 아니고 통계조사 전반에 대한 불신과 무용론의 대두로 볼 수 있다.

대부분의 언론에서 보도한 바와 같이 MBC가 23곳 KBS와 SBS가 21곳을 실제와 틀리게 예측 보도하여 국민들로부터 비난과 담당자들에 대한 문책이 이어졌다. 그러나 통계학적 측면에서 볼 때 당선자를 맞추지 못한 곳이 얼마나 되는가가 문제가 아니라 후보자들의 실제 득표율에 대한 예측값들의 차가 어느 정도 허용오차의 범위를 벗어났는지 그리고 허용오차의 범위를 벗어난 곳이 얼마나 되는지가 조사의 신뢰성을 판가름하는 기준이 된다

따라서 본 연구에서는 금번에 실시된 16대 총선에서 선거예측조사결과의 오류 발생의 원인들을 분석하고, 또한 법적 문제점, 조사방법상의 문제점 그리고 그 밖의 여러 사항들을 검토하여 향후 선거예측조사의 신뢰성증진을 위한 방안을 제시하고자 한다.

Ⅱ. 조사결과의 비교

우리 나라에서 실시된 선거에서의 여론조사는 14대 총선 이전에는 법적으로 할 수 없었다. 그러다 1992년 3월 24일에 실시된 14대 총선에서야 비로소 조사 자체는 허용되었지만 공표는 할 수 없었다. 그 후 14대 대선에서는 선거기간 이전에는 공표를 할 수 있게 되는 등 우리 나라에서 선거에 대한 여론 조사의 역사는 아주 미천한 상태이다. 그러나

국민들의 알권리에 대한 욕구가 거세지고 나라도 점차적으로 민주화되면서 국민들의 이러한 바램과 언론기관과 조사관련 단체 등에서의 계속적인 요구로 15대 총선에서야 방송사들과 조사기관들이 함께 참여하여 실시하는 여론조사가 실시되었다. 그러나 이러한 여론조사는 말만의 여론조사지 실제로는 여론을 제대로 반영할 수 없는 상황이었다.

1996년의 15대 총선에서는 투표당일 언론기관이나 조사기관이 투표결과를 예상하기 위해서 조사를 할 수 있었지만 투표소로부터 500m밖에서만 허용되었고 투표의 비밀이 침해되지 말아야 한다는 조항과 막대한 조사비용으로 인해서 조사다운 조사를 실시할 수 없었다. 이런 상황하에서 각 방송사와 조사기관은 연대해서 조사 일주일 전부터 전화조사를 실시하고 그것을 근거로 예상 득표율을 발표하였으나 많은 오류를 범하게 되었다. 이러한 문제가 발생한 후 많은 토론회와 연구에서 선진국에서 이미 실시되고 있는 출구조사의 필요성이 제기되었다. 만병의 치료제처럼 생각된 출구조사 방법을 병행한 이번의 조사도 만인의 기대를 여지없이 짓밟아 버렸는데 이는 출구조사의 실시를 위한 여건이 아직 충분히 마련되지 않았고 사전에 이에 대한 철저한 준비가 없었던 것이 큰 문제점이라 할 수 있다. 15대 총선과 이번 16대 총선에서의 차이를 비교해 보면 <표 1>과 같다.

<표 1> 15대와 16대 총선 비교

구 분	15대총선(1996. 4. 11)	16대총선(2000. 4. 13)
선거인 수	31,488,294	33,504,262
국회의원 수 (지역구/비례대표)	299(253/46)	273(227/46)
선 거 법	언론기관이 선거결과를 예상하기 위해서 투표소로부터 500m밖에서 투표의 비밀이 침해되지 않는 방법으로 질문이 가능	투표소로부터 300m밖으로 개정
투 표 율	63.9%	56.4%(최저)
1,000표 차 이내	14	15(300표차내:9, 20표차내:4)
최소 표 차	350	3
조사방법	전화조사(투표당일 포함 1주일간)	경합지역 80여 곳은 출구조사(면접, 무기명)를 하고 그 밖의 지역은 전화조사
조사비용	15억	45억 (MBC:22억,KBS-SBS:23억)

15대와 16대 총선에서의 차이는 우선 국회의원 수가 지역구에서만 26명이 줄었다. 500m이내에서 조사를 할 수 없었던 선거법이 300m 이내로 다소 완화되었지만 그밖에 법적으로 변화된 것은 없었다. 투표율은 16대가 사상 최저의 투표율(56.4%)을 기록하였고 당선자와 2위간의 차이가 1,000표 이내인 지역은 지난 15대와 비슷하나 박빙의 지역이 이번이 훨씬 많았다. 특히 300표 차 이하의 지역이 9곳이었고 20표 미만의 지역도 4곳이나 되었다. 예측조사에서의 큰 차이점은 금번에 처음으로 출구조사가 실시되었다는 점이다.

