18 대와 19 대 대선 데이타에서 나온 K 값에 대한 설명 (세 번째 자료)

작성자: 전희경, 현화신

작성일: 2017 년 7 월 30 일

<더플랜>에서 제시한 K 가 최근 2017 년 5 월에 치러진 19 대 대선 결과 분석에 적용되었습니다[1,2]. 통계적 분석 방법이 선거에 응용되는 것을 반갑게 생각하면서, 이러한 검증 과정이 지속적으로 이루어져 앞으로 공정하고 투명한 선거 제도를 정착시키는 데에 기여하게 되기를 바랍니다. 그런데 <더플랜>에서 미분류표 분석의 도구로 거론한 K 를 일부에서 오해하고 있는 것으로 보여 보충 설명을 하고자 합니다. 객관성을 유지하기 위하여 후보들의 이름보다는 당선자, 경쟁자 또는 후보 1, 2 로 표기합니다.

한국 18 대 대선에서 K=1.5 로 나타났고, 이 것은 당선자가 경쟁자보다 미분류표에서 상대적으로 더 많이 득표하여 선출된 것을 의미합니다. 또한 전국 251 개의 K 값이 1.5 로 모이는 현상을 발견하고 K=1.5 를 설명하는 방법의 하나로 전자개표기를 프로그램하는 시나리오를 세웠는데, 시뮬레이션 결과가 표 1 에 나타난 것과 같이 실제 선거 결과에 (251 개표소) 매우 가깝게 나타났습니다. 이 것이 의미하는 바는 **<더플랜>에서 제시한 시나리오에 의해 18 대 대선 결과를 재구성할 수 있다는 것입니다. 따라서 <더플랜>의 주제는 전자개표기 사용에 대한 경각심을 불러 일으키고, 더 나아가 개선 방법을 찾자는 것입니다.**

표 1: 18 대 대선과 시뮬레이션 결과 (6 개 개표소)

	18 대 대선 결과				시뮬레이션 결과			
district index	**P1_R**	**M1_R**	**P2_R**	**M2_R**	**P1_R2**	**M1_R2**	**P2_R2**	**M2_R2**
1	66%	33%	69%	22%	66%	34%	67%	23%
20	49%	51%	52%	38%	49%	51%	53%	37%
89	**88%**	**11%**	**82%**	**7%**	**86%**	**13%**	**83%**	**7%**
109	**7%**	**93%**	**9%**	**80%**	**9%**	**91%**	**9%**	**79%**
122	51%	49%	53%	35%	51%	49%	54%	35%
210	21%	78%	20%	71%	26%	74%	25%	61%
Mean of 251 districts	**52%**	**47%**	**53%**	**35%**	**52%**	**47%**	**54%**	**35%**

참고: 모든 지역에 대한 시뮬레이션 결과는 프로젝 부 사이트 "2012 아카이브"에 첨부되어 있습니다. http://www.projectboo.com/archive/153432

18 대와 19 대 대선 결과와 관련하여 K 를 이해하고 적용하는 방법을 아래와 같은 순서로 설명하겠습니다.

- K 란 무엇인가?
- K 값과 미분류표 생성에 대한 두 가지 가설 비교
- 18 대 대선과 가설 1 & 2 요약
- 19 대 대선과 가설 1 & 2 요약
- 제안과 결론
  - K=1 이 정상인가?
  - K=1 이 되는 선거 환경을 만들 수 있을까?
  - 18 대 대선의 K=1.5 값이 의도된 것인가?
- 자료 출처
- 부록 1 (K-test, 기대값과 분산)
- 부록 2 (K-impact, 승리방정식)

1. K 란 무엇인가?

<더플랜>에서 소개된 K 는 통계적 테스트를 할 수 있는 척도입니다. 예를 들어, 암을 치료하는 새로운 약이 나왔을 때, 이 약의 효능을 판단하기 위하여 두 그룹을 가지고 실험을 합니다. 이 때 통계적으로 정립된 t-test 는 두 그룹의 평균을 비교하면서 새로운 약의 효능이 있는지 또는 없는지 판단합니다. 이처럼 K 는 두 후보의 미분류율을 비교하면서 당선자가 미분류표에서 상대적으로 득표를 더 많이 했는지 또는 아닌지 여부를 판단하는 테스트 방법으로 이해하면 되겠습니다 (즉, two-sample t-test for equal means versus **two-candidate K-test for equal unclassified rates)**. K 는 한국 선거시스템을 고려하여 만들어진 통계적 테스트 방법이지만, 그림 1 에 나타나 있는 것처럼 한국과 같은 투개표 시스템을 가진 다른 나라에서도 활용될 수 있는 일반화된 테스트 방법입니다.

좀더 통계적 내용을 소개하자면, 두 가지의 test statistic 은 아래와 같이 정의됩니다.

$$T = (M1 - M2)/\sqrt{\frac{s1^2}{N1} + \frac{s2^2}{N2}},$$

$$K = (P2/M2)/(P1/M1).$$

여기에서 M1 & M2: two sample means (두 개의 샘플 평균),

$s1^2$ & $s2^2$: two sample variances (두 개의 샘플 분산) ,

N1 & N2: two sample sizes (두 개의 샘플 크기),

P1 & M1: two candidate received votes from the classified (두 후보 분류표),

P2 & M2: two candidate received votes from the unclassified (두 후보 미분류표).

Null hypothesis: T: 두 개의 샘플 평균이 같다.

K: 두 후보의 미분류율이 같다.

테스트 결론을 내리기 위해서는 위 두 가지 검정 통계량(test statistic)을 주어진 데이타에 적용하여 얻은 값을 귀무가설(null hypothesis)에서 나오는 값들과 비교하게 됩니다. 그런데 T 는 필요한 조건들을 만족하면 t-distribution 을 따르고 있으므로 이를 바탕으로 두 개의 샘플 평균이 같은지 판단합니다. 반면에 P1, M1, P2 & M2 는 각각 이항분포를 따르지만, K 는 특정한 분포를 따르지 않습니다. 따라서 두 후보의 득표수가 크다는 것을 가정하고, **두 후보의 미분류율이 같다는 귀무가설에서 나오는 기대값과 분산을 계산하였습니다 (부록 1 참고). 이러한 통계적 테스트 접근 방식에 따라 두 후보의 미분류율을 같다고 놓는 것이 K 의 통계적 특성을 찾는 데에 반드시 필요합니다.** 주목할 점은 이러한 귀무가설에서는 K 의 기대값이 지역 크기와 상관없이 항상 1 이고, 분산은 지역 크기에 따라 달라진다는 것입니다. 예를 들어 어떤 지역에서 K=1.05 또는 K=0.94 가 나왔을 때, 이 값들이 K 의 기대값 1 과 같은지를 판단하려면, 부록 1-(A)에 있는 분산을 적용하면 됩니다 (using log-normal distribution). 즉, 그 지역의 크기와 미분류율을 고려하여 판단하게 됩니다.

하지만 두 후보의 미분류율이 같지 않다면, K 의 기대값과 분산 모두 지역마다 달라지게 됩니다 (부록 1-(B) 참고) . 즉, 한국 18 대 대선에서 나온 251 개의 K 값은 하나의 같은 분포에서 나온 것이 아니라, 각각 다른 분포에서 나온 것이 됩니다.

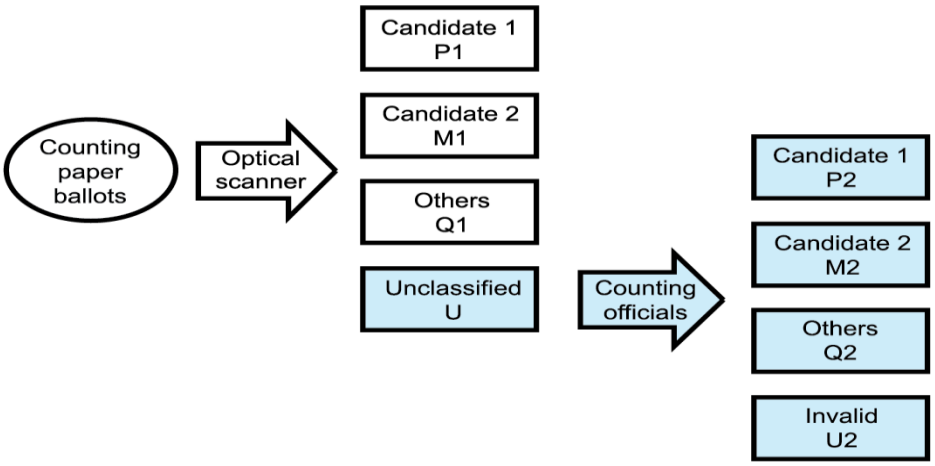

그림 1: 투표지 분류 과정: (1) 전자개표기에 (optical scanner) 의해 먼저 분류되고; (2) 미분류된 투표지들은 심사집계부에 의해 재분류된다.

