Evaluating Resistance to False-Name Manipulations in Elections

Bo Waggoner

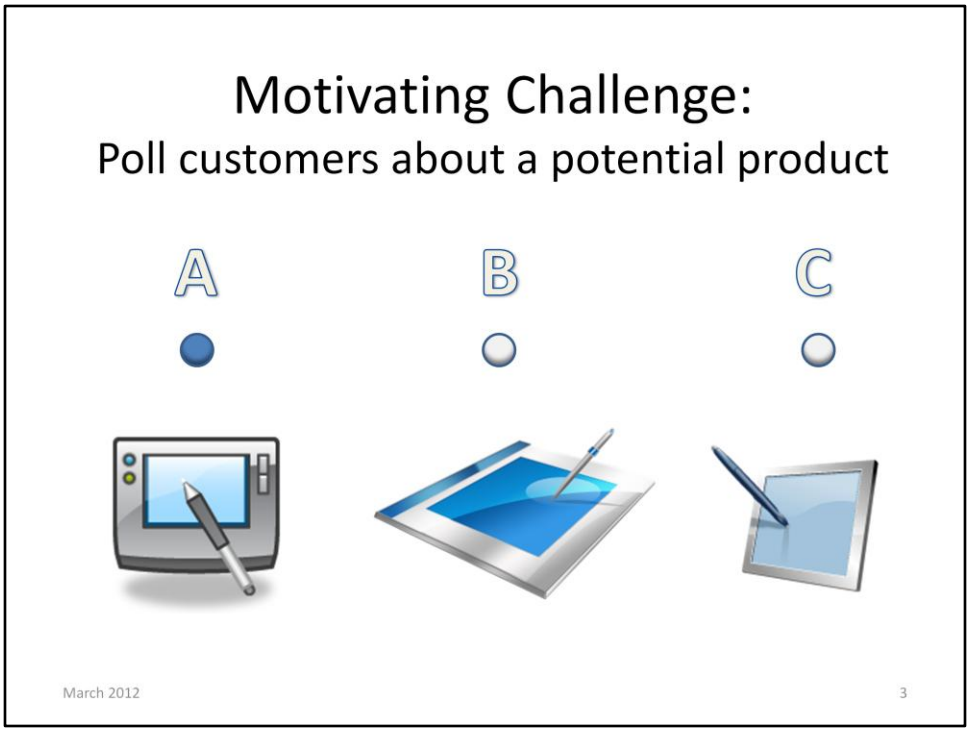
Lirong Xia

Vincent Conitzer

Thanks to Hossein Azari and Giorgos Zervas for helpful discussions!

Outline

- Background and motivation: Why study elections in which we expect false-name votes?
- Our model
- How to **select** a false-name-limiting method?
- How to **evaluate** the election outcome?
- Recap and future work



Suppose we wish to learn people's preferences using an election (possibly in an Internet setting). What challenges do we face in attempting to obtain accurate results?

Preventing strategic behavior

Deter or hinder misreporting

- Restricted settings (e.g., single-peaked preferences)
- Use computational complexity

Currently, there's quite a bit of literature on preventing misreporting of preferences – that is, I actually prefer A to B to C, but I report that I prefer B to C to A, for instance. Of course, the Gibbard-Satterthwaite Theorem shows that we cannot prevent misreporting in general, but we may be able to prevent it in certain settings. So we can only ask questions about which customers generally have single-peaked preferences, for instance. We might also try use a voting rule for which finding a manipulation is computationally hard.

False-name manipulation

- False-name-proof voting mechanisms?
- **Extremely** negative result for voting [C., WINE'08]
- Restricting to single-peaked preferences does not help much [Todo, Iwasaki, Yokoo, AAMAS'11]
- Assume creating additional identifiers comes at a cost [Wagman & C., AAAI'08]
- Verify some of the identities [C., TARK'07]
- Use social network structure [C., Immorlica, Letchford, Munagala, Wagman, WINE'10]

Overview article [C., Yokoo, AIMag 2010]

Common factor: false-name-*proof*

March 2012 6

But even if people are reporting true preferences, they might be reporting them multiple times. False-name manipulation is the creation of false identities in order to cast multiple votes.
Traditionally in voting, we assume that each agent participates exactly once. If we allow people to participate any number of times, we get an extremely negative result (the best we can do is the unanimity rule). Other prior work identifies assumptions we can make that allow us to recover false-name-proof voting mechanisms.
The problem is that, in general, we may find these assumptions too restrictive for a common voting setting.
But moreover, perhaps their goal is also too restrictive: We don't necessarily require false-name-proof rules, we just want to get the correct answer with high probability. And we think that perhaps we can do that if we just resist or limit false-name manipulations.

Let's at least put up some obstacles

140.247.232.88 jmhzdszx@sharklasers.com

Issues:

1. Some still vote multiple times
2. Some don't vote at all

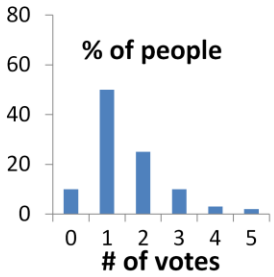
March 2012 7

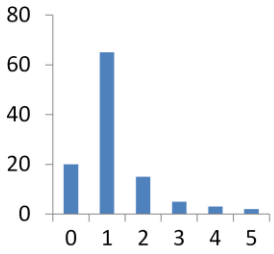
So in real elections, we tend to use false-name limiting methods such as CAPTCHAs, allowing only one vote per IP address, or so on. But these are false-name-limiting, not false-name-proof, methods.

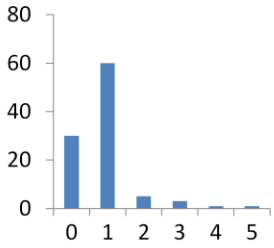
Approach

Suppose we can experimentally determine how many identities voters tend to use for each method.

