On polling and election prediction

Federico Fattorini

Università di Pisa

January 16, 2023

Simple prediction based on a poll

Introduction

I=" N people voting in the election, n people participating to the poll. 3 different parties: A, B, C. The one which gets more votes wins. "

n_A=" people participating to the poll and voting for A"
n_B=" people participating to the poll and voting for B"
n_C=" people participating to the poll and voting for A"
N_A=" people voting for A"
N_B=" people voting for B"
N_C=" people voting for C"

The aim is to compute the posterior probability for the victory for each party.

Prediction based on a poll

Posterior distribution

We compute the posterior distribution for N_a and N_b (N_c is constrained):

$$P(N_A, N_b | n_A, n_B, n_B, I) = \frac{P(N_A, N_B | I) P(n_A, n_B, n_C | N_A, N_B, I)}{P(n_A, n_B, n_C | I)}$$

- $P(N_A, N_B | I)$ is the prior. We choose it to be uniform over the integer values between $[0, N]$ both for N_A and N_B, constrained by $N_A + N_B \leq N$. The normalisation is $\frac{2}{(N+1)(N+2)}$
- $P(n_A, n_B, n_C | N_A, N_B, I) = mult(n_A, n_B, n_C; n, P_A, P_B, P_C)$ is the likelihood. $P_i = \frac{N_i}{N}$ is the probability of voting for party i. Note that $P_C = \frac{N - N_A - N_B}{N}$.
- $P(n_A, n_B, n_C | I) = \sum_{N_A=0}^{N} \sum_{N_B=0}^{N} P(N_A, N_b | I) P(n_A, n_B, n_C | N_A, N_B, I)$ is just a normalisation factor and is calculated numerically.

Prediction based on a poll

Probability of victory

Finally, the probability for victory of A (A_w) is the sum of the posterior calculated in all the N_A, N_B such that $N_A > N_C$ and $N_B > N_C$ (the * here stands for this constraint) :

$$P(A_w|n_A, n_B, n_C, I) = \sum_{N_A=0}^{N} \sum_{N_B=0}^{N} {}^{*} P(N_A, N_B|n_A, n_B, n_B, I)$$

Analogously for B_w and C_w, changing the constraints.

Prediction based on a poll

Results for $N = 1000$, $n_A = 20$, $n_B = 15$, $n_C = 10$

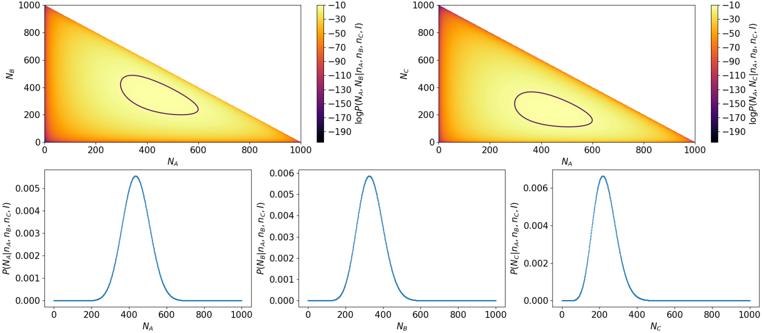


Figure: $P(A_w) = 0.78,\ P(B_w) = 0.19,\ P(C_w) = 0.02$
Most probable values:
$N_{A,max} = 445,\ N_{B,max} = 333,\ N_{C,max} = 222$

Prediction based on a poll

Results for $N = 10000$, $n_A = 180$, $n_B = 160$, $n_C = 150$

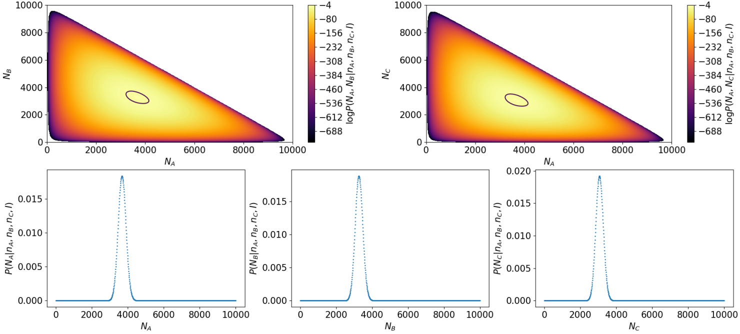


Figure: $P(A_w) = 0.83,\ P(B_w) = 0.13,\ P(C_w) = 0.03$
Most probable values:
$N_{A,max} = 3670, N_{B,max} = 3270,\ N_{C,max} = 3060$

