연구논문

출구조사를 위한 투표소 확률추출 방법*

Probability Sampling to Select Polling Places in Exit Poll

김영원** · 엄윤희***
Young−Won Kim · Yoon−Hee Uhm

출구조사에서 투표소 추출방법은 출구조사의 정확성을 결정하는 중요한 요소이다. 본 연구에서는 대표구 추출법을 대신할 수 있는 정렬계통추출법을 제안하고 그 활용 가능성 및 효율성을 분석한다. 아울러 제시된 정렬계통추출법을 사용하는 경우 추정량의 표본추출오차(sampling error)가 어느 정도 되며, 원하는 목표 오차를 만족하기 위한 표본크기를 결정하는 문제를 고려한다. 2004년 17대 총선 개표자료를 토대로 경험적인 분석을 통해 제시된 정렬계통추출법이 기존의 대표구 추출법에 비해 평균예측오차 관점에서 효율적이라는 사실을 규명하고, 기존의 출구조사에서 표본크기 및 추정오차를 해석하는 과정에서 발생하는 오류를 집락효과를 이용해 설명했다. 아울러 제안한 정렬추출법에서 얻어지는 추정량의 분산을 구하고, 설계효과 개념을 이용해 표본크기 결정문제를 다루었다.

주제어 : 계통추출법, 대표구 추출법, 설계효과, 출구조사

The accuracy of exit poll mainly depends on the sampling method of voting places. For exit poll, we propose a probability sampling method of selecting voting places as an alternative to the bellwether polling place sampling. Through an empirical study based on the 2004 general election data, the efficiency of the suggested systematic sampling from ordered voting places was evaluated in terms of mean prediction error and it turns out that the proposed sampling method outperformed the bellwether polling places sampling. We also calculated the variance of estimator from the proposed sampling, and considered the sample size problem to guarantee the target precision using the design effect of the proposed sample design.

* 본 연구는 숙명여자대학교 2004년도 교내연구비 지원에 의해 수행되었음.
** 교신저자(corresponding author): 숙명여자대학교 통계학과 교수 김영원.
E−mail: ywkim@sookmyung.ac.kr
*** 숙명여자대학교 통계학과 대학원생.

key words : bellwether sampling, design effect, exit poll, systematic sampling

I. 서 론

선거예측조사는 실제 모수의 참값과 표본조사를 통해 얻어진 추정값 간의 차이를 확인할 수 있는 거의 유일한 표본조사이다. 특히, 일반인들이 가장 흔히 접할 수 있고, 관심을 많이 갖게 되는 표본조사 중 하나이기 때문에 어찌 보면 표본조사의 위상을 제고하기 위해 표본이론 연구자들이 조금은 더 관심을 갖는 것이 필요한 영역이다. 하지만 현재까지 여러 가지 이유로 표본조사 연구자들마저도 이에 대한 구체적인 연구결과를 제시하지 못하고 있는 것이 현실이다. 이런 관점에서 선거조사 중 가장 관심이 높다고 볼 수 있는 출구조사(exit poll)에서 투표소 추출방법에 대한 문제를 본격적으로 다루어 보고자 한다.

우리나라에서 선거예측조사 또는 출구조사 등과 관련된 기존의 주요 연구결과들을 살펴보면, 류제복(2000)은 16대 총선에서 선거예측조사결과를 토대로 예상 득표율과 실제 득표율의 차이를 고찰함으로써 선거예측결과의 신뢰성을 검토하였으며, 또한 류제복(2003)은 우리나라 총선과 대선에서 적용된 출구조사 적용 현황과 조사방법, 무응답 처리, 가중치 조정 등과 관련된 개선방향을 제시하고 있다. 김정훈(2003)은 출구조사가 매우 성공적으로 이루어진 것으로 알려져 있는 16대 대선 출구조사와 관련해 특정 조사기관에서 적용한 표본추출, 실사진행, 예측치 산출과정 등에 대한 상세한 내용을 보여주고 있다.

한편 홍내리 · 허명회(2001)는 16대 총선 예측조사와 관련해 전화조사와 출구조사 현황 및 문제점을 정리하고 있다. 특히, 이들은 현행 출구조사의 경우 표본 투표소 수가 원하는 표본추출오차(sampling error)를 만족시키기에 매우 부족하다는 점과 투표소를 판단표본추출(judge-

ment sampling)함으로써 발생하는 문제, 또한 조사거절과 거짓 응답 때문에 발생하는 오차를 매우 체계적인 모의실험을 통해 구체적으로 지적하고 있다. 홍내리·허명회(2001)의 연구결과를 참고로 하면 우리나라에서 현재 수행되고 있는 출구조사와 관련해 표본추출이론 측면에서 보다 심층적인 후속연구가 절실히 필요함을 알 수 있다. 이런 관점에서 출구조사에서 흔히 사용되는 '대표구 추출법(bellwether polling place sampling)'과 '계통추출법(systematic sampling)'을 비교함으로써 확률추출법의 적용 가능성을 보여주고 있는 조성겸·김지연(2004)의 연구는 매우 의미 있는 결과를 보여주고 있다.

출구조사와 관련해 홍내리·허명회(2001)가 지적한 문제점과 조성겸·김지연(2004)이 제시한 투표소 추출방법 등은 우리나라 출구조사 수행 조사기관에서 반드시 고려해야 할 중요한 사항들을 보여주고 있으며, 본 연구에서는 이들 연구자들의 견해를 가능한 통계이론을 적용하여 보다 심층적으로 분석해 보고자 한다.

본 연구에서는 출구조사 투표소 추출문제와 관련하여 크게 두 가지 문제를 다룬다. 첫 번째, 조사기관에서 그 동안 사용해 왔던 대표구 추출법을 대신할 수 있는 확률추출법의 적용 가능성 및 확률추출법의 효율성 문제를 다룬다. 두 번째, 제시된 확률추출법을 적용하는 경우 추정량의 표본추출오차(sampling error)가 어느 정도 되며, 원하는 목표 오차를 만족하기 위해서는 과연 몇 명의 투표자를 표본으로 추출하는 것이 적절한지에 대한 문제를 다룬다. 물론 본 연구에서 이들 두 가지 질문에 대한 완전한 이론적인 해답을 제시하지는 못하지만, 본 연구를 통해 관련 후속 연구를 촉진하는 동시에 조사기관들이 출구조사와 관련해 통계이론적인 측면에 관심을 기울이게 되는 계기가 될 수 있기를 기대해 본다.

출구조사 문제를 이론적으로 접근하기 위해서는 상당히 현실적인 사항들을 고려해야 하고, 이에 따라 많은 제약이 따르게 된다. 또한 이론

적으로 모든 것들에 대한 답을 얻을 수 없다는 특징을 갖는다. 따라서 본 연구에서는 2004년 17대 총선 개표자료를 토대로 경험적인 분석을 통해 효율적인 투표소 확률표본추출법에 대한 연구결과를 제시하고자 한다. 우리나라 대선의 경우 대체적으로 출구조사가 성공적으로 이루어졌다고 평가받고 있지만, 총선의 경우 예측 정확성과 관련해 많은 문제점이 지적되고 있다는 점을 고려해 여기서는 총선을 위한 출구조사만을 연구대상으로 한정하기로 한다.

첫 번째 연구 목적을 위해 본 연구에서는 적절한 방식으로 투표소를 정렬하고 표본 투표소를 계통추출하는 확률추출법을 검토한다. 현재까지 우리나라 출구조사에서는 대부분 투표소를 대표구 추출법 방식으로 추출해 왔다. 여기서 대표구 추출법이란 과거 선거에서 특정 선거구의 최종결과와 가장 근접한 결과를 보인 투표소를 표본으로 추출하는 것이다. 물론 지역안배 등 다양한 형태의 변형이 가능하겠지만 이런 추출법은 본질적으로 판단추출법(judgement sampling) 또는 전문가선택(expert choice) 등으로 불리는 대표적인 비확률추출법(non-probability sampling)에 해당한다. 대표구 추출법의 대안으로는 적절한 기준에 따라 투표소를 정렬한 후 계통추출하는 방법(systematic sampling from ordered polling places)을 생각할 수 있다. 따라서 이런 추출법과 기존 대표구 추출법의 효율성을 2004년 총선 개표자료를 토대로 경험적인 분석을 통해 비교·검토하기로 한다. 이를 위해 선거구별 후보자의 예측 득표율과 실제 득표율의 차이로 설명되는 평균예측오차를 효율성 판단 기준으로 사용한다.

