Civic Participation II: Voter Fraud

Sharad Goel
Stanford University
Department of Management Science

March 5, 2018

These notes are based off a presentation by Sharad Goel (Stanford, Department of Management Science), for the section on "Civic Participation" in the Mechanism Design for Social Good Reading Group. The notes are taken by members of the reading group with some figures and texts taken from the accompanying paper, One Person, One Vote: Estimating the Prevalence of Double Voting in U.S. Presidential Elections [1]. Questions and comments from reading group members during the presentation are labeled as such. Please contact the reading group organizers with any questions or comments.

1 Estimating Voter Fraud

Disclaimer: The estimates given here are coarse, and should not be interpreted as precise.

- In the last election, there were many claims of voter fraud. Voter fraud might arise from a number of techniques:
  - Tampering with machines
  - Ballot destruction
  - Tampering with vote counts
  - Voter impersonation
  - Non-citizen voting
  - Double voting

  In this talk, we'll focus on double voting.
- Double voting occurs when a voter registers and votes in multiple states. It is not illegal to be registered in multiple states simultaneously. It is, however, illegal to vote more than once in the same election.
- Determining if someone has voted in multiple states is difficult since there is no unique national identifier that is recorded for each voter.

- Voter records are public, meaning that for each registered voter, we can look at first name, last name, date of birth, and whether that person voted.
- We'll adopt the following strategy to get an estimate for the number of double votes:
  - Start with the complete national record of voting.
  - Count records with matching first name, last name, and date of birth. This produces approximately 800k matches.
  - Adjust estimate by the number of matches expected to occur by chance, due to the birthday paradox. This reduces the number of matches from 800k $\rightarrow$ 29k.
- As a first approximation, the standard birthday paradox correction gives a reasonable estimate. However, this assumes a uniform distribution of births on each day throughout the year. In fact, fewer babies are born on weekends (see Figure 1). Taking this into account, the estimate drops 29k $\rightarrow$ 26k.

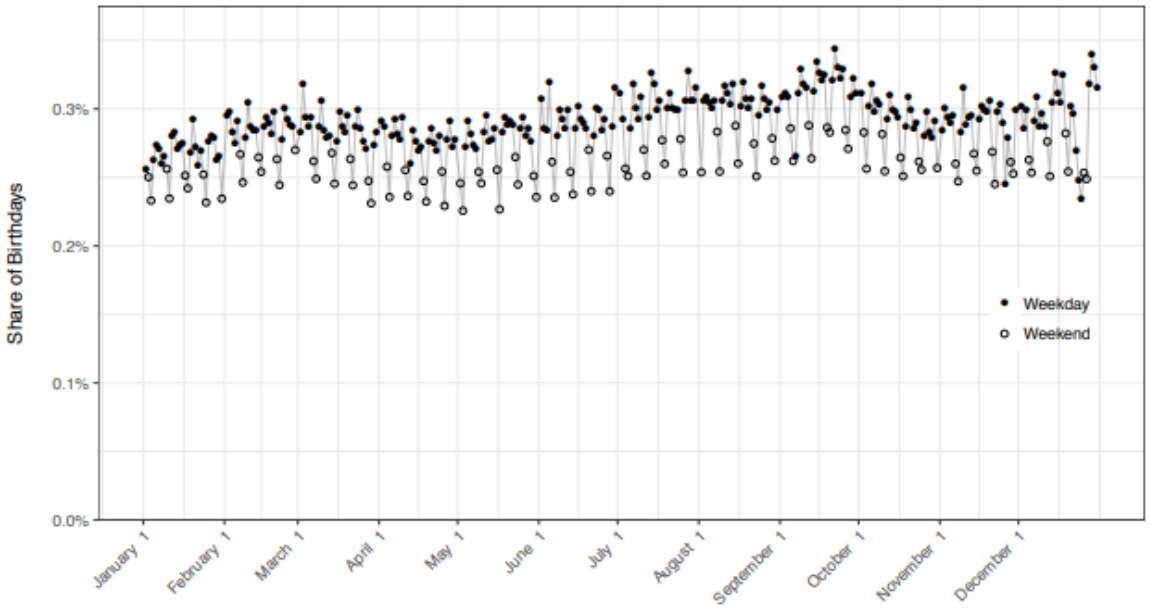

Figure 1: Fluctuation of birthday popularity in 1970

- Another non-uniformity is that first names are not uniformly distributed throughout a year, nor are they uniformly distributed across years. For example, the name June is most popular in June, and names rise and fall in popularity over time. To make this correction, we'll need three assumptions:

  Assumption 1: If years y_1 and y_2 have the same "weekday schedule," meaning they have the same number of days and begin on the same day of the week, assume that for a given name, a person's birthday has the same distribution over weekdays between y_1 and y_2.

Assumption 2: For any fixed day, the distribution of first names is independent of the day of the week. This isn't true, for example, if someone is named Wednesday. In general, however, conditioned on the date, first name doesn't contain much information about the day of the week.

Assumption 3: Birthday and day of the week are independent.

Let B, F, Y, and D be birth date, first name, year, and day of the week respectively. Under these assumptions, we get

$$\Pr[B = b \mid F = f, Y = y] \propto \Pr[B = b \mid F = f] \Pr[D = d_{b,y}] \tag{1}$$

where $\Pr[B = b \mid F = f]$ captures the first-name effect on the distribution over dates, and $\Pr[D = d_{b,y}]$ captures the effect that different days of the week have different likelihoods of birthdays.