16대 총선에서 227개 지역구에 대해 정당별 국회의원 의석수에 대한 방송사들의 예측은 <표 2>에 나타나 있다. 비례대표의원 45명에 대해서 MBC와 KBS-SBS의 예측은 민주당과 한나라당이 20석, 자민련이 5석으로 예측치가 같았으나 결과는 민주당이 19석 한나라당이 21석을 차지하였다.

<표 2> 16대 총선 정당별 지역구 의석 수

<table>
<tr><th>구 분</th><th>투표결과</th><th>MBC(한국갤럽)</th><th colspan="2">KBS-SBS
(미디어리서치 등 4개기관)</th></tr>
<tr><td>민주당</td><td>96</td><td>107(+11)</td><td colspan="2">112(+16)</td></tr>
<tr><td>한나라당</td><td>112</td><td>100(-12)</td><td colspan="2">95(-17)</td></tr>
<tr><td>자민련</td><td>12</td><td>12</td><td colspan="2">12</td></tr>
<tr><td>민국당</td><td>1</td><td>2(+1)</td><td colspan="2">1</td></tr>
<tr><td>한국신당</td><td>1</td><td>1</td><td colspan="2">1</td></tr>
<tr><td>민주노동당</td><td>0</td><td rowspan="2">5</td><td>2</td><td rowspan="2">6(+1)</td></tr>
<tr><td>무소속</td><td>5</td><td>4</td></tr>
<tr><td>계</td><td>227</td><td>227(±12, 5.3%)</td><td colspan="2">227(±17, 7.5%)</td></tr>
</table>

* ( )내의 숫자는 오차, %는 오차율을 나타냄

이번에 실시된 출구조사에서 각 방송사(조사기관들)들이 사용한 방법들에 대한 비교는 <표 3>에 있다. 출구조사는 앞에서 언급한 바와 같이 MBC, KBS-SBS 모두 경합(또는 혼선)지역 80여 곳을 대상으로 실시하였는데 각 선거구에서 6-10개의 투표소를 확률추출하고 추출된 투표소로부터 5-7명 주기의 계통추출을 사용하여 응답자를 선정하였다. 조사원은 대부분이 20대 여대생이었는데 이것이 조사원채용의 기본 원칙을 등한시한 것으로 이미 여러 곳에서 지적된 바와 같이 무응답 또는 응답거부를 줄이거

나 적절하게 대처하지 못하는 하나의 요인이 된 것으로 보여진다. 조사원들에 대한 교육은 다소 차이가 있으나 대체로 3-6시간 정도였고 조사원들을 위한 지침서를 마련하여 조사원교육에 활용하였다. 출구조사 시간은 한국갤럽이 오전 8시부터 오후 4시까지, 미디어리서치와 TN소프레스가 오전 6시부터 오후 4시, 한국리서치는 오전 6시부터 오후 3시, 그리고 코리아리서치가 오전 6시부터 오후 3시 30분까지 실시하여 조사시간에 약간의 차이가 있었다. 이번 오차의 가장 큰 요인으로 볼 수 있는 무응답자 또는 응답거부자들의 비율은 대략 30%정도로 추정되는 데 이 층의 대부분이 40대 후반 이후의 연령이고 그 중에서도 여성이 많았다고 보여진다. 무응답이나 응답거부의 주된 요인으로는 출구조사에 대한 이해부족과 출구조사시 개인의 비밀(누구에게 투표를 하였는지 여부)이 노출 될 것을 우려하였기 때문인 것으로 예상된다.

<표 3> 출구조사에 대한 비교

구 분	MBC(한국갤럽)	KBS-SBS(미디어리서치 등 4개 조사기관)
조사방법	경합지역 80여 곳은 당일 출구조사를 실시하고 나머지 지역은 사전, 당일 전화조사를 실시	좌 동
허용오차	±4.4%p	±3.1%p~ ±4.4%p
출구조사시 응답자 선정방법	투표인수에 따라서 5-7명당 1명씩 계통추출	좌 동(약간의 차이는 있음)
출구조사방법	직접질문->기록->전송	투표형식(용지주고->기입->봉투 또는 상자 투입)
조사원	여대생(20대)	좌 동(일부 남학생)
조사원 배치	조사구당 1팀 2명	조사구당 1팀 평균 5명
자료전송	개인휴대정보단말기(PDA)	ARS 등
무응답자 또는 응답거부자 조치방법	다음 사람으로 대체, 대체한 사람도 무응답이면 그 다음 사람(응답한 사람부터 다시 추출방법을 적용)	무응답자, 응답거부자들의 성별, 연령만을 기입하여 무응답 처리 하거나(2곳), 다른 응답자로 대체(2곳)
무응답 층과 출구조사의 이해 정도	40대 후반 이후의 연령이 무응답층의 주류를 이루고 있으면 그중 여성이 많음. 출구조사에 대한 이해가 부족하고 비밀보장이 안 된다고 우려	좌 동

* 본 자료의 일부는 2000년 4월 13일 오후 6시에 방송된 T·V자료를 참고하였음.