두 번째 문제에 대한 해답을 얻기 위해 표본 투표소 선정을 위해 계통추출법을 적용하는 경우, 추정량의 표본분산을 집락효과(cluster effect) 개념을 도입해 설명한다. 아울러 과거 결과에서 얻어진 설계효과(design effect)를 고려해 원하는 정도(precision)를 만족하기 위해 추출해야 하는 표본 투표자 수의 결정 문제를 다루기로 한다. 제시된 정렬

계통추출법은 각 선거구에서 투표소를 기준으로 투표소 집락(투표소의 모임)을 먼저 구성하고, 그 중 1개의 집락을 추출하는 경우에 해당하기 때문에 이론적으로는 표본 자료를 갖고 표본추출오차를 직접 계산할 수 없다는 점에 유의할 필요가 있다.

출구조사에서는 투표소가 결정된 후에는 투표소별로 일정 간격(k)을 두고 투표자를 추출해 조사하는 계통추출법을 거의 모든 조사기관에서 적용하고 있고, 현실적으로 이런 방법을 대신할 수 있는 보다 효과적인 방안이 없다고 판단되기 때문에 본 연구에서는 표본 투표소 선정 후 투표자의 추출 문제는 다루지 않기로 한다.

II. 투표소 추출방법

1. 대표구 추출법(bellwether polling place sampling)

현재 우리나라 출구조사 투표소 추출방법으로 가장 많이 사용하는 방법은 대표구 추출법이다. 대표구 추출법은 선거구 전체의 투표결과를 가장 잘 대표할 수 있을 것으로 판단되는 투표소를 표본으로 추출하는 비확률표본추출법이다.

대표 투표소를 선정하는 방법으로는 여러 가지 방안이 있을 수 있는데, 예를 들어 소득이나 연령, 성별, 교육수준 등 주요 유권자 특성이 전체 선거구를 대표할 수 있는 선거구를 추출하는 것도 생각해 볼 수 있을 것이다. 현재 우리나라 방송사 출구조사에서는 과거 선거에서 전체 선거구의 최종 개표결과와 가장 '유사한 결과'를 보인 투표소를 표본으로 선정하는 대표구 추출법을 사용하고 있다. 여기서 유사성은 일반적으로 개별 투표소의 각 후보별 득표율과 선거구 전체의 각 후보별 득표율 차이를 종합해서 판단하게 되며 흔히 '역대 선거 오차의 최소제곱법'이라고 부르고 있다. 각 조사기관에 따라 최소제곱법을 약간 변형한 대표구 추

출법을 사용하고 있다(홍내리 · 허명회 2001).

이런 투표소 추출방법은 주관적인 판단에 의해 모집단을 대표할 수 있는 표본을 추출하는 판단추출법의 일종이다. 따라서 이런 표본추출법은 확률추출법에 해당하지 않기 때문에 추정결과에 대한 표본추출오차를 산출하는 것이 이론상 불가능하다. 또한 표본크기에 해당하는 투표소 수가 증가한다고 해서 반드시 더 정확한 결과를 얻을 수 있다는 보장은 없다. 우리나라 총선 선거예측을 위한 출구조사의 경우, 비용 및 조사원 충원과 관리 등의 문제로 1개 선거구에서 6~10개 정도의 투표소를 표본으로 추출하고 있으며, 이런 경우 1차 추출단위(primary sampling unit; psu)에 해당하는 투표소 관점에서 보면 표본크기가 매우 작기 때문에 확률추출법에 비해 이런 판단추출법이 상당히 효과적일 수도 있다. 이런 관점에서 특히 총선을 위한 출구조사에서 대표구 추출법이 통계적 검증 없이 현재까지 사용되어 왔다고 생각된다.

한편 동일한 선거구내에서 지역에 따라 투표성향에 있어서 뚜렷한 차이를 보이는 경우 선거구내의 투표소를 먼저 층화한 후 대표구를 추출하는 '층화 대표구 추출법'을 활용하기도 한다. 층화 대표구 추출법은 선거구를 동부 · 읍부 또는 시 · 군 · 구 등으로 층화한 후 각 층별로 대표투표소를 추출하는 방법이다. 흔히 복합선거구의 경우 하나의 선거구를 구성하는 시 · 군 · 구를 자연스럽게 각각 층으로 처리할 수 있다. 이 방법은 총선의 경우 하나의 선거구내에서도 후보자의 출신지 등에 따라 득표율에 있어서 지역별로 큰 차이를 보일 수 있기 때문에 시 · 군 · 구 등의 지역을 층으로 처리함으로써 일반적인 층화추출에서 얻을 수 있는 효과를 기대하는 것이다(조성겸 · 김지연 2004).

2. 계통추출법(systematic sampling)

확률표본추출방법 중 투표소 추출에 무리 없이 손쉽게 적용할 수 있

는 방법은 계통추출법이다. 계통추출법을 실제 적용하는 데 있어서는 추출단위의 배열방식이 상당히 큰 영향을 주게 된다. 다시 말해, 구체적으로 어떻게 추출단위들을 추출틀에 배열하는지에 따라 얻어지는 계통추출표본의 성질이 달라지며, 이런 관점에서 투표소를 랜덤하게 나열한 후 계통추출법을 적용하는 '단순계통추출법(systematic sampling from random population)'과 투표소를 정해진 기준에 따라 인위적으로 정렬한 후 계통추출법을 적용하는 '정렬계통추출법(systematic sampling from ordered population)'으로 구분할 수 있다.

투표소 추출에 있어서 각 투표소의 투표성향을 잘 설명해 줄 수 있는 정보를 확보할 수 있다면 이런 정보를 활용해 투표소를 정렬하고 계통추출하는 정렬계통추출법이 매우 효과적인 것으로 알려져 있다. 물론 이런 정보를 활용하지 않고 선관위에서 제공하는 투표소 명부순서를 그대로 사용하는 경우, 이는 랜덤하게 배열된 상태에서 추출한 단순계통추출법으로 볼 수 있다. 단순계통추출법을 적용하면 이는 집락을 단순확률추출하는 것과 유사하기 때문에 1위와 2위 후보자간 경합도가 높은 선거구의 경우 출구조사를 통해 정확한 선거결과 예측을 위해서는 일반적으로 상당히 많은 수의 투표소를 추출해야 한다. 이런 사실은 홍내리·허명회(2004)의 연구결과에서도 확인할 수 있다.

실제 투표소 추출문제에 있어서는 각 투표소에서 나타난 과거 투표성향이 알려져 있고, 이 자료를 이용하면 과거 선거에서 나타난 각 투표소의 정당 또는 후보자 득표율에 따라 투표소를 정렬하는 것이 가능하다. 이런 접근방식은 실제 대표구 추출법에서 사용하는 과거 선거 관련 정보와 동일한 정보를 사용하는 것으로 볼 수 있다.

이와 같은 정렬계통추출법은 우리나라에서는 아직 실제 적용되지 않고 있지만 영국의 출구조사에서 이와 유사한 추출법이 활용되었으며(Moon 1999), 조성겸·김지연(2004)은 이런 추출법을 '종단적 계통추출법'이라 부르고, 이에 대한 의미 있는 연구결과를 제시하고 있다.

한편 투표소를 계통추출할 때, 각 투표소를 동일한 확률로 추출하는 단순계통추출법을 적용할 수도 있고, 투표소별 유권자수를 크기척도(measure of size)로 하는 크기비례확률(probability proportional to size; pps) 계통추출법도 고려할 수 있다. 이런 두 가지 경우 epsem (equal probability selection method)이 유지되도록 표본 투표자를 추출하기 위해서는 투표소를 단순계통추출하는 경우에는 각 투표소에서 동일 추출간격을 적용하는 것이 필요하다. 반면에 투표소를 pps계통추출하는 경우에는 각 투표소에서 동일한 수의 투표자를 표본으로 선정하는 것이 요구되는데, 출구조사에서 각 투표소의 실제 투표자 수를 사전에 정확히 예측할 수 없기 때문에 이 경우 epsem을 적용하는 데 어려움이 있다. 참고로 현재까지 우리나라 출구조사에서는 각 투표소에서 동일 추출간격을 적용하는 계통추출법을 사용하여, 결과적으로 각 투표자가 표본으로 추출될 확률이 동일해지도록 표본을 추출하고 있다. 이런 추출법을 적용하면 추정단계에서 가중치를 고려하지 않는 단순평균으로 추정이 가능하다는 장점이 있다.