The day of week effect is fairly straightforward to estimate. However, the first-name effect is more challenging, since a lot of names are relatively rare. The solution to this is to smooth the empirical distribution with a uniform prior over all dates. This is done by randomly generating $\beta = 10,000$ pseudo-voters according to the uniform distribution, though results don't vary significantly with β.[1]

This correction reduces 26k $\to$ 21k.

- Another potential source of error is in the process of recording that voters vote. We don't observe whether or not someone voted. Instead, we see whether they were recorded as having voted on the electronic record. Getting from the paper record to the electronic record requires manual scanning of barcodes. This could introduce some human errors.

- There are around 1.6 million people with registrations in multiple states (estimated using similar techniques as above). With this many double-registrations, a 1.3% scanning error rate explains nearly all potential double-voting.

- It's hard in practice to estimate the general rate. However, as part of the study, authors manually verified 30,000 vote records in Philadelphia. They compared the electronic vote record to poll books and found a 1% error rate. This isn't necessarily representative of the rest of the country, but if it's in the right ballpark for the error rate, that takes the number of double-votes from 21k $\to \varepsilon$ – nonzero, but very small.

[1] The solution to avoiding artifacts from the dataset when estimating D the distribution of $\Pr[B = b \mid F = f]$ is to use smoothing: Suppose you want to estimate some distribution D from a small number of samples, and you have some prior P on what you think D should be. Let S be the empirical estimate of D from the samples. Then, the smoothed version is $\lambda S + (1 - \lambda)P$ where $\lambda \in [0, 1]$.

The informal way to view smoothing is to sample $\beta = 10,000$ extra voters with first name f and birthdays drawn from $\Pr[B = b]$, but this introduces some noise from sampling. However, if you do this sampling over and over and then take expectations, you end up with the same thing as what smoothing gives you.

2 Potential for Disenfranchisement

- The *Interstate Crosscheck Program* was set up to check if voters are double-registered between multiple states. It returns potential double-voters to states, matching by first name, last name, and DOB.

- Authors obtained the information provided by Crosscheck to Iowa before the 2012 election. They also had access to data that included whether or not the last four digits of the social security numbers were a match. Merging the data, they could determine if there was two registrations with first name, last name, date of birth, and SSN-4 matches in multiple states. Ignoring people with no SSN data, this leaves 30-35k potential double voters with SSN data. Around 26k cases have SSN-4 matches, meaning they likely were the same person registered in multiple states. However, only 7 of these instances actually recorded votes in multiple states. (Of course there's still the possibility of the first 5 SSN digits not matching, or errors in recording who voted as discussed above.)

- On the other hand, people have considered using these matches to purge duplicate registrations. Extrapolating from this data, removing the registration that took place earlier (assuming it's no longer in use) would endanger around 300 votes for every double vote prevented. The takeaway here is that the majority of the effect of trying to prevent double-voting by purging registrations would actually fall on legitimate votes.

3 Stanford Policy Lab

Goal: Provide algorithmic solutions to effect policy change. Some existing projects:

- Stanford Open Policing Project: Collected 130 million traffic stop records from around 30 states. The hope is that this can be used to estimate the prevalence of things like racial disparity and discrimination in local police encounters.

- Developing statistical tests for discrimination. For example, what are the effects of enhanced sentencing laws (i.e. three strikes you're out)? How many years served are because of these laws? No one seems to have the answer to these questions, and analysis is tricky because records of sentencing are all free text. Understanding the effects of these laws requires building natural language models to extract sentencing enhancements from the data.

- Developing decision-making tools. One example is simple heuristics to help judges improve pre-trial decisions. Another is a tool to help with charging decisions. About half of people arrested in San Francisco aren't ultimately charged, but they still may end up waiting the night or weekend in jail waiting for a decision. An algorithmic tool might allow for people who are very unlikely to be charged to be released more quickly.

4 Q & A

- *Question:* Have there been instances in which it seems like agents are trying to behave strategically to game the system?

  Answer: In stop and frisk, we were worried that officers were marking that they suspected contraband only in cases where contraband was found, which could skew a lot of the analysis. In general, we don't find strong evidence of that, but it's a concern people have. We did find that when they started requiring a short narrative for each stop in New York, the number of stops dropped dramatically and the quality of stops improved. Why? Because slightly increasing the cost of stops and forcing officers to think about why they're making the stops changes the response of the officers. Information elicitation could be a useful point of intervention.

  Another instance is in cases where judges are given algorithmic recommendations and deviate from those. Some amount of deviation is good, but when the deviation is large, then the judge is circumventing the point of the risk assessment. One response might be to ask the judge to give a justification for why he or she chose to act against the recommendation. We could also consider scoring mechanisms, such as comparing a judge's record on flight risk vs. number of defendants released to some baseline and evaluating accordingly.

  One more thing we could consider is the way we label things. Algorithmic risk tools often have categories like "low risk," "medium risk," and "high risk," but "high risk" is usually around 20%. These should really be labeled something like "negligible risk," "very low risk," and "low risk." This will almost certainly change the way people will use these scores.

References

[1] Sharad Goel, Marc Meredith, Michael Morse, David Rothschild, and Houshman Shirani-Mehr. One person, one vote: Estimating the prevalence of double voting in us presidential elections. Technical report, Working Paper. Available at: https://www.dropbox.com/s/fokd83nn4x6wuw9/OnePersonOneVote.pdf, 2016.