140.247.232.88 jmhzdszx@sharklasers.com







March 2012 8

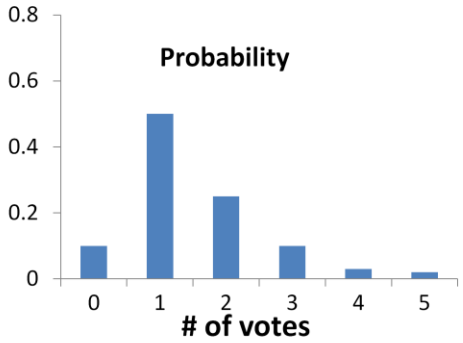
How can we study these methods theoretically?
Suppose that we can go out and run experiments to determine the effects of these false-name-limiting methods. Specifically, we try to discover how many false-name votes are being cast depending on the method used. For example, say hypothetically that, for method one, 10 percent of people who visit the page end up not bothering to vote at all; perhaps 50 percent choose to cast exactly one vote, and 25 percent cast exactly two votes, and so on. Then they come back to you with charts that look something like this.
So suppose we find out that, with a CAPTCHA, 10% of people don't bother to vote at all, 50% cast exactly one vote, and some people come back and cast 2 votes, 3 votes, etc.
More stringent methods might mean fewer people cast extra votes, but might also mean fewer people bother to vote at all. But now, we can figure out a theoretical handle on this problem.

Outline

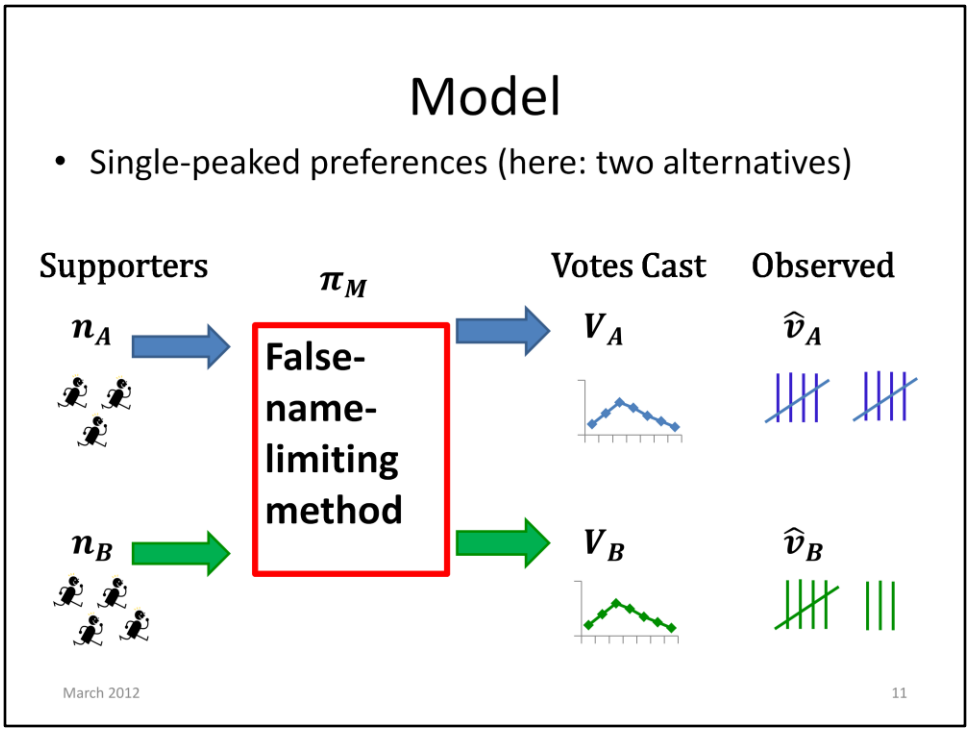
- Background and motivation: Why study elections in which we expect false-name votes?
- Our model
- How to **select** a false-name-limiting method?
- How to **evaluate** the election outcome?
- Recap and future work

Model

- For each false-name-limiting method, take the individual vote distribution π as given
- Suppose votes are drawn i.i.d.



First, we're going to assume that these individual vote distributions are inputs into our model. (Suppose for example that someone else has done the research or modeling to determine, for certain false-name-limiting method like CAPTCHAs or email registration, what the individual vote distribution looks like.) Furthermore, we'll assume that voters draw their number of votes i.i.d. from this distribution.



We're interested in nonstrategic voting settings: no matter how many false identities I use, I should always vote for the same alternative. So we've considered this in the case of single-peaked-preferences, and we present it here for the case of two alternatives.

Model: We have n_A supporters of alternative A and n_B of B. They draw votes i.i.d. according to the individual vote distribution for our chosen method, producing some number of total votes for each alternative. We'll let V_A and V_B be random variables for these total numbers of votes, and v_a-hat and v_b-hat are the actually observed numbers of votes.

We'll ask two questions. The first is, given a certain number of supporters of each alternative, which false-name-limiting method should we choose so that we get a "good" outcome? The second is, if we observe a certain outcome, how can we evaluate whether it is "good"; whether we can be confident that it reflects underlying preferences?

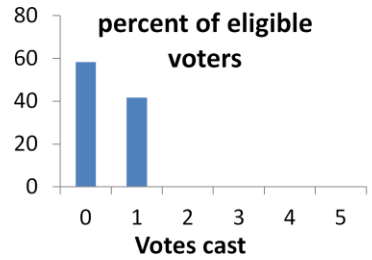
Outline

- Background and motivation: Why study elections in which we expect false-name votes?
- Our model
- How to **select** a false-name-limiting method?
- How to **evaluate** the election outcome?
- Recap and future work

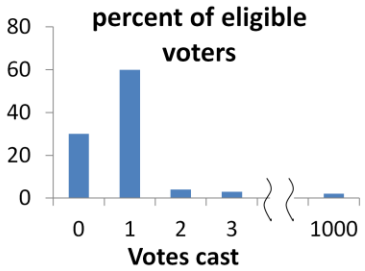
Example

- Is the choice always obvious?
- Individual vote distribution for 2010 U.S. midterm Congressional elections:

Actual (in-person)



Hypothetical (online)