선거예측조사의 성패는 후보자들에 대한 실제 득표율을 얼마나 정확하게 예측하느냐에 달려있는데 각 방송사에서 예측 발표한 예상 득표율과 실제 득표율과의 차이가 허용오차의 범위(지역별로 허용오차의 범위가 달라 여기서는 $\pm 5.0\%p$를 일차적으로 허용오차의 범위로 하였다)를 넘는 것이 상당수에 이르러 이번 선거예측결과는 통계학적으로 볼 때 신뢰성에 큰 문제가 있다고 평가할 수 있다. 당선자에 대한 예상 득표율과 실제 득표율과의 차이 그리고 1, 2위간 예상득표율의 차이와 실제 득표율간의 차이에 대한 비교가 <표 4>에 제시되어 있다.

<표 4> 지역별 예상 득표율과 실제 득표율과의 차이

지 역	$\lvert \hat{p}_1 - p_1 \rvert \times 100$				$\lvert (\hat{p}_1 - \hat{p}_2) - (p_1 - p_2) \rvert \times 100$			
	5.1 ~ 10.0 %p		10.1%p이상		10.0 ~ 19.9 %p		20.0%p이상	
	M	K-S	M	K-S	M	K-S	M	K-S
서울(45)	7	11	0	5	2	11	0	2
부산(17)	7	3	7	0	6	2	5	0
대구(11)	7	4	4	2	8	5	1	1
인천(11)	3	2	0	0	2	1	0	0
광주(6)	1	3	3	2	3	2	1	1
대전(6)	1	0	0	0	0	1	0	0
울산(5)	2	2	0	1	1	2	0	1
경기(41)	4	4	2	2	6	4	1	2
강원(9)	2	2	0	0	2	1	0	0
충북(7)	1	1	0	1	0	0	0	1
충남(11)	1	3	0	1	1	4	0	0
전북(10)	5	3	2	1	2	3	2	1
전남(13)	6	3	2	2	5	4	0	1
경북(16)	6	7	1	3	5	8	1	2
경남(16)	7	5	0	2	3	3	0	2
제주(3)	2	0	0	0	2	0	0	0
합계(227)	**62 (27.3%)**	**53 (23.3%)**	**21 (9.3%)**	**22 (9.7%)**	**48 (21.1%)**	**51 (22.5%)**	**11 (4.8%)**	**14 (6.2%)**

<표 4>로부터 당선자들에 대한 예상 득표율과 실제 득표율과의 차이를 볼 때 오차가 ±5.0%p를 초과한 지역이 전체 227개 선거구 중에서 MBC(한국갤럽)가 83개 지역으로 전체의 36.6%이고 KBS-SBS(미디어리서치외 3개회사)도 MBC의 경우보다는 다소 적지만 75개 지역으로 전체의 33.0%를 차지하고 있다. 오차가 ±10.0%p를 초과한 지역도 MBC와 KBS-SBS가 각각 21개 지역(9.3%)과 22개 지역(9.7%)이나 되었다. 예측치가 가장 크게 빗나간 경상도지역(부산, 대구, 울산, 경북, 경남)은 65개 선거구에서 오차가 ±5.0%p를 초과한 지역이 MBC가 41개 지역으로 경상도전체의 63.1%이고 KBS-SBS는 29개 지역으로 경상도전체의 44.6%에 이르고 있다. 또한 ±10.0%p를 초과한 지역도 MBC가 12개 지역(18.5%)이고 KBS-SBS가 8개 지역으로 12.3%를 차지하고 있다. 전라도(광주, 전북, 전남)지역도 전체 29개 지역 중에서 MBC는 ±5.0%p를 초과한 지역이 19개 지역으로 전라도전체의 65.5%에 이르고 있으며 ±10.0%p를 초과한 지역도 7개 지역(24.1%)이나 되었다. KBS-SBS도 전라도지역에서 ±5.0%p와 ±10.0%p를 초과한 곳이 각각 14개 지역과 5개 지역으로 전라도전체의 48.3%와 17.2%를 차지하였다. 특히 MBC는 부산과 대구지역에서 오차범위 ±5.0%p내에 있는 지역은 28개 선거구중에서 단 3곳에 불과했다. 반면에 KBS-SBS는 서울에서 오차범위 ±5.0%p를 초과한 지역이 16개 지역으로 45개 선거구의 35.6%가 되었다.