현재까지 논의된 바에 의하면 미분류표는 크게 두 가지로 분류되는데, 전자개표기 오류에 의한 "정상표"와 전자개표기가 판독할 수 없는 "도장표" (도장이 제대로 찍히는 않은 비정상표) 입니다. 분류표가 총 투표수의 96%를 차지하는 것에 비해 미분류표는 4%정도입니다. 따라서 박빙의 선거일 경우에 미분류표가 특정 후보의 당선 여부를 결정하는 데에 영향을 미칠 수 있습니다. 전자개표기가 정상적으로 작동하면, 분류표는 정상표임에 반해, 미분류표는 도장표로서 비정상표입니다. 그런데 미분류된 도장표는 선관위의 결정에 따라 유효 또는 무효표로 재분류 되는데, 선관위의 결정이 일정하지 않고 선거 때마다 바뀌는 부분이 있습니다. 따라서 박빙의 선거에서는 미분류표에 더 많은 관심을 갖게 되는 것이 당연합니다.

선거에서 K 를 활용하고 해석하는 방법은:
먼저 K=(**미분류표**에서 당선자/경쟁자)/(**분류표**에서 당선자/경쟁자)라 놓고 K-값을 구합니다.
(1) K = 1 이면, 두 후보의 미분류율이 공정하게 같게 나타난 경우이므로, 즉 정상적인 선거가 이루어졌음을 제시하는 것이고;
(2) K < 1 이면, 당선자가 미분류표에서 경쟁자보다 득표를 덜 하고도 선출된 것이므로, 즉 미분류된 비정상표에 의한 이득이 없으므로, 당선자 득표에 대한 의구심을 가질 필요가 없음을 제시하는 것이지만;
(3) K > 1 이면, 당선자가 미분류된 비정상표에서 경쟁자보다 득표를 더 많이 하여 선출된 것이므로, 즉 미분류표에 의한 이득이 있으므로, 그 원인을 찾는 것이 필요함을 제시하는 것입니다.

또한 K 의 특성을 다른 각도에서 찾을 수도 있습니다. 수학 식을 이용하여 K 가 당선자 확정에 미치는 영향을 분석해 보았습니다 (부록 2). 요약하자면 **18 대 대선에서 K=1.5 이라는 값은, 분류표에서 (총 투표수의 96% 해당) 후보 1 이 (당선자) 후보 2 의 (경쟁자) 99% 이상만 득표하면 선출될 수 있는 수치**라는 것을 승리 방정식 (winning equation)이 보여줍니다. 달리 말하자면, K 값은 상황에 따라 미리 결정될 수 있고, 이 것을 전자개표기가 프로그램에 의해 실행할 수 있다는 가능성을 보여주는 것입니다. 하지만 이러한 플랜은 분류표를 제대로 확인하지 않을 때에 이루어진다는 점을 밝힙니다.

2. K 값과 미분류표 생성에 대한 가설 비교

K 값에 대한 기대값은 미분류표의 특성과 관련이 있습니다. 현재까지 논의된 미분류표에 대한 가설은 두 가지인데, **미분류표는 모든 후보에게 똑같이 발생한다라는 가설 1** 과 (<더플랜> 가설) 또는 **노령층에 의한 도장표에 의해 미분류표가 발생한다는 가설 2** (이하 "노령층 가설" [3]) 있습니다. 지금까지 미분류표에 대한 공식 자료가 없기 때문에 두 가설 모두 검증이 필요하므로, 심정적 선택이 아니라 **개연성과 현실성** 여부를 비교하고자 합니다. 개연성은 선거 결과를 재구성할 수 있는 정도를 근거로 판단하는 것이 합리적이라 할 수 있겠습니다.

K 값의 기대값은 가설 1 에 의하면 후보와 상관없이 항상 1 이고 (이론적으로 증명됨), 가설 2 에 의하면 후보마다 달라지므로 1 보다 크거나 작을 수도 있습니다. 따라서 가설 1 은 K=1 이라는 기대값을 통해 두 후보의 미분류율이 같다 또는 다르다를 판단하는 근거를 제시합니다. 반면에 가설 2 는 판단의 기준이 없고 도장표 발생에 따라 달라지므로, 선거가 공정하게 이루어졌는지 그 여부를 판단할 수 있는 근거로 사용되지는 못합니다.

3. 18 대 대선과 가설 1 & 2 요약

가설 1 은 <더플랜>에서 보인 시나리오의 근거입니다. 가설 1 과 시나리오는 연결되어 있지만 같은 내용은 아닙니다. 18 대 대선을 재구성하기 위해 제안된 것이 시나리오인데, 이 시나리오가 가설 1 을 바탕으로 만들어진 것입니다. 그 결과 18 대 대선에서 나타난 K=1.5 를 바탕으로 계산한 확률들로 251 개표소에서 나온 두 후보의 분류표와 미분류표에서의 득표율 네 가지를 각각 매우 잘 재구성 할 수 있었습니다 (표 1). 심정적으로 받아들이기 어려운 부분이 있을 수도 있으나, 18 대 대선에서 전자개표기 프로그램이 절대적으로 불가능하다고 확인되지 않는 한, 가설 1 을 바탕으로한 시나리오를 틀렸다고 할 수 없겠습니다. 시뮬레이션과 실험으로 가설 1 의 개연성은 검증되었으나, 시나리오의 현실성에 있어서는 증명할 수 없는 부분이므로 가능성만 열어 놓습니다.

다른 한 편, 가설 2 는 심정적으로 받아들이기 쉬운 반면에 18 대 대선 251 개 개표소에서 나온 미분류율과 K 값들을 제대로 설명하지 못하고 있는 아쉬운 점이 있습니다. 먼저 간단한 예를 들자면, 18 대 대선에서 가장 미분류율이 높았던 곳이 12.5%이었는데 (50 대 이상이 60%인 상대적으로 노령층이 많은 지역), 이 곳의 K 값은 **1** 로 나타났습니다. 노령층이 많으면 미분류율이 높고, 또한 미분류율이 높으면 후보 1 에게 더 많은 표가 갈 것이므로 K>1 로 기대되는 점에 어긋나는 사례입니다. **가설 2 의 개연성은 전반적으로 미분류율을 어느 정도 설명할 수 있느냐에 달려 있습니다.**

가설 2 가 (노령층 가설) 18 대 대선의 미분류표를 충분히 뒷받침해주지 못하는 점을 크게 세 가지로 요약하겠습니다. 첫째, 가설 2 의 핵심적 내용은 미분류표가 노령층에 의해 만들어졌다는 것이므로, 251 개 지역에서 나온 미분류율을 노령층 분포[4]로 대략적인 설명을 할 수 있어야 합니다. 실제로는 미분류표에서 정상표들이 발견되었지만, **모든 미분류표가 도장표에 의해 전적으로 100% 만들어졌다고 가정하고, 총 투표지 3 천만표를 가지고** 확률을 계산해 보았습니다 (샘플링이 아니므로 오류가 없음). 50 대 이상의 투표자들이 (총 투표자의 44%) 미분류표를 만들 확률이 **6.2%**이고 50 대 미만에서는(총 투표자의 56%) **1.8%**로 나와 50 대 이상 투표자들의 미분류표 발생이 대략 3 배 정도 높은 것으로 나옵니다. (q=0.062 & r=0.018, 확률 계산 자세한 방법은 프로젝 부 참고 http://www.projectboo.com/archive/153432)

그런데 이 확률들을 251 개표소 각각의 노령층 분포에 (즉 50 대 이상의 구성 비율) 적용해보면, **미분류율의 실제값이 2%~13%**로 나온 반면에 가설 2 확률에 따른 **예상값이 1%~5%** 정도입니다. 절반에도 못미치는 예상값이므로, 후보별 득표율 예상값도 실제값과 차이가 큽니다. 따라서 가설 2 는 현실성은 있어 보이나 18 대 대선을 충분히 설명하지 못합니다.

