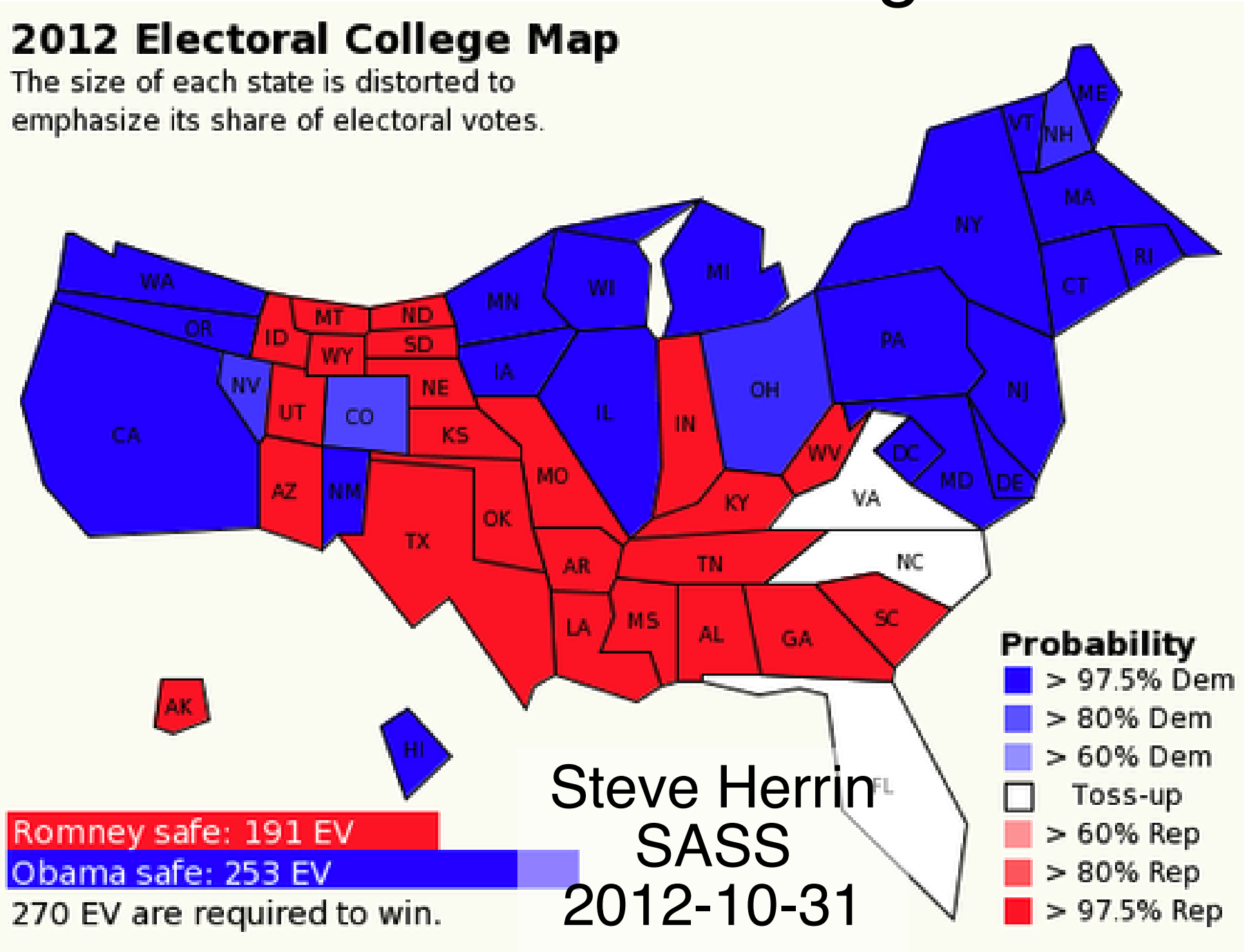
The Science Of Predicting Elections

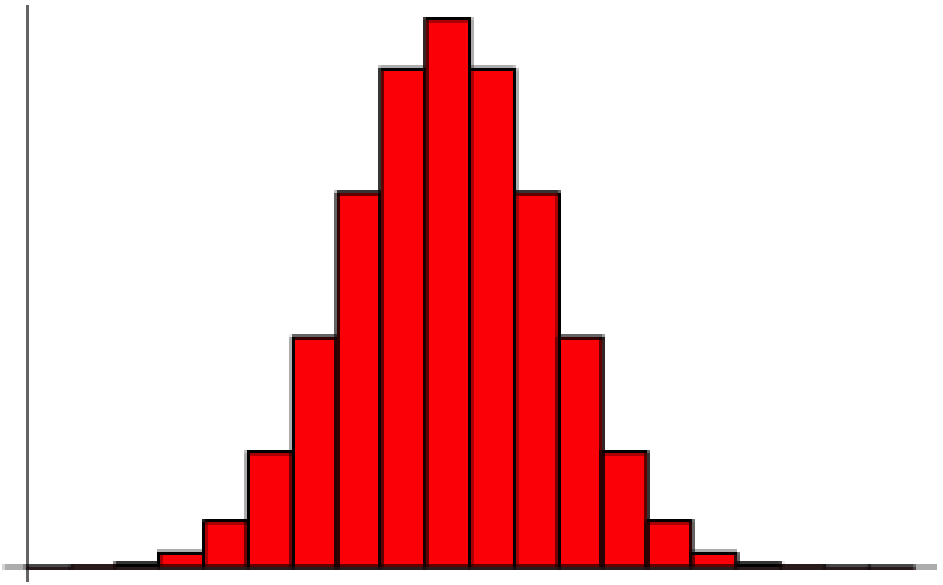


Steve Herrin
SASS
2012-10-31

Binomial Statistics

$$P(n; N, p) = \binom{N}{n} p^n (1-p)^{N-n}$$

Probability of getting *n* occurrences in *N* observations where probability of occurrence is *p*



