The Korean Journal of Applied Statistics (2013) DOI: http://dx.doi.org/10.5351/KJAS.2013.26.1.059
26(1), 59–70

Estimating the Number of Seats in Local Constituencies of a Party Using Exit Polls in the General Election

Ji-Hyun Kim[a,1]

[a]Department of Statistics and Actuarial Science, Soongsil University

(Received November 12, 2012; Revised December 8, 2012; Accepted December 13, 2012)

Abstract

Exit polls failed to estimate the number of seats in the National Assembly for each party in the 2012 General Election, even though they estimated it in interval. Three major broadcast companies jointly carried out exit polls, but made projections independently. The exact methods of projection were not publicly released. This paper proposes confidence intervals for the number of seats in local constituencies using the results of exit polls, and conducted simulation studies to assess the performance of the confidence intervals. The proposed confidence intervals were applied to the real data of 2012 General Election.

Keywords: Exit polls, multinomial distribution, Dirichlet distribution, p-value.

1. 서론

지난 2012년 4월 11일, 제19대 국회의원을 선출하는 총선이 있었다. 이번 총선에서 처음으로 지상파 방송 3사에서 공동으로 출구조사를 실시하였는데, 예측이 빗나가 "70억의 예산을 투입했음에도 불구하고 혼란만 키웠다"는 비난을 받았다(동아일보 2012.4.12 기사). 비난의 주된 이유 중의 하나는 정당별 예상의석수를 출구조사의 불확실성을 감안해 구간으로 추정하였음에도 불구하고 실제의석수가 이 구간을 벗어났다는 것이었다. 예를 들어 MBC는 새누리당과 민주통합당의 지역구 의석수를 각각 107∼127석, 108∼126석으로 예측했으나 실제 의석수는 각각 127석과 106석이었다. 정당별 의석수를 어떻게 예측하였을까? 출구조사에 의한 예측이 표집오차로 설명할 수 없을 정도로 잘못되었다면, 투표소와 투표자를 추출해 응답을 얻는 과정에 문제가 있었던 것일까, 아니면 구간 추정을 하는 분석 단계에 문제가 있었던 것일까? 본 연구는 이런 의문에서 출발하였다.

언론 보도에 의하면 지상파 3사가 출구조사는 공동으로 실시하였지만 지역구와 비례대표를 합한 정당별 의석수 예측은 각자 하였다고 한다. 정당별 지역구 의석수를 예측하는 문제는 지역구별로 당선자를 예측하는 문제와는 다른 문제인데, 사후분석을 통해 출구조사의 문제점을 지적한 연구들은 있었으나 (Kwak과 Kim, 2010; Lee, 2004; Hyun, 2005) 정당별 의석수를 예측하는 구체적 방법을 명시한 문헌은 찾을 수 없었다. 본 연구의 목적은 정당별 지역구 의석수를 예측 또는 추정하는 적절한 방법이 무엇인지를 살펴보는 데 있다. 어떤 가능한 방법들이 있고 이론적 근거는 무엇이며 성능은 어떤가에 대해 알아본다. 출구조사의 개선을 위해서는 정당별 의석수의 구간 추정 방법이 적절했는지를 먼저 따져봐야 한다. 적절한 근거에 의해 추정했음에도 불구하고 실제의석수가 신뢰구간을 벗어났다면 그 원인을 밝히고 대처 방안을 제시하는 것은 그 다음 단계의 작업이 되어야 할 것이다.

[1]Professor, Department of Statistics and Actuarial Science, Soongsil University, 369 Sangdo-Ro, Dongjak-Gu, Seoul 156-743, Korea. E-mail: jxk61@ssu.ac.kr

2. 정당별 실제의석수에 대한 점추정과 구간추정

출구조사에서 얻은 결과로부터 정당별 의석수를 구간으로 추정하는 방법에 대해 서술한다. 먼저 필요한 기호에 대해 설명한다.

$\hat{p}_1 =$ 지역구 i에서 관심 있는 정당 (관심 있는 정당을 정당 A라고 하고 첫 번째 범주에 두기로 한다) 후보의 출구조사에 의한 관측득표율. 지역구를 나타내는 첨자 i를 생략함.

$\hat{p}_M =$ 지역구 i에서 정당 A 후보를 제외한 후보들의 관측득표율 중에서 최댓값
$= \max(\hat{p}_2, \ldots, \hat{p}_{K_i})$, K_i는 지역구 i에서 출마한 입후보자의 수.

$n =$ 해당 지역구의 출구조사 표본크기. 역시 지역구를 나타내는 첨자 i를 생략함.

246개 지역구 각각에 대해 관측값 $\hat{p}_1$과 $\hat{p}_M$을 얻으면 다음 확률변수의 값을 계산할 수 있다.

$X_i =$ 지역구 i에서 정당 A의 후보가 당선 예측이면 (즉, $\hat{p}_1 > \hat{p}_M$이면) 1, 아니면 0인 이항변수. 정당 A가 후보를 내지 않는 지역구가 있을 때는 $X_i = 0$으로 정의하며 X_i의 분산도 0이다.

$S = \sum_{i=1}^{246} X_i$: 정당 A의 지역구 의석수에 대한 점추정값 (예상의석수라 부르기로 한다).

이제 정당 A의 실제의석수에 대한 추정 문제는 S의 기댓값인 $E(S)$의 추정 문제로 귀결된다. $E(S)$의 정확한 해석 또는 의미에 대해서는 뒤에 다시 언급하기로 한다.