문제를 단순화하기 위해 본 연구에서는 계통추출법을 적용하는 경우 표본 투표소는 정렬계통추출하고, 각 투표소에서 동일한 추출간격(일반적으로 7명 중 1명)을 적용하는 표본추출법을 고려하기로 한다. 다시 말해 계통추출에서 pps추출법을 적용하는 경우는 다루지 않기로 한다.

본 연구에서는 2004년 실시된 17대 총선 개표결과 자료를 바탕으로 제시된 추출법의 효율성을 비교해 보고자 한다. 출구조사에서 기존의 대표구 추출법의 대안으로는 단순계통추출법보다 정렬계통추출법이 관심대상이 된다. 따라서 여기서는 다음과 같은 방식의 정렬계통추출법을 주 연구 대상으로 한다.

1) 정당 득표율에 따른 정렬계통추출법(계통추출법 Ⅰ)

이 방법은 과거 선거에서 나타난 특정 정당이나 후보자의 득표율을

기준으로 투표소를 정렬한 다음 투표소를 계통추출하는 방법을 말한다. 따라서 정당 득표율이란 측면에서 투표성향이 각기 다른 다양한 투표소들이 표본으로 추출되기 때문에 선정된 표본은 전체 선거구의 특성을 잘 반영해 줄 수 있다. 즉, 대표구 방식은 선거구 전체의 평균과 가장 유사한 투표성향을 보이는 투표소들만을 뽑는 반면, 정당득표율에 따라 투표소를 정렬한 후 계통추출방법을 적용하면 여당지지가 강한 투표소, 여야의 득표율이 비슷한 투표소, 야당지지가 강한 투표소 등을 모두 표본으로 추출해 모집단의 전반적인 특성을 반영하게 된다.

이런 계통추출법의 적용에 있어서 중요한 사항은 어떤 기준을 적용하여 투표소를 정렬할 것인지가 추정 결과의 정확성에 큰 영향을 주게 된다. 가능하면 투표소별 실제 개표결과와 상관계수가 높은 변수를 활용하여 투표소를 정렬하는 것이 요구된다. 총선의 경우 그 이전에 실시된 대선 또는 총선의 정당 득표율을 기준으로 활용할 수도 있고, 만약 과거 당선자가 이번 선거에도 다시 후보자로 출마하는 경우 해당 후보자의 지난 선거 득표율에 따라 투표소를 정렬하는 방법 등을 고려할 수 있다. 어떤 방법이 효과적인 방법이 될 수 있는지에 대해서는 이론적인 규명은 불가능하지만 실제 자료를 토대로 한 경험적인 분석을 통해 검증해 볼 수 있다.

2) 시·군 및 정당 득표율에 따른 정렬계통추출법(계통추출법Ⅱ)

몇 개의 시·군이 합쳐져서 하나의 선거구를 구성하는 복합선거구의 경우 각 시·군별 투표성향에 상당한 차이가 있을 수 있어, 이런 지역구분을 투표소 추출에 반영하는 것이 효과적일 수 있다. 예를 들어, 총선의 경우 후보자가 특정 시·군출신인 경우 후보자의 소속 정당 득표율만으로 이 후보자에 대한 유권자의 투표성향을 설명하는 데는 한계가 있다. 이와 같은 상황은 특히 복합선거구의 경우 심각하게 나타날 수 있

다. 따라서 복합선거구의 경우 전체 투표소를 우선 시·군으로 분류하고 각 시·군내의 투표소를 정당 득표율 등 특정 기준에 따라 정렬한 후 계통추출하는 방안을 고려할 수 있다. 실제 이런 추출방법은 시·군별로 투표소를 층화한 후 각 층에서 투표소를 추출하는 조성겸·김지연(2004)이 제안한 '층화 후 종단적 추출법'과 유사한 방식이지만, 층별 배분 문제 등을 고려할 필요가 없고, 층화추출의 추정식 대신 단순평균을 사용해 추정결과를 얻을 수 있다는 장점을 갖고 있다. 이런 정렬계통추출법은 대표구 추출법에서 복합선거구의 경우 지역(시·군·구)별로 층화한 후, 각 층에서 대표구를 추출하는 층화 대표구 추출법과 유사한 방식을 적용한 것이다.

III. 투표소 추출방법의 예측오차 비교

1. 연구대상 선거구

조성겸·김지연(2004)에 정리된 것과 같이 2004년에 실시된 17대 총선의 경우 16대에 비해 지역구 국회의원 수가 227개에서 243개로 늘어나면서 선거구 구역 조정이 일어난 곳이 상당히 많다. 또한 출구조사의 경우 사전 선거여론조사에서 후보자들간의 득표율 차이가 근소한 것으로 나타난 지역만을 대상으로 한다. 따라서 본 연구에서는 2004년 총선의 243개 지역구 중에서 1위와 2위 후보자간의 득표율 차이가 8% 이상이거나 투표소 변동이 심해 과거 선거자료 활용이 불가능한 선거구는 연구대상에서 제외했다.

이에 따라 17대 총선에서 1위와 2위 후보자간의 득표율 차이가 8% 이내인 80개 지역 중, 선거구내의 투표소 변동이 비교적 적은 74개 지역을 대상으로 추출법의 예측오차를 비교·분석했다.

17대 총선의 선거구를 2002년에 실시된 16대 대선과 2000년에 실시

된 16대 총선과 비교하였을 때, 투표소 수가 증가된 선거구 현황을 정리하면 <표 1>과 같다. 16대 대선과 16대 총선에 비해 17대 총선에서 투표소 수가 감소한 경우도 있지만 이들 선거구의 경우 일부 통합 투표소를 동명을 기준으로 정리하면 과거 선거결과를 활용하는 데 큰 문제가 없기 때문에 이런 선거구들은 분석대상에 그대로 포함했다. 또한 단순히 일부 '동'의 행정명(동명)이 변경된 선거구의 경우도 변경된 동명을 확인하여 서로 대응시켜 분석대상에 포함한 것이다.

<표 1> 17대 총선 대비 각 선거구 투표소 수 증가 현황

투표소 수 증가	0	+1	+2	+3	+4	+5 이상	전체
16대 대선	39	14	6	7	4	4	74
16대 총선	22	16	6	11	2	17	74

2. 추출방법의 효율성 비교 방법

선거예측에서는 특히 1위와 2위 득표자의 득표율 예측의 정확도가 중요한 관심대상이 되기 때문에 본 연구에서는 각 표본추출방법에 따라 얻어진 추정결과의 정확도를 비교하기 위해 조성겸·김지연(2004)[1] 가 사용한 다음과 같은 예측오차를 기준으로 효율성을 비교하기로 한다.

$$\text{예측오차} = |\,(\text{1위 득표자 예측 득표율} - \text{2위 득표자의 예측 득표율}) - (\text{1위 득표자의 최종 득표율} - \text{2위 득표자의 최종 득표율})\,|$$

정렬계통추출방법은 특정 정당이나 후보자의 득표율을 기준으로 투표소를 정렬한 다음, 일정한 간격으로 투표소를 추출한다. 따라서 2004년 17대 총선을 연구대상으로 하는 경우, 활용 가능한 과거 선거결과로

1) 조성겸과 김지연(2004)에서는 | |가 누락되어 있음.

는 2000년 16대 총선결과와 2002년 16대 대선결과 등이 있다. 한편 16대 총선, 16대 대선, 17대 총선까지 변함없이 그대로 유지된 정당은 한나라당 밖에 없었다. 따라서 계통추출법의 경우 16대 대선과 16대 총선에서 한나라당의 득표율을 기준으로 투표소를 정렬한 다음, 표본 투표소의 수가 8 또는 10이 되도록 계통추출하는 방안을 고려했다.