Prediction based on a poll

Results for $N = 100000$, $n_A = 250$, $n_B = 300$, $n_C = 280$

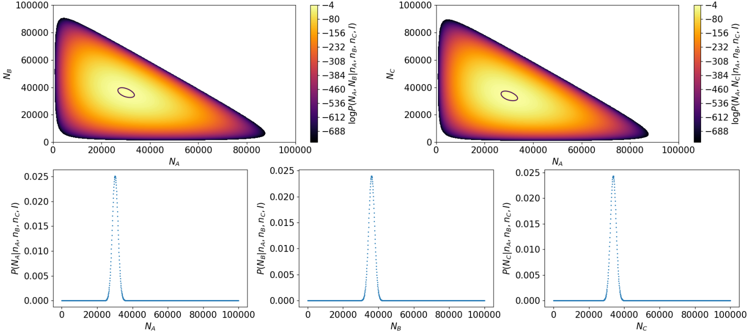


Figure: $P(A_w) = 0.01,\ P(B_w) = 0.79,\ P(C_w) = 0.20$
Most probable values:
$N_{A,max} = 30100,\ N_{B,max} = 36200,\ N_{C,max} = 33700$

Prediction based on a poll, with liars

Introduction

Now we want to account for the possibility of people not saying the truth answering the poll. Let's call f the probability of saying the truth. We compute the posterior in the same way of the first case, but changing the likelihood according to the new scenario.

New definition for n_A, n_B, n_C

n_A="people participating to the poll and saying A"
n_B="people participating to the poll and saying B"
n_C="people participating to the poll and saying C"

We left everything unchanged from the previous case, except the probabilities of getting a "A", "B" and "C" as answers in the poll.

Prediction based on a poll, with liars

Probability for the answers of the poll

A_t=''the person actually votes for A in the election''
A_p=''the person answers A in the poll''
T=''the person says the truth''
F=''the person lies''

Logically we have:

$$A_p = (A_t T) + (B_t F A_p) + (C_t F A_p)$$

Then:

$$P(A_p|N_a, N_b, I, f) = P(A_t T|N_a, N_b, I, f) + P(B_t F A_p|N_a, N_b, I, f) +$$

$$+ P(C_t F A_p|N_a, N_b, I, f)$$

Prediction based on a poll, with liars

Probability for the answers of the poll

$$P(A_t T|N_a, N_b, I, f) = P(A_t|T, N_a, N_b, I, f)P(T|f) = \frac{N_A}{N} f$$

$$P(B_t F A_p|N_a, N_b, I, f) = P(A_p|B_t, F, N_a, N_b, I, f)P(B_t|F, N_a, N_b, I, f)P(F|I)$$

$$P(C_t F A_p|N_a, N_b, I, f) = P(A_p|C_t, F, N_a, N_b, I, f)P(C_t|F, N_a, N_b, I, f)P(F|I)$$

Assuming there are no preferences in lying, and considering that a person votes for a party regardless of whether he is a liar or not:

$$P(A_p|C_t, F, N_a, N_b, I, f) = P(A_p|B_t, F, N_a, N_b, I, f) = \frac{1}{2}$$

$$P(B_t|F, N_a, N_b, I, f) = \frac{N_B}{N}$$

$$P(C_t|F, N_a, N_b, I, f) = \frac{N_C}{N}$$

Prediction based on a poll, with liars

Probability for the answers of the poll

Finally we have:

$$P(A_p|N_a, N_b, I, f) = \frac{N_A}{N} f + \frac{1}{2}(1 - f)\frac{N_C + N_B}{N}$$

Analogously:

$$P(B_p|N_a, N_b, I, f) = \frac{N_B}{N} f + \frac{1}{2}(1 - f)\frac{N_A + N_C}{N}$$

$$P(C_p|N_a, N_b, I, f) = \frac{N_C}{N} f + \frac{1}{2}(1 - f)\frac{N_A + N_B}{N}$$

In the limit $f = 1$ we get the same result as before.