또한 19 대 대선과 비교하기 위해 세 가지 연령 그룹으로 나누어 확률을 계산하여 보면, 60 세 이상 (총 투표자의 23%), 50 대 (21%), 50 미만 (56%) 세 가지 연령 그룹에서 미분류표를 만들었을 확률이 각각 **5.5%, 6.4%, 1.8%**로 계산됩니다. 즉 50 대가 60 세 이상 그룹보다 미분류표를 더 많이 만든 것으로 나타납니다. 이 점도 가설 2 를 뒷받침하지 않는다고 볼 수 있습니다.

둘째, 가설 2 는 미분류표가 연령에 의해서 발생한다고 했는데, 18 대 대선 결과를 보면 지지하는 후보에 따라 미분류표 발생 비율이 달라지는 것으로 나타났습니다. 예를 들면, 후보 1 의 지지자들이 후보 2 의 지지자들보다 미분류표를 더 많이 만들어냈습니다. 이 현상은 상대적으로 젊은 지역 (50 대 이상 구성 비율이 20~40%인 지역들)에서도 나타났으므로, 미분류표는 연령뿐만 아니라 후보에 따라 발생한 것으로 이해하는 것이 적절합니다. 이 것은 서울 지역의 25 개 개표소뿐만 아니라 (그림 2) 전국적으로 나타난 현상으로서 (그림 3), 같은 연령 분포를 가진 지역이라도 지지하는 후보자에 따라 미분류표 발생율이 달라진 것을 볼 수 있습니다.

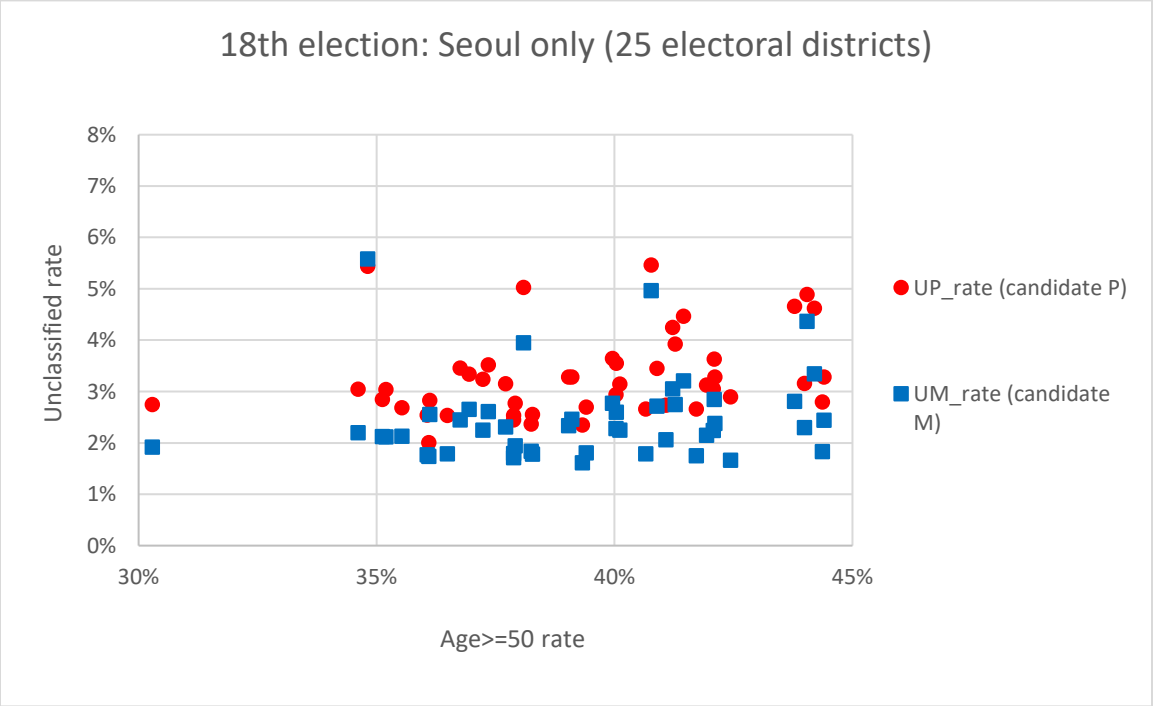

그림 2: 18 대선 서울지역에서 두 후보 사이의 미분류율 비교: (1) x 값: 각 지역에서 50 세 이상 투표자들의 비율 (%); (2) y 값: 후보별 미분류율=P2/(P1+P2) or M2/(M1+M2).

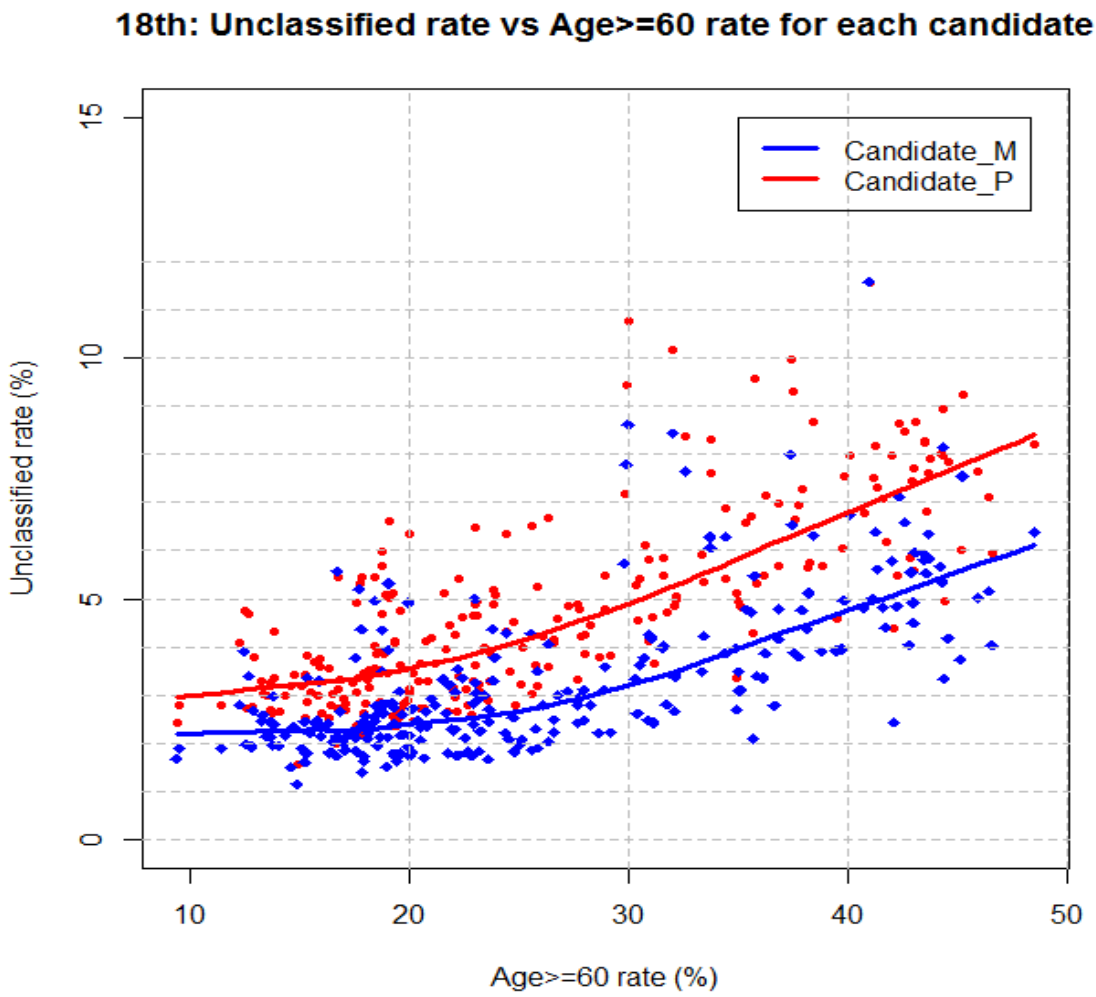

그림 3: 18 대선 전국에서 미분류율과 60 세이상 노령층 비율 사이의 관계: (1) x 값: 각 지역 60 세 이상 투표자들의 비율 (%); (2) y 값: 후보별 미분류율=**P2/(P1+P2)** or **M2/(M1+M2).**

그림 3 은 미분류율이 노령층의 비율이 높아짐에 따라 커지는 것을 보여줌과 동시에 노령층 분포만으로는 설명하지 못하는 점도 나타냅니다. 즉, 연령 분포와 상관없이 후보 1 지지자들은 (빨간색) 후보 2 지지자 (파란색) 보다 비정상표인 미분류표를 더 많이 만들었다는 것을 뚜렷하게 보여줍니다. 그런데 이 점을 가설 2 는 설명하지 못하고 있습니다. 이 부분을 "보수"로 설명하려면, 보수 투표자들은 보수가 아닌 투표자들보다 비정상적인 도장표를 더 많이 만들어 낸다는 사실을 인정해야 합니다.