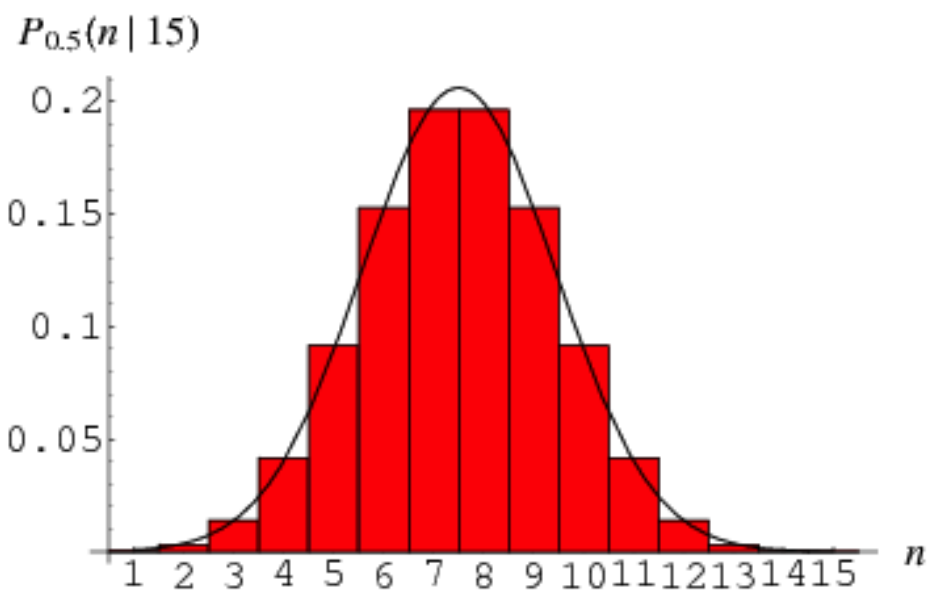
Example:

- If a random person has probability *p* to support a candidate, what is the probability that *n* people say they do when you ask *N*?
- You want to estimate the probability *p* that a person in the whole population supports a candidate. You ask *n* people if they support that candidate, and *k* say they do.

The Normal Approximation

If *np* and *n(1-p)* are both large, then the distribution becomes approximately Gaussian:

$$P(n;N,p)=\frac{1}{\sqrt{2\pi N p(1-p)}}\exp\left[\frac{-(n-Np)^2}{2Np(1-p)}\right]$$



$$\mu = Np$$
$$\sigma = \sqrt{Np(1-p)}$$

Confidence Intervals

Suppose you poll *N* people and *n* say they support candidate A. Then you estimate

$$p = \frac{n}{N}$$

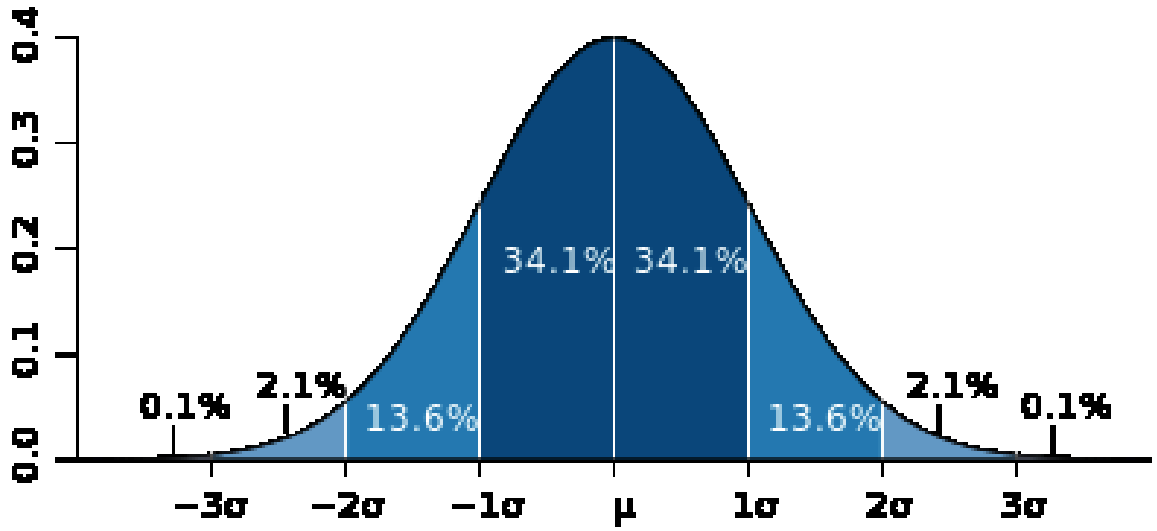
But since you didn't sample the entire population, you need to quantify your uncertainty

$$p = \frac{n}{N} \pm \Delta p$$

Choose Δp such that the probability that the interval correctly contains the true value of p is β

Confidence Intervals

σ	Probability (%)
1	68.3
1.64	90.0
1.96	95.0
2	95.4
2.58	99.0
3	99.7



In the Gaussian approximation, use the standard deviation (σ) to compose confidence intervals