Prediction based on a poll, with liars

Likelihood and probability of victory

The likelihood is again a multinomial, but the probabilities are the ones just calculated, substituting $N_C = N - N_A - N_B$

$$P(n_A, n_B, n_C | N_A, N_B, I) = mult(n_A, n_B, n_C; n, P(A_p), P(B_p), P(C_p))$$

Finally, the probability for the victory of A (A_w) is the sum of the posterior calculated in all the N_A, N_B such that $N_A > N_C$ and $N_B > N_C$ (the * stands for this constraint) :

$$P(A_w | n_A, n_B, n_C, I, f) = \sum_{N_A=0}^{N} \sum_{N_B=0}^{N} {}^* P(N_A, N_B | n_A, n_B, n_B, I, f)$$

Analogously for B_w and C_w .

Prediction based on a poll, with liars

Results for $N = 10000$, $n_A = 180$, $n_B = 160$, $n_C = 150$, $f = 0$

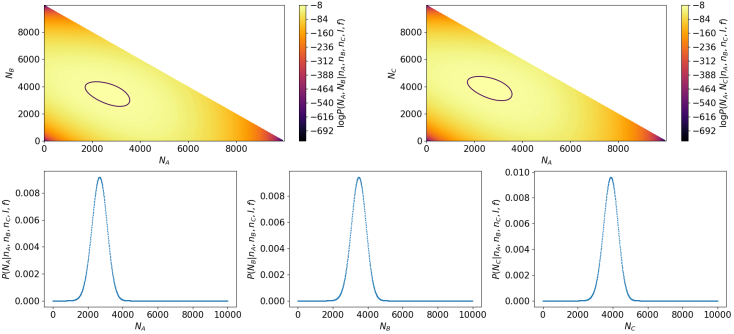


Figure: $P(A_w) = 0.02,\ P(B_w) = 0.27,\ P(C_w) = 0.70$
Most probable values:
$N_{A,max} = 2650,\ N_{B,max} = 3470,\ N_{C,max} = 3880$