실제의석수에 대한 신뢰구간을 구하는 방법에 대해 알아보자. 먼저 X_i의 분포는 베르누이 분포로서 $X_i \sim b(1, P_i)$로 표현한다. 이 때 $P_i = E(X_i) = P(X_i = 1) = P(\hat{p}_1 > \hat{p}_M)$로서 P_i는 0과 1 사이의 상수이다 (아예 P_i를 추정해야 할 상수가 아닌 확률변수로 간주하여 의석수를 추정하는 베이즈(Bayes) 접근법도 가능하다. 베이즈 접근법은 출구조사 이전에 이루어진 선거여론조사의 결과를 사전분포를 통해 추정에 반영할 수 있다는 점에서 매력적이지만, 본 연구에서는 베이즈 접근법을 시도하지 않았다.). 만약 지역구 i에서 정당 A가 후보를 내지 않았다면 $P_i = 0$으로 정의한다.

$$E(S) = \sum_{i=1}^{246} P_i, \qquad V(S) = \sum P_i(1 - P_i)$$

이며

$$\frac{S - E(S)}{\sqrt{V(S)}} \approx N(0, 1)$$

임을 이용하여 (Casella와 Berger, 2002, 연습문제 4.36 참조) $E(S)$에 대한 신뢰수준 $100(1-\alpha)\%$의 근사적인 신뢰구간을 구할 수 있다. 즉,

$$S \pm z_{1-\frac{\alpha}{2}} \sqrt{\hat{V}(S)}, \qquad \hat{V}(S) = \sum \hat{P}_i \left(1 - \hat{P}_i\right). \tag{2.1}$$

이 때 P_i의 추정이 중요한데 P_i의 추정량에 대해 알아본다.

먼저 P_i에 대해 살펴보면, $P_i = P(\hat{p}_1 > \hat{p}_M)$로서, 기호 p_1을 관심 있는 정당 후보의 실제득표율, p_M을 이 후보를 제외한 나머지 후보들의 실제득표율 중에서 최댓값이라고 할 때, P_i는 p_1과 p_M의 값에 따라 결정된다. 만약 $p_1 \gg p_M$이면 (즉, $p_1 - p_M$의 차가 크면) P_i는 1에 가까운 값을, $p_1 \ll p_M$이면 0에 가까운 값을 갖게 될 것이다. P_i의 추정을 위해 다음과 같은 가설 검정 문제를 생각해 보자.

$$H_0 : p_1 - p_M = 0, \qquad H_1 : p_1 - p_M > 0.$$

이 때 p값이 작을수록 H_0를 기각하고 H_1을 채택할 증거가 많음을 나타낸다. 따라서 '$1-p$값'은 H_1을 지지하는 증거의 정도이며, $\hat{p}_1 - \hat{p}_M$이 0이면 0.5, $\hat{p}_1 - \hat{p}_M$이 양의 값을 가지면서 커지면 1에 수렴하고, $\hat{p}_1 - \hat{p}_M$이 음이면서 절대 값이 커지면 0에 수렴하므로 P_i에 대한 합리적 추정량의 기본 조건을 갖추었다고 볼 수 있다. P_i의 추정값으로 위와 같은 가설검정 문제의 $1-p$값을 쓰면 되는 근거에 대해 다음과 같이 베이즈 추론(Bayesian inference)으로 생각해 볼 수도 있다.

두 확률변수 Y와 θ의 분포가

$$Y|\theta \sim N\left(\theta, \sigma^2\right), \quad \theta \sim N\left(\mu, \tau^2\right)$$

이며 θ를 제외한 σ^2, μ, τ^2가 상수라면, θ의 사후분포는

$$\theta|Y = y \sim N\left(\mu_p, \sigma_p^2\right), \quad \mu_p = \frac{\tau^2}{\tau^2+\sigma^2}y + \frac{\sigma^2}{\sigma^2+\tau^2}\mu,\ \sigma_p^2 = \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}$$

임이 알려져 있다. 이 때 θ에 대한 사전정보가 전혀 없어 $\sigma^2/\tau^2 = 0$로 두면, $\mu_p = y$, $\sigma_p^2 = \sigma^2$이 되어

$$\theta|Y = y \sim N\left(y, \sigma^2\right)$$

이 된다. $Y = \hat{p}_1 - \hat{p}_M$, $\theta = p_1 - p_M$으로 치환하면

$$p_1 - p_M|\hat{p}_1 - \hat{p}_M \sim N\left(\hat{p}_1 - \hat{p}_M, \sigma^2\right) \tag{2.2}$$

이다. 위 식에서 σ^2은 $p_1 - p_M$의 값이 주어졌을 때 $\hat{p}_1 - \hat{p}_M$의 분산이다. 따라서 출구조사 결과가 주어졌을 때 당선 사후확률을 나타내는 $P(p_1 - p_M > 0|\hat{p}_1 - \hat{p}_M)$은 식 (2.2)에 의해

$$P\left(p_1 - p_M > 0|\hat{p}_1 - \hat{p}_M\right) = P\left(Z > \frac{-(\hat{p}_1 - \hat{p}_M)}{\sigma}\right) = P\left(Z < \frac{(\hat{p}_1 - \hat{p}_M)}{\sigma}\right)$$

이 되는데 (Z는 표준정규 확률변수), 위 식의 마지막 항은 빈도론자(frequentist)의 시각으로 봤을 때 $H_0 : p_1 - p_M = 0$, $H_1 : p_1 - p_M > 0$라는 가설검정 문제에서의 p값과 관련이 있으며, σ를 H_0가 참일 때 $\hat{p}_1 - \hat{p}_M$의 표준오차로 간주하면 정확히 $1-p$값이 된다. 즉, $p_1 - p_M$의 사전분포와 $\hat{p}_1 - \hat{p}_M$의 분포가 정규분포라면 사후확률 $P_i = P(p_1 - p_M > 0|\hat{p}_1 - \hat{p}_M)$가 빈도론자의 시각에서 본 $1-p$값이 됨을 알 수 있다.