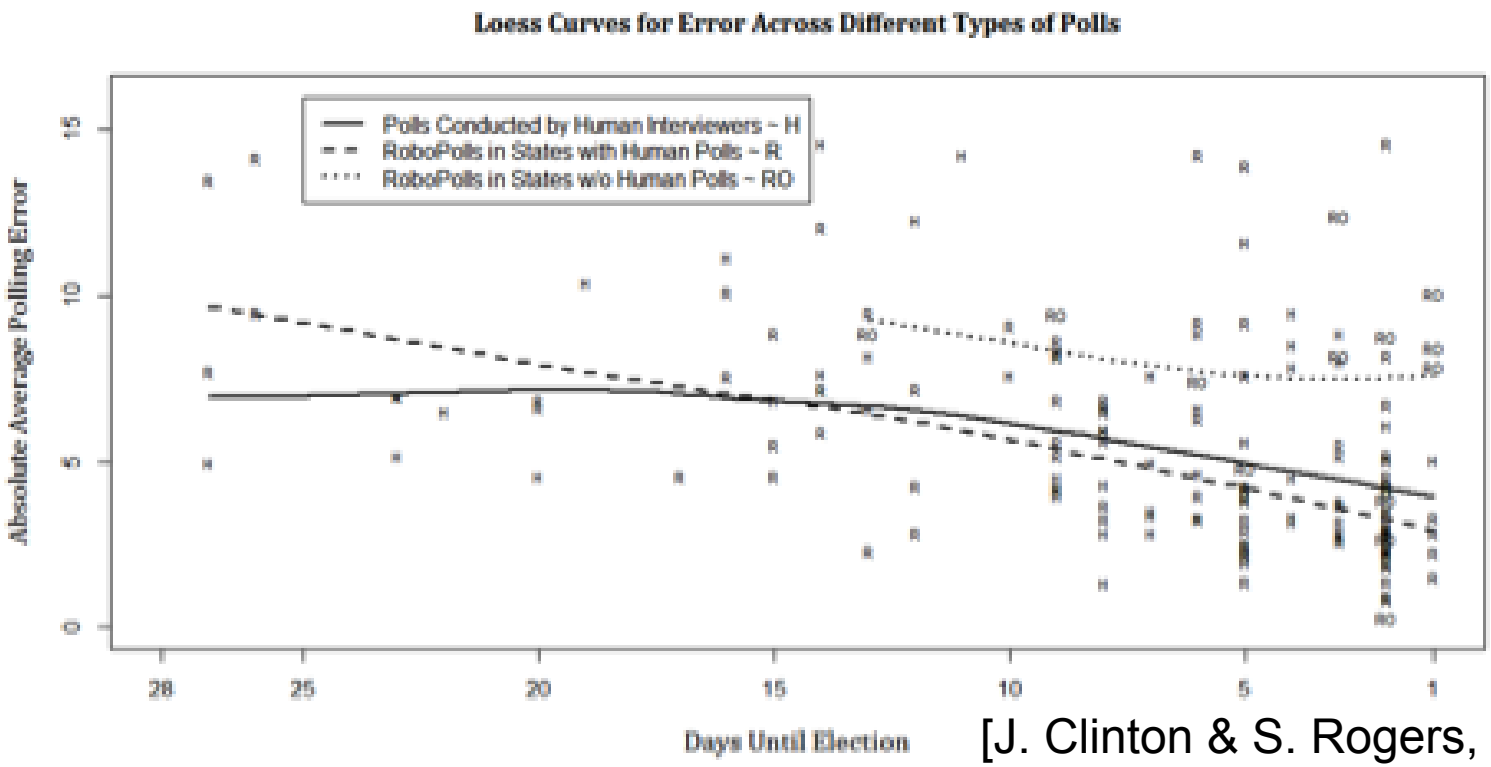
$$p = \frac{n}{N} \pm 1.96 \sqrt{\frac{p(1-p)}{N}}$$

N	p (×100)	1.96 σ (×100)
30	50	17.9
100	50	9.80
300	50	5.65
1000	50	3.10
3000	50	1.79
10000	50	0.98

Doing a Poll

You know how big you want your confidence interval to be, and you can guess at *p* (for an election, probably around 50%), so it's easy to work out what *N* needs to be

- Robo-calling vs. live pollster
  - Bradley effect: are people going to tell a human pollster the more socially acceptable answer, whereas they'll be more honest with a robot?
  - Are robo polls less accurate?



[J. Clinton & S. Rogers, 2012]

Doing a Poll

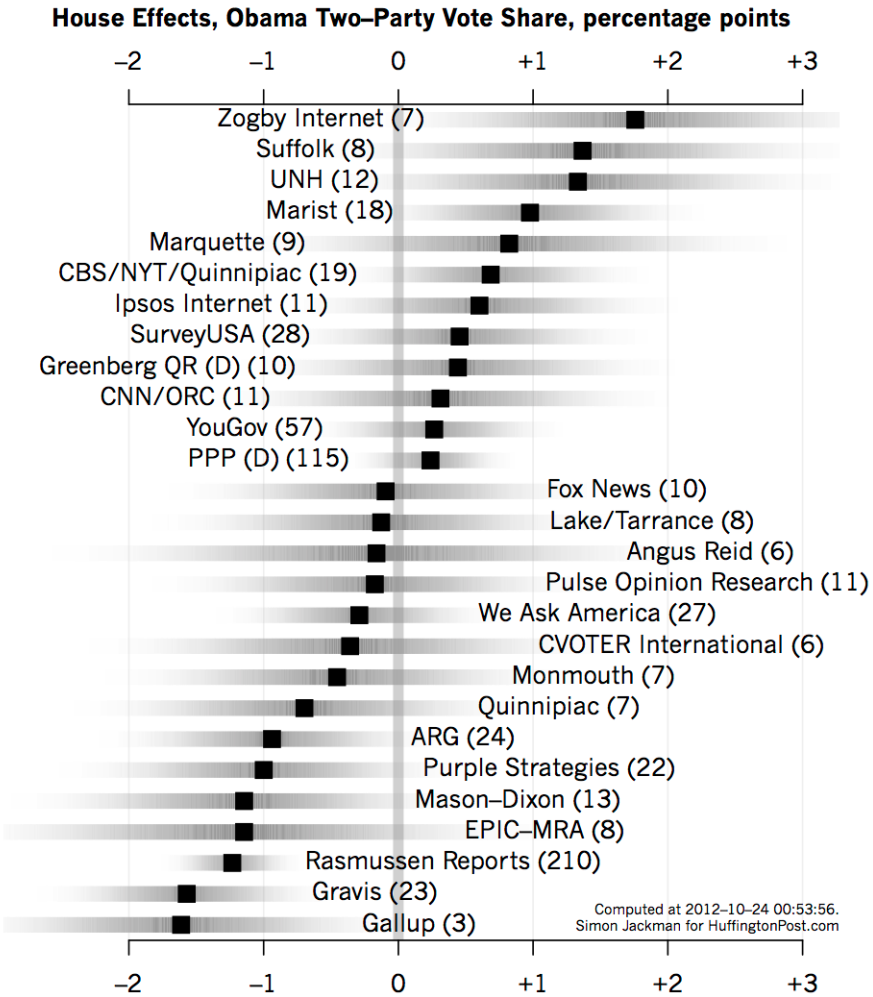
- Weighting:
  - Suppose you know the population you poll is 50% female, but your sample is only 45% female. You can correct for this.
  - “Unskewed Polls” – Party identification is not stable (people tend to say they're democratic if they're planning to vote for that candidate), so you can't correct for that.
- Likely Voter:
  - You probably end up sampling the population with telephones, but this is not the same as the population that votes. So you can ask questions to determine who is likely to vote and weight accordingly
- This is where polling firms tend to differ most