March 2012

13

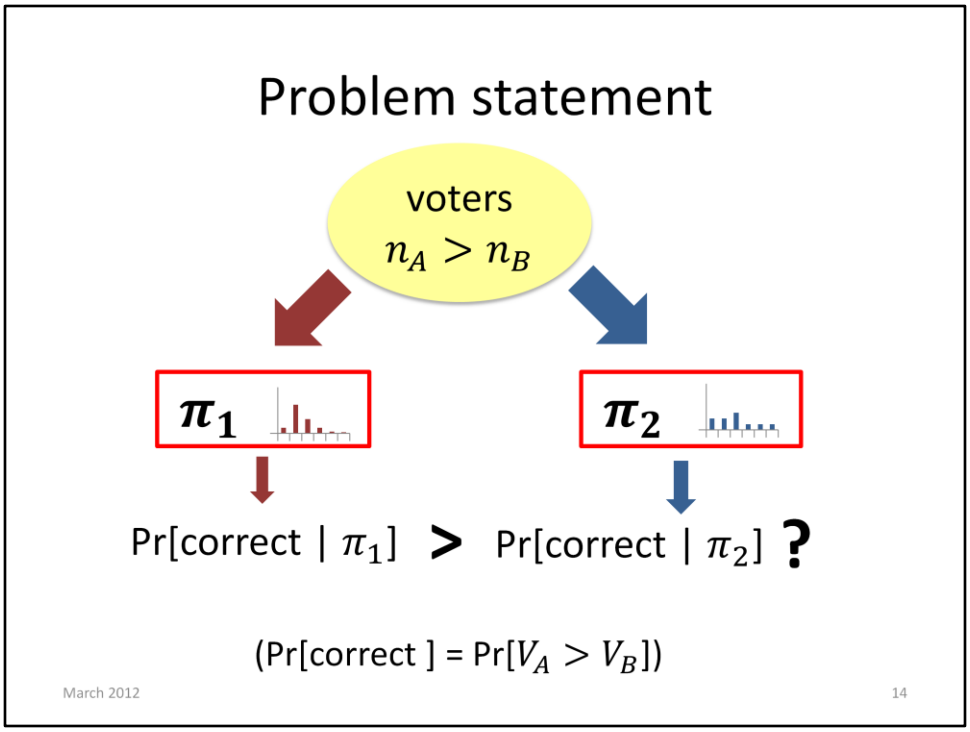
OK, we're now ready to address the first big question: How do we select among different false-name limiting methods?

It seems clear that we want to select a method that doesn't allow people to vote too many times, but I want to quickly convince you that the choice is not obvious in general. So here's a real-life example. This is the individual vote distribution for the US Congressional elections in 2010. 41.7% of the eligible voters cast a vote, and 58.3% did not. (Source: http://elections.gmu.edu/Turnout_2010G.html)

Notice that we had a quite restrictive false-name-limiting method in place here. Registering to vote is somewhat time consuming, requires identification, and so on, and then to actually vote you need to take time off of work or whatever during the day, transport yourself to a polling station, wait in line, and cast your vote.

OK, so what's the alternative? Well, say I proposed that Americans vote online. We can speculate that we'd get a much higher voter turnout rate, but critics would worry that people would manage to cheat the system and vote multiple times (imitating someone else, for instance). You might even get a tiny fraction of people who manage to steal people's identities and cast, say, 1000 votes. So perhaps the individual vote distribution would look like this.

Where we have a lot more people participating, but also some cheaters, and maybe some really serious cheaters. And even if you have hard numbers, it's not really clear a priori which of these distributions is more likely to select an alternative that is actually preferred by the most people.



OK, suppose we will run an election in which more voters support alternative A. We want to know which is better: method 1, which has this individual vote distribution; or method 2, which has this. We can compute probability of a "correct" outcome under each distribution. (Correct just means that A gets more votes than B.)
And the question is, which is better? Well, this is easy for any given supporter profile. If I tell you there are 10 supporters of A and 8 supporters of B, and give you two distributions, you can run this calculation and it's no problem. But how do we compare two distributions in general?

Our results

- We show: which of π_1 and π_2 is preferable as elections grow large
- Setting: sequence of growing supporter profiles (n_A, n_B) where:
  1. $n_A - n_B \in O(\sqrt{n})$ (elections are "close")
  2. $n_A - n_B \in \omega(1)$ (but not "dead even")

We show how to compare these probabilities of correctness as the size of the elections grows large.
So we'll have these supporter profiles where n_A people prefer alternative A and n_B prefer B, and we'll take a sequence of these profiles that get larger and larger – more and more people. We're going to suppose that more people prefer A than B. So as we make the election larger, A should be preferred by a few more people. And as we make it even larger, A should be preferred by a few more people. And we just need a few reasonable bounds on this growth.
First, we'll want these elections to be reasonably close. So if n is the total number of supporters – n_A plus n_B – then the margin of victory should grow no faster than square root of n. If the margin of victory is growing faster than that, then in some sense we expect *any* false-name-limiting method to work, because alternative A will have so many more supporters.
But second, we do still need there to be some margin of victory. If the margin of victory is staying constant as the number of supporters diverges – say A always wins by 10 votes as the number of voters goes to infinity – we think of that as a somewhat pathological case, and in elections that are this close, we can't really say which method is better because both of them are going to give about a 50-50 chance that you get the right answer. (Because about half the people will prefer A and half will prefer B.)
So assuming these conditions are satisfied, we have the following theorem.

Selecting a false-name-limiting method

Theorem 1.

Suppose $\frac{\mu_1}{\sigma_1} > \frac{\mu_2}{\sigma_2}$. *Then eventually*

$\Pr[\text{correct} \mid \pi_1] > \Pr[\text{correct} \mid \pi_2]$.

"For large enough elections, the ratio of mean to standard deviation is all that matters."

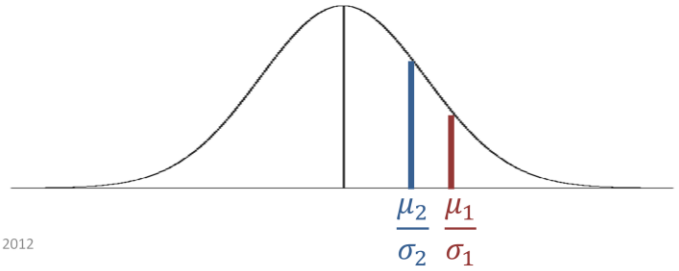
Let's suppose that method one has a higher mean over standard deviation than method two. Then for sufficiently large elections, method one is always more likely to produce a correct outcome.

So what this result shows is that, if we have an individual vote distribution, what really matters is the ratio of the mean to standard deviation. So we want people on average to cast a lot of votes each, but we want everyone to be casting about the same number of votes, and this ratio tells us exactly how to trade off those characteristics.