이제 p값을 추정하는 방법에 대해 생각해 보자. $\hat{p}_1 - \hat{p}_M$의 분포가 정규분포라면

$$\frac{(\hat{p}_1 - \hat{p}_M)}{\sqrt{\hat{V}(\hat{p}_1 - \hat{p}_M)}} \approx N(0,1) \tag{2.3}$$

일 것이며, 이로부터 p값을 추정할 수 있다. 이 때 $\hat{p}_1 - \hat{p}_M$의 표준오차의 식으로 두 가지를 생각해 볼 수 있다. 먼저 $\hat{p}_1$과 $\hat{p}_M$을 단순히 다항분포의 두 범주의 관측확률로 간주해서

$$\sqrt{\hat{V}(\hat{p}_1 - \hat{p}_M)} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1) + \hat{p}_M(1-\hat{p}_M) + 2\hat{p}_1\hat{p}_M}{n}} \tag{2.4}$$

의 식을 쓸 수 있고, 다음으로 가설 H_0이 참일 때 $p_1 = p_M$이므로 식 (2.4)에서 $\hat{p}_1$과 $\hat{p}_M$ 대신에 $\bar{p} = (\hat{p}_1 + \hat{p}_M)/2$을 치환하여 다음과 같은 식을 쓸 수도 있다.

$$\sqrt{\hat{V}(\hat{p}_1 - \hat{p}_M)} = \sqrt{\frac{2\bar{p}(1-\bar{p}) + 2\bar{p}^2}{n}} = \sqrt{\frac{2\bar{p}}{n}}\,. \tag{2.5}$$

뒤에서 설명할 모의실험을 실시한 결과 두 표준오차의 식 (2.4)와 (2.5)에 따른 차이는 무시할 수 있을 정도로 작았다. 따라서 표준오차 식 (2.5)에 의한 결과만 보고하기로 한다.

다항분포에서 순서통계량이 포함된 $\hat{p}_1 - \hat{p}_M$의 정확한 분포를 해석적으로 밝히기가 힘들다. 이런 이유로 식 (2.3)과 같이 $\hat{p}_1 - \hat{p}_M$이 정규분포를 따른다는 가정을 하지 않고 p값을 몬테칼로 방법에 의해 구하는 방법도 가능하다. 그 절차를 설명하면 다음과 같다.

(1) 각 지역구의 출구조사에서 관측한 득표율(예상득표율 또는 관측득표율)을 참득표율(또는 실제득표율)로 간주하되 관심 있는 정당의 득표율과 그 정당을 제외한 정당의 득표율 중에서 최댓값을 H_0를 만족하는 값인 $(\hat{p}_1 + \hat{p}_M)/2$으로 같이 둔다.

(2) 단계 (1)에서 정해진 참득표율과 출구조사의 표본크기 n명을 시행횟수로 하는 다항분포 난수를 생성하고, 여기서 $p_1 - p_M$의 값을 구한다.

(3) 단계 (1)과 (2)를 N번 반복한다.

(4) 단계 (3)에서 얻어진, H_0를 만족하는 N개의 $p_1 - p_M$값 중에서 $\hat{p}_1 - \hat{p}_M$ 보다 큰 값의 비율을 p값으로 간주한다 (N의 크기를 200부터 1000 사이의 값으로 변화시켜가며 실험해 보았는데 그 차이는 무시할 수 있을 정도로 작았다).

이상 논의를 요약하면, 정당별 실제의석수에 대한 신뢰구간은 식 (2.1)에 의해 구해지며, $\hat{P}_i$을 구하는 방법으로, $\hat{p}_1 - \hat{p}_M$의 정규분포를 가정하고 구하는 방법과 정규분포를 가정하지 않고 구하는 방법이 있다. 이 두 방법의 성능을 알아보기 위해 모의실험을 실시하였다.

3. 신뢰구간의 성능을 알아보기 위한 모의실험

식 (2.1)은 $E(S)$에 대한 신뢰구간인데 $E(S) = \sum_{i=1}^{246} P_i$의 정확한 해석에 대해 생각해 보자. 앞 절에서 지적했듯이, $P_i = P(\hat{p}_{1i} > \hat{p}_{Mi})$로서 0에서 1 사이의 값을 갖는다 (첨자 i는 지역구를 나타낸다). 한편, 실제의석수를 $\sum_{i=1}^{246} I(p_{1i} > p_{Mi})$로 표현할 수 있는데 (여기서 I는 표시함수로서 괄호 안에 주어진 조건이 참이면 1, 거짓이면 0의 값을 가짐) 실제득표율 p_{1i}와 p_{Mi}는 확률변수가 아닌 상수이므로 $\sum_{i=1}^{246} P(p_{1i} > p_{Mi})$로 표현할 수도 있다. $P(\hat{p}_{1i} > \hat{p}_{Mi}) \neq P(p_{1i} > p_{Mi})$이므로 $E(S)$가 정확히 실제의석수가 되지는 않는다. 따라서 식 (2.1)의 $E(S)$에 대한 신뢰구간이 실제의석수를 제대로 구간 추정하는지에 대해 실증적으로 살펴볼 필요가 있다. 또한 P_i를 추정하는 두 가지 방법에 따라 성능에 차이가 있는지에 대해서도 모의실험을 통해 살펴본다. 모의실험의 절차는 다음과 같다.

(1) 디리클레(Dirichlet) 분포로부터 K명의 후보의 참득표율을 생성하고 정당 A후보의 당선여부를 판단한다. 이 때 후보자(또는 범주)의 수 K를 지역구마다 달리 할 수 있으나 이 절의 모의실험에서는 고정시켰다.

(2) 단계 (1)에서 생성한 참득표율과 출구조사 표본크기인 n번의 시행횟수를 모수로 갖는 다항분포로부터 K명의 후보의 출구조사 득표수를 생성하고 관측득표율을 계산한다. 출구조사의 표본크기 n도 지역구마다 달리 할 수 있으나 같은 값으로 고정시켰다.

(3) 단계 (2)에서 생성한 관측득표율로부터 2절에서 설명한 두 가지 추정방법으로 P_i를 추정한다.

(4) 단계 (1)–(3)을 246개 지역구에 대해 차례대로 적용하고 그 결과로부터 참의석수(또는 실제의석수)와 예상의석수를 계산하고 신뢰구간을 구한다.

(5) 단계 (1)–(4)를 1000번 반복해서 얻은 결과로부터 실제신뢰수준과 신뢰구간의 평균 길이를 계산한다.

