연구노트

RDD 표본 대 전화번호부 표본: 2007년 대통령 선거 예측 사례

RDD Sample versus Directory - Based Sample for Telephone Surveys: The Case of 2007 Presidential Election Forecasting in Korea

허명회* · 김영원**
Myung - Hoe Huh. Young - Won Kim

이제까지 우리나라에서 전화조사를 위한 표본목록은 거의 대부분 전화번호부로부터 나왔다. 그러나 전화번호부의 모집단 포함률이 너무 떨어진다는 지적이 있어 대응수단으로 국제적 기준인 RDD(random digit dialing, 임의번호걸기)가 구현된 바 있다. 2007년 12월의 17대 대통령 선거에 대한 예측을 위해 투표일을 5~6일 앞서 실시된 KBS · MBC 전화조사는 표본을 반씩 나누어 절반은 RDD로, 나머지 절반은 전화번호부에서 응답자 표본목록을 추출하였다. 이 사례연구는 KBS · MBC 전화조사의 RDD 표본과 전화번호부 표본을 대비시켜 공통점과 상이점을 살펴본 것이다. 향후 수년 동안 전화번호부 표본과 RDD 표본이 공존할 것으로 예상되는 상황에서, 이 연구결과가 두 방식의 비교에 시사점을 제시할 것으로 기대한다.

주제어: 선거예측, 전화조사, 전화번호부, 임의번호걸기(RDD), 다항 로짓 모형

In most telephone surveys in Korea, telephone numbers are selected from the directories. Inevitably, such samples may lack representativeness due to poor coverage rate. To resolve the problem, Kang et al.(2008) implemented RDD(random digit dialing) method for nationwide sampling in Korea. The aim of this study is to compare an RDD sample with a traditional telephone quota sample that were collected independently by two survey institutes commissioned

* 교신저자(corresponding author): 고려대학교 통계학화 교수 허명회.
E - mail: stat420@korea.ac.kr

** 숙명여자대학교 수학통계학부 교수.

by the KBS-MBC consortium for the 2007 Presidential Election of Korea.

Key words : election prediction, telephone survey, telephone directory, RDD(random digit dialing), multinomial logit model

Ⅰ. 2007년 대통령 선거 예측을 위한 전화조사 사례

전화조사는 1980년대 후반 이래 우리나라에서 사회여론조사의 주된 방법론으로 활용되고 있다. 신속성이 가장 큰 장점이고 비용 측면에서도 경제적인 편이지만 수년 전부터 비판이 끊이지 않고 있다(허명회 2007). 조사 방법론적 측면에서 우리나라 전화조사의 취약점은 다음 두 가지이다. 첫째는 지역·성·연령대 할당 표본추출(quota sampling)을 하고 있는데 그럼으로써 각종 편향이 개입할 수 있다는 점이다(조성겸 1997; 허명회·황진모 2006). 둘째는 전화번호부 기반 표본추출(directory-based sampling)을 함으로써 전화번호 비등재자가 조사대상에서 원초적으로 제외된다는 점이다. 이 사례연구는 후자에 초점을 맞추어 전화번호부 기반 표본과 이에 대한 대안인 RDD(random digit dialing, 임의번호걸기) 표본이 어느 정도로 어떻게 다른가를 실제 사례를 통하여 살펴볼 것이다.

2007년 12월 19일의 17대 대통령 선거를 위한 예측을 위해 KBS와 MBC는 공동으로 투표일 직전 일주일에 두 차례의 전화조사를 하였는데, 그 중 1차 조사(12월 13일, 14일)에서는 다음과 같이 서로 다른 방식으로 독립적인 표본이 추출되었다.

- RDD 표본: 지역·성·연령대 할당추출을 하되 RDD 방식으로 전화번호를 추출하였다. RDD로 추출한 전화번호의 등재여부는 사후에 전화번호부 데이터베이스에서 확인하였다. 1,522명의 응답자 중 등재번호 응답자가 742명(48.8%)이었고 비등재번호 응답자가 780명(51.2%)이었다.
- 전화번호부 표본: 지역·성·연령대 할당추출을 하되 전화번호부 데이터베이스에서 전화번호를 추출하였다. 1,503명이 응답하였다.