Selecting a false-name-limiting method

Intuition.

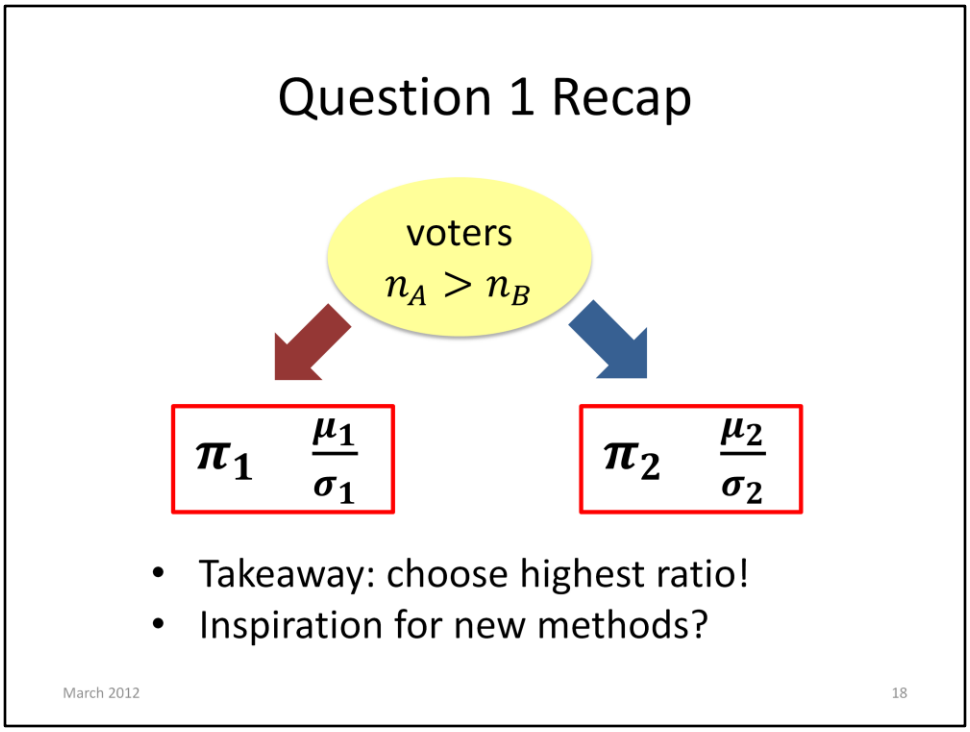
- Distributions approach Gaussians
- Pr[correct] = Pr[$V_A > V_B$] = Pr[$V_A - V_B > 0$] approaches $\Phi\left(\frac{\mu}{\sigma}\frac{n_A - n_B}{\sqrt{n}}\right)$.



March 2012

17

The proof is a little tedious, but the intuition is very straightforward. First, the distributions of the number of votes cast approach Gaussians, and we can use a Berry-Esseen bound to find out how quickly they approach Gaussian distributions. So we know that the probability that an election outcome is correct, which by assumption is the chance that A gets more votes than B, approaches the standard normal distribution of this value. So the picture looks like this:
Here's the approximate probability of a correct outcome under method 2, and here under method 1. And as long as the election is reasonably close, then when n is large enough, this difference is large enough to ensure that method 1 always gives a higher chance of correctness.



So the takeaway is that, it seems we should prefer methods with a high "signal-to-noise" ratio. And possibly this could inspire new methods, because current methods seem focused on pushing everything down – both mean and standard deviation. But this shows that we can do better if we can actually increase the mean without increasing the standard deviation too much.

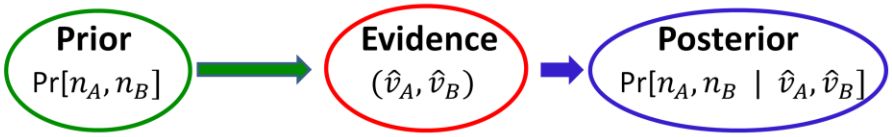
Outline

- Background and motivation: Why study elections in which we expect false-name votes?
- Our model
- How to **select** a false-name-limiting method?
- How to **evaluate** the election outcome?
- Recap and future work

OK, so suppose we select some particular false-name-limiting method and we run our election. All we observe is the number of votes cast. So now we have the inverse problem – given the outcome, how likely is it to be correct?

Analyzing election results

- Observe votes $\hat{v}_A > \hat{v}_B$
- One approach: Bayesian



Requires a prior, which may be
- costly/impossible to obtain
- biased or open to manipulation

- Our approach: statistical hypothesis testing

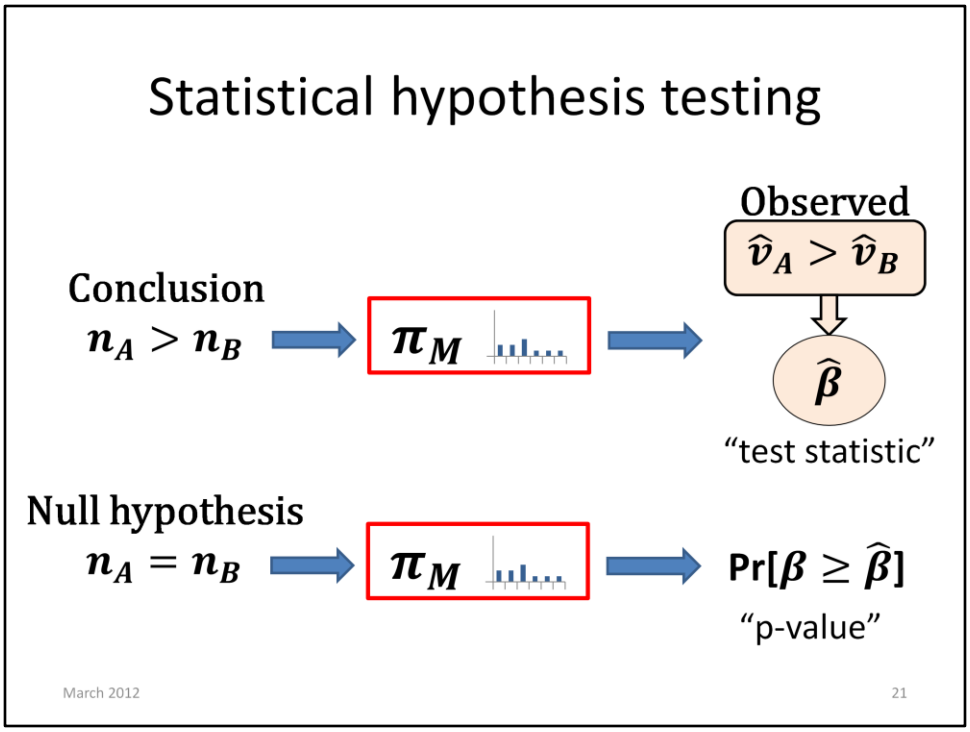
March 2012 20

So we've used a particular false-name limiting method to run an election, and we've observed a certain number of votes for A, and a certain number for B. Suppose A got more votes; how sure can we be that more people actually prefer A?