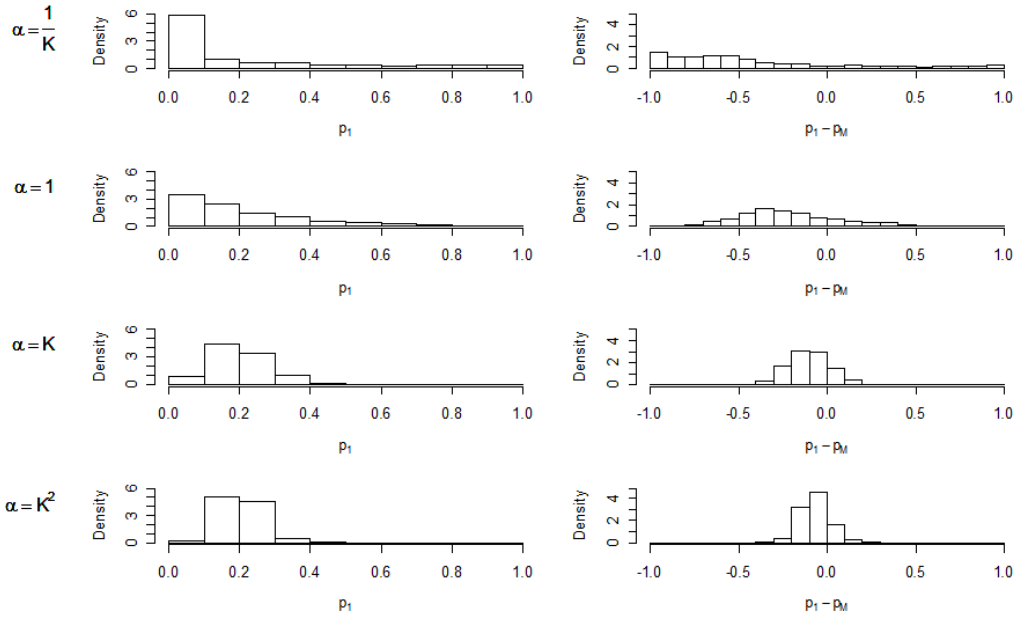

Figure 3.1. The distributions of p_1 and $p_1 - \max(p_2, \ldots, p_K)$ where $(p_1, \ldots, p_K)$ denotes a random number from a symmetric Dirichlet distribution with parameter α: $K = 5$

(6) 단계 (1)–(5)를 다양한 상황에서 실행한다. 각 지역구의 후보자 수 K의 값을 3, 4, 5, 6으로, 대칭형 디리클레 분포의 모수 α의 값을 $1/K$, 1, K, K^2으로, 출구조사의 각 지역구 표본크기 n도 500과 2800으로 변화시켜 다양한 모의실험 상황을 만든다.

다항분포의 확률 $(p_1, \ldots, p_K)$을 나타내는 난수를 생성하기 위해 대칭형 디리클레 분포(symmetric Dirichlet distribution)를 이용하였다. 디리클레 분포의 확률밀도함수는

$$f(p_1, \ldots, p_{K-1}) = \frac{1}{B(\alpha_1, \ldots, \alpha_K)} \prod_{i=1}^{K} p_i^{\alpha_i - 1}, \quad p_i > 0,\ p_1 + \cdots + p_K = 1,\ \alpha_i > 0$$

로서 $B(\alpha_1, \ldots, \alpha_K) = \prod_{i=1}^{K} \Gamma(\alpha_i) / \Gamma(\sum_{i=1}^{K} \alpha_i)$이다. 대칭형 디리클레 분포는 $\alpha_1 = \cdots = \alpha_K = \alpha$인 특수한 디리클레 분포로서 확률변수 p_i에 대한 주변분포(marginal distribution)가 모든 i에 대해 동일하다. 모수 α의 값이 커질수록 p_i의 분산이 작아져 평균인 $1/K$ 근처에 밀집하게 되며, 따라서 확률변수 $p_1 - \max(p_2, \ldots, p_K)$의 값이 0 근처에 많이 분포하게 되어 경합이 치열해지는 상황이 된다. 모의실험의 조건들을 쉽게 파악하기 위해 α의 값의 변화에 따른 p_1과 $p_1 - \max(p_2, \ldots, p_K)$의 분포를 Figure 3.1에 히스토그램으로 나타내어 보았다.

통계량 $\hat{p}_1 - \max(\hat{p}_2, \ldots, \hat{p}_K)$의 분포가 정규분포라는 가정 (2.3)은 $\hat{P}_i$를 구할 때 필요한 가정이다. 2000개의 $(p_1, \ldots, p_K)$ 난수로부터 구한 2000개의 $p_1 - \max(p_2, \ldots, p_K)$의 값으로 정규성 검정을 해 보면 정규분포라는 가설이 모수 α의 값에 따른 네 가지 경우 모두 기각되며 (R의 함수 shapiro.test() 적용 결과 p값은 모두 0.01보다 작았다), 500개의 난수로 정규성 검정을 하면 $\alpha = K^2 = 25$인 경우를 제외하고는 모두 기각되어 정규분포 가정은 하기 어렵다는 사실을 알 수 있다. Figure 3.1은 2000개의 난수로 그린 것이다.

Table 3.1. The performance of two confidence intervals: $K = 5$, $n = 2800$

명목 신뢰수준	α	RMSE: S	RMSE: $\sum \hat{P}_i$	신뢰구간1의 실제신뢰수준	신뢰구간1의 길이	신뢰구간2의 실제신뢰수준	신뢰구간2의 길이
0.95	0.2	0.89	0.77	0.898	2.84	0.899	2.83
		(0.03)	(0.02)	(0.010)	(0.03)	(0.010)	(0.03)
	1	1.51	1.29	0.916	5.04	0.913	5.00
		(0.06)	(0.05)	(0.009)	(0.03)	(0.009)	(0.03)
	5	2.32	2.22	0.917	8.05	0.910	7.88
		(0.11)	(0.11)	(0.009)	(0.03)	(0.009)	(0.03)
	25	3.58	5.23	0.918	12.41	0.909	11.81
		(0.23)	(0.29)	(0.009)	(0.02)	(0.009)	(0.02)
0.99	0.2	0.88	0.73	0.970	3.78	0.972	3.78
		(0.03)	(0.02)	(0.005)	(0.04)	(0.005)	(0.04)
	1	1.50	1.26	0.969	6.56	0.970	6.50
		(0.06)	(0.05)	(0.005)	(0.03)	(0.005)	(0.03)
	5	2.32	2.18	0.983	10.55	0.980	10.33
		(0.12)	(0.10)	(0.004)	(0.03)	(0.004)	(0.03)
	25	3.60	5.32	0.979	16.37	0.971	15.59
		(0.22)	(0.30)	(0.005)	(0.03)	(0.005)	(0.03)

각 지역구의 출구조사 표본크기를 알 수 없어 동일하게 두기로 하였는데, 총 70만 명을 조사하였다고 하므로 246개 지역구로 나누었을 때 가까운 값인 2800으로 두어 실험하였고, 표본크기에 따른 차이를 보기 위해 500으로 두어 추가적으로 실험하였다.