본 사례의 RDD 표본에는 전국의 4,615개 '지역번호+국번'의 모든 4자리 가입자 번호 중 업종·상호 편에 수록된 번호를 제외한 나머지 41,654,678개 번호에서 표본번호를 임의로 추출하는 방식이 적용되었고(강현철 외 2008), 전화번호부 표본(이하 '번호부' 표본으로 약칭함)에는 전화번호부 데이터베이스에서 표본번호를 선택하는 전화번호부 기반 추출방식이 적용되었다. 2007년 12월 현재 대다수의 조사회사들이 전화번호부로부터 표본번호를 추출하고 있는데, 전화번호부의 모집단 포함률(population coverage)이 60% 이하로 추정되므로 이런 방식으로 추출된 표본의 경우 대표성 확보에 상당한 문제가 있을 수 있다(강현철 외 2008).

RDD 표본과 번호부 표본은 각기 다른 조사회사에 의해 수집되었다. 따라서 논리적으로는 두 표본의 결과적 차이가 추출방식에 의한 차이에 기인하는지 아니면 조사기관 효과(house effect)에 의한 것인지가 구분되지 않는다. 그러나 두 조사가 KBS·MBC 선거방송 팀에 의해 최대한 동일하도록 기획되었으므로 조사기관 간 차이는 없다고 가정하기로 한다.

〈표 1〉은 KBS·MBC 선거예측 조사에서 나온 지지후보 분포를 표본별로 보여 준다. 이명박 후보와 정동영 후보간 지지율 차이는 RDD 표본에서 27.0%p이고 번호부 표본에서는 27.2%p로 나타나 핵심 사항에서 표본 간 차이는 거의 없는 것으로 보인다.[1] RDD 표본과 번호부 표본 간 후보별 지지율 차이를 카이제곱 검증했으나 유의하지 않았다. 다만, 상대적으로 RDD 표본에서 무응답 비율이 높았다.

〈표 2〉는 RDD 표본에서 응답자를 등재번호 응답자와 비등재번호 응답자로 구분하여 지지후보 분포를 본 것이다. 등재번호 그룹과 비등재번호 그룹 간 차이를 카이제곱 검증한 결과 유의하지 않았다. 그러나 비등재번호 응답자가 등재번호 응답자에 비해 무응답 비율이 낮았다.

---

1) 실제로는 지역·성·연령대 가중치를 적용하여 A 표본에서 정동영 후보와 이명박 후보에 대한 지지율이 각각 15.0%와 42.2%로 조정되었고 B 표본에서는 각각 17.1%와 43.9%로 조정되었다. 이에 따라 A 표본과 B 표본에서 두 후보 간 차이는 각각 27.2%p와 26.8%p가 되었다.

〈표 1〉 RDD 표본과 번호부 표본에서의 지지후보 분포

	정동영	이명박	이회창	문국현	기타	무응답	합계
RDD 표본: 1,522명	15.2%	42.2%	12.6%	7.2%	4.7%	18.1%	100.0%
번호부 표본: 1,503명	17.2%	44.4%	12.6%	6.0%	4.7%	15.2%	100.0%
합계 : 3,025명	16.2%	43.3%	12.6%	6.6%	4.7%	16.7%	100.0%

* 피어슨 카이제곱 8.44, 자유도 5, p-값 0.134.
* 기타후보: 권영길, 이인제, 허경영, 정근모, 전관, 금민.

〈표 2〉 RDD 표본의 등재번호 응답자와 비등재번호 응답자의 지지후보 분포

	정동영	이명박	이회창	문국현	기타	무응답	합계
등재 : 742명	14.7%	42.5%	12.3%	6.9%	3.5%	20.2%	100.0%
비등재: 780명	15.6%	41.9%	12.9%	7.6%	5.8%	16.3%	100.0%
합계 : 1,522명	15.2%	42.2%	12.6%	7.2%	4.7%	18.1%	100.0%

* 피어슨 카이제곱 8.27, 자유도 5, p-값 0.141.