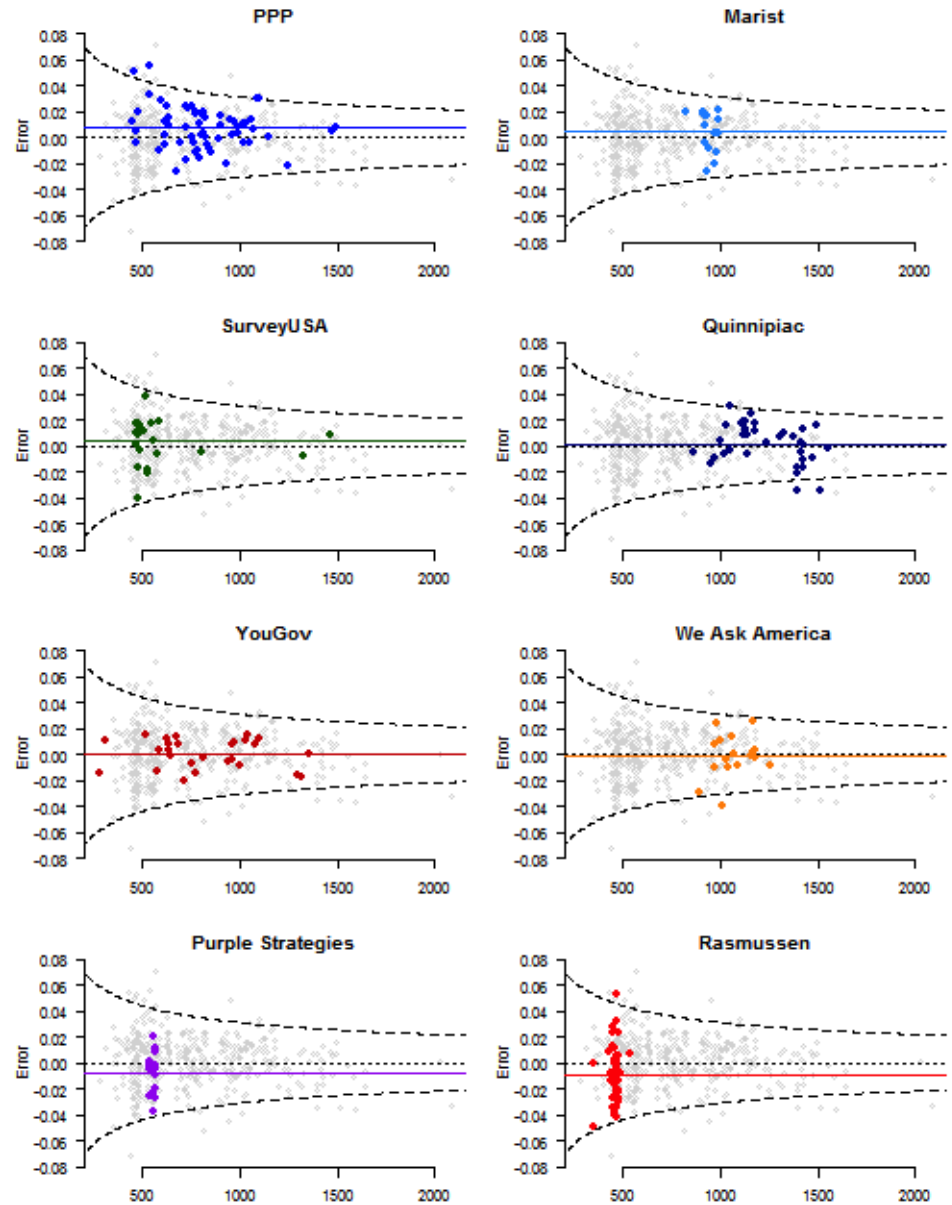
Gallup:

1. Thought given to election (quite a lot, some)
2. Know where people in neighborhood go to vote (yes)
3. Voted in election precinct before (yes)
4. How often vote (always, nearly always)
5. Plan to vote in 2012 election (yes)
6. Likelihood of voting on a 10-point scale (7-10)
7. Voted in last presidential election (yes)

House Effects

[http://huff.to/QG4k2y]