모의실험의 결과 중 일부를 Table 3.1에 요약하였다. 두 종류의 신뢰구간의 성능에 초점을 맞추었다. 두 신뢰구간의 중심인 점추정량은 공통적으로 $S = \sum_{i=1}^{246} X_i$이지만 신뢰구간의 폭을 결정하는 분산 $\hat{V}(S) = \sum \hat{P}_i(1 - \hat{P}_i)$의 식에서 $\hat{P}_i$를 구하는 두 가지 방법에 따라 달라진다. 첫 번째 신뢰구간은 2절에서 설명한 몬테칼로 방법으로 구한 것이고, 두 번째 신뢰구간은 가정 (2.3)과 식 (2.5)에 의해 구한 것이다.

모의실험에 따른 변동을 알아보기 위해 Table 3.1에서 각 추정량의 표준오차를 구해서 괄호 안에 표시하였다. 대부분의 통계량은 평균값이므로 1000번의 반복실험에서 얻어진 추정량의 값의 표준편차를 반복실험횟수인 1000의 제곱근으로 나누어 표준오차를 구할 수 있다. 하지만 RMSE(Root Mean Square Error; 제곱근평균제곱오차)는 제곱의 평균을 구한 다음 다시 제곱근을 취한 형태로서 표준오차를 추정하기 위해 다른 방법이 필요한데 다음과 같이 구했다. $W_j = Y_j - \mu_j$, Y_j는 j번째 반복실험에서의 추정의석수이고 μ_j는 개표 완료 후 알 수 있는 실제의석수, n을 반복실험횟수라고 할 때

$$V\left(\widehat{\mathrm{MSE}}\right) = V\left(\sum_{j=1}^{n} W_j^2/n\right) = \frac{1}{n^2}\sum V\left(W_j^2\right)$$
$$= \frac{1}{n^2}\sum\left[E\left(W_j^4\right) - E^2\left(W_j^2\right)\right] = \frac{1}{n}\sum\left[E\left(W^4\right) - E^2\left(W^2\right)\right]$$

이므로

$$\hat{V}\left(\widehat{\mathrm{MSE}}\right) = \frac{1}{n}\left[\frac{\Sigma W_j^4}{n} - \left(\frac{\Sigma W_j^2}{n}\right)^2\right] = \frac{1}{n}\left[\frac{\Sigma W_j^4}{n} - \left(\widehat{\mathrm{MSE}}\right)^2\right]$$

이며, 따라서 델타 방법(delta method)에 의해

$$\hat{V}\left(\sqrt{\widehat{\mathrm{MSE}}}\right) \approx \left(\frac{1}{2\sqrt{\widehat{\mathrm{MSE}}}}\right)^2 \hat{V}\left(\widehat{\mathrm{MSE}}\right)$$

이다.

모의실험은 Table 3.1의 조건 외에도 K가 3부터 6까지, 그리고 각 지역구의 출구조사 표본크기가 500일 때에 대해서도 실시하였는데, 분량이 많아 별도로 보고하기로 한다 (http://bayes.ssu.ac.kr/~jhkim/Publication/exitpoll.txt). 모의실험 결과를 종합해서 보면 α의 값이 커질수록 점추정량 S의 RMSE와 두 신뢰구간의 폭이 모두 커지는데, 이는 α의 값이 커질수록 경합지역이 많아져서 정확한 추정이 어렵기 때문이다. RMSE(정확히 말하면 RMSE의 추정값)는 1000번의 반복실험에서 관측된, 점추정량 S와 참의석수 μ의 평균거리로서 $\sqrt{\sum_{j=1}^{1000}(S_j-\mu_j)^2/1000}$로 계산하였다. α의 값이 커질수록 몬테칼로 방법에 의한 첫 번째 신뢰구간의 실제신뢰수준이 명목신뢰수준에 조금 더 가까운 경향을 보이지만 차이는 크지 않았다. 신뢰구간의 폭은 두 번째 신뢰구간이 조금 더 작은 경향이 있지만 차이가 크지 않았으며, 실제신뢰수준이 같지 않으므로 두 신뢰구간의 폭을 단순히 비교할 수는 없다. 통계량 $\hat{p}_1 - \max(\hat{p}_2, \ldots, \hat{p}_K)$의 분포가 정규분포가 아님에도 불구하고 두 신뢰구간의 성능에는 별 차이가 없다는 점이 특이하다.

두 신뢰구간에 공통적이면서 중요한 사실은 실제신뢰수준이 명목신뢰수준보다 낮다는 것인데, 이는 앞 절에서 지적했듯이 $E(S) = \sum_{i=1}^{246} P_i$가 참의석수인 μ와 일치하지 않기 때문이 아닌가 짐작된다. 하지만 95% 신뢰구간은 90% 정도의 실제신뢰수준을, 99% 신뢰구간은 96% 이상의 실제신뢰수준을 모든 모의실험 조건에서 안정적으로 나타내므로 큰 문제는 없으며, 95% 이상의 실제신뢰수준을 확보하고 싶으면 명목신뢰수준을 95%가 아닌 99%로 하는 것이 안전하다.