한 사례에서 RDD 표본과 번호부 표본 간 그리고, RDD 표본 내에서 등재번호 그룹과 비등재번호 그룹 간 특정 사안(후보별 지지율)에 대해 통계적으로 유의한 반응 차이가 발견되지 않았다고 해서 두 표본추출방법에 의해 얻어지는 표본이 별 차이가 없다고는 당연히 말할 수 없다. 두 표본 간에 인구사회적 특성에서 차이가 있고 각 인구사회적 그룹에서 고유한 반응 차이가 있는데도 불구하고 얼마든지 표본 간 반응차이가 나타나지 않을 수 있기 때문이다. 이 연구에서 다루는 사례가 바로 그런 경우이다.

Ⅱ. RDD 표본과 전화번호부 표본의 인구사회적 특성 비교

이 절에서 다음 두 질문에 답하고자 한다: 1) RDD 표본과 번호부 표본은 어떻게 다른가? 2) RDD 표본에서 등재번호 응답자 그룹과 비등재번호 응답자 그룹은 어떻게 다른가? 결론부터 말하면, RDD 표본은 번호부 표본에 비해 자영업이 많은 반면 기타무직이 적고, 또한 100만원 이하의 저소득자가 적은 반면 무응답자가 많다. 그리고 RDD 표본에서 비등재번호 응답자는 등재번호 응답자에

〈표 3〉 RDD 표본의 지역 · 성 · 연령대별 비등재번호 응답자 비율

	빈도	비등재율		빈도	비등재율
서울	328	62%	남성	746	52%
부산	114	54%	여성	776	51%
대구	80	46%	전체	1522	51.5%
인천	78	60%			
광주	43	54%	20대-	310	56%
대전	45	67%	30대	358	60%
울산	31	45%	40대	347	58%
경기	333	57%	50대	234	44%
강원	45	33%	60대+	273	32%
충북	45	51%	전체	1522	51.2%
충남	63	30%			
전북	58	43%			
전남	60	37%			
경북	87	35%			
경남	96	38%			
제주	16	25%			
전체	1522	51.2%			

비해 화이트칼라와 자영업이 많은 반면 농림어업과 기타무직이 적고 소득이 다소 높다는 것이다.

우선 RDD 표본의 인구사회적 특성을 살펴보기로 한다. 〈표 3〉은 RDD 표본의 지역 · 성 · 연령대별 비등재번호 응답자 비율인데 비등재 비율이 전국적으로 51.2%였다. 지역적으로는 대전, 서울, 인천 등에서 60%대로 높았고 제주의 경우 25%, 충남, 강원, 경북 등에서 30%대로 낮았다. 즉, 대도시(대구 · 울산 제외) 지역에서는 비등재번호 응답자가 등재번호 응답자보다 많았고 대도시가 아닌 지역에서는 등재번호 응답자가 비등재번호 응답자보다 많았다. 성별로는 차이가 없었으나 연령대별로는 30~40대에서 비등재 비율이 가장 높았고 60대 이상에서 가장 낮았다. 이로부터 대도시와 30~40대에서 비등재 성향이 큼을 짐작할 수 있다.

이제 RDD 표본의 사회적 특성을 보기로 하자. <표 4>는 RDD 표본의 등재번호 그룹과 비등재번호 그룹의 교육수준·직업·소득 분포인데, 두 그룹 간 차이가 교육수준과 직업, 소득 모두에서 통계적으로 유의하였다. 교육수준 측면에서 볼 때, 비등재자 그룹은 상대적으로 고학력자가 많았고 저학력자가 상대적으로 적었다. 직업 측면에서는 비등재자 그룹은 상대적으로 화이트칼라와 자영업이 많았고 농림어업과 기타무직이 적었다. 또한 비등재자 그룹은 상대적으로 고소득자가 많고 저소득자와 소득무응답이 적었다.