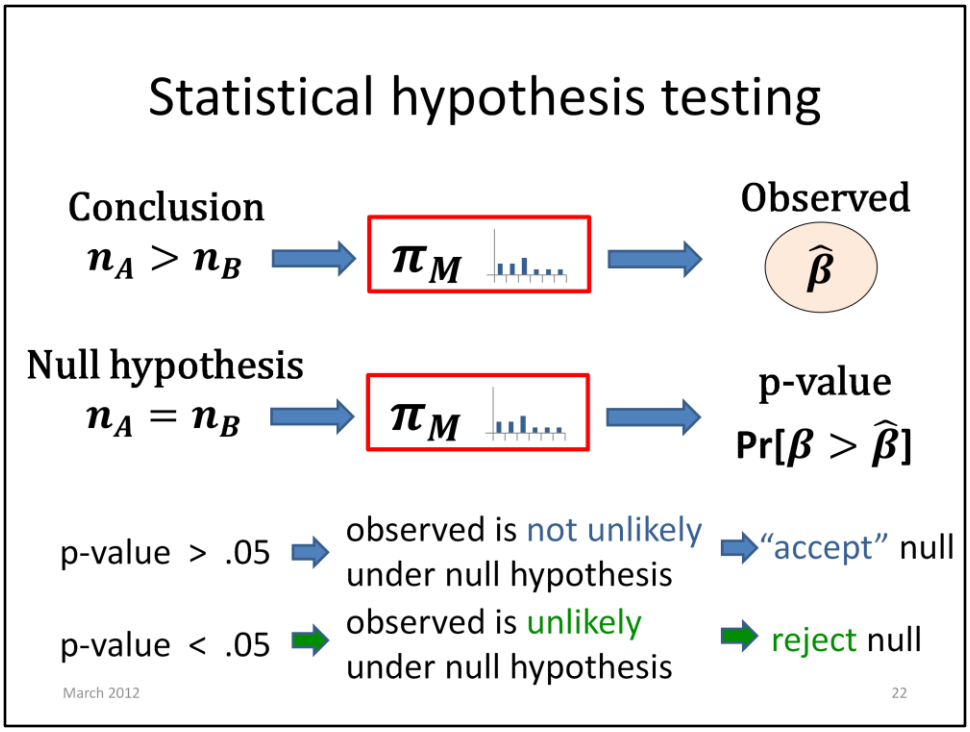
We could take a Bayesian approach to the question. We have a joint prior on the parameters, we observe some evidence, and obtain a posterior over how many people support each alternative. We still wouldn't be done – we'd have to compute the probability that A has more supporters than B, and determine some significance level above which we accept the election results and below which we say the election is inconclusive.

We didn't like this approach for several reasons. First, a prior doesn't seem to help unless we actually have prior information. In general, this information may be costly to obtain or impossibly to obtain. Second, in many settings we might value fairness. So we don't want any possibility of manipulating the outcome. Even if we require that the prior be neutral between A and B, it might still be biased towards, say, close elections, with effects on the confidence we later compute in the outcome. And there doesn't seem to be a standard "neutral" or "fair" prior that we can use in this setting.

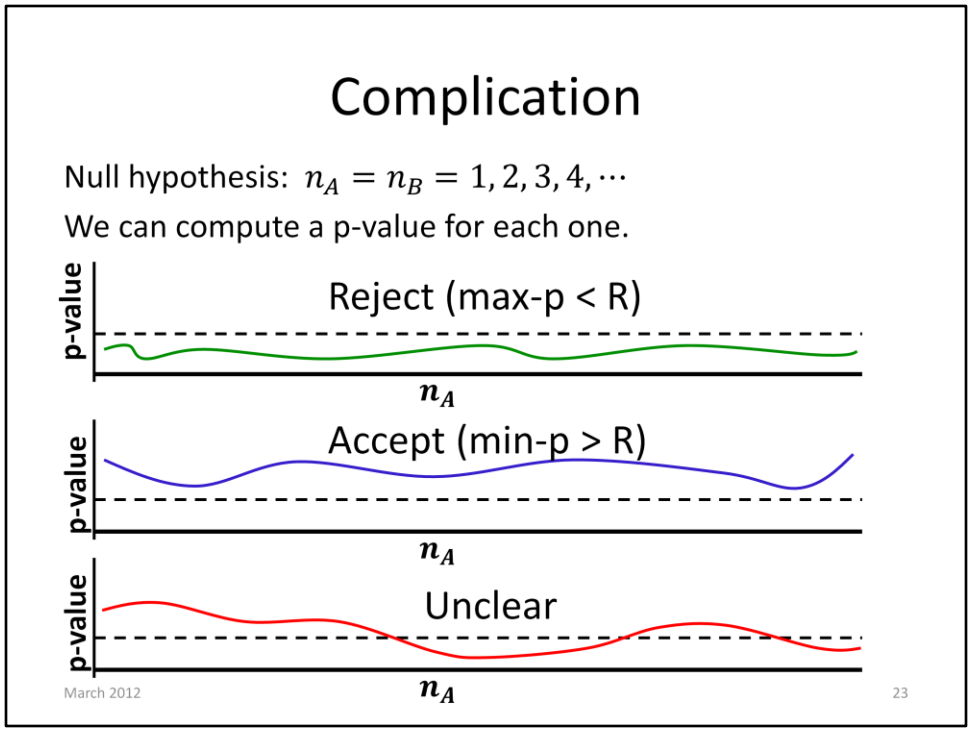
So we decided to use statistical hypothesis testing. So now I'll briefly explain this approach in general, and show how we applied it.