참의석수의 점추정량으로 $S = \sum_{i=1}^{246} X_i$와 함께 $\sum_{i=1}^{246} \hat{P}_i$도 같이 고려하여 보았다. 디리클레 모수인 α의 값이 아주 크지 않을 때는 $\sum \hat{P}_i$는 S에 비해 RMSE가 작은 경향이 있으나, K의 값이 5 이상이고 $\alpha = K^2$으로 클 때 $\sum \hat{P}_i$의 RMSE가 S에 비해 터무니없이 커져 안정적이지 않고, 또 $\sum \hat{P}_i$의 분산에 대한 적절한 (즉, 이론적 근거가 있으면서도 지나치게 커지지 않는) 식이 없다. 이런 이유로 점추정량 $\sum \hat{P}_i$에 근거한 신뢰구간에 대해서는 고려하지 않았다.

이 절에서 대칭적 디리클레 분포의 범주의 수와 변동의 정도를 변화시켜가며 신뢰구간의 성능을 알아보는 모의실험을 실시하였다. 실제의석수에 대한 두 신뢰구간의 성능에 별 차이가 없어서 통계량 S의 분산을 구하는 방법으로 어떤 것을 써도 무방함을 알 수 있었다. 다만 두 신뢰구간 모두에서 실제신뢰수준이 명목신뢰수준을 밑도는 문제점이 있어 주의가 요구되지만 다양한 상황에서 안정적인 성능을 보였다. 하지만 이 모의실험에서는 정당별 득표율의 분포에 차이를 두지 않는 대칭적 디리클레 분포를 가정했다는 한계가 있는데, 5절에서 실제에 보다 가까운 상황에서 모의실험을 추가적으로 실시하였다.

4. 실제자료

2012년 4월 11일에 치러진 총선에서 지상파 3사가 전국 246개 지역구 전체에 대해 공동으로 출구조사를 실시하였는데, 2484개 투표소에서 약 70만 명을 대상으로 조사하였다고 한다. 출구조사 결과를 보면 전체 246개 지역구 중에서 '경합'으로 판단된 지역구가 60개로서 정당별 지역구 실제의석수를 정확하게 추정하기가 어려움을 짐작해 볼 수 있다. 이 절에서는 실제의석수를 2절에서 서술한 방법으로 추정해 보고자 한다.

Table 4.1. The actual number of seats and the estimated confidence intervals in 2012 general election

		새누리당	민주통합당	통합진보당
실제의석수		127	106	7
예상의석수 S		114	119	9
95% 신뢰구간	$n = 2800$:	109∼119	114∼124	7∼11
	$n = 500$:	107∼121	112∼126	6∼12
99% 신뢰구간	$n = 2800$:	107∼121	113∼125	6∼12
	$n = 500$:	104∼124	110∼128	5∼13

신뢰구간: 정수값을 얻기 위해 하한은 버림을, 상한은 올림을 함

지상파 3사가 공동으로 실시한 출구조사의 자료 중 각 지역구의 후보별 예상득표율이 인터넷에 공개되어 있다 (http://vote2012.imbc.com/livevote/exit/candidate.aspx). 실제의석수를 추정하기 위해서는 이 예상득표율 외에 지역구별 표본크기가 필요하다. 경합이 예상되는 지역과 그렇지 않은 지역의 표본크기가 달랐겠지만 지역구별 정확한 표본크기를 알 수 없어 전체 70만 명을 총 지역구수로 나눈 값인 2800을 지역구별 공통 표본크기 값으로 두고 추정하였고, 표본크기가 달라지는 데에 따른 효과를 알아보기 위해 2800보다 훨씬 작은 값인 500으로 두고 추가적으로 추정하여 보았다.

정당별 지역구 의석수의 실제 결과와 출구조사 자료로부터 추정한 결과를 Table 4.1에 요약하였다. 95%와 99% 신뢰수준, 2800개와 500개 두 표본크기의 모든 조건에서 새누리당과 민주통합당의 실제 의석수는 신뢰구간을 벗어난다. 이 사실은 출구조사가 잘못 이루어졌을 가능성이 높음을 시사한다. 즉, 출구조사에서 얻어진 투표 결과가 실제 투표 결과를 대표할 수 있는 확률 표본이 아닐 가능성이 높다는 것이다. 출구조사에서 추출된 표본의 편향성에 대한 증거는 충분하다. 만약 출구조사에서 얻어진 표본이 확률 표본이라면 246개 지역구의 출구조사에서 얻어진 예상득표율과 개표 완료 후 얻어진 실제득표율에는 평균적으로 차이가 없어야 한다. 하지만 여당인 새누리당의 예상득표율과 실제득표율의 차이를 246개 지역구에 대해 구한 값으로 t검정을 해보면 '평균적으로 차이가 없다'는 가설은 기각되며, 차이에 대한 95% 신뢰구간은 $(-0.94, -0.31)$이다 (단위는 백분율). 이는 여당의 숨은 표가 많았음을 의미한다. Figure 4.1에서 예상득표율과 실제득표율의 차이에 대한 분포를 지역별로 비교해 보았다. 전체 의석수의 거의 절반(49.2%)을 차지하는 수도권과 강원도에서 새누리당의 실제득표율이 과소 추정되었음을 알 수 있다.

지상파 3사는 출구조사를 공동으로 실시했지만 의석수 예측은 각자 하였다. 세 방송국에서 구체적으로 어떤 추정 방법을 썼는지는 모르지만 실제의석수를 제대로 예측하지 못한 점에 있어 별 차이가 없었는데, 본 연구의 결과로 추정방법에 문제가 있는 것이 아니라 출구조사에서 얻은 표본의 대표성에 문제가 있을 가능성이 높다는 것을 알 수 있다. 따라서 신뢰구간의 추정 방법보다는 표본의 대표성을 확보하는 데에 출구조사의 정확성 개선을 위한 노력의 초점이 맞추어져야 할 것이다.