RDD 표본에서 등재자 그룹과 비등재자 그룹 간 교육수준, 직업, 소득 등 사회적 변수의 분포 차이가 발견되지만 이것이 전화번호 등재 여부가 갖는 고유한 효과인지 아니면 지역·성·연령대의 등재자 비율 차이에 기인하는지를 따져 볼 필요가 있다. <표 5>는 교육수준·직업·소득에 대한 다항 로짓 모형(multinomial logit model)의 요약을 보여준다.[2] 이에 따르면 교육수준에 대하여 등재번호 여부가 고유한 영향을 주지 않는다. 즉 등재자 그룹과 비등재자 그룹 간 교육수준에서 차이가 있지만 이것은 지역·성·연령대의 등재자 비율 차이로 충분히 설명된다. 그러나 직업에 대하여는 등재여부가 고유한 영향을 가지며 그 영향은 앞서 포착한 바와 같이 비등재자의 확보에 따라 화이트칼라와 자영업이 많아지고 농림어업과 기타무직이 적어지는 결과를 초래하는 것으로 나타난다. 소득에 대하여도 등재여부가 고유한 영향을 미친다(단측 p-값 = 0.034). 따라서 RDD 표본추출로 비등재자가 표본에 포함됨에 따라 고소득자가 늘어나고 소득무응답이 줄어든다.

---

2) 다항 로짓 모형은 다변량 속성 $x=(x_1,\cdots,x_p)$의 개체가 범주 $j\,(=1,\cdots,J)$에 반응할 확률 $\pi_j(x)$이 다음과 같다고 가정한다 (SPSS 2007).

$$\pi_j(x) = \frac{\exp(\beta_j' x)}{1+\sum_{j=1}^{J-1}\exp(\beta_j' x)},\quad j=1,\cdots,J-1.$$

여기서 $\beta_j=(\beta_{j1},\cdots,\beta_{jp})'$이고 J는 기준범주(reference category)이다. 이에 따라

$$\log\frac{\pi_j(x)}{\pi_J(x)} = \beta_j' x,\quad j=1,\cdots,J-1.$$

로 표현된다. 로짓 모형은 $J=2$인 다항 로짓 모형이다.

〈표 4〉 RDD 표본의 등재자 그룹과 비등재자 그룹의 교육수준 · 직업 · 소득 분포

교육수준	빈도	중졸-	고졸	대재+				
- 등재자	742	27.5%	31.3%	41.2%				
- 비등재자	780	14.0%	32.7%	53.3%				
직업	빈도	화이트	블루	자영업	농업	주부	학생	무직
- 등재자	742	12.5%	7.3%	16.4%	8.0%	36.5%	8.5%	10.8%
- 비등재자	780	18.3%	7.8%	19.4%	0.4%	37.8%	9.4%	6.9%
소득	빈도	100-	100+	200+	300+	400+	500+	무응답
- 등재자	742	19.9%	18.6%	19.1%	10.5%	7.7%	8.0%	16.2%
- 비등재자	780	7.3%	16.2%	23.1%	16.4%	10.1%	12.1%	14.9%

* 교육수준 무응답자 31명(2.1%)는 '중졸-'로 자료값 대체를 하였음.
* 교육수준: 피어슨 카이제곱 45.8, 자유도 2, p-값 0.000
  직업: 피어슨 카이제곱 70.6, 자유도 6, p-값 0.000
  소득: 피어슨 카이제곱 68.3, 자유도 6, p-값 0.000

〈표 5〉 RDD 표본에서 교육수준 · 직업 · 소득에 대한 다항 로짓 모형의 요약

종속변수	요인	카이제곱	자유도	p-값
교육수준	지역	87.4	30	0.000
	성	32.8	2	0.000
	연령대	591.9	8	0.000
	등재여부	1.0	2	0.618

종속변수	요인	카이제곱	자유도	p-값
직업	지역	164.9	90	0.000
	성	1180.0	6	0.000
	연령대	851.7	24	0.000
	등재여부	31.5	6	0.000