셋째, 가설 2 는 미분류표가 노령층에 의해서 발생한다고 했는데, 특히 50 대 (50 세~59 세)의 분포만 보면 이를 뒷받침하지 않습니다. 50 대의 분포가 18-22%인 지역에서는 50 대의 비중이

커져도 미분류율이 전반적으로 비슷하여 차이가 없는 것으로 나타났습니다 (그림 4). 참고로 이 현상은 18 대뿐만 아니라 19 대 대선에서도 나타났는데, 가설 2 를 뒷받침하지 않는 점을 19 대 대선에서 더욱 뚜렷하게 보여줍니다 (그림 5).

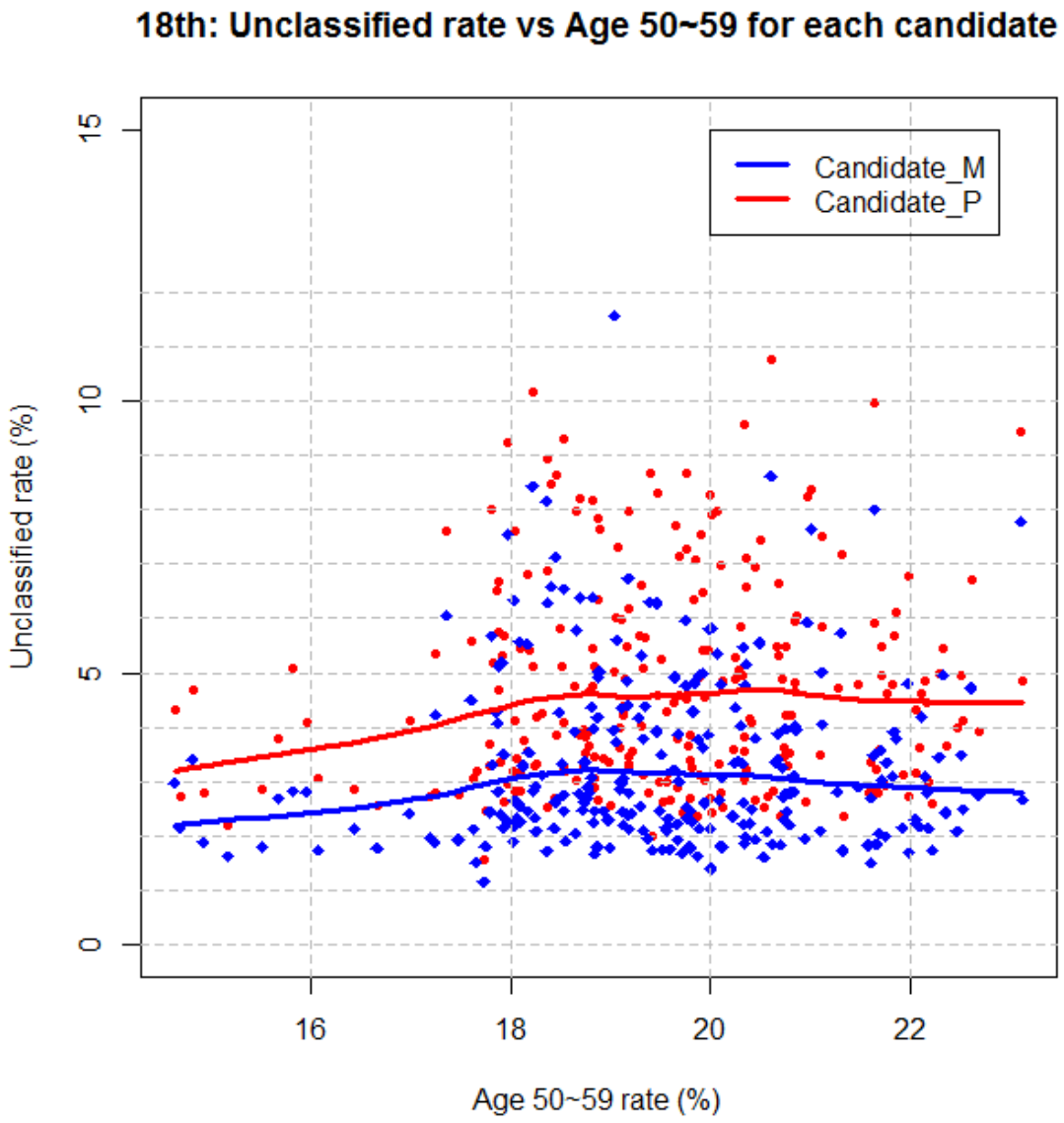

그림 4: 18 대선에서 미분류율과 50 대 연령층 비율 사이의 관계: (1) x 값: 각 지역에서 50 대 (50 세~59 세) 투표자들의 비율 (%); (2) y 값: 후보별 미분류율=P2/(P1+P2) or M2/(M1+M2).

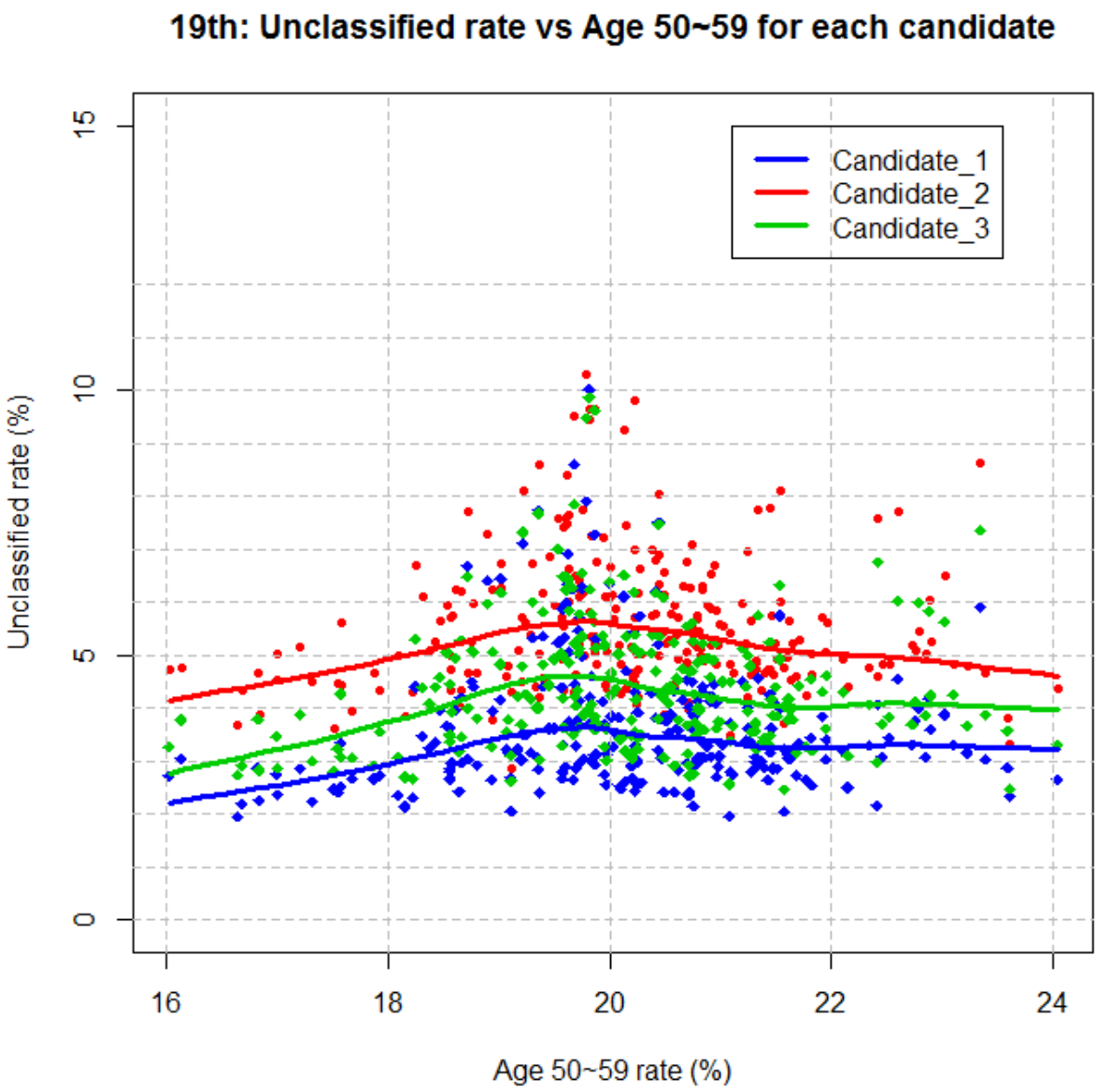

그림 5: 19 대선에서 미분류율과 50 대 연령층 비율 사이의 관계: (1) x 값: 각 지역에서 50 대 (50 세~59 세) 투표자들의 비율 (%); (2) y 값: 후보별 미분류율= M2/(M1+M2) in blue, H2/(H1+H2) in red or A2/(A1+A2) in green.