정렬계통추출법과 비교 대상이 되는 대표구 추출법으로는 16대 대선과 16대 총선 때 선거구 전체의 투표결과와 가장 비슷한 결과를 보인 투표소를 선정하는 경우를 고려하도록 한다.

참고로 1위와 2위 후보자간의 득표율 차이가 8% 이내인 74개 지역 중에서 강진·완도 지역은 16대 총선 때 한나라당 후보가 존재하지 않아 16대 총선 자료를 대상으로 하는 정렬계통추출을 사용하는 경우에는 73개 지역만을 대상으로 계통추출법을 적용했다.

결과적으로 본 연구에서는 다음 4가지 방법에 따른 표본추출방법의 효율성을 예측오차를 기준으로 비교한 것이다. 한편 각 추출법에 따라 추출되는 표본 투표소의 수는 선거구당 8개의 투표소를 추출하는 경우와 10개의 투표소를 추출하는 경우를 함께 검토하기로 한다.

- 16대 총선자료 기준 대표구 추출법
- 16대 대선자료 기준 대표구 추출법
- 16대 총선자료 기준 정렬계통추출법
- 16대 대선자료 기준 정렬계통추출법

정렬계통추출법의 경우 연구대상 모든 선거구에서 제시된 추출법에 따라 각 선거구별로 추출이 가능한 모든 표본을 대상으로 예측오차를 각각 산출하고, 이렇게 얻어진 예측오차들을 토대로 모든 연구대상 선거구별로 평균 예측오차를 산출한 후, 이를 종합한 결과를 기준으로 추출법의 효율성을 비교한 것이다. 반면에 대표구 추출법의 경우 최소제곱법을 기준으로 표본 투표소를 선정하기 때문에 각 선거구에서 단지 하나

의 표본이 추출된다.

예를 들어, 용산구의 경우 59개 투표소가 있고, 여기서 8개의 투표소를 표본으로 추출하는 경우를 고려해 보자. 대표구 추출법의 경우, 16대 총선 또는 16대 대선 자료를 활용하여 8개의 투표소를 추출하는 경우, 각각 1개의 표본이 추출되고 이 표본에서 얻어진 예측오차를 효율성 비교 대상으로 한다. 결과적으로 대표구 추출법의 경우 각 추출법에 따라 선거구별로 1개의 예측오차가 계산된다.

한편, 계통추출법의 경우 표본크기가 8인 경우 7개의 가능한 표본이 있고, 이들 모든 가능한 표본에서 예측오차를 산출한 후 이들의 평균을 구한 것을 해당 선거구에서 계통추출법의 평균예측오차라고 하고, 이를 기준으로 효율성을 비교한다. 이 과정에서 투표소 수가 8의 배수가 아니기 때문에 발생하는 문제를 처리하기 위해, 전체 투표소 중 3개 투표소를 랜덤하게 삭제하고 56개의 투표소를 대상으로 하면 표본 투표소 수가 8개인 경우 용산구에서는 7개의 표본이 추출 가능하고, 이들 7개의 표본에서 각각 예측오차가 산출될 수 있다. 따라서 계통추출의 경우 이들 7개 예측오차의 평균을 구해 효율성을 비교한다.

참고로, 조성겸・이지연(2004)의 연구에서도 16대 대선 자료를 활용하는 경우 17대 총선자료를 기준으로 대표구 추출법과 정렬계통추출법(그들은 '종단적 계통추출법'이라고 했음)의 효율성을 비교하고 있다. 그들의 연구에서는 계통추출법의 경우 각 선거구에서 1개의 표본을 랜덤하게 추출하여 추출방법의 효율성을 비교하고 있기 때문에 특정 선거구에서 가능한 계통 표본 중 어떤 표본이 선택되는지에 따라 계통추출법의 효율성이 결정된다는 한계를 갖고 있다. 이런 측면을 고려해 본 연구에서는 계통추출의 경우, 각 선거구에서 추출 가능한 모든 계통 표본을 고려해 효율성을 비교한 것이다.

3. 정렬계통추출법의 예측오차

본 연구에서 연구대상으로 하고 있는 74개 선거구에서 얻어진 결과를 모두 종합하여 예측오차를 비교한 결과는 〈표 2〉와 같다. 여기서 괄호안의 수는 표본 투표소 수를 나타내며, 계통추출의 경우 과거 16대 대선 또는 16대 총선 자료를 사용하여 투표소를 정렬한 후 계통추출한 결과이다. 한편 복합선거구의 경우 지역(시/군) 개념도 동시에 반영한 계통추출법이 보다 더 효과적일 것으로 예측되기 때문에, 18개 복합선거구를 제외하고 일반 선거구만을 대상으로 했을 때 얻어진 평균예측오차를 분리해 함께 보여주고 있다.

〈표 2〉를 보면 대표구 추출법의 경우, 16대 대선 자료를 이용하는 것보다 16대 총선 자료를 이용하는 것이 예측오차를 기준으로 더 정확성이 높은 것으로 나타났다. 참고로 조성겸·이지연(2004)에서는 정당의 이합집산에 따른 변동과 2002년에 이루어진 16대 대선이 2000년에 실시된 16대 총선보다 시차가 적다는 점 등을 고려해 16대 대선 자료를 토대로 한 분석을 하고 있다. 한편, 선거구를 정렬한 후 계통추출법을 적용하는 경우에는 대선 자료나 총선 자료 중 어느 자료를 이용하여도 큰 차이가 없는 것으로 나타났다.

〈표 2〉 전체 선거구 대상 평균 예측오차 비교

	74개 선거구 대상		56개 선거구 대상 (복합선거구 제외)	
	대표구	계통추출 I	대표구	계통추출 I
대선(8)	4.5465	**2.1980**	4.0131	**2.0845**
대선(10)	4.2309	**2.6278**	3.9690	**1.7983**
총선(8)	**2.6644**	3.0998	2.2225	**2.1977**
총선(10)	**2.4538**	2.5104	1.9134	**1.7997**

전체 연구 대상인 74 선거구에서 얻어진 결과에서 보면 대선 자료를 이용하는 경우 계통추출법 I 이 상당히 효율적인 것으로 나타나고 있으나, 총선 자료를 이용하는 경우에 대표구 추출법이 차이가 크지는 않지만 더 좋은 것으로 나타났다. 하지만 전체 선거구 중 복합선거구에 해당하는 18개 선거구를 제외한 56개 선거구만을 대상으로 하면 과거의 총선 또는 대선 어떤 자료를 이용하든지 계통추출법 I 이 정확성이 높은 것으로 나타났다. 또한 대표구 추출방법과 정렬계통추출방법 모두 각 선거구에서 표본 투표소 수를 8개에서 10개로 증가시키는 경우 정확성이 높아지고 있다.

복합선거구 지역에 대해서는 계통추출법 I 보다 지역(시·군·구)을 고려한 계통추출법 II 가 효율적인 방법이 될 수 있으며 이에 대해서는 다음에 좀 더 자세히 설명하기로 한다.

한편, 연구대상 선거구 중 지역적으로 투표소가 밀집되어 지역개념을 정렬과정에 반영할 필요가 없다고 볼 수 있는(다시 말해, 복합선거구와 반대 속성을 갖는) 가장 대표적인 선거구들에 해당하는 서울의 23개 선거구만을 분리해 효율성을 비교해 본 결과는 〈표 3〉과 같다. 〈표 3〉을 보면, 모든 경우 대표구 추출법 대신 계통추출법 I 을 적용하게 되면 예측오차가 대폭 감소한다는 사실을 확인할 수 있다.

〈표 3〉 서울특별시 23개 선거구 대상 평균 예측오차 비교

	대표구	계통추출 I
대선(8)	3.5145	**1.8289**
대선(10)	3.3767	**1.3302**
총선(8)	2.5269	**1.8021**
총선(10)	2.2068	**1.7826**

4. 복합선거구에서 정렬계통추출법의 예측오차

복합선거구는 태백·영월·평창·정선과 같이 2개 이상의 시·군이 합쳐져 하나의 선거구를 구성하는 경우를 의미한다. 이런 복합선거구의 경우 전체 투표소가 자연스럽게 시·군 개념으로 그룹화 될 수 있으며, 이에 따라 조성겸·이지연(2004)은 투표소를 시·군별로 층화한 후 각 층에서 계통추출법을 적용하는 방안을 검토하기 위해 철원·화천·양구·인제 선거구에 대한 분석결과를 제시하고 있다. 하지만 이들이 제시한 층화 계통추출법을 적용하는 경우 층별 표본 투표소 배분문제 및 불균등 추출 확률을 반영하기 위한 가중치 적용문제들이 고려되어야 하기 때문에 실제 출구조사에 활용하기에는 현실적으로 어려움이 있다.