Prediction based on a poll, with liars

Results for $N = 10000$, $n_A = 180$, $n_B = 160$, $n_C = 150$, $f = 1/3$

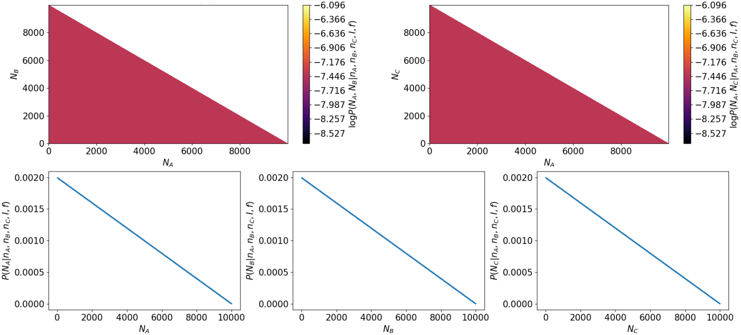


Figure: $P(A_w) = 0.33,\ P(B_w) = 0.33,\ P(C_w) = 0.33$

Prediction based on a poll, with liars

Results for $N = 10000$, $n_A = 180$, $n_B = 160$, $n_C = 150$, $f = 0.5$

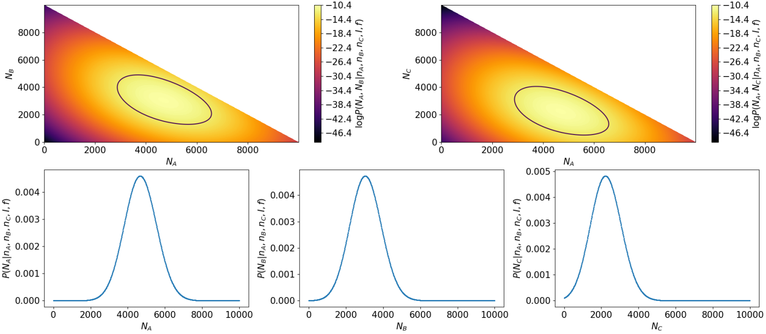
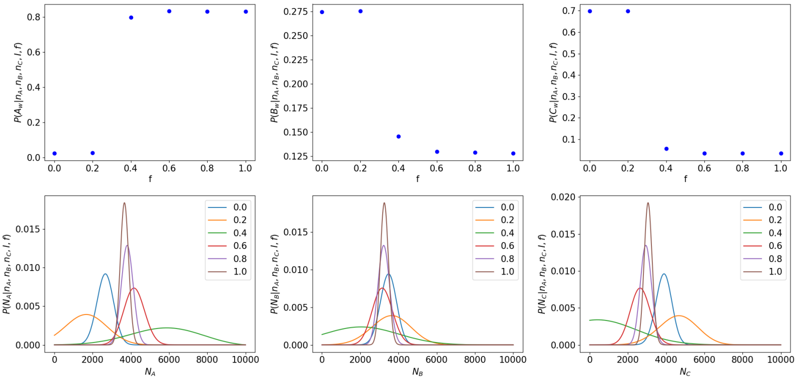


Figure: $P(A_w) = 0.83,\ P(B_w) = 0.13,\ P(C_w) = 0.03$
Most probable values:
$N_{A,max} = 4690,\ N_{B,max} = 3060,\ N_{C,max} = 2250$

Prediction based on a poll, with liars

Comparisons for different f, with $N = 10000$, $n_A = 180$, $n_B = 160$, $n_C = 150$



Prediction based on a poll, with liars

Comments

The plots shows different behaviours for the values of f:

- In the range $0 \leq f \leq \frac{1}{3}$ the poll gives wrong informations (the majority lies), so the actual distribution of votes will be in contrast to the poll's results.
- In the range $\frac{1}{3} < f < \frac{1}{2}$ the informations of the poll are still mainly wrong, but not enough to make the victory of A unlikely.
- In the range $f > \frac{1}{2}$ the poll essentially gives the right trend, but part of the answers is false. Note that this penalizes all the parties in the same way.

Two stages election

Introduction

Now we update the information "I" accounting for a two stages election. At the first step only the two parties with the most votes pass to the next stage. The electors of the losing party follow these rules for the second vote:

Group of people that votes for A on the first round will vote for B if A loses the first round

Group of people that votes for B on the first round will vote for C if B loses the first round

Group of people that votes for C on the first round will vote for A if C loses the first round

Two stages election

Definitions

First, we want to compute the posterior distribution for the parties through the second stage. We reduce to the two cases where A passes the first stage.

N'_A=" N'_A people voting for A in the second stage, in the case in which A and B pass the first one"

N'_B=" N'_B people voting for A in the second stage, in the case in which A and B pass the first one"

N''_A=" N'_A people voting for A in the second stage, in the case in which A and C pass the first one"

N''_C=" N''_C people voting for A in the second stage, in the case in which A and C pass the first one"

$A_n B_n$=" A and B pass the first round"

$A_n C_n$=" A and C pass the first round"

Two stages election

Posterior distribution

$$P(N'_A A_n B_n | n_A, n_B, n_C, I) = \sum_{N_A}^{N} \sum_{N_B}^{N} P(N'_A A_n B_n, N_A, N_B | n_A, n_B, n_C, I) =$$

$$= \sum_{N_A}^{N} \sum_{N_B}^{N} P(N'_A A_n B_n | N_A, N_B, n_A, n_B, n_C, I) P(N_A, N_B | n_A, n_B, n_C, I)$$