4. 19 대 대선과 가설 1 & 2 요약

<더플랜>에서 제시한 K 는 이항 분포 (Binomial distribution)를 적용하여 체계를 갖춘 것으로, 각 후보의 분류표와 미분류표 득표수가 각각 모두 크다고 가정한 후 기대값을 구한 것입니다. 그런데 19 대 대선의 13 명의 후보자 중에서 전국 250 개 지역 모두에서 큰 득표를 했다고 볼 수 있는

후보는 세 후보이므로 K 값을 후보 1, 2, & 3 에게만 적용합니다. 실제로는 득표 1, 2 위를 한 당선자와 (후보 1) 경쟁자 (후보 2) 두 후보에게 적용하는 것이 선거 감시의 의도이지만, 가설을 검증하기 위해 세 후보를 비교합니다.

먼저 18 대 대선처럼 하나의 예를 들겠습니다. 19 대 대선에서 가장 미분류율이 높았던 곳이 11%이었는데, 이 곳은 60 세 이상이 30% 정도인 지역입니다. 그런데 이 곳에서 나온 두 개의 K 값은 **1.06** (후보 1/후보 2) & **1.02** (후보 1/후보 3)로 각각 1 에 가깝게 나타났습니다. 미분류율이 높은데도 두 개의 K 값이 1 에 가깝게 나온 점이 가설 2 의 예상과는 다름을 보여주는 또 하나의 사례입니다. 전체적인 면에서 19 대 대선 결과를 세 가지로 요약하겠습니다.

첫째, 19 대 대선에서 세 후보 사이의 K 값을 구해보면, 당선자 후보 1 과 경쟁자 후보 2 사이에서 K=0.6, 그리고 후보 1 과 후보 3 사이에서 K=0.8 로 나타났습니다. 이 것은 가설 1 로 설명이 되지 않습니다. 즉, 세 후보의 미분류율이 같지 않다는 판단에 이르게 됩니다. 그러나 어떻게 미분류율이 후보마다 다르게 나오게 되었는지는 가설 1 로 설명할 수 없으므로 의문으로 남습니다. 반면에 가설 2 는 이러한 미분류율 차이를 노령층의 도장표로 설명하려고 시도합니다.

18 대 대선에서는 K>1 로서 당선자가 상대 후보보다 미분류표에서 더 많이 득표함으로써 선출되었지만, 19 대 대선에서는 K<1 로서 당선자가 상대 후보보다 미분류표에서 덜 득표하였는데도 당선되었습니다. 즉, <더플랜>의 K 는 선출된 후보가 상대 후보와 비교해서 미분류표를 통해 상대적 득표를 더 많이 또는 적게 했는지를 보여주는 척도로 이해하면 되겠습니다.

둘째, 가설 2 로 19 대 대선의 미분류율을 제대로 설명할 수 있는지 살펴봅니다. 이를 위해서는 연령별 투표율과 후보 지지율 두 가지가 필요합니다. 18 대 대선과는 달리 표 2 에 나타나 있듯이 60 대 이상과 50 대 (50~59)의 투표 성향이 달랐으므로 투표자들의 연령 그룹을 60 세 이상, 50 대, 50 세 미만 세 그룹으로 조정합니다. **모든 미분류표가 도장표라고 가정하면**, 즉 미분류표에 정상표가 전혀 없다고 가정하면, 세 가지 연령 그룹에서 미분류표를 발생시키는 확률이 각각 **6.6%, 2.9%, 2.5%**로 계산됩니다.

여기에서 유의할 점은 이러한 **확률값이 표 2 에 나와있는 연령별 투표율과 연령별 후보 지지율 두 가지에 의해 결정된다**는 사실입니다. 달리 말하자면, 연령별 후보 지지율에 대한 여론 또는 출구 조사 결과에 따라 확률값들이 바뀔 수 있고, 또한 이에 대한 신뢰할만한 자료가 없다면 확률값들을 구할 수 없게 된다는 뜻입니다. 즉, 노령층 가설은 수량화 검증이 어려운 내용으로서 미분류표들을 미지의 영역으로 남게 하는 면이 있습니다. 이 점은 미분류표의 투명성을 약화시켜 공정성 의혹의 근거가 될 수 있으므로, 도장표 방지가 공정성을 지키는 데에 절대적으로 필요하다고 봅니다.

표 2:연령별 투표율과 후보 지지율 (출구 조사 결과 [5])

19 대 대선	60 이상	50 대	50 미만
투표율	**24.4%**	**19.9%**	**55.7%**
후보 1	**23.4%**	**36.9%**	**52.3%**
후보 2	**48.4%**	**26.8%**	**9.4%**
후보 3	**23.1%**	**25.4%**	**19.4%**

그런데 이 확률들을 250 개 지역에 각각 적용해보면, 18 대 대선처럼 각 지역의 미분류 비율을 제대로 설명하지 못합니다. 예를 들면, 19 대 대선에서는 미분류표 비율이 실제로 3%~11% 사이로 나타난 반면, 가설 2 에 따른 연령별 확률로는 3%~5% 사이의 예상값이 나옵니다. 가설 2 가 심정적으로 받아들이기 쉽고 현실성이 있는 것으로 여겨지나, 충분한 설명을 하지 못하는 것으로 판단하게 됩니다.

셋째, 가설 2 는 60 세 이상과 50 대 투표자들이 50 대 미만 투표자들보다 미분류표를 상대적으로 더 많이 만들어 낸다고 설명하고 있으며, 앞에서 살펴본 확률로도 차이가 있음을 보았습니다. 가설 2 가 옳고 또한 출구 조사 연령별 지지율이 (표 2) 신뢰할만 하다고 가정하면, **19 대 대선에서 나타난 후보 1 과 후보 3 의 미분류율을 이해하기 어렵게 됩니다**. 표 2 에 의하면 후보 3 은 50 세 이상 (60 세 이상 포함) 연령 그룹에서 후보 1 보다 지지를 덜 받은 것으로 나타났기 때문에 미분류율이 후보 1 보다 작게 나타나야 하는데 실제로는 더 높게 나타났습니다 (그림 6). 따라서 이 현상도 가설 2 로 설명되지 않습니다.