따라서 본 연구에서는 복합선거구의 경우에도 투표소를 적절한 기준으로 정렬한 후 계통추출법을 적용하는 방법을 그대로 적용하기로 한다. 하지만 시·군별 투표성향을 반영하기 위해 우선 전체 투표소를 시·군별로 구분하고 각 시·군내에서 투표소를 과거 선거의 정당 득표율에 따라 정렬하여 계통추출하는 계통추출법Ⅱ를 적용하기로 한다. 이런 추출법을 흔히 내재적 층화(implicit stratification) 방법이라고 부르기도 하며, 이론상 층화추출에서 비례배정을 하는 경우와 거의 동일한 추출법에 해당하고 결과적으로 epsem에 해당하게 된다.

이런 방법을 통해 얻을 수 있는 효율성 증대 효과를 검토하기 위해 18개 복합선거구를 대상으로, 한나라당 득표율만을 고려해 정렬한 후 계통추출하는 방법(계통추출법Ⅰ)과 우선 시·군별로 투표소를 구분한 후 각 시·군내에서 투표소를 한나라당 득표율에 따라 정렬한 후 계통추출하는 방법(계통추출법Ⅱ)를 비교해 본 결과는 〈표 4〉와 같다.

복합선거구의 경우 계통추출법Ⅱ가 16대 대선 또는 16대 총선 자료를 사용하는 모든 경우에 있어서 정확성이 높은 것으로 나타났다. 한편 〈표 2〉의 결과와 비교해 보면 전반적으로 복합선거구의 경우 어떤 추출

〈표 4〉 18개 복합선거구 대상 계통추출의 평균 예측오차 비교

	대표구	계통추출 I	계통추출 II
대선(8)	6.2060	4.4707	**3.7048**
대선(10)	5.0457	4.1665	**3.7844**
총선(8)	4.1201	4.8503	**3.7463**
총선(10)	4.2339	3.8887	**3.1222**

법을 적용하던지 일반 선거구에 비해 대체적으로 예측오차가 상대적으로 크게 나타나고 있음을 볼 수 있다. 따라서 향후 복합선거구에서의 예측오차 감소방안에 대해서는 좀더 심층적인 연구가 필요할 것으로 판단된다.

IV. 표본추출오차와 표본크기의 결정

1. 현행 출구조사에서 오차의 한계 및 표본크기의 결정

대표구 추출법은 비확률추출법에 해당하기 때문에 이론적으로 표본추출오차를 산출할 수 없다.[2] 그럼에도 불구하고 대표구 추출법으로 추출된 최종 투표자 표본을 단순확률추출법(simple random sampling)으로 추출된 것으로 간주하여 출구조사의 표본추출오차를 산출하고, 이런 잘못된 가정하에서 목표 오차의 한계를 만족시키는 표본 투표소 수를 결정하는 경우 큰 오류가 발생하게 된다. 일반적인 여론조사에서 흔히 사

2) 대표구 추출법을 적용하는 현행 방송사 출구조사에서 표본추출오차 산출방법은 공개되고 있지 않기 때문에 구체적으로 파악할 수 없다. 출구조사에 참여한 연구자의 설명에 따르면 해당 방송사의 경우 과거 선거에서 대표구 추출법을 적용했을 때 얻어진 선거구별 오차를 토대로 한 경험적인 결과와 단순확률추출법의 표본추출오차 산출방법을 종합적으로 반영하여 출구조사의 표본추출오차를 개략적으로 산출하고 있다고 한다.

용하는 단순확률추출을 전제로 한 표본크기 결정 과정은 다음과 같이 요약될 수 있다.

선거구의 전체 투표자 중에서 단순확률추출로 투표자를 표본추출하여 출구조사가 된다고 가정하는 경우, 표본크기(최종적으로 조사된 표본 투표자 수)가 n이라고 하면, 신뢰수준 95%에서 오차의 한계는 $1.96\times\sqrt{\frac{pq}{n}}$이다. 우리나라 여론조사에서는 보통 사전에 p를 알 수 없기 때문에 일반적으로 $p=0.5$라고 가정하고 표본크기를 결정하고 있다. p가 0.5 일 때 오차의 한계는 최대가 되므로, 최대 오차의 한계는 다음과 같이 근사적으로 계산된다.

$$1.96\times\sqrt{\frac{pq}{n}} \doteq 2\times\sqrt{\frac{0.25}{n}} = \frac{1}{\sqrt{n}}$$

따라서 출구조사에서 얻어진 표본이 단순확률추출된 것으로 가정하는 경우, 출구조사에서 조사된 표본 투표자 수가 1,000명이라면 최대 오차의 한계는 3.16%인 것으로 간주된다. 조성겸・김지연(2004)에 의하면 총선 출구조사의 경우 개략적으로 각 선거구별로 10개의 투표소를 추출하고 각 투표소에서 200명 내외를 조사하고 있다. 따라서 최종 표본크기는 2,000명 정도가 되며, 이 경우 최대 오차의 한계는 2%내외가 될 것으로 예상하고 있다. 다시 말해, 표본설계시에 95% 신뢰수준에서 목표 최대 오차의 한계를 4%, 3%, 2%로 설정하는 경우, 단순확률추출을 전제로 하면 필요한 표본크기는 각각 625명, 1,111명, 2,500명인 것으로 계산된다. 하지만 이 표본크기는 단순확률추출일 때만 성립한다는 점에 유의해야 한다.

김재광(2004)이 지적한 것과 같이 출구조사의 경우 투표소를 추출하고 표본 투표소내에서 투표자를 추출하는 집락추출법이 적용되기 때문에 집락효과를 오차의 산출 과정에 반영하는 것이 필요하다. 이런 집락효과를 적절하게 고려하면 실제 오차의 한계는 단순확률추출을 전제로

산출된 결과보다 훨씬 더 커지게 된다. 하지만 현행 출구조사에서는 이런 사실이 표본크기 산출 및 오차의 한계 산출에 있어서 충분히 반영되지 않고 있다.

2. 단순계통추출에서 설계효과 및 표본크기 결정

1) 단순계통추출에서 설계효과

만약 출구조사에서 투표소를 랜덤하게 나열하고 계통추출하는 경우, 즉 단순계통추출하는 경우, 이는 투표소를 단순확률추출하는 경우와 거의 유사한 것으로 볼 수 있다. 또한 각 표본 투표소에서 투표자를 다시 1/k 계통추출로 선정한다고 하면[3], 이런 출구조사는 이론적으로 전형적인 2단 집락추출법에 해당된다. 따라서 이 경우 표본분산의 산출 및 표본크기의 결정 문제는 집락추출법에서 사용되는 방법으로 계산될 수 있다.

단순집락추출에서 목표오차를 만족시키는 표본크기는 개략적으로 $n \times deff$으로 설명된다. 여기서 n은 단순확률추출에서 목표오차를 만족시키는 표본크기이고, $deff$는 집락추출에서의 설계효과(design effect)를 나타낸다.

투표소를 집락으로 처리하는 단순집락추출에 따른 설계효과를 검토하기 위해서는 우선 집락내 단위들의 유사성 정도를 파악해야 한다. 집락추출에서 집락내의 단위들의 유사성을 나타내는 척도로는 Sarndal (1992)이 제시하고 있는 다음 척도를 사용할 수 있다. 유사성계수 (homegeneity coefficient)인 δ는 다음과 같이 정의된다.