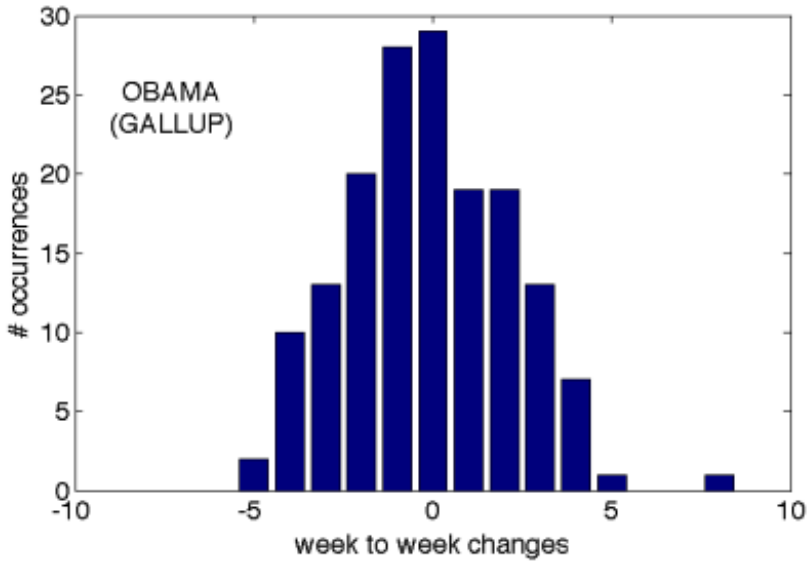


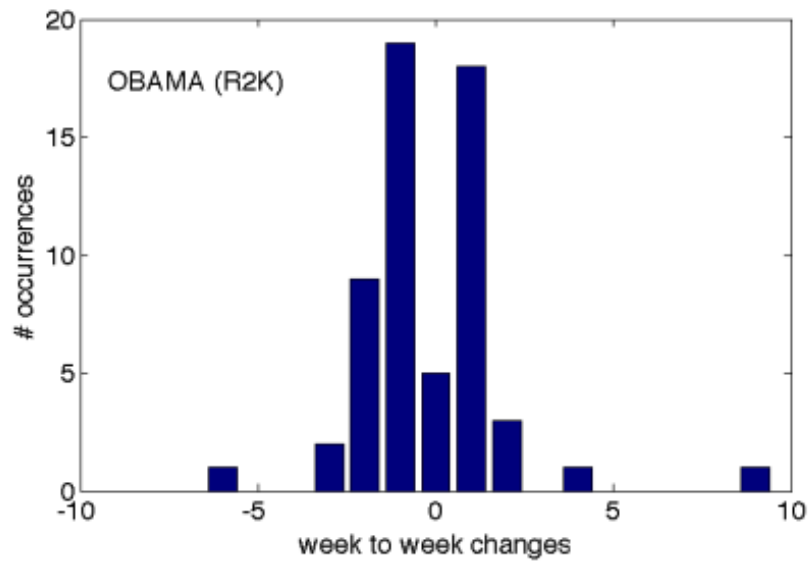


Keep in mind that for 95% confidence intervals, 1 in 20 polls will be wrong

[http://votamatic.org/looking-for-house-effects/]

Or you can just make everything up





6/3/10	Favorable		Unfavorable		Undecided	
Question	Men	Women	Men	Women	Men	Women
Obama	43	59	54	34	3	7
Pelosi	22	52	66	38	12	10
Reid	28	36	60	54	12	10
McConnell	31	17	50	70	19	13
Boehner	26	16	51	67	33	17
Cong. (D)	28	44	64	54	8	2
Cong. (R)	31	13	58	74	11	13
Party (D)	31	45	64	46	5	9
Party (R)	38	20	57	71	5	9

[http://www.dailykos.com/story/2010/06/29/880179/-Research-2000-Problems-in-plain-sight]

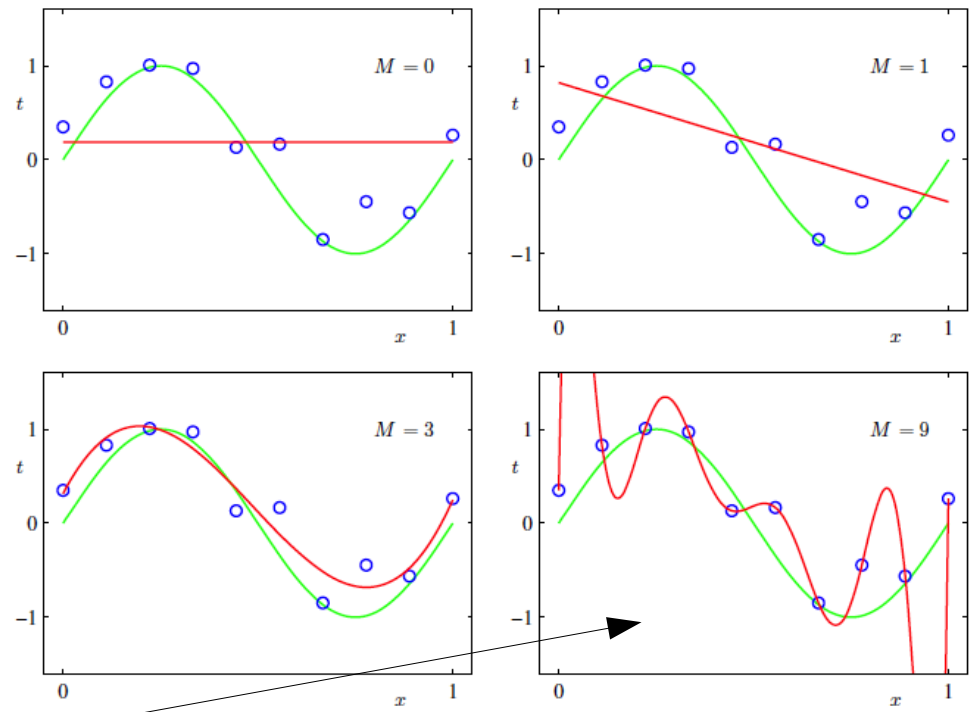
Economic Models

e.g. Ray Fair Model

$$V_P = 48.39 + 0.672\,G - 0.654\,P + 0.990\,Z$$

V_P = Democratic share of the presidential vote
G = growth rate of real per capita GDP in 1st 3 quarters of 2012
P = growth rate of the GDP deflator in the first 15 quarters of the Obama administration
Z = number of quarters in the first 15 of the Obama administration in which the growth rate of real per captita GDP is greater than 3.2 %