종속변수	요인	카이제곱	자유도	p-값
소득	지역	142.5	90	0.000
	성	7.5	6	0.270
	연령대	495.4	24	0.000
	등재여부	11.8	6	0.068

이렇듯 등재자와 비등재자 간 차이가 있으므로, RDD 표본과 번호부 표본 간에도 동일한 패턴의 차이가 생길 것으로 예상할 수 있다. 그러나 두 표본이 할당추출로 얻어졌으므로 실제 상황은 그렇게 간단하지 않다. 〈표 6〉는 RDD 표본과 번호부 표본을 교육수준 · 직업 · 소득 측면에서 비교한 것인데, 예상대로 교육수준에서 RDD 표본과 번호부 표본이 통계적으로 유의한 차이가 없지만, 직업과 소득에서는 통계적으로 유의한 차이가 발견되었다. 직업 분포에서 RDD 표본이 번호부 표본에 비하여 자영업이 많고 기타무직이 적다는 것은 예상과 일치하였으나 화이트칼라는 오히려 적고 가정주부가 많다는 사실은 예상과 불일치하였다. 소득 측면에서는 RDD 표본이 번호부 표본에 비하여 100만원 이하의 저소득자가 적은 반면 무응답자가 많았다.[3)]

〈표 6〉 RDD 표본과 번호부 표본 간 교육수준 · 직업 · 소득 분포의 비교

교육수준	빈도	중졸-	고졸	대재+				
- RDD 표본	1522	20.6%	32.0%	47.4%				
- 번호부 표본	1503	21.4%	31.3%	47.4%				
직업	빈도	화이트	블루	자영업	농업	주부	학생	무직
- RDD 표본	1522	15.5%	7.6%	17.9%	4.1%	37.2%	8.9%	8.8%
- 번호부 표본	1503	17.6%	8.6%	14.2%	4.9%	34.3%	8.6%	11.8%
소득	빈도	100-	100+	200+	300+	400+	500+	무응답
- RDD 표본	1522	13.5%	17.3%	21.2%	13.5%	8.9%	10.1%	15.5%
- 번호부 표본	1503	17.2%	17.3%	21.0%	14.2%	8.8%	11.2%	10.4%

* RDD 표본의 교육수준 무응답자 31명(2.1%)과 번호부 표본의 무응답자 17명(1.1%)는 '중졸-'로 자료값 대체를 하였음.

* 교육수준: 카이제곱 0.38, 자유도 2, p-값 0.829.
직업: 카이제곱 19.4, 자유도 6, p-값 0.004.
소득: 카이제곱 23.3, 자유도 6, p-값 0.001.

3) 직업에서 RDD 표본이 번호부 표본에 비하여 화이트칼라가 적고 가정주부가 많다는 사실과 소득무응답이 많다는 사실 등 일부 예상과 다른 점들은 조사기관 효과(house effect)가 작용하였기 때문일 수도 있다.

〈표 7〉 전체 표본(=RDD+번호부)에서 지역·성·연령대 별 후보지지율(%)