---

3) 투표소내에서 투표자를 추출간격 k로 계통추출해서 얻어진 결과를 단순확률표본으로 간주하기로 한다. 이는 투표자들이 투표성향에 상관없이 투표소에 랜덤하게 도착한다고 가정하는 것과 같다.

$$\delta = 1 - \frac{S_{yW}^2}{S_{yU}^2}$$

여기서,

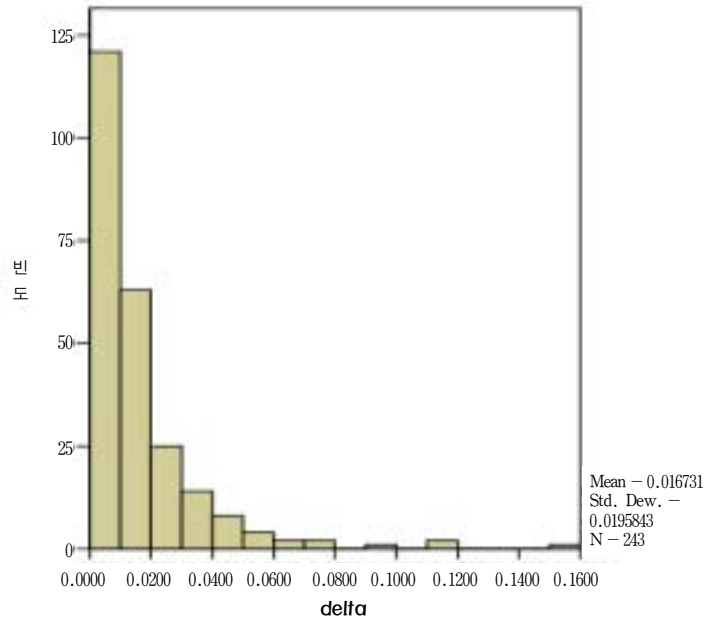
$$S_{yW}^2 = \frac{\sum (N_i - 1) S_{yU_i}^2}{\sum (N_i - 1)} \quad ,$$

$S_{yU_i}^2 = \frac{1}{N_i - 1} \sum_{U_i} (y_k - \overline{y_{U_i}})^2 = P_i(1 - P_i)$ 이고, N_i는 투표소 i의 투표자 수, P_i는 투표소 i에서 후보자의 득표율을 나타내며, $S_{yU}^2 = P(1-P)$ 이고, P는 전체 투표소에서 후보자의 득표율이다.

17대 총선 투표소별 개표결과 자료를 바탕으로 전체 243개 선거구를 대상으로 집락내 유사성을 검토해 보기로 한다. 집락내 유사성을 나타내는 척도로는 앞에 제시된 동질성 계수인 δ를 이용한다. 243개 선거구에서 얻어진 값들에 대한 히스토그램은 <그림 1>과 같고, 요약 통계는 <표 5>와 같다.

<표 5> 유사성 계수 δ의 요약 통계

요약 통계	
평균	0.0167
표준편차	0.0196
최소값	0.0014
1사분위수	0.0060
중위수	0.0100
3사분위수	0.0198
최대값	0.1579



〈그림 1〉 유사성 계수 δ의 히스토그램

단순확률추출에서 표본크기가 n인 집락표본과 동일한 정도를 제공하는 유효표본크기(effective sample size: neff)는 $neff = n/deff$ 로 설명된다. 여기서 $deff$ 는 집락추출의 설계효과를 나타낸다. 예를 들어, 95% 신뢰수준에서 목표 오차의 한계가 4%가 되기 위해서는 단순확률추출에서 표본크기는 625이므로, 목표오차를 만족하는 단순집락추출법에서 표본크기는 $625 \times deff$이다.

하지만 실제 출구조사의 경우 각 표본 투표소에서 $1/k$ 계통추출법을 적용하여 투표자를 조사하기 때문에 2단 집락추출법을 고려한 설계효과를 계산할 필요가 있다.

2) 단순계통추출에서 표본크기 결정

현재 우리나라 출구조사에서는 흔히 6~10개 정도의 투표소를 추출

하고 각 투표소에서 7명 중 1명을 추출하는 계통추출을 하고 있다는 점을 고려하여, 여기서는 $k=7$인 경우, 표본 투표자 수를 결정하는 문제를 다루기로 한다. 각 표본 투표소에서 추출간격이 k인 계통추출을 적용하면 설계효과는 근사적으로 다음과 같이 계산된다.

$$deff\,(Cluster,\ SRS) \doteq 1 + (\frac{\overline{N}}{k} - 1)\,\delta$$

여기서 $\overline{N}$는 투표소별 평균 투표자수이다.

이를 이용하면 단순계통추출로 투표소를 추출하는 경우 목표오차를 만족시키기 위해 조사해야 할 총 표본 투표자 수를 산출할 수 있다. 여기서 사용되는 동질성 계수 δ는 <그림 1>과 <표 5>에 제시된 것과 같이 각 선거구별로 상당한 차이를 보이고 있다. 따라서 각 투표소에서 산출되는 $deff$ 역시 많은 차이를 보이게 된다. <표 6>은 243개 선거구에서 얻은 동질성 계수들의 사분위수를 기준으로 얻은 목표오차를 만족시킬 수 있는 표본크기를 산출한 것이다. 여기서 사분위수만을 고려한 이유는 극단적인 경우들은 일단 연구대상에서 제외하고, 대체적으로 선거구의 동질성 계수가 취하게 되는 범위를 고려하기 위한 것이다.

<표 6> 1/7 계통추출의 경우 표본 투표자 수

목표 오차의 한계	단순 확률추출	정렬계통추출		
		1사분위수 (*deff*=2.5500)	중위수 (*deff*=3.2474)	3사분위수 (*deff*=4.7163)
4%	625	1,594	2,030	2,948
3%	1,111	2,833	3,608	5,240
2%	2,500	6,375	8,119	11,791

〈표 6〉은 표본 투표소를 단순계통추출하고, 각 표본 투표소에서 7명 중 1명의 투표자를 조사하는 경우에 95% 신뢰수준에서 4%, 3%, 또는 2%의 오차한계를 만족시키기 위해서 조사해야 하는 최종 표본 투표자 수를 보여주고 있다. 여기서 $deff$의 1사분위수는 2.5500, 중위수는 3.2474, 3사분위수는 4.7163이다. 예를 들어, 목표오차가 4%라면 단순확률추출에서는 625명을 추출하는 것이 요구되지만, 집락추출에서는 중위수에 해당하는 $deff$값인 3.2474인 경우를 고려하면 625×3.2474로 2,030명을 표본으로 추출해야 한다. 따라서 출구조사에서 오차의 한계 및 표본크기를 단순확률추출로 간주하는 경우 상당히 큰 오류를 범할 수 있다.

〈표 6〉에 제시된 결과는 투표소를 단순계통추출하는 것을 전제로 한 것이다. 따라서 본 연구에서 관심대상으로 하고 있는 정렬계통추출법을 적용하는 경우 추정량의 분산은 감소하게 되고, 투표소내에서 1/7 계통추출을 가정하는 경우 목표오차를 만족시키는 표본 투표자 수는 〈표 6〉의 결과보다 작아지게 된다. 하지만 정렬계통추출법을 적용하는 추정량의 분산은 앞에서 언급한 유사성 계수 등으로 간단히 설명될 수가 없다. 이에 대한 이론적인 해답은 향후 과제로 남겨두기로 하고, 정렬추출법을 적용하는 경우 표본추출오차는 대체적으로 어느 정도 발생하는지 다음 절에서 살펴보기로 한다.

3. 정렬계통추출에서 표본분산 및 표본크기 결정

1) 정렬계통추출에서 표본분산

제안된 정렬계통추출법 표본 투표소 추출방식은 우선 각 선거구에서 과거 정당 득표율 등 정해진 기준에 따라 투표소를 정렬한 후, 일정 간격을 건너뛰면서 전체 투표소를 B개의 그룹(집락)으로 묶은 다음, 이 중 하나의 집락을 추출하는 것으로 정리될 수 있다. 그런 다음 선정된 하나

의 집락내의 모든 투표소에서 1/k 계통추출로 투표자를 선정하는 2단 집락추출에 해당한다. 이는 2단 집락추출에서 psu로 하나의 집락만을 추출하기 때문에 일반적인 방법으로는 표본 자료에서 표본분산을 추정하는 것이 불가능하다. 하지만 17대 총선 개표 자료를 이용하면, 비록 사후적이지만 모집단 전체 자료를 확보하고 있는 상황이기 때문에 추정량의 표본분산을 계산하는 것이 가능하다.