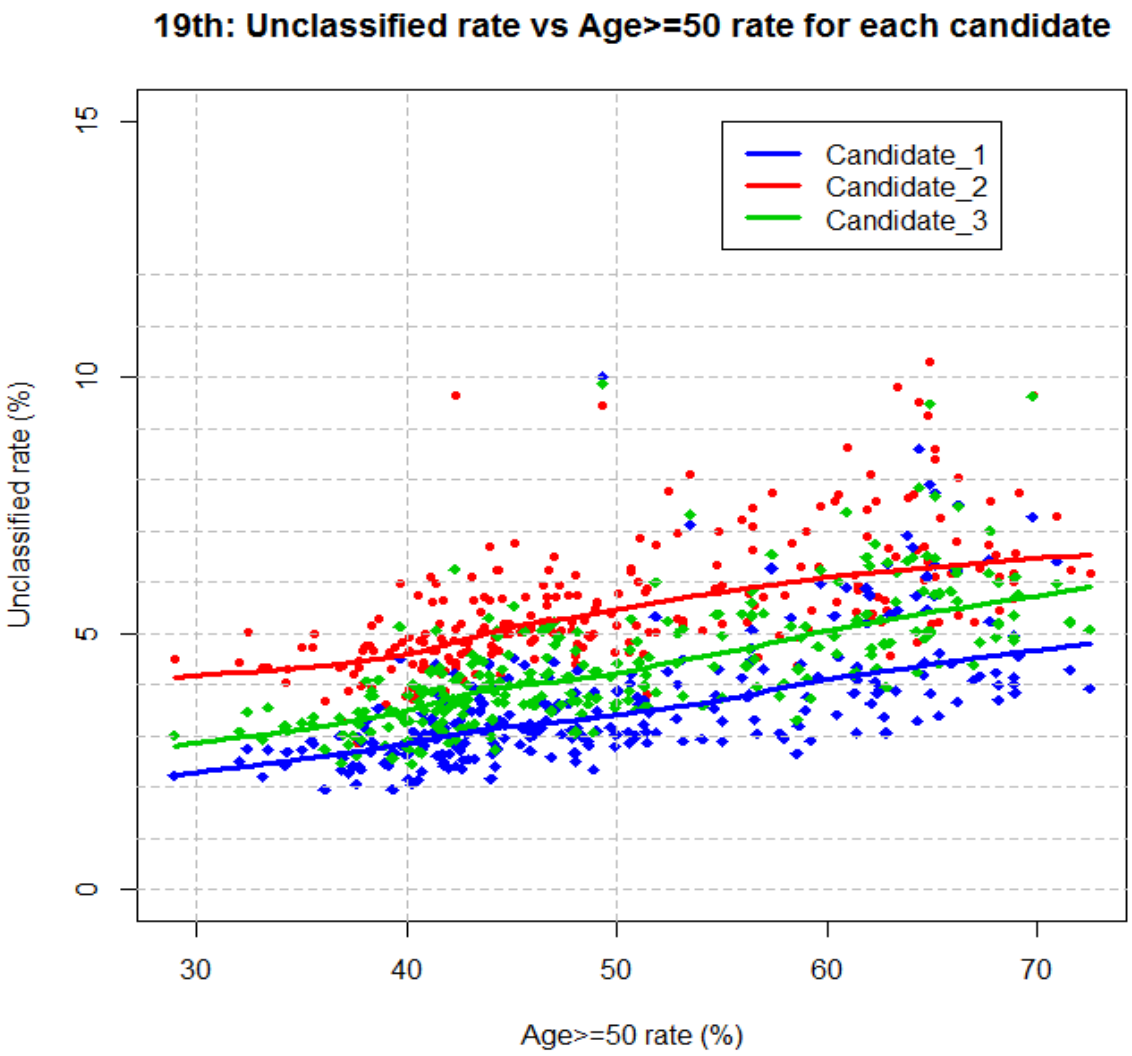

그림 6: 19 대선에서 미분류율과 50 세 이상 노령층 비율 사이의 관계: (1)x 값: 각 지역에서 50 세 이상 투표자들의 비율 (%); (2) y 값: 후보별 미분류율= M2/(M1+M2) in blue, H2/(H1+H2) in red or A2/(A1+A2) in green.

5. 제안과 결론

1) K=1 이 정상인가?

일반적으로 공정한 선거라고 할 때, 연령이 높다고 미분류표를 더 많이 만들어 내고, 또는 지지하는 후보에 따라 미분류표 비율이 달라지는 것을 정상적이라고 생각하기 어렵습니다. 공정한 선거와 공정한 선거시스템을 구별할 필요가 있습니다. 18 대와 19 대 대선 결과를 보면, 한국의 투표 방법이 (투표지 디자인과 도장 크기 등등) 미분류표 발생 과정에 많은 영향을 미치는 것으로 보입니다. 이것은 선거시스템의 문제입니다. 공정한 선거를 견인할 수 있는 공정한 선거시스템이

아니라고 볼 수 있습니다. 따라서 연령이나 후보에 따라 미분류율이 달라지는 것을 당연하다고 받아들이기 전에, 왜 달라지고 있는지 알아보고 그 원인을 제거함으로써 모든 후보의 미분류율이 같아지도록 개선하는 것이 절대적으로 필요합니다. 즉 K=1 이 정상이 되는 투개표 환경을 만들어 나가면서, 전자개표기 사용을 감독할 수 있는 여건을 만들어 나갈 때에만 공정한 선거를 통계적으로 확인할 수 있게 됨을 강조합니다.

예를 들어 전자개표기를 사용하는 한국의 투개표 시스템과 비슷한 선거 시스템을 운영하는 다른 나라가 있다고 가정합니다. 그 나라에서는 투표지 디자인과 투표 표시 방법을 한국과 다르게 사용하고 있습니다. 이 선거에 κ 를 적용하면 어떠한 결과가 나올까요? κ 의 어떤 값이 정상이 될까요? 한국뿐만 아니라 다른 나라의 선거까지 고려하면서 일반적 선거 검증 도구로 κ 가 적용된다면, K=1 이 정상이라는 전제 아래 선거 결과를 분석하는 것이 적절할 것입니다.

다른 예를 들자면, κ 는 전체 미분류율과 상관이 없이 적용될 수 있습니다. 18 대와 19 대 대선에서 전국 합하여 대략 4% 정도의 미분류율이 나왔는데, 만약 이 미분류율이 20%로 커진다고 하더라도 κ 값이 1 에 가까이 나온다면 정상적 선거가 이루어졌다고 판단할 수 있습니다. 전자개표기가 낙후되어 또는 준비가(calibration) 잘못되어 미분류율이 높아진다하더라도 각 후보별 미분류율이 같다면 (즉 K=1), 기계적 문제일 뿐 개표의 공정성 문제는 아닙니다. 이 점 역시 한국에서 K=1 이 정상이 되는 선거 환경을 만들어가야 하는 이유가 됩니다.

2) K=1 이 되는 선거 환경을 만들 수 있을까?

결론부터 말하자면 **가능합니다**. 노령층 가설은 두 가지를 가정하고 있습니다. (1) 노령층이 도장표를 만든다; (2) 노령층이 특정 후보를 지지한다. 여기에서 (2)는 투표권을 가진 누구에게나 주어지는 권리이므로 반론의 여지가 없습니다. 그런데 (1)이 뜻하는 것은 "노령층은 도장을 제대로 찍지 못한다"라는 의미입니다. 노령층 가설과는 별도로 도장표를 만드는 투표자들의 특성에 따르자면, "보수는 도장을 제대로 찍지 못한다"[2], 또는 "특정 후보의 지지자들은 도장을 제대로 찍지 못한다"라는 해석이 나오기도 합니다.

그런데 (1)의 도장표는 방지할 수 있습니다. 즉 노령층도, 보수도, 또는 특정 후보 지지자들도 불필요한 책임에서 벗어날 수 있습니다. 투표지 디자인과 도장찍는 방법 등등 투표 방식의 개선을 통해서 가능합니다. 현재 사용되는 투표지와 도장을 그대로 사용한다면, 그림 7 에 있는 도구를 사용하여 지지하는 특정 후보에 도장을 제대로 찍을 수 있도록 도움으로써 도장표를 방지할 수 있습니다. 이 외의 여러가지 방법으로 도장표 방지가 가능할 것입니다.

다른 한 편, 도장표를 방지하여 K=1 이 정상이 되는 선거 환경을 만든다면, 사전투표 감시 또는 개표 과정 감시에 따르는 비용과 수고를 줄이게 되는 장점이 생깁니다. 즉 K-값만 개표 후 확인하고, K-값이 1 보다 큰 개표소에 대한 검증만 하면 됩니다.