Generic statistical hypothesis testing:
-observe an outcome – in our case, the number of votes received by A and by B
-assume A gets more votes than B
-we use this to compute a test statistic beta hat. We'll try to figure out exactly how later, but we think of it as a number describing how much A won by, in some sense.
-we draw a conclusion from this observed outcome, and here the conclusion to be drawn is that there were more supporters of A than of B
-Problem: What if this outcome or effect was just due to chance, not to a difference in parameters?
-formulate a null hypothesis: suppose that actually, it was neutral between A and B and the effect was just due to chance
-Question: then what is the probability that we observe an outcome as or more extreme than what we did see? (This is the p-value.)



If p-value is above some threshold, like .05, then we think it's actually likely that we observe beta hat under the null hypothesis, so we must accept the null hypothesis – that our results may be due to chance. And that means we can't confidently draw this conclusion that A has more supporters.
If p-value is below the threshold, then it's unlikely to observe beta hat under the null hypothesis, so we reject the null hypothesis and we can confidently draw this conclusion.



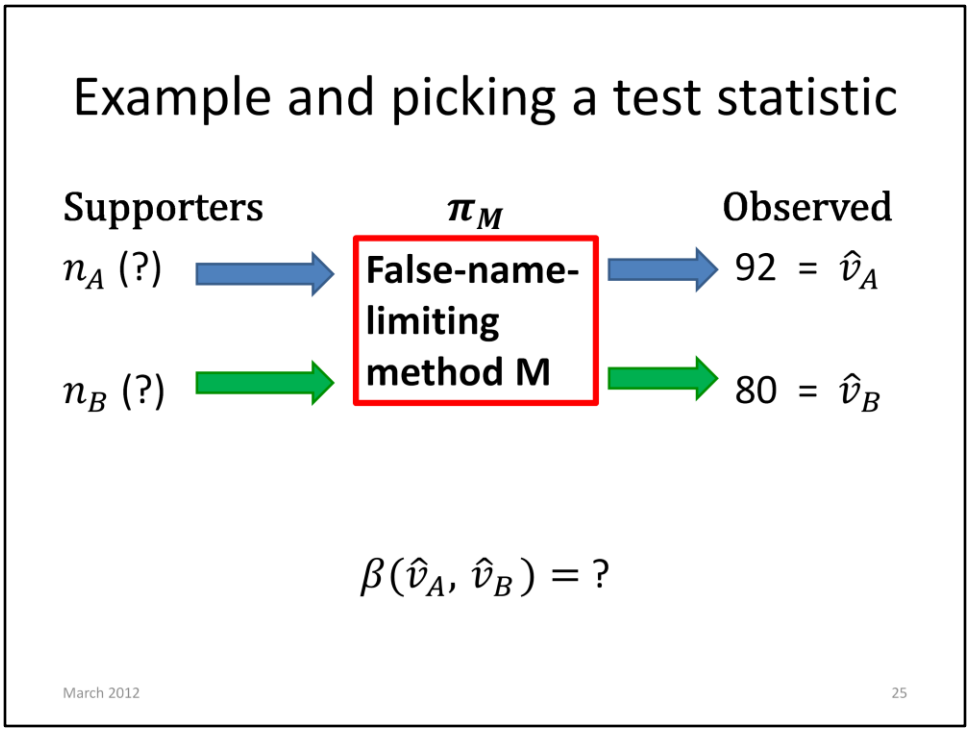
Complication: We actually have many null hypotheses – it could be that A and B each have one supporter, or each have 2, or so on.
Solution: we start by computing a p-value for each one
-Suppose that all the p-values are below our threshold
-Then the max-p-value is below the threshold
-Then definitely reject null hypothesis – unlikely in all circumstances.
Conversely, suppose all p-values are above the threshold
-Then the min-p value is above the threshold
-Then definitely accept null hypothesis – likely in all circumstances.
But, we have to worry about when it's unclear whether to accept or reject. We'll try to pick a test statistic that minimizes the chances of this happening and always falls into one of those categories.
There's a second complication. In our case, we only observe one data point – one election. Most statistical tests, as far as we know, tend to assume or require more data than that (for example, so we can compute a sample variance). So we don't know of a statistical test that really applies to this setting. Instead, we'll propose our own, which will have similarities with common tests.

Our statistical test

Procedure:

1. Select significance level R (e.g. 0.05).
2. Observe votes $\hat{v}_A > \hat{v}_B$.
3. Compute $\hat{\beta}$.
4. If $\max_{n_A=n_B}$ p-value < R, reject.
5. If $\min_{n_A=n_B}$ p-value > R, don't reject.
6. Else, inconclusive whether to reject or not.

So this is our statistical hypothesis test.
Select some significance or "threshold" level R.
Observe the outcome of the election.
Compute a test statistic.
If the max p-value over all n_A equal to n_ B is below the threshold, reject the null hypothesis. We can confidently conclude that the more people support A than B.
If the min p-value is above the threshold, don't reject the null hypothesis. We cannot be confident that our outcome is correct.
If neither of those happens, our test is inconclusive.
So now we'll show what test to use to compute beta-hat.



So here's an example. We have some unknown numbers of supporters who cast votes according to some distribution pi, and we observe 92 votes for A and 80 votes for B. Now we have to compute a test statistic and calculate our p-values.
Suppose we just take the difference in votes. So we say that A's margin of victory is 12.
We'll compute a test statistic called A's adjusted margin of victory. So what should it be?

Selecting a test statistic

Observed: $\hat{v}_A = 92, \quad \hat{v}_B = 80.$

Difference rule: $\hat{\beta} = \hat{v}_A - \hat{v}_B = 12$

Percent rule: $\hat{\beta} = \frac{\hat{v}_A - \hat{v}_B}{\hat{v}} \approx 0.07$

General form: $\hat{\beta} = \frac{\hat{v}_A - \hat{v}_B}{\hat{v}^\alpha} = \frac{12}{172^\alpha}$

(Adjusted margin of victory)

March 2012 26

One natural choice is to just use the margin of victory and take the difference in votes.
Another is to take the percent of votes that A won by, so 12 votes out of the 172 that were cast.
We'll consider the general form of both of these rules, which is where the winning margin is scaled by some power of the number of votes cast. In the difference rule, alpha is zero; in the percent rule, alpha is one.
And the question to answer, then, is what value of alpha gives a good test, if any.

