Ph11 First Hurdle (2023)
Errors in Voting

Enoch Guo
eguo@caltech.edu

October 23, 2023

Abstract

In this hurdle, the goal is the determine sources of error in the US presidential election process and estimate how they contribute to the uncertainty of the final outcome of who is president. To do this, I split the population of voters into two cases: those that vote and those that do not. For those that do vote, I attempt to better count the total number of true ballots for a candidate by modeling it as a time discrete stochastic process and incorporating in an uncertainty term that represents random error of ballot machines, voters, etc., where each time interval actually represents a singular voter. I model this using an Itô stochastic differential equation and run simulations. For those that do not vote, I model voter disenfranchisement by a Bernoulli distribution. I show that random error of miscounted ballots contributes very little to the final outcome's uncertainty, and thus more error comes from people who did not vote. Then in the discussion, I generalize to account for intentional voter fraud by adjusting certain parameters in my model.

1 Introduction

First, I consider what ground truth means in this problem. I define it to be the proportion of all eligible American voters that intended to vote for some candidate A. This differs from the measured proportion due to factors like those that did not vote, or faulty ballot machines, or whatever other reason. Using stochastic differential equations, I will estimate the total number of votes that actually reflected the voter's true intention. I will consider the election process as Bernoulli distributed with $p =$ the measured proportion of voters for a Candidate A. Then, using the calculated number of true votes and the actual measured number of people that voted for a particular candidate and the total population of voters, I can establish a range on how the variance of the distribution changes when accounting for faulty votes. This change in variance will reflect the uncertainty of the election process.

2 Assumptions

I make the following assumptions throughout my method:

1. Each state's electors vote according to the overall popular vote of their state. That is, the outcome of the election is uniquely determined by the voters themselves and is truly a contest of "the will of the people." (In practice, only Maine and Nebraska's electors do not vote by popular vote). This is to avoid the effects of gerrymandering and other occurrences where the people's vote does not reflect the winner.

2. The election is a purely two-party affair, meaning there are only two viable candidates to vote for.

3. The ballot error incorporates the random error of all voters, election officials, ballot tallies, voting machines, etc. (Had I more time, I would have expressed each of these categories of error as a distinct random variable and expressed the total error as a sum or convolution of *i.i.d.* variables.)

4. The effect of voter disenfranchisement on the final outcome is negligible in all states except historically swing states, where the probability of voting for Candidate A is Bernoulli distributed with $p = 0.5$.

3 Ballot Error

First I will explain why it is reasonable to estimate the total number of true ballots using a stochastic process. A stochastic process is a collection of random variables that is indexed by some set, usually representing time. In this case, our random variables are indexed by the set of voters, $\tau = \{1, 2, ..., n\}$, where the n is the total number of eligible voters that actually voted. For each voter n, there are three cases: the vote was counted correctly, the vote was counted incorrectly, or the vote was not counted at all, and the probability of these depend on the random Brownian motion/Wiener process we assign to model the chance of a faulty voting machine/voter where the sampling distributions of the uncertainty term $\sim \mathcal{N}(0, \sigma^2)$. We can then interpret the stochastic process of all voters as the solution to an SDE. We start with a common linear homogeneous stochastic differential equation often used to model

population data. This makes sense because for the population model, with every year the population either increases or decreases with some random probability. Similarly, in the voting analogy, with each successive voter, the true count of votes for Candidate A will either go up, down, or stay the same. We center the Brownian motions around mean 0 because ideally with every voter the vote count should increase by exactly one and not rely on any fluctuation. The equation is given as

$$dX(t) = aX(t)dt + \sqrt{bX(t)}\, dW(t)$$

for $t \in \tau$ and $X(0) = 1$, where $X(t)$ is the true count of votes for a candidate and a and b are are constants to be chosen. The constant a is the drift of the process and b is the diffusion coefficient. The second term in this equation is what differentiates it from a standard differential equation; W is a Wiener process that adds a stochastic element into the equation. It simulates the randomness of ballot error. By Itô's lemma, we can verify that this equation does indeed have a solution, which is given in the explicit form

$$X(t) = X(0)\, exp((a - \frac{1}{2}b^2)t + bW(t))$$

Given prior data samples, we can fit our model and find values for a and b using maximum likelihood estimation, but for now we will choose them arbitrarily and manually. Because we want the true count of votes to correlate almost linearly with voter number, we will set them both to 1.

4 Disenfranchisement Error

Voters refrain from voting for a number of reasons. They might fail to register, or be busy on a particular day, or not have access to polling locations. Whatever, the reason eligible voters that do not vote affect the outcome of an election. We consider only disenfranchised voters in swing states to simplify the problem, as in states that lean very heavily towards one candidate these people that do not vote have much smaller impact. The assumption is that in all swing states, the eligible voters that have not voted would have voted for a Candidate A with expected value of 0.5. Therefore, to see the effect of abstained voters, all I do is to find the number of non-voters for a swing state, assign half of them to the predicted (see above) amount of true votes for a candidate. I can then see if this number makes a significant difference in the resulting electoral vote and if it affects the overall outcome. Given more time, I would perform a significance test here.

5 Example Analysis of the 2020 Election

According to US Census data, 168.31 million US citizens were above the legal age of 18 years and thus eligible to vote. Also, in the 2020 election, 81,282,916 people voted for candidate Joe Biden. Thus the proportion of people voting for Biden was $81282916/168310000 = 48.29\%$ After iterating an Euler method approximation for the SDE for each of the 81,282,916 voters many times, the model converged to a true amount of voter for Biden of approx. 82,000,000, which equates to a proportion of 48.72%. Using the variance formula for a Bernoulli distribution, the measured variance was $0.4829(1-0.4829) = 0.2497$, while my estimated true variance was 0.2498.

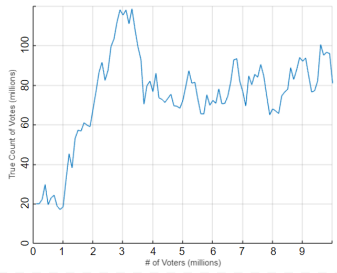

Figure 1: Sample simulation of the SDE for the 2020 election.

If we consider the swing state of Wisconsin in 2020, approx. 563140 people did not vote. Wisconsin ended up favoring Biden, but even if half of these people voted for Trump it would still not have been

sufficient to turn the electoral vote. Therefore Wisconsin contributed minimal uncertainty to the election result.

6 Conclusion

These results show that the effect of purely random ballot error contributes minimally to the overall uncertainty of the outcome of an election. This makes sense, because any random errors negatively affecting one candidate (e.g. under-counting by a voting machine) ought to also affect the other candidate in a nearly equal manner.

7 Discussion

Now to adjust the model for intentional fraud, we have only to change a couple parameters, namely, the constant b and the sampling distribution of the Wiener process. In the case of intentional fraud, there is almost surely a high probability of votes being over-counted for a particular candidate, and so we can pretend that votes for that candidate are never uncounted or miscounted but only counted extra. This equates to centering the normal distribution of the Wiener process around a mean of 1 rather than 0, and also increasing the coefficient b to demonstrate that the effect of the high probability of over-counting votes for one candidate are significantly amplified by the fraud. When I graph simulations for this scenario, the graphs still show a stochastic trend, except it trends sharply upwards at the same time, demonstrating that voter fraud drastically effects the uncertainty of the outcome of an election. To model more extreme cases of fraud, all one needs to do is determine the coefficient b.