이번 16대 총선에 대한 선거예측조사에서의 특징은 전체적으로 오차범위를 넘은 지역이 많았고 지역별로 편차가 극심하였다. 전국227개 선거구 중에서 경상도와 전라도에는 전체 선거구의 41.4%인 94개 선거구가 있다. 그러나 오차가 ±5.0%p를 초과한 지역이 MBC의 경우 전체 83개 지역 중에서 경상도가 41개 지역(49.4%) 전라도가 19개 지역(22.9%)으로 두 지역이 전체의 72.3%를 차지하고 있으며 ±10.0%p를 초과하는 지역은 전체 21개 선거구 중에서 경기도지역 2곳을 제외한 19개 지역이 모두 경상도와 전라도지역이다. 한가지 흥미로운 것은 오차가 ±10.0%p를 초과하는 21개 지역 중에서 경기도 2곳과 경북 1곳을 제외하고는 모두 작게 예측하였다. 한편 KBS-SBS도 허용오차가 ±5.0%p를 초과한 전체 75개 지역 중에서 경상도 29개 지역(38.7%), 전라도 14개 지역(18.7%)으로 전체의 57.3%나 되고 ±10.0%p를 초과하는 지역은 전체 22개 선거구 중에서 경상도(8)와 전라도(5)가 13개 지역(59.1%)이었다. 그러나 MBC와는 반대로 17개 지역을 높게 예측하였다.

당선자에 대한 예측치의 차가 가장 크게 난 지역은 MBC가 부산 사하갑의 엄호성(한나라당)후보에 대해 실제보다 20.7%나 낮게 예측하였고 KBS-SBS는 전북의 진안-무주-장수 지역의 정세균(민주당)후보를 18.8%나 높게 예측하였다.

1위와 2위간의 차에 대한 예측도 실제와의 차이가 ±10.0%p이상인 지역이 MBC가 59개 지역((26.0%)이고 KBS-SBS가 65개 지역(28.6%)이나 되었다. 차이가 가장 큰 지역은 MBC의 경우 역시 부산 사하갑에서 실제보다 40.77%나 낮게 예측한 반면에

KBS-SBS는 전북의 진안-무주-장수 지역에서 37.2%나 높게 예측하였다.

위의 결과로 볼 때, 예측결과의 오류에 큰 영향을 주는 요인으로 지역적 특성을 들 수 있다. 따라서 향후 예측에서는 이러한 지역적 특성에 의한 조사방법의 개발과 가중치 조정문제에 많은 심혈을 기울여야 할 것이다.

MBC와 KBS-SBS가 당선자들에 대해 각각 발표한 예상 득표율의 차이가 ±5.0%p가 넘는 지역이 100개 지역이나 되었고 1위와 2위간의 차이에 대한 예상득표율과 실제 득표율간의 차가 ±10.0%p이상 나는 지역도 80곳이나 됨을 <표 5>로부터 알 수 있다. 1위 예측차이가 10.1%p이상 나는 39지역 중에서 2곳을 제외한 37개 지역에 대해서 KBS-SBS가 MBC보다 크게 예측하였고 1위와 2위와 예측차이에 대해서도 1곳을 제외한 15개 지역에 대해서 KBS-SBS가 크게 예측하였다. 예측에 있어서 전체적으로 KBS-SBS가 MBC보다 크게 예측하였다. 예측한 방송사들간의 차이를 볼 때도 예측에 있어서의 심각한 오류가 있음을 알 수 있다.

<표 5> 방송사들 간의 예측치 차이

구 분	$\lvert KS(\hat{p}_1) - M(\hat{p}_1) \rvert$		$\lvert KS(\hat{p}_1 - \hat{p}_2) - M(\hat{p}_1 - \hat{p}_2) \rvert$	
	5.1 ~ 10.0%p	10.1%p이상	10.0 ~ 19.9%p	20.0%p이상
서울(45)	11	7	9	2
경기인천(52)	11	4	9	3
경상도(65)	21	19	27	8
전라도(29)	7	5	10	1
충청도(24)	8	4	6	2
강원도(9)	1	0	2	0
제주도(3)	2	0	1	0
계(227)	**61**	**39**	**64**	**16**

Ⅲ. 표본크기의 계산

표본조사를 실시하기 전에 주어진 여건을 감안하여 표본 크기를 정하여야 한다. 표본 크기는 조사의 신뢰성 및 조사비용과 직접적으로 관련이 있으므로 이점을 고려하