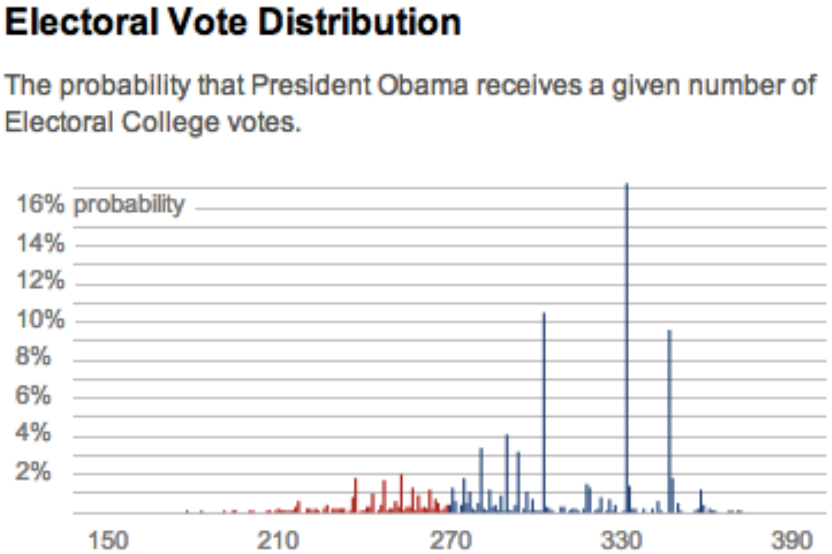
[http://fairmodel.econ.yale.edu/vote2012/index2.htm]



Issues:

- Few data points, so tendency to overfit
  - e.g. Colorado model with 14 parameters fit to 8 elections
  - [M. Berry & K. Bickers, Political Science & Politics 45.04, October 2012]
- Coefficients drift over time
  - Models get updated when they get new results, but what use is that predictively?
  - Voters priorities change over time

FiveThirtyEight



- Methodology:
  - Take weighted average of polls in a state
    - Exponentially decaying weight for older polls
    - Larger weight for polls with larger sample size
    - Certain pollsters receive larger weighting based on their observed error in previous elections
  - Adjust average
    - National polling (so that a state is less wrong if it hasn't been polled recently)
    - House effect (some pollsters have a systematic bias)
    - Likely voter correction applied if the poll only reports registered voters
  - Add in a regression component based on non-poll factors
  - Extrapolate average to what it's likely to be on Election Day
  - Use information from previous elections to estimate error on projection
  - Run Monte Carlo simulations of the election, based on the projection and estimated error

[http://fivethirtyeight.blogs.nytimes.com/]

Monte Carlo Simulation

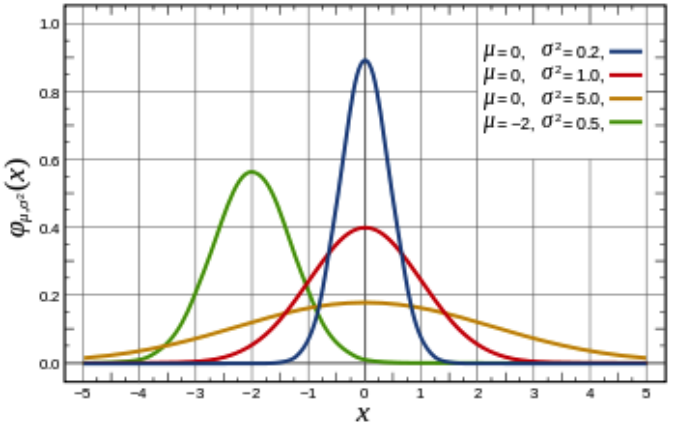
There are many outcomes

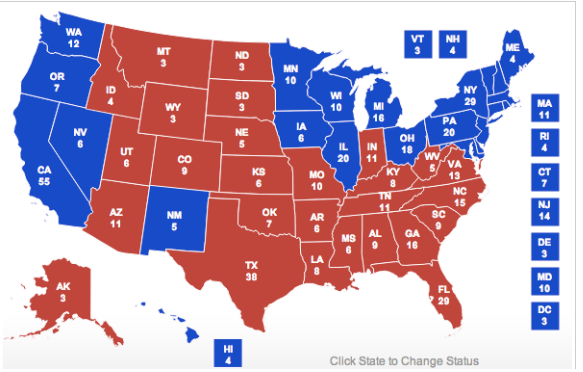
- $2^{51} = 2.3\text{x}10^{15}$(50 states + DC)
- Working out probability for each one is computationally infeasible

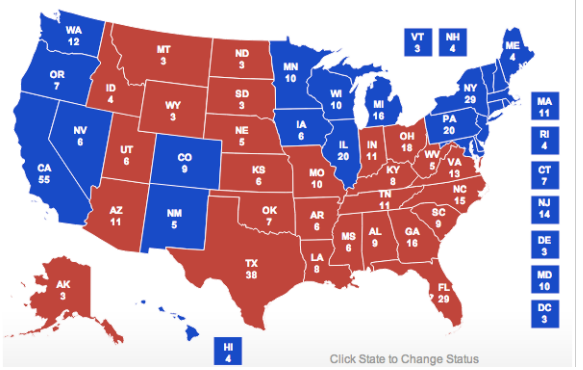
So how do you estimate it?

Monte Carlo method:

- Repeat N times:
  - For each state, determine outcome randomly according to distribution
  - Using the state outcomes, determine the overall outcome
- Then just count different outcomes







FiveThirtyEight

Monte Carlo allows you to look at specific scenarios:

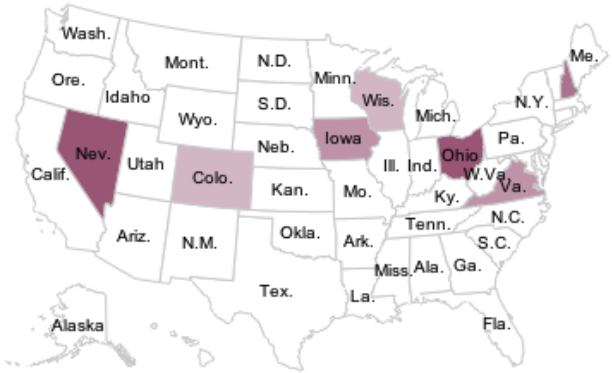
Scenario Analysis

How often the following situations occurred during repeated simulated elections.