Using the product rule, the first term is:

$$P(N'_A | A_n B_n N_A, N_B, n_A, n_B, n_C, I) P(A_n B_n | N_A, N_B, n_A, n_B, n_C, I)$$

Now we note that the second term is null when $N_A < N_C$ or $N_B < N_C$, and is equal to one in all the other cases. Hence we can omit it and constrain the sum $(^*)$

Two stages election

Posterior distribution

$$\sum_{N_A}^{N} \sum_{N_B}^{N} {}^{*} P(N_A'|N_A, N_B, n_A, n_B, n_C, I) P(N_A, N_B|n_A, n_B, n_C, I)$$

Since $N_B = N_B'$, $N = N_B + N_A'$, then $P(N_A'|N_A, N_B, n_A, n_B, n_C, I)$ is null if $N_B \neq N - N_A'$ and equal to 1 otherwise. Finally:

$$P(N_A' A_n B_n|n_A, n_B, n_C, I) = \sum_{N_A}^{N} \sum_{N_B}^{N} {}^{*} \delta(N_B + N_A' - N) P(N_A, N_B|n_A, n_B, n_C, I)$$

Note that the second term is just the posterior of the first scenario.

Two stages election

Probability of victory

With the same calculation, but considering $N''_A = N_A$ (and the different constraint) in the case in which A and C pass the first stage, we have:

$$P(N''_A A_n C_n | n_A, n_B, n_C, I) = \sum_{N_A}^{N} \sum_{N_C}^{N}{}^{*} \delta(N_A - N''_A) P(N_A, N_C | n_A, n_B, n_C, I)$$

The probability for A winning the second stage ($A'_w A_n B_n$ or $A''_w A_n C_n$) is :

$$P(A'_w A_n B_n | n_A, n_B, n_C, I) = \sum_{N_A = N/2+1}^{N} P(N'_A A_n B_n | n_A, n_B, n_C, I)$$

And since the two cases are mutually exclusive:

$$P(A_w | n_A, n_B, n_C, I) = P(A'_w A_n B_n | n_A, n_B, n_C, I) + P(A''_w A_n C_n | n_A, n_B, n_C, I)$$

Two stages election

Results for $N = 1000$, $n_A = 20$, $n_B = 15$, $n_C = 10$

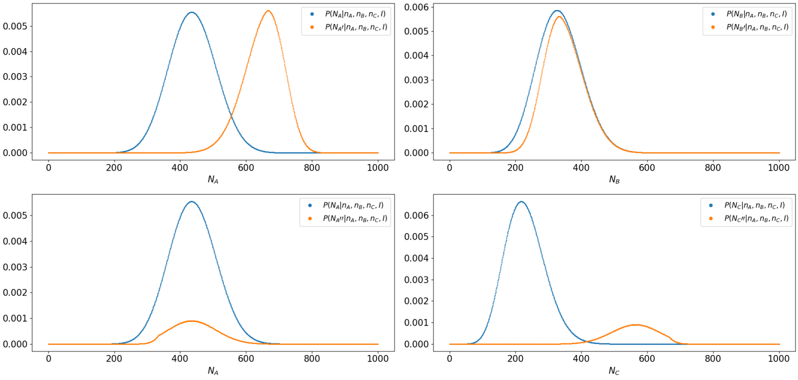


Figure: $P(A'_w) = 0.81,\ P(A''_w) = 0.03,\ P(A_w) = 0.84$
Most probable values:
$N'_{A,max} = 667,\ N'_{B,max} = 333,\ N''_{A,max} = 511,\ N''_{C,max} = 489$