Test statistics that fail

Theorem 2.

Let the adjusted margin of victory be

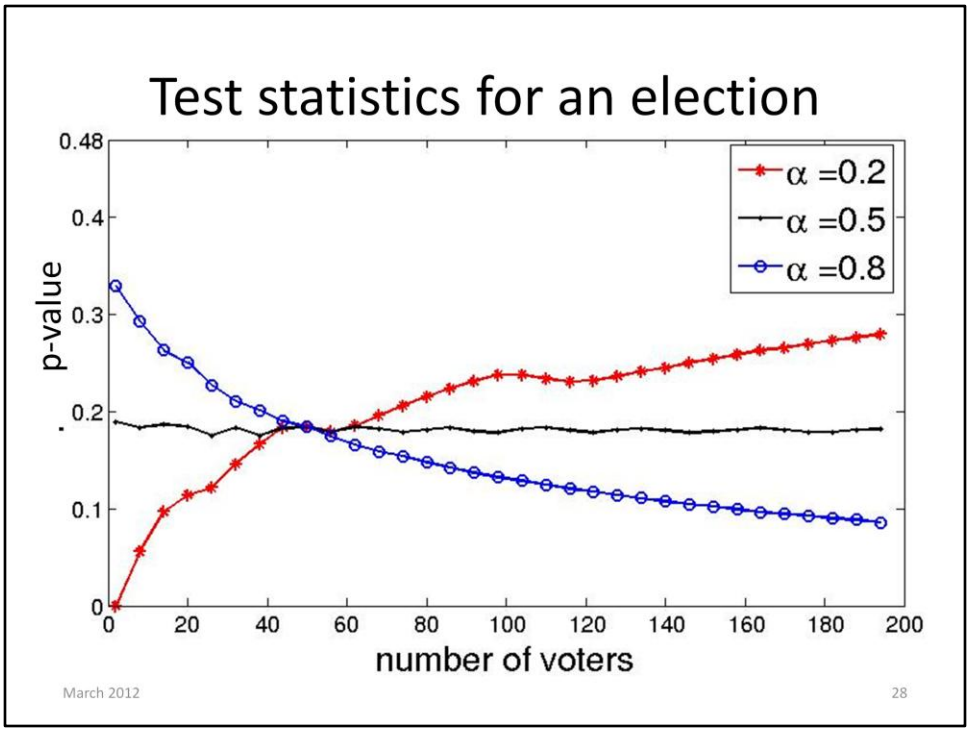
$$\beta = \frac{\hat{v}_A - \hat{v}_B}{\hat{v}^{\alpha}}.$$

Then

1. *For any* $\alpha < 0.5$, max-p = ½: *we can never be sure to reject. (Type 2 errors)*
2. *For any* $\alpha > 0.5$, min-p = 0: *we can never be sure to "accept". (Type 1 errors)*

March 2012 27

Here, we see which values won't work. Suppose our adjusted margin of victory is the difference in votes, scaled by some power of the total number of votes. Notice that the difference rule is alpha equals zero, and the percent rule is alpha equals 1. We say that small values of alpha are always susceptible to Type II errors, where we ought to reject the null hypothesis and conclude that our election outcome is correct, but we don't. Similarly, large values of alpha are susceptible to Type I errors, where we shouldn't reject, but we do and run the risk of erroneously concluding that our election outcome is correct.



Here, we're plotting p-values for a particular election outcome using different adjusted margin of victory formulae. So first, let's look at a small alpha, the asterisks. Here we see that, as n is increasing, the p-value is going to one-half. So we can set any significance level we want – pick any R between zero and 0.5 – and we will still accept the null hypothesis at some point.

Similarly, for large alpha, the circles, we can see that, even for a very stringent R-value – like .001 – we will eventually reject the null hypothesis.

But when alpha is one-half, we always get about the same value, which we kind of think of as the "right" p-value.

The "right" test statistic

Theorem 3.

Let the adjusted margin of victory formula be

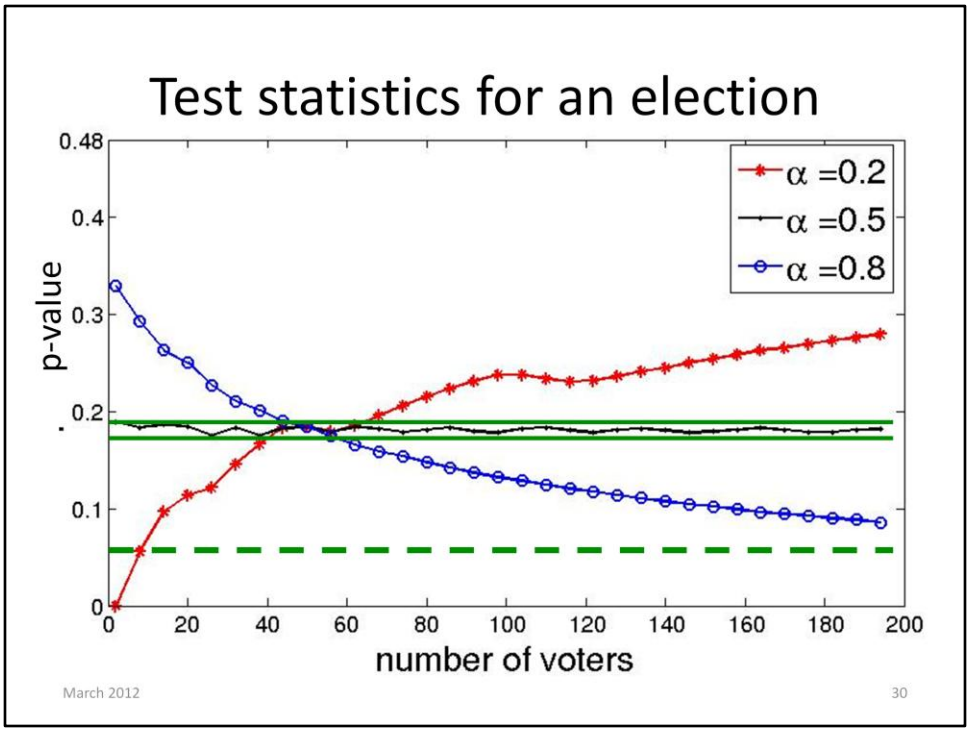
$$\beta = \frac{\hat{v}_A - \hat{v}_B}{\hat{v}^{0.5}}.$$