여 결정하게 된다. 통상적으로 선거예측조사에서 사용되는 표본크기의 결정은 특정후보의 득표율을 예측하기 위한 것이므로 비율추정의 분산공식을 사용한다. 주어진 신뢰수준과 허용오차의 범위 하에서 특정후보의 지지율을 예측하기 위한 표본 크기는 식(1)을 사용한다. 지난 15대 총선의 전화조사나 이번 16대 총선의 전화조사와 출구조사의 경우도 마찬가지지만 지금까지 대부분의 선거조사에서 이러한 공식을 사용하였다고 보여진다. 그러나 선거예측조사에서 관심의 초점은 당선이 확실한 지역보다는 1위와 2위 후보들간의 당선 가능성이 백중한 곳이 되고 이 지역에 대한 예측에서의 차이가 여론조사의 신뢰성에 큰 영향을 미치게 된다. 이번의 선거예측조사에서도 MBC의 경우 혼선지역(1위와 2위간의 예측차이가 $\pm 5.0\%p$이하의 지역) 53개 중에서 실제와 예측치 간의 차이가 이 범위를 벗어난 지역이 19개 지역(35.8%)이고 KBS-SBS의 경우 경합지역(MBC와 달리 1위와 2위의 예상득표율의 차가 $\pm 5.0\%p$이하가 아니라 허용오차 범위내의 지역) 51개중에서 주어진 범위를 벗어난 지역이 13개(25.5%, 허용오차의 범위가 각 지역별로 달라 일정하지 않아 MBC와 같이 $\pm 5.0\%p$를 기준으로 하였다)이나 되었다. 이러한 결과로 볼 때 사전조사에 의해 혼선 또는 경합지역으로 분류되는 지역에 대해서는 타 지역과 동일한 방식을 사용하기보다는 두 후보간의 예상득표율 차의 분산공식을 사용하여 표본크기를 정하는 것이 바람직하다고 본다. 이때는 비율 차가 다항분포를 하게 되므로 식(2)를 사용하여 표본크기를 정한다. <표 6>과 <표 7>에 의하면 신뢰수준과 허용오차의 범위가 같다하여도 표본크기가 두 후보간의 비율 차를 사용하였을 때가 커지게 된다.

1. 특정 후보에 대한 득표율 추정($\hat{p}$)

i번째 응답자(X_i)가 특정 후보에 투표를 했으면 1이고 그렇지 않으면 0이라 하면 $X_i \sim b(1, p)$ 인 베르누이분포를 한다. 단순확률표본 n으로부터 특정후보의 득표율에 대한 추정량과 추정량의 분산은 각각 다음과 같다.

$$\hat{p} = \frac{1}{n}\sum_{i=1}^{n} x_i,$$

$$Var(\hat{p}) = \frac{p(1-p)}{n}.$$

95% 신뢰수준 하에서 허용오차를 B로 했을 때 다음 관계식으로부터 필요한 표본크기의 식(1)을 얻는다

$$|\hat{p} - p| \le z_{0.025}\sqrt{\frac{p(1-p)}{n}} = B(\text{허용오차}),$$

$$n = \frac{z^2_{0.025}\, p(1-p)}{B^2} \leq \frac{2^2}{4B^2} \qquad (1)$$

여기서 $z_{0.025} \doteq 2$.

<표 6> 당선자 득표율 예측에 필요한 표본크기

B	0.01	0.02	0.025	0.03	0.04	0.05
n	10,000	2,500	1,600	1,111	625	400

2. 두 후보에 대한 득표율 차의 추정 ($\hat{p}_1 - \hat{p}_2$)

두 후보의 득표율 차에 대한 추정량의 분산은 다음과 같이 된다.

$$Var(\hat{p}_1 - \hat{p}_2) = Var(\hat{p}_1) + Var(\hat{p}_2) - 2Cov(\hat{p}_1, \hat{p}_2)$$

$$= \frac{p_1(1-p_1)}{n} + \frac{p_2(1-p_2)}{n} + \frac{2p_1p_2}{n}$$

95% 신뢰수준 하에서 허용오차를 B로 했을 때 다음 관계식으로부터 필요한 표본크기의 식(2)을 얻는다

$$|(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)| \leq z_{0.025}\sqrt{\frac{1}{n}\{p_1(1-p_1) + p_2(1-p_2) + 2p_1p_2\}}$$

$$= B(\text{허용오차})$$

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

여기서 $z_{0.025} \doteq 2$

<표 7> 두 후보의 득표율 차를 예측하는데 필요한 표본크기

B	0.01	0.02	0.025	0.03	0.04	0.05
n	40,000	10,000	6,400	4,444	2,500	1,600

Ⅳ. 편향의 측정

선거예측조사의 관건은 당선자에 대한 예상 득표율과 실제 득표율과의 차이가 어느 정도가 되는가가 문제가 된다. 그러나 선거 환경이나 국민들의 의식에 따라 차이가 발생하게 된다. 편향이 생기는 요인은 여러 가지가 있을 수 있겠으나 우리 나라에서 선거예측조사를 실시한 결과를 놓고 볼 때 다음과 같은 이유로 편향이 발생한다고 볼 수 있다.