		정동영	이명박	이회창	문국현	기타	무응답	합계
지역	서울	12.2	51.6	10.9	7.2	2.8	15.3	100.0
	부산	9.3	46.0	13.7	5.8	6.2	19.0	100.0
	대구	9.9	55.6	15.9	4.6	2.0	11.9	100.0
	인천	13.3	47.5	10.1	7.0	6.3	15.8	100.0
	광주	41.7	19.0	6.0	13.1	4.8	15.5	100.0
	대전	11.6	37.2	22.1	10.5	3.5	15.1	100.0
	울산	6.3	50.0	17.2	3.1	6.3	17.2	100.0
	경기	16.9	44.0	11.8	7.5	5.1	14.7	100.0
	강원	13.0	41.3	18.5	2.2	4.3	20.7	100.0
	충북	12.9	33.3	25.8	4.3	5.4	18.3	100.0
	충남	10.6	33.3	22.8	5.7	8.1	19.5	100.0
	전북	49.6	9.6	5.2	6.1	4.3	25.2	100.0
	전남	51.2	8.1	4.1	8.1	14.6	13.8	100.0
	경북	4.2	61.9	12.5	4.8	0.6	16.1	100.0
	경남	10.4	43.8	14.6	7.3	5.2	18.8	100.0
	제주	11.8	32.4	14.7	5.9	5.9	29.4	100.0
성	남자	17.3	43.8	14.0	6.7	5.3	12.8	100.0
	여자	14.8	42.3	11.8	6.8	4.2	20.1	100.0
연령대	20대-	12.8	34.5	18.5	14.0	5.0	15.1	100.0
	30대	17.3	40.3	11.8	8.9	6.6	15.0	100.0
	40대	19.4	44.1	11.5	5.2	4.9	14.9	100.0
	50대	12.7	50.6	11.8	2.8	4.0	18.1	100.0
	60대+	16.8	48.8	10.8	0.9	2.4	20.3	100.0
전체	-	16.1	43.0	12.9	6.7	4.8	16.5	100.0

* 지지율 산출시 각 표본에 독자적인 지역·성·연령대 가중치를 적용하였다.

Ⅲ. 지지후보 선택모형

대통령 선거에서 후보 지지율은 여러 인구사회적 변수들과 연관되어 있다. 〈표 7〉과 〈표 8〉에서 보듯, 지역·성·연령대·교육수준·직업·소득에 따라 후보지지율에서 크고 작은 차이가 있다. 예컨대 이명박 후보의 경우 가장 넓은 편차(=최대값-최소값)를 보인 요인은 지역으로 53.8%p(최대: 경북 61.9%, 최소: 전남 8.1%)이고, 그 다음으로 직업 18.7%p, 연령대 15.9%p, 소득 15.4%p, 교육수준 7.1%p, 성 1.5%p의 순서이다. 그러나 이와 같은 자료기술적 분석에는 여러 요인들의 효과가 중첩되어 있어 각 요인의 중요도를

〈표 8〉 **전체 표본(=RDD+번호부)에서 교육수준·직업·소득별 후보지지율(%)**

		정동영	이명박	이회창	문국현	기타	무응답	합계
교육	중졸-	19.3	38.3	9.9	2.0	3.8	26.6	100.0
	고졸	16.7	45.4	13.8	4.8	5.1	14.2	100.0
	대재+	14.3	43.4	13.6	9.9	4.9	13.8	100.0
직업	화이트	16.6	44.1	12.1	8.5	4.9	13.8	100.0
	블루	16.3	36.6	21.4	3.5	8.9	13.2	100.0
	자영업	20.0	46.0	12.0	6.3	2.9	12.8	100.0
	농업	25.2	32.1	13.0	1.5	8.4	19.8	100.0
	주부	14.9	43.9	10.9	5.8	3.7	20.9	100.0
	학생	11.0	35.8	17.7	16.7	6.4	12.4	100.0
	무직	14.2	50.8	10.9	3.0	4.0	17.2	100.0
소득	100-	18.8	39.7	10.4	1.1	4.5	25.4	100.0
	100+	18.0	42.5	13.9	5.1	5.7	14.9	100.0
	200+	16.2	45.1	12.7	8.6	5.0	12.4	100.0
	300+	17.1	45.9	12.7	8.9	4.1	11.3	100.0
	400+	15.3	47.4	13.1	9.7	4.9	9.7	100.0
	500+	11.4	49.1	12.3	9.8	3.8	13.6	100.0
	무응답	13.4	33.7	14.9	5.6	4.9	27.4	100.0
전체		16.1	43.0	12.9	6.8	4.8	16.5	100.0

* 지지율 산출시 각 표본에 독자적인 지역·성·연령대 가중치를 적용하였다.

제대로 파악하기 어렵다. 대안으로서 지지후보를 종속변수로, 성·연령대·교육수준·직업·소득·등재여부 또는 표본구분(RDD vs 번호부) 지시변수를 설명변수로 하는 다항 로짓 모형을 생각하기로 한다.