Then

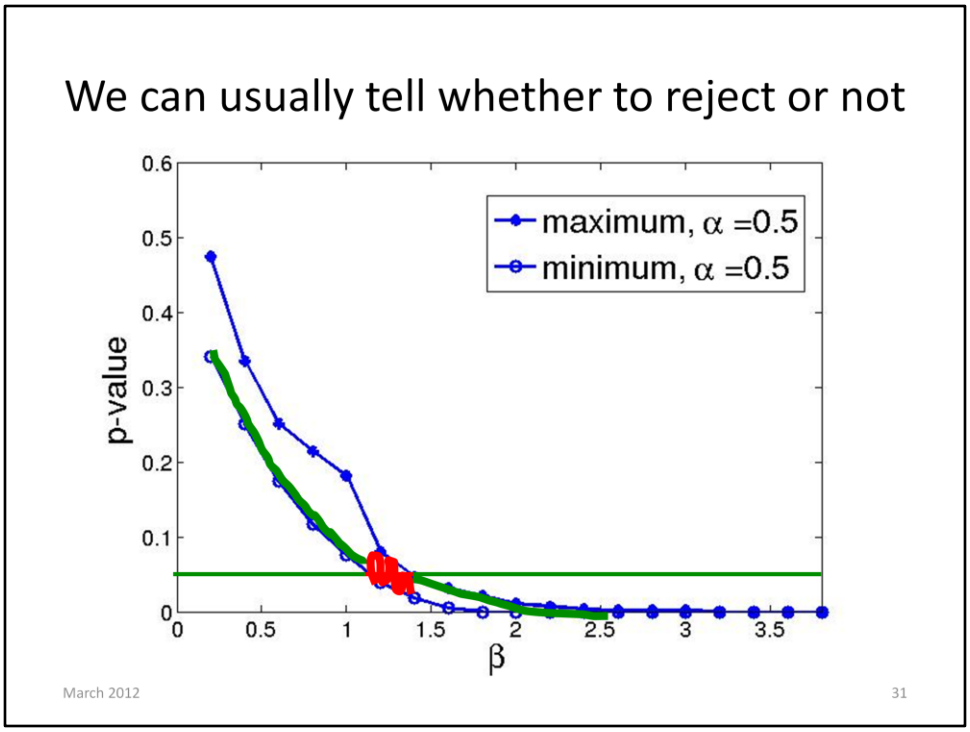
1. *For a large enough $\hat{\beta}$, we will reject. (Declare the outcome "correct".)*
2. *For a small enough $\hat{\beta}$, we will not reject. (Declare the outcome "inconclusive".)*

March 2012 29

This is just the positive result that corresponds to the previous theorem. When we select alpha = 0.5, then if we observe a big enough margin of victory, we will reject the null hypothesis, for all values of n, and conclude that our outcome is good. Conversely, if we observe a small enough margin of victory, then we will not reject the null hypothesis for any value of n, and so we can certainly conclude that the margin of victory is not significant.
This holds for any significance level R between zero and 0.5.



And we think, due to simulation results, that this test actually gives a result very often – that it's actually quite rare to be inconclusive. This image gives the intuition: Usually, we think from simulations that these max-p and min-p values are very close together, which means that, for a given significance level, we probably can tell whether to accept or reject.



Here's another example (again, for a particular false-name-limiting method). This graph tells us the min-p and max-p values you'll get for a given outcome beta-hat. So if we look at a significance level, it's only for a small range that beta-hat will be inconclusive. Usually, the min-p will be above the significance, or the max-p will be below.

Use this test!

1. Select significance level R (e.g. 0.05).
2. Observe votes $\hat{v}_A > \hat{v}_B$.
3. Compute $\hat{\beta} = \frac{\hat{v}_A - \hat{v}_B}{\hat{v}^{0.5}}$.
4. If $\max_{n_A = n_B}$ p-value < R, reject: high confidence.
5. If $\min_{n_A = n_B}$ p-value > R, don't: low confidence.
6. Else, inconclusive whether to reject or not. (rare!)

Recap of this section's results.

Outline

- Background and motivation: Why study elections in which we expect false-name votes?
- Our model
- How to **select** a false-name-limiting method?
- How to **evaluate** the election outcome?
- Recap and future work

Summary

- Model: take π as given, draw votes i.i.d.
- How to **select** a false-name-limiting method?

  A: Pick the method with the highest $\frac{\mu}{\sigma}$.
- How to **evaluate** the election outcome?

  A: Statistical significance test with

  $$\hat{\beta} = \frac{\hat{v}_A - \hat{v}_B}{v^{0.5}}$$

  using max p-value and min p-value.

So the key assumptions we made were to take the individual vote distribution as given, and to assume votes were drawn iid. We used our model to address two key questions.

The first major question is how to design such elections. In our setting, the answer is that, as elections grow large, all that really matters is the “signal-to-noise” ratio – mean over standard deviation.

The second major question is how to evaluate the outcome of an election – how do we know whether to rely on the results. And in our setting, we proposed this statistical significance test where we scale the margin of victory by one over the square root of the number of total votes, and show that this gives us the right kind of test.

Future Work

- Single-peaked preferences (done)
- Application to real-world problems
- Other models or weaker assumptions
- How to actually produce distributions π?
  - Experimentally
  - Model agents and utilities

Thanks!

So for future work. One thing we have done is extend these results to single-peaked preference domains.
An obvious step is to see if we can apply these findings in real-world elections, especially online elections. Ideally, we should be able to use these insights to design better false-name-limiting methods and to better understand the tradeoffs involved when we do something like require email registration to vote. So we'd like to apply these results and even verify them on datasets if we can.
Another question is whether we can weaken some assumptions in our model, or whether there are other models that can be proposed to answer this same question.
And possibly the most interesting direction is to ask, how do we actually get these individual vote distributions? One way is to look at methods like CAPTCHAs experimentally and try to empirically produce them. Another is to develop an agent model for utilities on each outcome, and model false-name-limiting methods as imposing costs on casting each vote.
Thanks!