1. 무응답에 따른 편향

1996년 15대 총선과 1997년 15대 대선에서 실시한 전화조사에서 무응답률이 무려 50%(응답거부, 응답자 접촉 불능 포함; 박무익(1998), 박재수(1996), 21세기 방송연구소(1996))를 넘었고, 이번 16대 총선에서도 여전히 무응답률이 문제가 되었다. 특히 이번에 처음으로 실시된 출구조사에서도 정확한 수치는 얻지 못했으나 조사원들과 조사기관의 예측으로 볼 때 무응답률이 대략 30%선에 이른다니 선거예측 추정에 편향(bias)이 생기지 않을 수 없다.

이제까지 사용했던 전화조사의 단점을 보완하기 위해서 갈망했던 출구조사는 확률표본을 얻을 수 있고 이에 대한 통계적 방법의 적용이 다양하다는 점이다. 그런데 대면조사로 실시한 출구조사에서도 상당수의 무응답이 발생하므로 이점을 고려한 추정이 필요하게 된다.

1) 특정 후보에 대한 득표율 추정($\hat{p}$)

무응답편향을 측정하기 위해서 모집단 비율을 다음과 같이 나타낸다.

$$p = \lambda p_n + (1-\lambda) p_r$$

여기서, λ는 모집단을 무응답그룹과 응답그룹으로 나누었을 때의 무응답그룹의 비율이고 무응답그룹과 응답그룹에서 특정후보에 투표한 비율을 각각 p_n와 p_r이라 하자. 표본추출과 무응답으로 인한 오차이외에는 조사오차가 없는 경우를 가정하고, 응답자들만을 기반으로 한 p의 불편추정량을 $\hat{p}_r$이라 할 때,

$$\begin{aligned} Bias(\hat{p}_r) &= E(\hat{p}_r) - p \\ &= p_r - \lambda p_n - (1-\lambda) p_r \\ &= \lambda(p_r - p_n) \end{aligned}$$

가 된다. 무응답률 λ와 응답집단과 무응답집단에서 특정후보에 투표한 비율의 차에

따른 편향은 <표 8>에 있다. λ와 비율 차($p_r - p_n$)가 클수록 편향이 증가하는 것을 쉽게 알 수 있다. 두 후보의 예상득표율 차에 대한 경우도 유사한 결과를 얻는다.

<표 8> 무응답 발생에 따른 편향

λ \ $p_r - p_n$	0.05	0.10	0.15	0.20
0.3	0.015	0.030	0.045	0.060
0.5	0.025	0.050	0.075	0.100
0.7	0.035	0.070	0.105	0.140
0.9	0.045	0.090	0.135	0.180

2) 두 후보에 대한 득표율 차의 추정 ($\hat{p}_1 - \hat{p}_2$)

특정 후보에 대한 득표율 추정에서와 동일한 가정 하에서 편향은 다음과 같다.

$$Bias(\hat{p}_{1r} - \hat{p}_{2r}) = E(\hat{p}_{1r} - \hat{p}_{2r}) - (p_1 - p_2)$$
$$= \lambda\{(p_{1r} - p_{2r}) - (p_{1n} - p_{2n})\}$$

2. 거짓응답에 따른 편향

무응답과 함께 예측조사에서 오류발생의 큰 요인으로 거짓응답을 들 수 있다. 이것도 출구조사에서 문제점으로 지적되고 있는 사항 중의 하나로 응답자들이 솔직한 응답을 하지 않았을 것이라는 점이다. 그렇다면 응답자들의 거짓응답으로 인한 편향이 어느 정도인가를 살펴보는 것도 중요하다.

실제 투표소에서 특정후보(A 후보라 가정한다)에게 투표를 한 사람이 나와서 출구조사에 응할 때 거짓 응답의 경우를 살펴본다. A후보에 투표한 사람들의 비율을 p_A(실제 득표율)라 하면 단순확률표본인 n명의 응답자들 중 n'명이 A 후보에게 투표를 했다고 솔직하게 응답하면 A후보의 실제 득표율 p_A의 추정량 $\hat{p}_A$는 다음과 같이 된다.

$$\hat{\pi}_A = \frac{n'}{n}$$

이때 $\hat{p}_A$는 p_A의 최우추정량이고 추정량의 분산은 $p_A(1-p_A)/n$ 가 된다.