〈표 9〉는 RDD 표본에서 지지후보 선택모형(다항 로짓 모형)의 요약을 등재여부 지시변수가 포함된 경우와 포함되지 않은 경우로 나누어 보여준다. 등재여부가 개인의 사생활과 개인정보 보호 등에 대한 심리적 성향을 나타내고 이러한 심리적 성향이 지지후보 선택에 영향을 줄 가능성이 있다. 하지만 등재여부 지시변수가 통계적으로 유의하지 않은 것으로 나타났는데, 이는 지역·성·연령대·교육수준·직업·소득이 설명변수로 모형에 포함되는 경우 등재번호 여부가 지지후보 선택에 직접적인 영향을 주지는 않음을 뜻한다. 즉, 등재여부가 모든 설명요인들과 연관되어 있지만 지지후보 선택에 고유한 영향을 미치는 것은 아니다. 이에 따라, RDD로 추출된 RDD 표본 응답자와 전화번호부로부터 추출된 번호부 표본 응답자도 지지후보 선택에서는 구분되지 않은 성향을 가질 것으로 기대된다.

〈표 10〉은 전체 표본(=RDD+번호부)에서 지지후보 선택모형(다항 로짓 모형)의 요약을 표본구분(RDD vs 번호부) 지시변수가 포함된 경우와 포함되지 않은 경우로 나누어 보여준다. 전자의 모형에서 표본구분(RDD vs 번호부) 지시변수는 통계적으로 유의하지 않은 것으로 나타났다. 이는 지역·성·연령대·교육수준·직업·소득이 설명변수로 모형에 포함되는 경우 RDD 표본 응답자와 번호부 표본 응답자가 지지후보 선택에서 비차별적인 성향을 보인다는 것을 의미한다.

〈표 9〉 RDD 표본에서 지지후보 선택모형 요약

종속변수	지지후보			종속변수	지지후보		
요인	카이제곱	자유도	p-값	요인	카이제곱	자유도	p-값
지역	268.3	75	0.000	지역	267.1	75	0.000
성	1.1	5	0.951	성	1.1	5	0.951
연령대	85.9	20	0.000	연령대	87.9	20	0.000
교육수준	21.5	10	0.018	교육수준	21.7	10	0.017
직업	41.2	30	0.084	직업	40.2	30	0.102
소득	86.3	30	0.000	소득	85.9	30	0.000
등재여부	7.7	5	0.173				

〈표 10〉 전체 표본(=RDD+번호부)에서 지지후보 선택모형 요약

종속변수	지지후보			종속변수	지지후보		
요인	카이제곱	자유도	p-값	요인	카이제곱	자유도	p-값
지역	452.6	75	0.000	지역	452.6	75	0.000
성	6.0	5	0.309	성	5.3	5	0.380
연령대	129.0	20	0.000	연령대	131.2	20	0.000
교육수준	31.1	10	0.001	교육수준	31.4	10	0.001
직업	46.4	30	0.028	직업	46.4	30	0.028
소득	81.8	30	0.000	소득	83.2	30	0.000
표본추출방식	5.9	5	0.319				

〈표 9〉의 우측 모형과 〈표 10〉의 우측 모형에서 각 요인의 카이제곱을 자유도로 나눈 값이 요인의 중요도를 나타낸다고 볼 때, 〈표 9〉의 RDD 표본에서 중요도 순서는 연령대, 지역, 소득, 교육수준, 직업, 성의 순이고, 〈표 10〉의 전체표본에서는 중요도 순서가 연령대, 지역, 교육수준, 소득, 직업 그리고, 성의 순이다. 2007년 12월 대통령 선거에서 지지후보 선택에 연령대와 지역이 두드러진 영향을 주었으나 이외에 교육수준, 소득, 직업도 의미 있는 영향을 주었음을 알 수 있다.