Two stages election

Results for $N = 1000$, $n_A = 15$, $n_B = 15$, $n_C = 10$

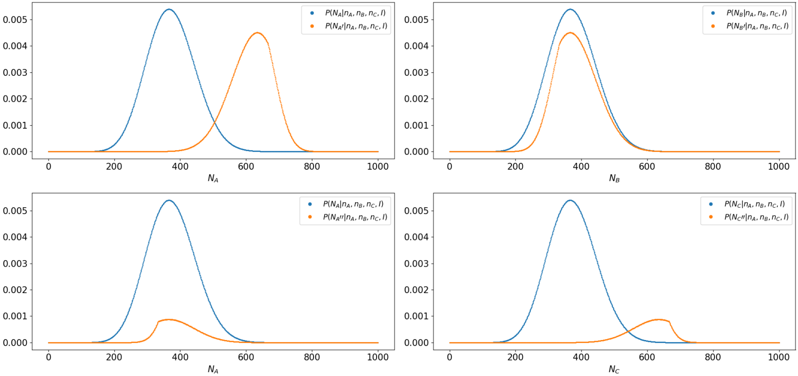


Figure: $P(A'_w) = 0.70,\ P(A''_w) = 0.01,\ P(A_w) = 0.71$
Most probable values:
$N'_{A,max} = 635,\ N'_{B,max} = 365,\ N''_{A,max} = 365,\ N''_{C,max} = 635$

Two stages election, with irresolute losers

Introduction

We consider a two stage election, with these new rules for the second vote of the losing electors.

Group of people that vote for A on the first round will vote for B with probability p_1 if A loses the first round

Group of people that vote for B on the first round will vote for C with probability p_2 if B loses the first round

Group of people that vote for C on the first round will vote for A with probability p_3 if C loses the first round

In the following we will use the same notation as the previous case.

Two stages election, with irresolute losers

Posterior distribution

We are still interested in the posterior distribution. We use then the same procedure as the previous case.

$$\sum_{N_A}^{N}\sum_{N_B}^{N}{}^{*} P(N'_A|N_A, N_B, n_A, n_B, n_C, I, p_3)P(N_A, N_B|n_A, n_B, n_C, I)$$

In this case $P(N'_A|N_A, N_B, n_A, n_B, n_C, I, p_3)$ is not "deterministic" anymore. It is composed by a certain component (N_A), but the remaining $N'_A - N_A$ is the result of a binomial process with N_C ($= N - N_A - N_B$) trials and probability of success p_3. Hence, the posterior is:

$$\sum_{N_A}^{N}\sum_{N_B}^{N}{}^{*} bin(N'_A - N_A, N - N_A - N_B, p_3)P(N_A, N_B|n_A, n_B, n_C, I)$$

Two stages election, with irresolute losers

Other posteriors and winning probability

We do the same for party B and for the distribution of N''_A, changing the probabilities and the number of trials (N_B in the latter case):

$$P(N'_B A_n B_n) = \sum_{N_A}^{N} \sum_{N_B}^{N}{}^{*} bin(N'_B - N_B, N - N_A - N_B, 1 - p_3) P(N_A, N_B | n_A, n_B, n_C, I)$$

$$P(N''_A A_n C_n) = \sum_{N_A}^{N} \sum_{N_C}^{N}{}^{*} bin(N''_A - N_A, N - N_A - N_C, 1 - p_2) P(N_A, N_C | n_A, n_B, n_C, I)$$

As before, the probabilities for A winning the second stage ($A'_w A_n B_n$ or $A''_w A_n C_n$) and for winning the election are:

$$P(A'_w A_n b_n | n_A, n_B, n_C, I) = \sum_{N_A = N/2+1}^{N} P(N'_A A_n b_n | n_A, n_B, n_C, I)$$

$$P(A_w | n_A, n_B, n_C, I) = P(A'_w A_n b_n | n_A, n_B, n_C, I) + P(A''_w A_n b_n | n_A, n_B, n_C, I)$$

Two stages election, with irresolute losers

Results for $N = 1000$, $n_A = 20$, $n_B = 15$, $n_C = 10$, $p_2 = 0.7$, $p_3 = 0.2$