그러나 응답자들이 진실되게 응답하지 않을 경우, 류제복(1993)에서 가정한

바와 같이 응답확률변수 $R_i(i=1,2,\cdots,n)$를 i번째 응답자가 "A후보에게 투표를 했다" 라고 응답하면 1이고 "A후보에게 투표를 하지 않았다"라고 응답하면 0이라 정의하고 모든 응답자에 대하여 다음과 같은 가정을 한다.

① $\delta = P(R = 0 \mid A)$

② $\beta = P(R = 1 \mid \overline{A})$

③ 응답은 각 표본단위에 대하여 독립이다.

이러한 가정으로부터 응답자들이 A후보에게 투표를 했다고 응답할 확률은

$$\lambda = P(R = 1)$$

$$= p_A(1-\delta) + (1-p_A)\beta$$

가 된다. 그런데 가정③으로부터 $\sum_{i=1}^{n} R_i$ 는 $b(n, \lambda)$가 되므로 거짓응답으로부터 p_A의 최우추정량 $\hat{p}_A$를 얻는다.

$$\hat{p}_A = \frac{1}{n}\sum_{i=1}^{n} R_i$$

$\hat{p}_A$를 p_A의 추정량으로 사용하면 편향이 발생한다. 즉,

$$Bias(\hat{p}_A) = (1 - p_A)\beta - p_A\delta$$

만약 $\delta = \beta(=\theta)$라면 편향은 다음과 같이 간단하게 된다.

$$Bias(\hat{p}_A) = p_A + \theta(1-2p_A)$$

3. 무응답 교체에 따른 편향

금번 출구조사에서 응답자들이 응답을 거부하거나 또는 응답을 하지 않은 경우가 상당히 높은 것으로 나타났다. 이에 대해 조사기관에서는 응답거부자나 무응답자들을 다른 사람들로 교체(substitution)해서 응답을 얻는 방법을 사용하였다. 그러나 이러한 교체방법을 사용할 경우에도 편향이 발생하게 된다. 가장 간단한 경우의 예를 들어보면, 모집단을 응답집단과 교체집단으로 나누고 교체집단의 비율을 4.1절에서와 같이 λ로 두고 교체집단으로부터 얻은 특정후보의 지지율을 p_s라 두면 특정후보를 지지하는 비율은 다음 같이된다.

$$p = \lambda p_s + (1-\lambda)p_r$$

표본추출과 무응답으로 인한 오차이외에는 조사오차가 없는 경우를 가정하고, 응답

자들과 교체응답자들을 기반으로 한 p의 불편추정량을 $\hat{p}_{rs}$이라 할 때,

$$Bias(\hat{p}_{rs}) = E(\hat{p}_{rs}) - p$$

$$= \lambda(p_s - p_n)$$

이 된다. 따라서 무응답자를 교체하는 경우의 편향을 줄이기 위해서는 4.1.절에서와 같이 무응률을 줄이는 것과 교체표본들의 특정후보를 지지하는 비율이 무응답자들의 지지비율과 같도록 교체하는 것이 바람직하다. 예를 들어, 50대 이후의 여성들이 무응답 또는 응답거부자일 때는 이들과 성향이 비슷한 표본을 교체하는 것이 편향을 줄이는 방법이 될 것이다.

V. 결론 및 토의

법적으로 우리 나라에서 선거예측조사가 1992년부터 가능하였으므로 채 10년도 되지 않은 시점에서 완벽한 조사결과를 기대한다는 것은 지나친 욕심이지만 어쨌든 선거예측조사가 실시되고 있는 시점에서 보다 정확한 조사결과를 내놓는 것이 조사기관뿐만 아니라 이 분야에 종사하는 모든 사람들의 책임이라 할 수 있다.

지금까지의 선거예측조사에 대한 내용을 검토해 볼 때 향후 선거예측조사의 신뢰성을 높이기 위해서는 다음과 같은 몇 가지 사항에 대한 개선 및 연구가 필수적이다.

<법적인 문제>

지난 15대 총선과 대선에서는 법적으로 선거일에 투표소로부터 500m이내에서의 조사가 금지되어 전화조사만으로 선거예측조사를 하였다. 이러한 법적 제한이 2000년 2월 16일에 개정되어 종전의 500m이내 제한이 300m 이내로 완화되기는 했지만 이러한 법적 제한 속에서는 실질적인 출구조사가 불가능하다. 이번의 출구조사에서 일부 지역은 법적 제지로 출구조사시 문제가 발생하기도 했다. 따라서 제대로 된 출구조사로 신뢰성 있는 조사결과를 얻기 위해서는 이러한 법적 제한은 철폐해야 한다. 아울러 선거기간동안 조사결과의 발표금지 조항도 너무 지나친 항목으로 이도 전반적으로 삭제 또는 수정되어야 한다. 왜냐하면 법적으로 금지하고 있다는 것 자체가 일반 국민들에게 해가 된다는 인상을 줄 수 있기 때문이다.