제안된 정렬계통추출법을 적용하는 경우, 각 선거구에서 특정 후보자 득표율 추정량에 대한 분산은 다음과 같다.

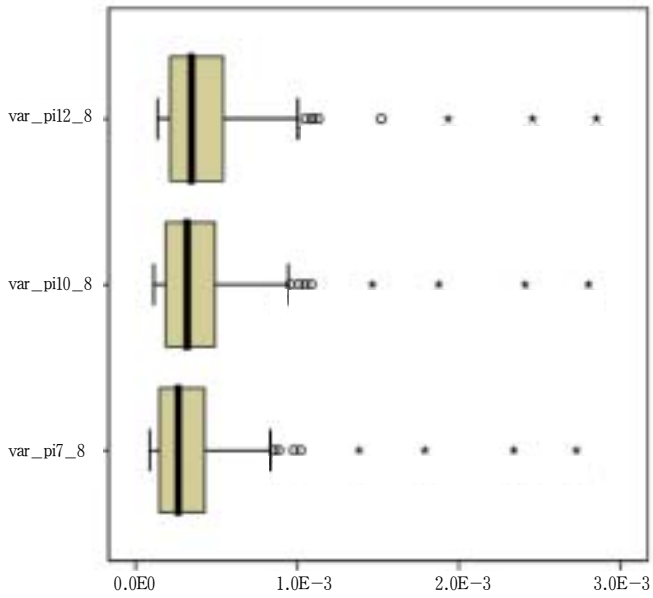
$$
\begin{aligned}
Var(\hat{p}_{sys}) &= \left(\frac{1}{B}\right)\times\frac{1}{\overline{M}^2}\times\frac{1}{B-1}\sum_{i=1}^{B}(t_i - pM_i)^2 \\
&\quad + \frac{1}{B}\times\frac{1}{\overline{M}^2}\sum_{i=1}^{B}M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{S_i^2}{m_i} \\
&= \frac{1}{B}\times\frac{1}{\overline{M}^2}\left[\sum_{i=1}^{B}(t_i - pM_i)^2 + \sum_{i=1}^{B}M_i^2\left(1-\frac{m_i}{M_i}\right)\frac{S_i^2}{m_i}\right]
\end{aligned}
$$

여기서 $S_i^2 = \frac{\sum(y_{ij}-p_i)^2}{M_i - 1} = p_i(1-p_i)$ 이고, B는 각 선거구에서 추출 가능한 모든 정렬계통표본의 수, $\overline{M} = \frac{1}{B}\sum_{i=1}^{N}M_i$ 은 각 집락(투표소의 모임)에서 평균 투표자 수, M_i는 각 집락의 총 투표자 수, t_i는 각 표본에서 후보자의 득표수, p는 선거구 전체에서 후보자의 실제 득표율, $p_i = \frac{t_i}{M_i}$ 는 각 표본에서 후보자 득표율이다. 표본으로 선택된 각 투표소에서 전수조사를 하지 않으므로 m_i는 집락에 따라 달라진다. 예를 들어, 각 투표소에서 7명중 1명이 추출되면 $m_i = M_i / 7$이다.

제시된 추정량의 분산은 모집단 자체를 알고 있는 상태에서는 계산할 수 있지만, 실제 출구조사에서는 표본으로 선정된 하나의 집락에 대한 자료만을 확보할 수 있기 때문에 분산 추정식을 구할 수는 없다는 점에 유의할 필요가 있다.

〈표 7〉 표본 투표소 수가 8일 때 $Var(\widehat{P_{sys}})$

추출간격(k)	최소값	1사분위수	중위수	3사분위수	최대값
7	0.00008	0.00015	0.00026	0.00043	0.00273
10	0.00011	0.00019	0.00032	0.00049	0.00280
12	0.00013	0.00021	0.00034	0.00054	0.00285



〈그림 2〉 표본 투표소 수가 8일 때 $Var(\widehat{P_{sys}})$의 상자그림

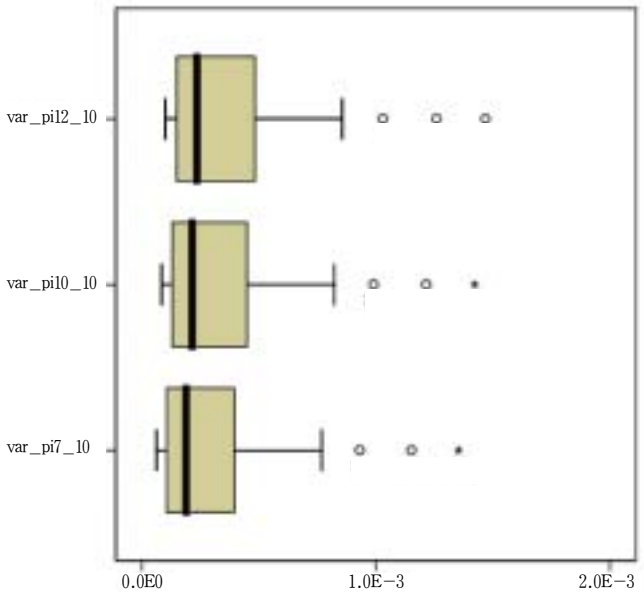
본 연구에서 정렬계통추출의 효율성 검토를 위해 연구대상으로 고려한 74개 선거구를 대상으로, 표본으로 8개의 투표소를 추출하는 것을 가정하고 제시된 정렬계통추출법을 적용하는 경우 산출되는 1위 후보자의 득표율 추정량의 분산들을 구하여 정리한 결과는 〈표 7〉과 같다. 여기서

얻어진 결과들은 일반 선거구의 경우 16대 총선 한나라당 득표율로 정렬하여 계통추출하는 경우(계통추출Ⅰ)를 고려한 것이고, 복합선거구의 경우는 지역 및 한나라당 득표율을 함께 적용하여 정렬한 후 계통추출한 경우(계통추출Ⅱ)에 얻은 결과들을 종합한 것이다. 여기서 k는 투표소에서 투표자의 추출간격을 나타낸다. <그림 2>는 각 선거구에서 얻어진 분산을 k값의 변화에 따라 상자그림으로 정리한 것이다. <표 7>과 <그림 2>를 보면 투표소에서 투표자의 추출간격이 커질수록 분산이 커지고 있는데, 이는 표본 투표소 수를 8로 고정하고 추출간격을 증가시키면 결과적으로 실제 조사대상 표본 투표자 수가 감소하게 됨으로써 나타나는 당연한 현상이다.

<표 7>을 보면, 표본으로 8개 투표소를 정렬계통추출하고 표본 투표소에서 k=7인 계통추출로 투표자를 표본조사하면, 이 때 추정량의 표본분산은 연구대상 선거구 중 중위수에 해당하는 경우를 고려하면 0.00026이 된다. 따라서 95% 신뢰수준에서 오차의 한계는 이 경우 $1.96\times\sqrt{.00026} = 0.0316$, 즉 3.16%p라는 것이다. k=7인 계통추출로 투표자를 추출하면서 표본 투표소 수를 달리 하는 경우에 대해서도 추정량의 분산을 계산할 수 있다. 투표소를 10개 표본 추출하는 경우 얻어지는 결과는 <표 8>과 <그림 3>에 제시되어 있다.

<표 8> 표본 투표소 수가 10일 때 $Var(\widehat{P}_{sys})$

추출간격(k)	최소값	1사분위수	중위수	3사분위수	최대값
7	0.00006	0.00011	0.00019	0.00039	0.00135
10	0.00008	0.00013	0.00021	0.00045	0.00142
12	0.00010	0.00015	0.00023	0.00048	0.00147



〈그림 3〉 표본 투표소 수가 10일 때 $Var(\widehat{P_{sys}})$의 상자그림

2) 정렬계통추출에서 설계효과

본 연구에서 고려한 정렬계통추출법을 적용하는 경우 목표오차를 만족하는 표본 투표자 수를 결정하는 문제를 고려해 보도록 하자. 정렬계통추출법의 경우 주어진 분산식에서 목표오차를 만족하는 표본 투표소 수와 투표자 추출간격을 직접 산출할 수 없다. 따라서 정렬계통추출법을 적용하는 경우 얻어지는 추정량의 분산과 동일한 표본크기로 단순확률추출하는 경우 얻어지는 추정량의 분산의 상대적인 차이를 설명해 주는 정렬계통추출의 설계효과를 이용하여 표본크기를 결정하는 방법이 하나의 대안이 될 수 있다.