Ⅳ. 맺음 말

이제까지의 분석 결과를 묶고자 하면 다음 두 질문에 부딪힌다: 1) 2절에서 RDD 표본과 번호부 표본이 다른 교육수준, 직업, 소득 분포를 갖는다고 했고, 3절에서는 지역, 연령대, 교육수준, 직업, 소득이 지지후보 선택에 유의한 영향을 주는 요인이라고 했다. 그런데 1절에서는 RDD 표본과 번호부 표본의 지지후보 분포가 다르지 않다고 했다. 어떻게 그럴 수 있는가? 2) 2절에서 지역, 연령대에 따라 비등재율이 다르고 전화번호 등재자 그룹과 비등재자 그룹이 다른 교육수준, 직업, 소득 분포를 보인다고 했으며, 3절에서는 지역, 연령대, 교육수준, 소득이 지지후보 선택에 유의한 영향을 주는 요인이라고 했다. 그런데 1절에서는 등재자 그룹과 비등재자 그룹의 지지후보 분포가 다르지 않다고 했다. 어떻게 그럴 수 있는가?

그 직접적인 이유를 명확히 밝히기는 어려울 것으로 보인다. 지역, 성, 연령대, 교육수준, 직업, 소득의 조합 칸이 모두 23,520개(=16×2×5×3×7×7)이고 각 후보의 전국 지지율이란 각 칸의 크기와 지지율을 곱하여 모두 더한 것인데, 그 과정에서 실제 수많은 개별 효과들이 우연히 상쇄되는 방향으로 합해지는 경우가 있을 것이다. 이 연구가 다룬 사례가 바로 그런 경우로 생각된다. 즉, 풀어보면 다음과 같다.

1) RDD 표본이 번호부 표본에 비해 자영업이 많고 기타무직이 적은데 이명박 지지율은 자영업과 기타무직에서 모두 높다. 또한 RDD 표본이 번호부 표본에 비해 100만원 이하 소득이 적고 소득무응답이 많은데 이명박 지지율은 이 두 범주에서 모두 낮다. 지역, 성, 연령대 그리고, 교육수준의 측면에서는 RDD 표본과 번호부 표본이 별 차이가 없다. 따라서 두 표본에서 이명박 지지율은 거의 같게 된 것이다.
2) RDD 표본의 등재자 그룹과 비등재자 그룹의 이명박 지지율이 비슷하게 된 이유는 다음과 같다. 이명박 지지율은 대구, 울산 그리고, 경북과 50대와 60대 이상에서 높은데 이들 범주에서는 등재자가 많다. 그러나 100만원 이하 소득과 소득무응답 범주에 등재자가 많으며 이들 범주에서 이

명박 지지율이 낮고, 또한 이명박 지지율이 낮은 전북과 전남에서 등재자가 많다. 그리고 등재자가 많은 무직 그룹에서는 이명박 지지율이 낮고, 등재자가 적은 자영업 그룹에서는 이명박 지지율이 높다. 따라서 등재자 그룹과 비등재자 그룹의 이명박 지지율에 차이가 없어진다.

그러면 앞으로도 RDD 조사와 전화번호부 조사가 유사한 결과를 낼 것으로 기대할 수 있을까? 항상 그렇지는 않을 것이다. 요인조합 칸들의 크기와 지지율의 곱이 상쇄 방향이 아니라 배증 방향으로 합해지는 경우가 있을 것이기 때문이다.

참고문헌

강현철 · 한상태 · 김지연 · 정용찬 · 허명회. 2008. "RDD 전화조사와 주요결과." 《조사연구》 9(1): 1-22.

조성겸. 1997. "대통령선거 여론조사와 할당표집 방법의 문제점." 《언론과 사회》 18(12): 29-53.

허명회 · 황진모. 2006. "전화조사를 위한 시간균형할당표본추출." 《조사연구》 7(2): 39-52.

허명회. 2007. "여론조사 방법론: 과제와 전망." 대한통계협회 《통계》 33(1): 27-36.

SPSS. 2007. *SPSS 16.0 Algorithms*. 506-522. SPSS Inc., Chicago.

[접수 2008/6/14, 수정 2008/7/29, 게재확정 2008/8/10]