<조사방법 상의 문제>

1. 표본크기의 선정 : 통상적으로 선거예측조사에서 표본크기는 한 특정후보의 득표율을 추정량의 분산식을 이용한 식(1)로 구하고 있다. 그러나 앞에서 지적하고 있는

바와 같이 경합지역 또는 혼선지역의 경우는 두 후보간의 차의 비율 공식을 이용한 식(2)를 사용하는 것이 바람직하다. 따라서 항상 동일한 식을 사용하기보다는 필요에 따라 적절한 공식을 사용하여 표본크기를 결정하는 것이 바람직하다.

2. 무응답이나 거짓응답을 줄이기 위해서 응답자들의 신분을 보장해주는 방법(확률화응답기법, 등)들을 적용하고 무응답률과 거짓응답률에 대한 추정의 문제도 고려하여야 한다. 또한 4.3절에서 언급한 바와 같이 무응자 교체시 편향을 줄이기 위해서는 무응답층에 대한 정확한 분석과 무응답층과 유사한 사람으로 교체하여야 한다. 그리고 응답자 교체가 여의치 않을 경우 대체(imputation)추정방법을 사용하도록 한다.

3. 예측치와 실제치간의 오류발생이 지역적 특성에 상당히 많은 영향을 받고 있으므로 향후 선거예측조사에서는 이런 지역적 특성에 따른 조사방법의 개발과 가중치조정문제에 많은 심혈을 기울여야 할 것이다.

<기타 문제>

1. 조사원 채용시 응답자들의 여건과 비슷한 사람들을 조사원으로 채용하여 응답자들의 부담을 덜어주어 무응답과 응답거절 및 거짓응답의 가능성을 줄인다.

2. 한국의 정치 사회적 여건상 50대 이후의 연령층들이 출구조사에 대한 이해부족과 응답자들의 신분 보장이 안 된다는 생각에 조사에 응답을 하지 않거나 거부하며, 심지어 거짓 응답의 가능성이 있다. 그러므로 조사 전에 출구조사에 대한 이해와 응답자들의 신분보장에 대한 충분한 설명 등이 있어야 한다.

3. 언론보도에서도 마치 점쟁이가 점을 치는 것과 같은 형태로, 예상했던 후보가 실제로 얼마나 당선되었는 가로 예측조사의 신뢰성를 평가할 것이 아니라 좀더 과학적이고 통계적인 방법으로 이를 평가하고 아울러 국민들에 선거예측결과에 대한 올바른 판단을 할 수 있도록 유도하는 것이 공영방송으로서의 의무라 본다. 그래야만 선거예측조사의 미천한 역사 속에서도 한시 빨리 정확하고 신뢰성 있는 여론조사의 문화가 정착될 것이다.

이상과 같은 점을 수정하고 보완하여 조사를 설계한다면 향후 선거예측조사의 신뢰성을 증진시킬 수 있을 것으로 확신한다.

<참고문헌>

류제복, 홍기학, 이기성 공저. 1993. ≪확률화응답모형≫ 자유아카데미

이시윤편. 1993. ≪1993년도판 판례소법전≫ 청림출판사

이택규외 편저. 2000. ≪2000년 대한민국 신헌법≫ 법률출판사

21세기 방송연구소. 1996. "투표자조사, 어떻게 할 것인가." ≪제9회 토론회 결과보고서≫

노규형. 1998. "선거예측조사의 이론과 실제." ≪1998년 한국통계학회 춘계학술발표회 논문집≫ 10-16

동아일보. 2000년 4월 14일, 15일자 신문

류제복. 1993. "대체 확률화응답기법." ≪응용통계연구≫ 6(2): 311-318

박무익. 1998. "한국의 제15대 대통령선거와 선거예측조사." ≪1998년 한국통계학회 춘계학술발표회논문집≫ 1-9

박재수. 1996. "선거예상조사에 대한 비판(15대 총선투표자 전화조사를 중심으로)." ≪통계≫ 22(2): 2-17

조선일보. 2000년 4월 14일, 15일자 신문

중앙일보. 2000년 4월 14일, 15일자 신문

MBC, KBS-SBS, 2000년 4월 13일, 선거예측조사 T·V방송보도자료

Backstrom, C. H. and Hursh-César. 1981. *Survey Research*(2nd. ed.): John Wiley & Son, Inc.

Lessler, J. T. and Kalsbeek, W. D. 1992. *Nonsampling Error in Surveys*: John Wiley & Son, Inc.

Scheaffer, R. L., Mendallhall, W. III, and Ott, R. L. 1996. *Elementary Survey Sampling*(5th. ed.): Duxbury Press

http:\\www.nec.go.kr

http:\\www.scort.go.kr