Electoral College tie (269 electoral votes for each candidate)	0.4%
Recount (one or more decisive states within 0.5 percentage points)	9.5%
Obama wins popular vote	74.9%
Romney wins popular vote	25.1%
Obama wins popular vote but loses electoral college	2.1%
Romney wins popular vote but loses electoral college	4.6%
Obama landslide (double-digit popular vote margin)	0.5%
Romney landslide (double-digit popular vote margin)	<0.1%
Map exactly the same as in 2008	<0.1%
Map exactly the same as in 2004	<0.1%
Obama loses at least one state he carried in 2008	99.7%
Obama wins at least one state he failed to carry in 2008	5.4%

Return on Investment Index

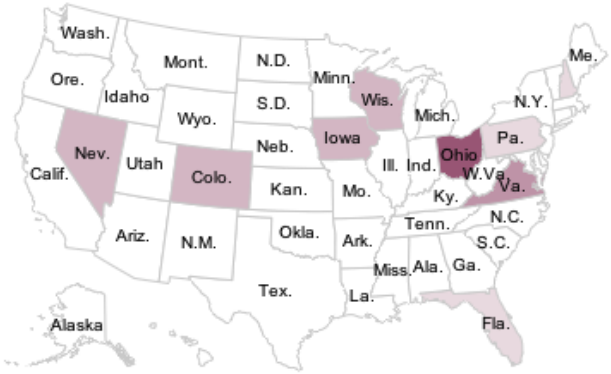
The relative likelihood that an individual voter would determine the Electoral College winner.



1	Ohio	11.5	6	Wis.	3.7	11	N.M.	0.3
2	Nev.	9.5	7	Colo.	3.5	12	Minn.	0.2
3	N.H.	6.7	8	Pa.	0.7	13	Ore.	0.2
4	Iowa	5.1	9	Me. Dist. 2	0.5	14	N.C.	0.1
5	Va.	4.8	10	Fla.	0.3	15	Mich.	0.1

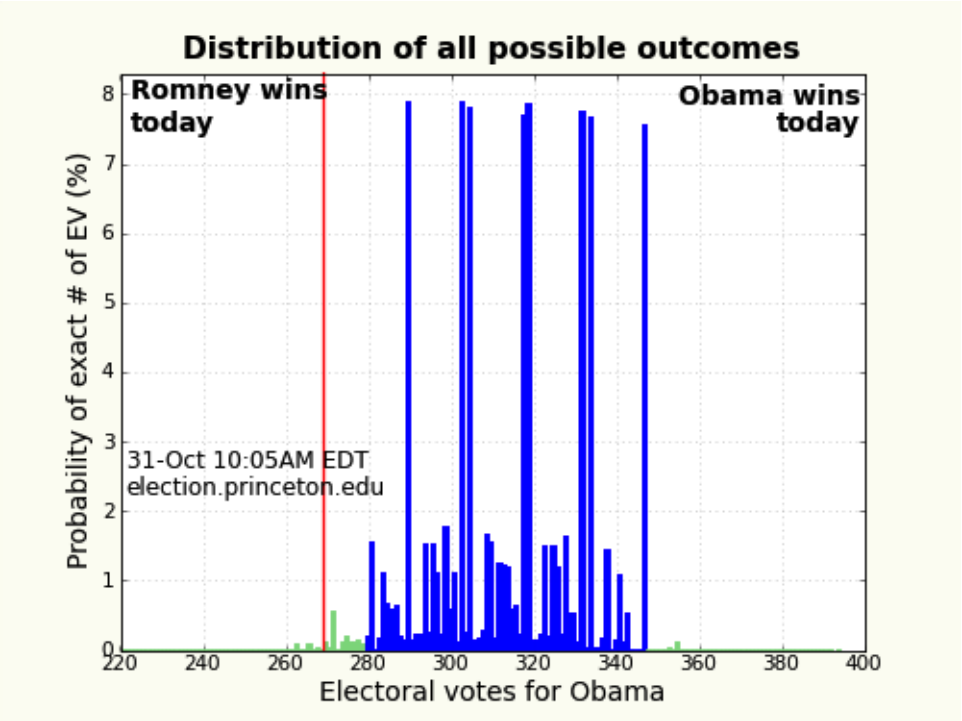
Tipping Point States

The probability that a state provides the decisive electoral vote.



1	Ohio	48.6%	6	Iowa	5.8%	11	Minn.	0.3%
2	Va.	13.5%	7	N.H.	3.4%	12	Ore.	0.3%
3	Wis.	8.2%	8	Pa.	3.0%	13	Mich.	0.3%
4	Nev.	7.1%	9	Fla.	2.1%	14	N.M.	0.2%
5	Colo.	6.4%	10	N.C.	0.4%	15	Me. Dist. 2	0.1%

Princeton Election Consortium



$$f(x) = \prod_{i=1}^{51} \left((1 - p_i) + p_i x^{N_i} \right)$$

Probability of getting N electoral votes is just the coefficient of x^N

p_i = prob. of winning in state i

N_i = number of electoral votes for state i

Methodology:

- Take median (and estimated error on median) of last 3 polls or last week's worth of polls for a state
- Compute probability of winning a state by integrating a Gaussian distribution formed from the median and error
- Compute the polynomial and read off coefficients
- (Can be used for Senate, House election, too)

[http://election.princeton.edu]