5. 실제에 가까운 상황에서의 모의실험

2절에서 제안한 신뢰구간의 성능을 알아보기 위해 3절에서 모의실험을 실시하였다. 경합의 정도와 후보자의 수, 표본의 크기라는 여러 조건을 동시에 고려해야 하므로 경합의 정도를 하나의 모수로 간결하게 표현할 수 있는 대칭 디리클레 분포라는 특수한 분포를 3절에서 이용하였다. 하지만 실제 상황은 지역구마다 경합의 정도가 다르고 후보자의 수도 다르므로 이런 실제 상황에서 신뢰구간의 성능을 추가적으로 살펴볼 필요가 있다. 신뢰구간의 성능을 알아보기 위해서는 반복 실험이 필요한데, 따라서 실제에 가까운 상황을 반복적으로 만들어 신뢰구간의 성능을 살펴보는 모의실험을 실시해 볼 필요가 있다. 실제

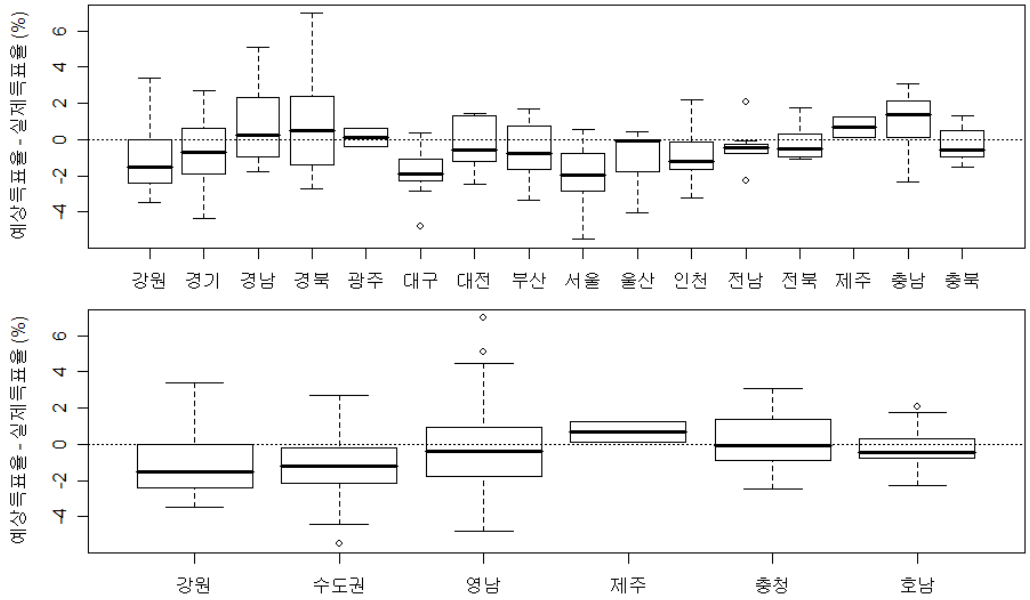

Figure 4.1. Differences between the actual rate of votes earned and its estimate

에 가까운 상황에서의 모의실험의 절차는 다음과 같다.

(1) 지역구 i에 대한 출구조사에서 얻어진 K_i명 후보의 관측득표율을 $(\hat{p}_{1i}, \ldots, \hat{p}_{Ki})$라고 할 때, 디리클레 분포 $D(\alpha_{1i}, \ldots, \alpha_{Ki})$, $\alpha_{ji} = c_i \cdot \hat{p}_{ji}$, $j = 1, \ldots, K_i$에서 난수를 발생하여 참득표율로 간주한다. 이 때 관심 있는 정당 후보의 득표율이 첫 번째 범주에 오도록 하며, 그 후보의 당선여부를 판단한다. (앞의 모의실험과는 달리 각 지역구의 후보자의 수가 고정되지 않고 K_i로서 지역구마다 다른 값을 가지며 대칭적 디리클레 분포가 아니다.)

(2)–(5) 단계는 3절의 모의실험의 절차와 같다.

(6) $c_i(= \sum_{j=1}^{K} \alpha_{ji})$의 값을 1, K_i, K_i^2, K_i^3, K_i^5으로 변화시켜가며 모의실험을 실시해 득표율의 변동에 따른 효과를 살펴본다. c_i가 커질수록 변동 폭이 적어져 실제자료에 가까워진다. c_i가 K_i^5일 때, 모의실험에서의 참득표율을 나타내는 난수 p_{1i}의 표준편차를 디리클레 분포의 표준편차 식 $\sqrt{\alpha_{1i}(c_i - \alpha_{1i})/(c_i^2(c_i + 1))}$을 이용해 구해 보면, 새누리당의 경우 246개 지역구에 걸쳐 계산되는 표준편차 값 중에서 제일 큰 값이 0.087이 되어, 출구조사에서 얻은 관측득표율에 가까운 참득표율 난수를 얻게 된다. 반대로 c_i가 1이 되면 표준편차는 최소 0.099, 최대 0.354가 되어 변동이 아주 크게 된다. 한편, 표본크기도 2800과 500으로 변화시켜 그 효과를 살펴보았다.

모의실험 결과의 일부를 Table 5.1에 보고하였다. 통계량 S의 분산을 추정하는 두 가지 방법에 따라 달라지는 두 신뢰구간의 성능은 구별할 수 없을 정도로 비슷했다. 95% 신뢰구간의 실제신뢰수준이 90% 정도로서 다소 낮은 경향을 보인 점과, 99% 신뢰구간의 실제신뢰수준이 95% 이상을 보인 점은 3절의 모의실험 결과와 비슷했다.