앞에서 각 선거구별로 계산된 정렬계통추출법의 추정량의 분산을 이

용하면 정렬계통추출의 설계효과(*deff*)는 $\frac{Var(\hat{p}_{sys})}{Var(\hat{p})}$ 으로 계산할 수 있다. $Var(\hat{p}_{sys})$는 정렬계통추출법에서 득표율 추정량의 분산으로 이에 대해서는 앞 절에서 자세히 설명했다. 한편 $Var(\hat{p})$은 각 선거구에서 정렬계통추출법과 동일한 표본크기의 단순확률표본을 추출해 얻어지는 득표율 추정량의 분산이다.

〈표 9〉는 표본투표소가 8과 10인 경우에 추출간격을 7, 10, 12로 했을 때, 각 선거구에서 얻은 *deff*의 사분위수를 정리한 것이다.

〈표 9〉 *deff*의 사분위수

표본 투표소 수	추출간격(k)	1사분위수	중위수	3사분위수
8	7	1.5325	1.9957	3.3887
	10	1.3499	1.6667	2.6441
	12	1.2787	1.5408	2.3572
10	7	1.4521	2.1243	2.9485
	10	1.2856	1.7351	2.2972
	12	1.2216	1.5860	2.0476

3) 정렬계통추출에서 표본크기 결정

제시된 정렬계통추출의 설계효과를 이용하면 목표오차를 만족시키기 위해 출구조사에서 조사해야 할 표본 투표자 수를 개략적으로 파악할 수 있다. 여기서는 과거 총선 자료를 기준으로 표본 투표소의 수가 8이 되도록 정렬하여 계통 추출하는 경우 얻어지는 〈표 9〉에 제시된 *deff*를 기준으로 목표 오차한계를 만족시킬 수 있는 표본투표자 수를 산출한 것이다.

〈표 10〉 목표오차에 따른 정렬계통추출 표본 투표자 수

추출간격(k)	목표 오차의 한계	1사분위수	중위수	3사분위수
7	4% p	958	1,247	2,095
	3% p	1,703	2,216	3,724
	2% p	3,832	4,987	8,380
10	4% p	844	1,040	1,641
	3% p	1,500	1,848	2,917
	2% p	3,375	4,159	6,564
12	4% p	800	962	1,466
	3% p	1,421	1,710	2,605
	2% p	3,197	3,848	5,862

〈표 10〉을 보면 목표오차가 4%라면, 단순확률추출에서는 625명이 표본으로 필요하지만 중위수를 기준으로 하는 경우 각 투표소에서 추출간격을 7로 하면 1,247명, 10으로 하면 1,040명, 12로 하면 962명을 표본으로 선정하여야 하고, 목표오차가 2%라면 단순 확률추출에서는 2,500명을 표본으로 필요하지만 각 투표소에서 추출간격을 7로 하면 4,987명, 10으로 하면 4,159명, 12로 하면 3,848명을 표본으로 추출해야 한다.

〈표 10〉은 상당히 제한적인 결과만을 보여주고 있다는 점에 유의할 필요가 있다. 우선 실제 출구조사에서 최종 표본 투표자 수는 표본 투표소 수와 투표자 추출간격에 의해 결정된다는 사실이 반영되지 않은 결과이다. 예를 들어, 투표소를 8개 추출하고 추출간격을 7로 한다면 어떤 투표소들이 실제 추출되는지에 따라 최종 표본 투표자 수가 자동으로 결정되게 된다. 따라서 〈표 10〉의 결과는 정렬계통추출법을 적용할 때 투표소와 추출간격을 결정하기 위한 하나의 가이드라인 역할은 할 수 있지만, 표본크기를 결정해 주는 이론적인 해답은 아니다. 아울러 향후 다른 선거예측 출구조사에서는 실제 선거구별 투표소 상황이 달라지기 때문

에 17대 총선결과를 토대로 얻어진 〈표 9〉의 결과도 그대로 유지될 수 없다. 따라서 위에 제시된 결과는 향후 다른 선거예측조사에 그대로 적용될 수는 없고, 단지 참고자료로서 의미가 있을 것이다. 이런 계산과정은 각 선거구별로 별도로 이루어져 얻어진 결과라는 점에도 유의할 필요가 있다.

하지만 정렬계통추출법을 출구조사에 적용하게 되는 경우 〈표 10〉의 결과는 표본 투표소 수를 결정하기 위한 참고자료로 매우 유용한 정보를 제공해 줄 수 있다고 판단된다. 특히 제시된 결과를 활용하는 것이 단순확률추출을 전제로 잘못된 방식으로 목표오차를 설정하고 표본 투표소 수와 투표자 추출간격을 결정하는 것에 비하면 표본추출이론 측면에서 훨씬 설득력이 있는 접근방식이라고 볼 수 있다.

V. 결 론

본 연구에서는 우리나라 출구조사에서 많이 사용되고 있는 대표구 추출법과 이에 대한 대안으로 제시한 정렬계통추출법의 효율성을 비교하였다. 17대 총선에서 1위와 2위의 격차가 8% 이내인 74개 선거구의 개표 자료를 사용하여 평균 예측오차를 기준으로 효율성을 비교해 본 결과 정렬계통추출법이 기존의 대표구 추출법에 비해 효율적이란 사실을 확인할 수 있었다. 아울러 제시된 정렬계통추출법은 선거구 구역 변동이 심하게 발생하여 대표구 추출법이 현실적으로 적용 불가능한 경우에도 매우 효과적인 방안이 될 수 있다는 점을 지적하고자 한다.

또한 집락효과 개념을 도입해 출구조사에서 목표 오차의 한계를 만족하는 표본크기를 결정하는 방안을 제시했으며, 효과적인 것으로 규명된 정렬계통추출법을 적용하는 경우 얻게 되는 추정량의 분산을 구하고, 그에 따른 설계효과를 기초로 비록 상당히 제한적이지만 목표오차에 따

른 표본 투표자 수를 결정하기 위한 가이드라인을 제시하고 있다.

한편 출구조사의 투표소 추출을 위해, 크기비례확률추출법(pps sampling)의 활용 가능성 등, 본 연구에 다루지 못한 다양한 형태의 확률추출법들이 적용 가능할 것으로 판단된다. 여기서 다룬 추출법 이외에 어떤 확률추출법의 적용이 가능할 것인지? 또한 분산 추정 문제가 연계된 표본 투표소 수의 결정 문제를 어떤 방식으로 접근하는 것이 이론적으로 타당할 것인지? 등과 같이 과학적인 출구조사를 수행하기 위해서 규명되어야 할 많은 문제들이 방치되어 있다. 따라서 본 논문에서 얻어진 기초적인 결과들이 출구조사와 관련된 문제들에 대한 명확한 해결책을 제시할 수 있는 국내 연구를 촉진하는 계기가 될 수 있기를 기대해 본다.

참고문헌

김정훈. 2003. "선거예측과 출구조사: 16대 대선을 중심으로." 《조사연구》 4(2): 87−102.

김재광. 2004. "제 17대 총선 예측조사와 관련된 통계적 이슈 고찰." 《조사연구학회 춘계학술 대회 발표논문집》 203−207.

류제복. 2000. "선거예측조사의 신뢰성 증진방안 −16대 총선을 중심으로." 《조사연구》 1(2): 15−34.

류제복. 2003. "출구조사의 역사와 개선방향." 《조사연구》 4(1): 31−48.

조성겸 · 김지연. 2004. "출구조사의 투표소 표집방안 비교." 《조사연구》 5(2): 3−29.

홍내리 · 허명회. 2001. "제16대 국회의원 선거의 예측조사에 대한 사후적 검증." 《조사연구》 2(1):1−36.

Moon, N. 1999. *Opinion Polls: History, Theory and Practice*. Manchester University Press. Chapter 7. Exit polling.

Sarndal, C. 1992. *Model Assisted Survey Sampling*. New York, Springer−Verlag.