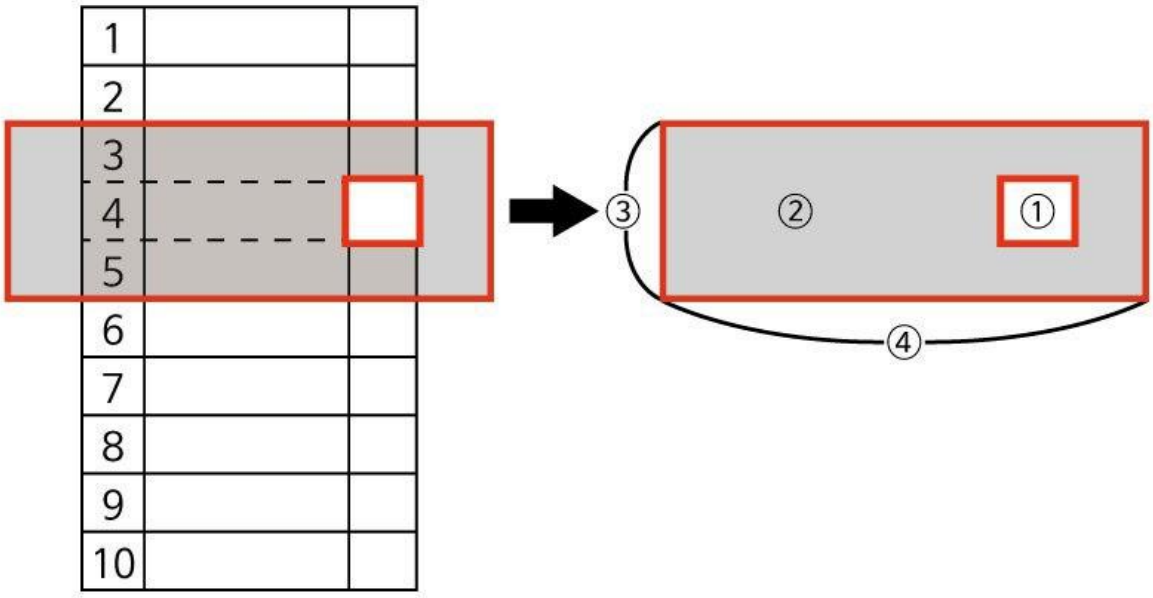

그림 7: 투표지 도장 제대로 찍기를 돕는 도구

1) 도장을 찍을 수 있는 구멍.
2) 투명한 카버로 투표지를 볼 수 있으나 도장을 찍을 수 없음.
3) 높이는 후보 3 명을 카버할 수 있을 정도.
4) 넓이는 손으로 만질 여백을 위해 투표지보다 넓게.

(아이디어: Kim, MH, 디자인: 프로젝 부)

3) 18 대 대선의 K=1.5 값이 의도된 것인가? (Intentional or Unintentional?)

K=1.5 를 설명하는 방법으로 <더플랜>에서 제시한 방법과 (시나리오) 가설 2 사이에 큰 차이점이 있습니다. <더플랜>에서는 정상표가 미분류로 간 것으로 간주하고 플랜에 의해 **의도적으로 K=1.5** 이라는 결과가 나온 것으로 설명한 반면에, 가설 2 는 어떠한 **의도 없이 노령층 도장표로 K=1.5** 를 설명하려고 합니다. 여기에서 비교 판단하는 근거는 두 가지 방식이 얼만큼 실제 대선 결과를 잘 반영할 수 있느냐는 것입니다. 왜냐하면 가설을 세울 때의 궁극적 목적은 선거 결과를

재구성하는 것입니다. 가설 2 로는 18 대와 19 대 대선을 제대로 재구성하지 못함을 앞에서 밝혔습니다. 반면에 <더플랜>에서 제시한 방법은 18 대 대선 재구성을 제대로 했습니다 (표 1).

<더플랜>에서 제시한 것은 18 대 대선이 의도적이었다고 단정적 결론을 내린 것이 아니라, **K 값은 승리방정식에 의해 미리 설정될 수 있고 (<더플랜>에서 언급되지 않았음),** 이러한 플랜에 의해 전자개표기를 프로그램하면 **K=1.5** 가 만들어질 수 있음을 시뮬레이션을 통해 보였고 동시에 그것이 가능함을 실험을 통해 보였습니다. 또한 미분류율에 영향을 미치는 다른 요인이 있을 수 있으므로, **K=1.5** 가 만들어지는 방법이 다양할 수 있다고 <더플랜>에서 언급하였습니다. 통계적 분석으로는 증상만 판단할 뿐 **(K=1.5)** 원인을 확정할 수 없다고 그 한계도 밝혔습니다. 누가 어떻게 시도했느냐를 밝히는 것은 통계적 분석 범위를 넘어서는 것이므로, 전자개표기를 사용하는 한국 선거 시스템의 잠재적 문제성을 인식하고 개선해 나가자는 것이 <더플랜>의 핵심임을 다시 한 번 밝힙니다. 따라서 어떠한 의도도 없이 선거 결과를 제대로 재구성 할 수 있도록 가설 2 가 보완이 되든지 또는 다른 가설이 나오지 않는 한, <더플랜>에서 제시한 의도성은 여전히 남아있다고 할 수 있겠습니다. 즉, 어떤 가설로도 **K=1.5 가** 자연스러운 현상이라고 제대로 설명할 수 없다면, <더플랜>에서 제시한 방법이 마지막 카드로 남는 것이 되겠습니다.

하지만 가설 2 로 18 대 또는 19 대 대선 결과를 전체는 아닐지라도 일부를 설명할 수 있음을 확률을 통해 보았습니다. 만일 가설 2 가 전적으로 옳다면, 즉 미분류표에 정상표는 소수이고 도장표가 대부분 이라면, 한국 사회의 고령화에 따라 점점 미분류표는 많아질 수밖에 없고, 이러한 면이 공정한 개표를 검증할 수 없게 만드는 원인이 될 수 있습니다. 현재의 투표 시스템을 정밀하게 분석하여, 앞으로 노령층에 의한 도장표가 사라지도록 개선해 나가야 할 것입니다. 미분류표가 만들어지는 책임을 노령층에 돌리지 말고, 투표 방식을 개선해 연령에 상관없는 투표 시스템을 구축해나가야 공정한 선거를 치를 수 있게 됩니다.

<더플랜>에서 제시한 시나리오에서는 연령별 후보 지지율을 고려하지 않았습니다. 이에 대한 정보는 방송사 또는 신문사 조사 결과에만 전적으로 의지해야 하는데 일관되지 않은 결과들이 나오므로 신뢰도에 문제가 있다고 판단했고 또한 지역별 자료는 없거나 샘플 수가 작으므로 시뮬레이션에 반영하지 않았습니다. 즉, 확실히 신뢰할 수 있는 자료만 참고하였습니다. 그럼에도 불구하고 **<더플랜> 시나리오를 구상할 당시 연령별, 성별 분포와 K 값들을 살펴보았으나, 연령별 후보별 지지율을 고려하지 않았던 점은 본 연구의 한계였습니다.**

끝으로 <더플랜>의 K 를 5 년마다 실시하는 대선에 계속 활용해 나가기를 기대합니다. 당선자와 상대 후보를 비교 판단하는 기준으로 적용하고, 특히 K>1 인 경우 노령층 가설 또는 다른 가설들로 충분히 설명할 수 있어야만 선거 분석이 완성된다고 할 수 있겠습니다. 다른 한 편, <더플랜>의 K 를 대선마다 적용하면, 전자개표기를 통해 시도하려는 선거 부정을 미리 방지하는 예방 효과도

있습니다. 앞에서 살펴본 것들을 모두 종합하자면, K=1 이 될 수 있는 선거시스템으로 개선하여, 한국의 투개표 방식을 감시하고 공정한 선거 민주주의를 이끌어 가는 것이 바람직하다는 결론을 내립니다.

6. 자료 출처

1) 김어준, 최진성, 더플랜, Available at https://youtu.be/aGGikPMNn2w
2) 뉴스타파, 19 대선 문-홍 K 값은 1.6…정규분포, Available at http://newstapa.org/40171
3) 중앙선거관리위원회, 제 18 대 대통령선거 개표부정 의혹 영화 (더 플랜)에 대한 입장 Available at http://nec.go.kr/portal/bbs/view/B0000342/35474.do?searchYear=2017&menuNo=200470
4) 행정자치부, 연령별 인구통계, Available at http://rcps.egov.go.kr:8081/ageStat.do?command=month
5) 경향신문, 방송 3 사 출구조사

참고 1: 19 대 대선 자료를 공유해주신 뉴스타파에 감사드립니다.
참고 2: <더플랜> 연구팀이 18 대 대선 자료 검증을 하였으나, 19 대 대선 자료 검증은 하지 않았습니다.

부록 1 : K-test 에 필요한 기대값과 분산

Here B represents a binomial distribution, and all equations are meant for each electoral district.

(A) Null hypothesis: This assumes an equal unclassified rate between two candidates, i.e. r= $Pr(U_M)=Pr(U_P)$.

(1) Expectation of K

- P1 ~ B(P, 1-r) and P2 ~ B(P, r).
- M1 ~ B(M, 1-r) and M2 ~ B(M, r).

Since not all votes are sorted to the classified (r=0) or unclassified (r=1), we have $0 < r < 1$.