실제의석수에 대한 또 다른 점추정량 $\sum_{i=1}^{246} \hat{P}_i$의 RMSE가 모든 실험조건에서 S보다 작아 더 나은 추정량이 될 수 있는 가능성을 보였으나, 3절에서 본 바와 같이 경합이 더 치열해질 경우 아주 부정확해지며

Table 5.1. Simulations near to real situation for the number of seats of Saenuri Party: $n = 2800$

명목 신뢰수준	c	RMSE: S	RMSE: $\sum \hat{P}_i$	신뢰구간1의 실제신뢰수준	신뢰구간1의 길이	신뢰구간2의 실제신뢰수준	신뢰구간2의 길이
0.95	1	0.98	0.82	0.904	3.26	0.904	3.26
		(0.03)	(0.03)	(0.009)	(0.03)	(0.009)	(0.03)
	K	1.41	1.18	0.915	4.71	0.916	4.72
		(0.06)	(0.04)	(0.009)	(0.03)	(0.009)	(0.03)
	K^2	1.84	1.51	0.906	6.10	0.906	6.11
		(0.08)	(0.06)	(0.009)	(0.02)	(0.009)	(0.02)
	K^3	2.17	1.78	0.896	7.14	0.899	7.14
		(0.10)	(0.07)	(0.010)	(0.02)	(0.010)	(0.02)
	K^5	2.59	2.23	0.890	8.40	0.891	8.41
		(0.13)	(0.10)	(0.010)	(0.02)	(0.010)	(0.02)
0.99	1	1.03	0.88	0.957	4.23	0.960	4.23
		(0.04)	(0.03)	(0.006)	(0.04)	(0.006)	(0.04)
	K	1.51	1.26	0.962	6.24	0.960	6.24
		(0.07)	(0.05)	(0.006)	(0.03)	(0.006)	(0.03)
	K^2	1.87	1.58	0.961	7.95	0.963	7.96
		(0.09)	(0.07)	(0.006)	(0.03)	(0.006)	(0.03)
	K^3	2.26	1.85	0.962	9.42	0.962	9.42
		(0.10)	(0.08)	(0.006)	(0.03)	(0.006)	(0.03)
	K^5	2.54	2.12	0.963	11.03	0.963	11.04
		(0.14)	(0.1)	(0.006)	(0.03)	(0.006)	(0.03)

분산에 대한 적절한 식이 없다는 한계가 있다.

이 실험을 통해, 첫째, 지역구별 출마자수와 득표율 등 실제에 가까운 상황에서도 두 신뢰구간은 비슷한 성능을 보이며, 둘째, 실제신뢰수준이 명목신뢰수준을 밑도는 문제점이 있지만 여러 변동 조건에서 안정적인 성능을 보인다는 것을 알 수 있다. 따라서 4절에서 실제의석수가 신뢰구간을 벗어난 것은 신뢰구간이 아니라 표본의 대표성에 문제가 있어서 그럴 가능성이 높다는 것을 이 모의실험을 통해 다시 한 번 확인할 수 있다.

6. 결론 및 요약

방송에서 실제로 출구조사 결과를 발표할 때 지역구 의석수만 따로 발표하지 않고 비례대표 의석수까지 합한 정당별 총의석수를 발표한다. 비례대표 의석수의 추정은 다른 방법으로 이루어져야 하며 두 종류의 의석수를 합할 때의 문제점도 고려되어야 할 것이다. 본 연구에서는 의석수가 더 많은 지역구 의석수의 추정 문제를 다루었으며, 표본의 대표성이 보장될 때 적용 가능한 방법을 제시하고 모의실험을 통해 성능을 살펴보았다.

2절에서 정당별 지역구 의석수에 대한 점추정량을 제안하였고 신뢰구간의 길이를 결정하는 분산을 구하는 가능한 방법을 몇 가지 제시하였다. 3절과 5절의 모의실험을 통해 분산을 구하는 방법에 따른 차이는 무시할 수 있다는 것을 알 수 있었으나, 신뢰구간의 실제신뢰수준이 명목신뢰수준보다 다소 낮아지는 문제점을 발견하였다. 99% 명목신뢰수준을 쓰면 다양한 조건에서 적어도 95% 이상의 실제신뢰수준을 달성할 수 있다는 사실도 모의실험을 통해 알 수 있었다.

19대 총선 출구조사의 실제 자료를 이용해 지역구 의석수를 추정해 보았는데 정당별 실제의석수가 신뢰

구간을 벗어났다. 이렇게 제대로 추정되지 못한 이유는 출구조사의 표본의 대표성에 문제가 있기 때문으로 판단된다.

정당별 지역구 의석수를 추정할 때, 신뢰구간의 종류보다는 (즉, 추정량으로 S와 $\sum \hat{P}_i$ 중 어떤 것을 쓰느냐, 추정량의 분산 식으로 어떤 것을 쓰느냐의 문제보다는) 표본의 대표성이 더 중요하다. 하지만 신뢰구간에 대한 연구는 기본적인 것으로서 그 중요성을 간과할 수 없다. 따라서 추정 방법이 공개되어 검증을 받고 더 나은 방법이 제안될 수 있도록 해야 하는데, 본 연구가 그 출발점이 되기를 바란다.

References

Casella, G. and Berger, R. L. (2002). *Statistical Inference*, 2nd ed., Duxbury.

Hyun, K. B. (2005). A Study on the Election Poll and its Accuracy in the 17th General Election, *Journal of Korea Regional Communication Research Association*, **5**, 301–336.

Kwak, E. S. and Kim, Y. W. (2010). A Total Survey Error Analysis of the Exit Polling for General Election 2008 in Korea, *Survey Research*, **11**, 33–55.

Lee, J. W. (2004). Problems of the Election Forecasting in the 2004 Korean General Election, *Journal of Communication Research*, **41**, 110–135.

R Development Core Team (2008). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

총선 출구조사에서 정당별 지역구 의석수 추정

김지현[a,1]

[a]숭실대학교 정보통계보험수리학과

(2012년 11월 12일 접수, 2012년 12월 8일 수정, 2012년 12월 13일 채택)

요 약

2012년 4월 11일 총선 당일 이루어진 출구조사에서 정당별 의석수를 구간으로 예측했음에도 불구하고 예측이 빗나갔다. 지상파 3사가 출구조사는 공동으로 실시하였지만 정당별 의석수 예측은 각자 하였다고 하는데 구체적 예측 방법은 공개하지 않았다. 이 논문에서 정당별 지역구 의석수를 구간으로 추정하는 방법을 제안하고 그 성능을 모의실험을 통해 알아보았다. 그리고 제19대 총선 출구조사의 실제자료에 적용해 보았다.

주요용어: 출구조사, 다항분포, 디리클레 분포, p-값.

[1](156-743) 서울시 동작구 상도동 511, 숭실대학교 자연과학대학 정보통계보험수리학과, 교수.
E-mail: jxk61@ssu.ac.kr