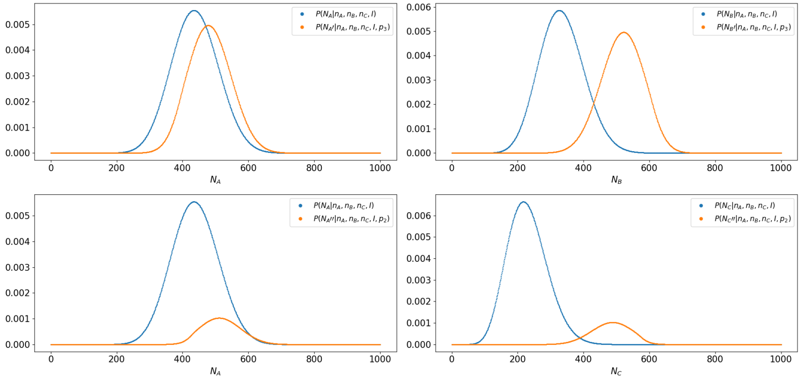


Figure: $P(A'_w) = 0.31,\ P(A''_w) = 0.09,\ P(A_w) = 0.40$
Most probable values:
$N'_{A,max} = 478,\ N'_{B,max} = 522,\ N''_{A,max} = 511,\ N''_{C,max} = 489$

Two stages election, with irresolute losers

Results for $N = 1000$, $n_A = 20$, $n_B = 15$, $n_C = 10$, $p_2 = 0.5$, $p_3 = 0.5$

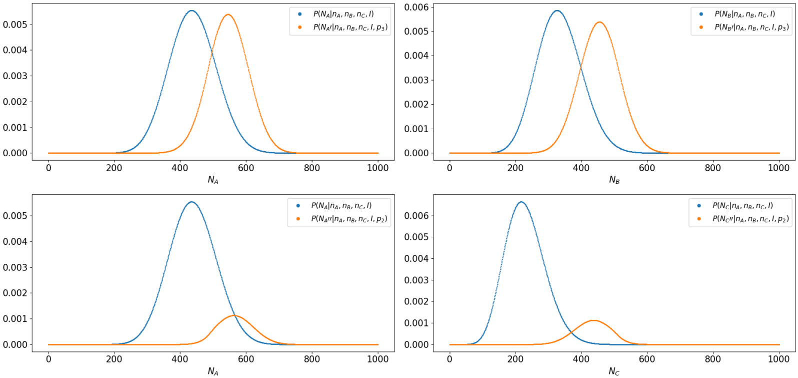


Figure: $P(A'_w) = 0.63,\ P(A''_w) = 0.14,\ P(A_w) = 0.77$
Most probable values:
$N'_{A,max} = 545,\ N'_{B,max} = 455,\ N''_{A,max} = 563,\ N''_{C,max} = 437$

Two stages election, with irresolute losers

Results for $N = 1000$, $n_A = 20$, $n_B = 15$, $n_C = 10$, $p_2 = 0.2$, $p_3 = 0.7$

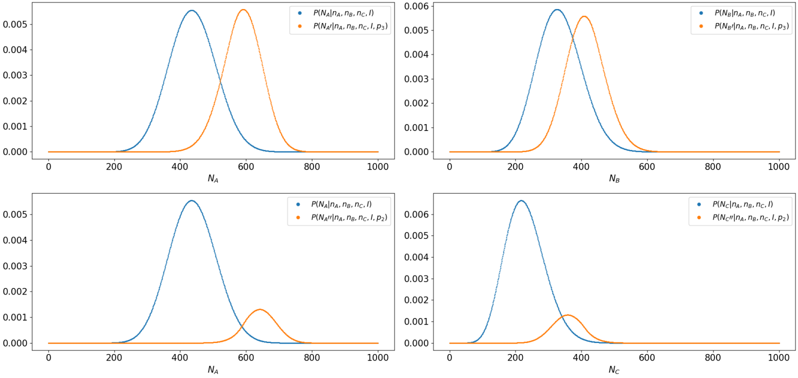


Figure: $P(A'_w) = 0.76,\ P(A''_w) = 0.15,\ P(A_w) = 0.91$
Most probable values:
$N'_{A,max} = 592,\ N'_{B,max} = 408,\ N''_{A,max} = 642,\ N''_{C,max} = 358$