It is known that if X ~ B(n, p), $E\left[\frac{1}{X+a}\right] = \int_0^1 t^{a-1} \cdot P_x(t)dt$, for $0 < p < 1$, where *E* represents expectation and $P_x(t)$ is the probability generating function. We thus have

$$E\left[\frac{1}{X+1}\right] = \int_0^1 t^0 \cdot (q+pt)^n dt = \frac{1-q^{n+1}}{(n+1)p}, \text{ where } q = (1-p) \text{ [A1]}.$$

Therefore $E[K] = 1$ asymptotically. (More details are available in the Appendix at http://www.projectboo.com/archive/153353.

(2) Variance of K

In general, there are no closed-form expressions for $E\left[\frac{1}{(X+1)^a}\right]$ for a binomial variable X and a constant a>1, but asymptotic results are available. Cribari-Neto et al. [A2] suggested an approximation:

$$E\left[\frac{1}{(X+1)^a}\right] = E\left[\frac{1}{(X)^a}\right] = (np)^{-a} + \left(\frac{a-1}{2p} - \frac{a+1}{2}\right)\frac{\Gamma(a+1)}{\Gamma(a)}\frac{1}{n^{a+1}p^a} \text{ for large X.}$$

Since P1 ~ B(P, 1-r) and M2 ~ B(M, r), we get

$$E\left[\frac{1}{(P1)^2}\right] \cong \frac{1}{\left(P(1-r)\right)^2}\left(1 + \frac{1}{P(1-r)} - \frac{3}{P}\right),$$

$$E\left[\frac{1}{(M2)^2}\right] \cong \frac{1}{(Mr)^2}\left(1 + \frac{1}{Mr} - \frac{3}{M}\right).$$

Applying the above to the K, we get asymptotically

$$\text{Var(K)} \cong \frac{(P+M)r(1-r)+1}{PM(r(1-r))^2} \cong \frac{(P+M)}{PMr(1-r)}.$$

(B) Generalized hypothesis: If there exist external factors, which could lead to different unclassified rates of the two candidates, the binomial distributions in the null hypothesis change to:

- P2 ~ B(P, α) and P1~B(P, 1- α) for $0 < \alpha < 1$.
- M2 ~B(M, β) and M1~B(M, 1- β) for $0 < \beta < 1$.

$E[K] = \frac{\alpha(1-\beta)}{(1-\alpha)\beta}$ asymptotically, which can be approximated by $\frac{\alpha}{\beta}$ if $\alpha \ll 1$ & $\beta \ll 1$.

References

[A1] M.T. Chao, and W.E. Strawderman, *Negative moments of positive random variables*, J. Amer. Statist. Soc. 67 (1972), pp. 429-431.

[A2] F. Cribari-Neto, N.L. Garcia, and K.L.P. Vasconcellos, *A note on inverse moments of binomial variates*, Braz. Rev. Econom. 20 (2000), pp. 269-277.

부록 2 : K-impact, 승리방정식 (Impact of the national model parameter (K) on winning)

National Model: E[P2/M2]= K* (P1/M1), where K=1.5 for the 18th election (1)

As model (1) assigns more votes to candidate 1 than candidate 2 in the unclassified proportionally to the classified regardless of who wins in the classified, we face a critical question: ***what is the impact of the model parameter on election outcome?***

To answer the question, we let **L=P1/M1** and P2/M2=K*(P1/M1), where K ≥1 and L >0. Then P2/M2=K*L, and thus P1=L*M1 and P2=K*L*M2. For simplification without loss of generality, we take candidate 1's perspective of winning the election. If candidate 1 gets more votes than candidate 2, it clearly means that (P1+P2) > (M1+M2), and thus we set up candidate 1's winning condition as follows:

(L-1)*M1+(K*L-1)*M2>0. (2)

The only non-trivial case for candidate 1 winning via the unclassified ballots is as follows: If L<1 (i.e. P1<M1), then (K*L-1)*M2 > (1-L)*M1. Since M2>0, we have
(K*L-1) > (1-L)* (M1/M2). To simplify this inequality, we set **x=M1/M2**, where x is a ratio of two votes for candidate 2 between the classified (M1) and the unclassified (M2). Then equation (2), which is the candidate 1's winning condition, becomes a function of x and K such that
L > (1+x)/(K+x). This is the interesting case to be discussed further.

Candidate 1's winning depends on two variables, x and K, even if K >1. We fix K = 1.5 and examine the effect of the model parameter (K=1.5), varying the values of x. Since the votes for candidate 2 from the classified are larger than the unclassified (i.e. M1>M2 from the 18th election), the minimum value of x is 1 (x≥1).

As demonstrated in Figure A1, model (1) increases candidate 1's winning as the number of the unclassified ballots goes up. Since the unclassified are to be generated by the op-scan counters, the model parameter K value is linked to the op-scanner's operation. The impact of the model parameter (K) on candidate 1's winning would be maximized when candidate 1 lost in the classified ballots.

In the 18th election, x was observed close to 35 and thus L=0.99, which indicates candidate 1 could win the election with less votes from the classified (at least 99% of candidate 2) but more votes from the unclassified. This could be achieved by setting K=1.5 and thus elucidates the meaning of the model parameter 1.5 in the 18th presidential election.

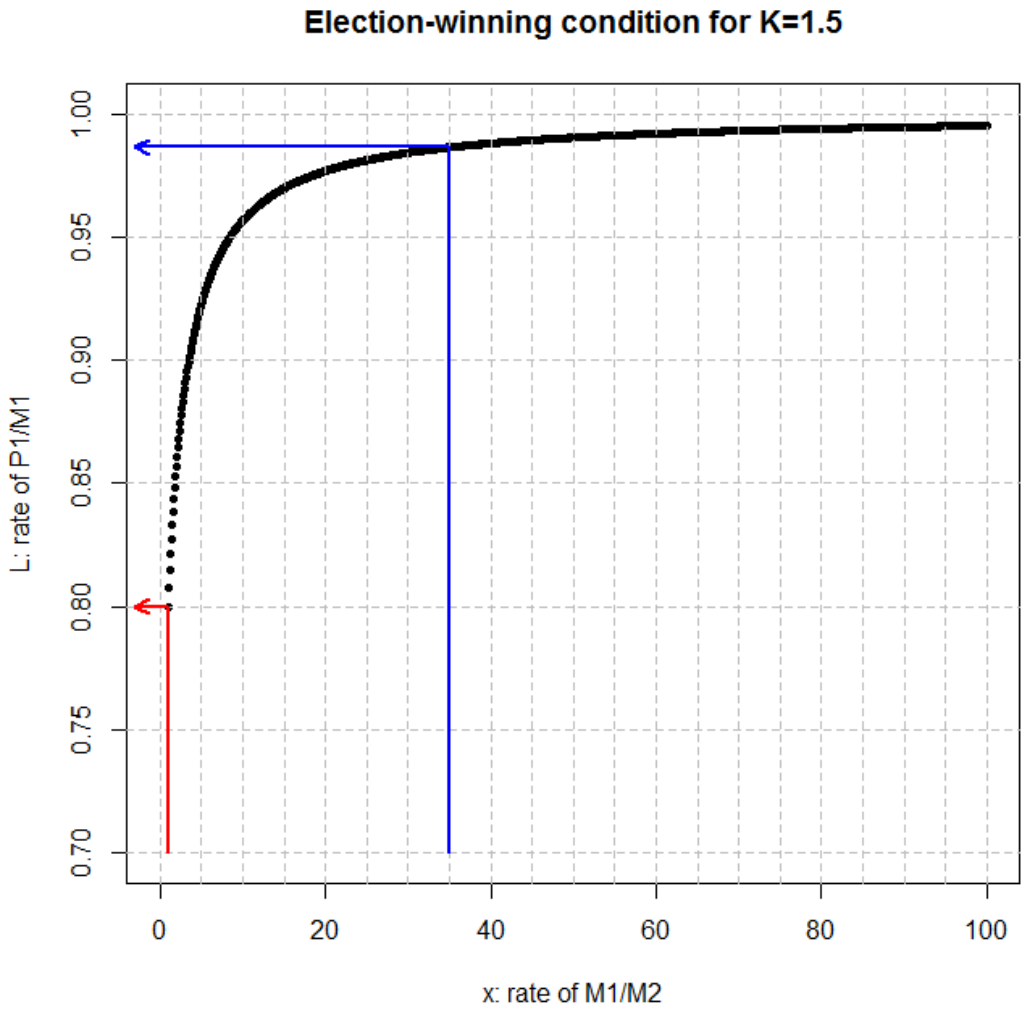

Figure A1: A plot on the impact of parameter 1.5 in model (1), satisfying the winning equation (2).