Aside: Bayesian Inference

Prior distribution:

$$f_X(x)$$

- Expresses your initial beliefs about the outcome

$$f_{X|Y=y}(x) = \frac{f_X(x)\, L_{X|Y=y}(x)}{\int f_X(x)\, L_{X|Y=y}(x)\, dx}$$

You observe *Y*

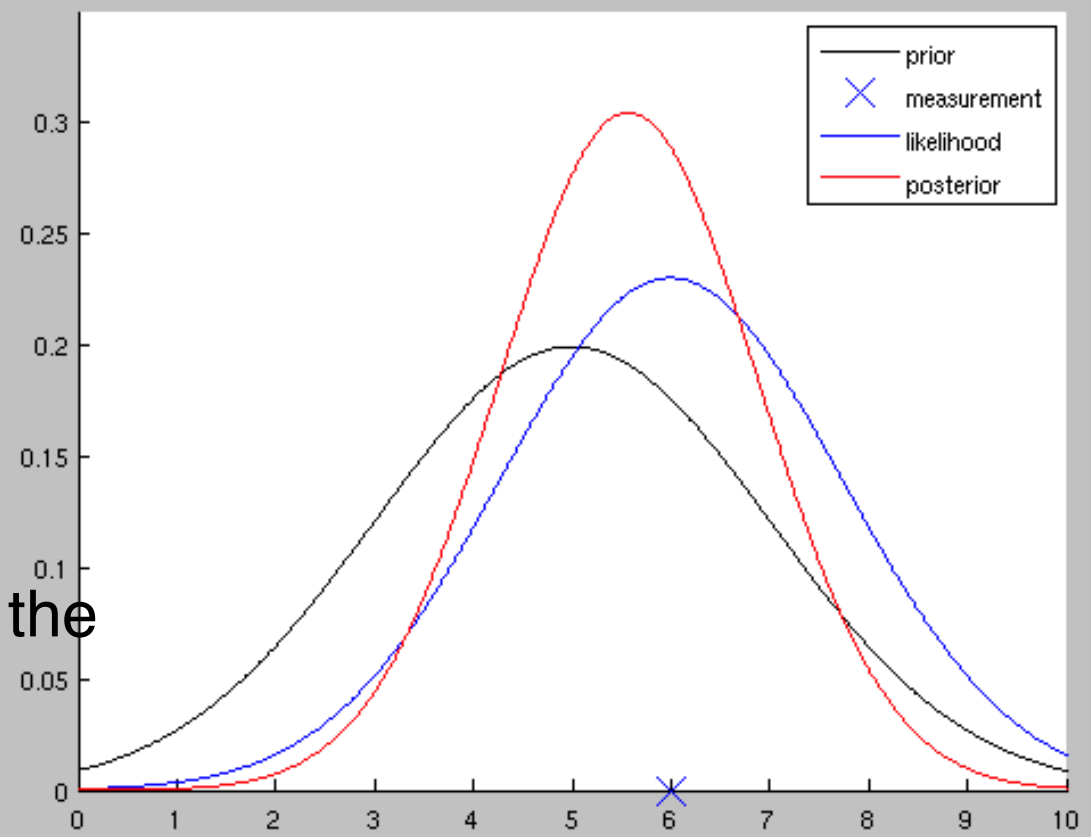
- Likelihood

$$L_{X|Y=y}(x) = f_{Y|X=x}(y)$$

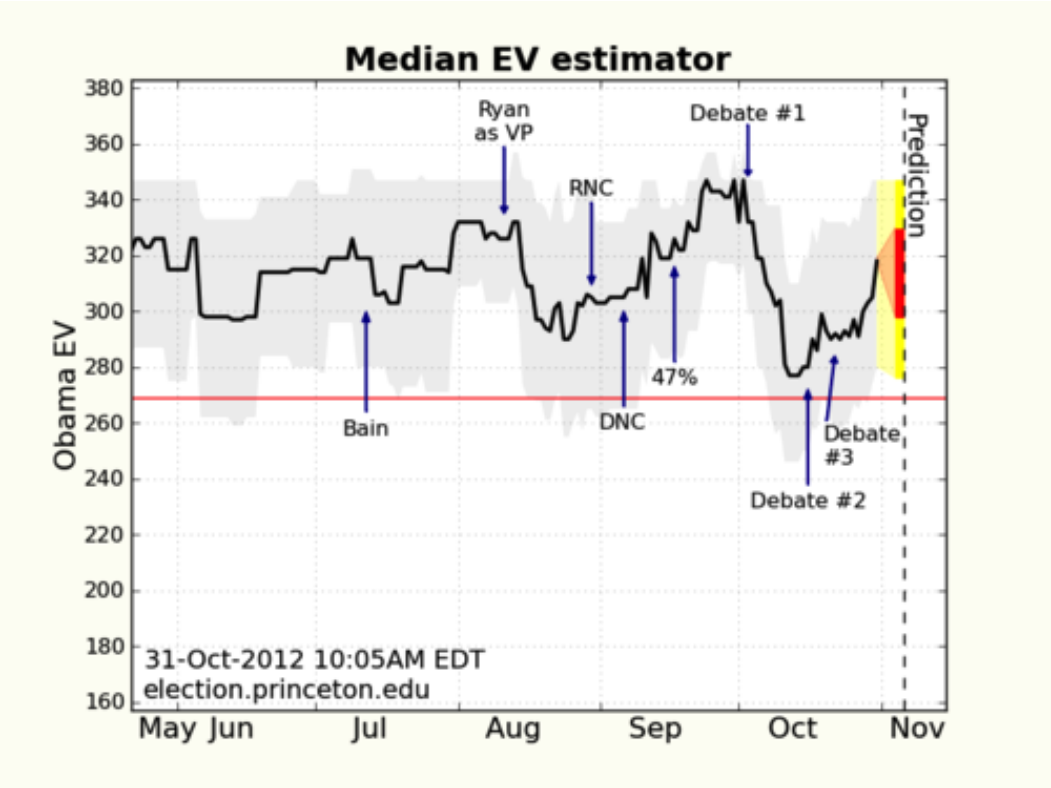
Form posterior distribution

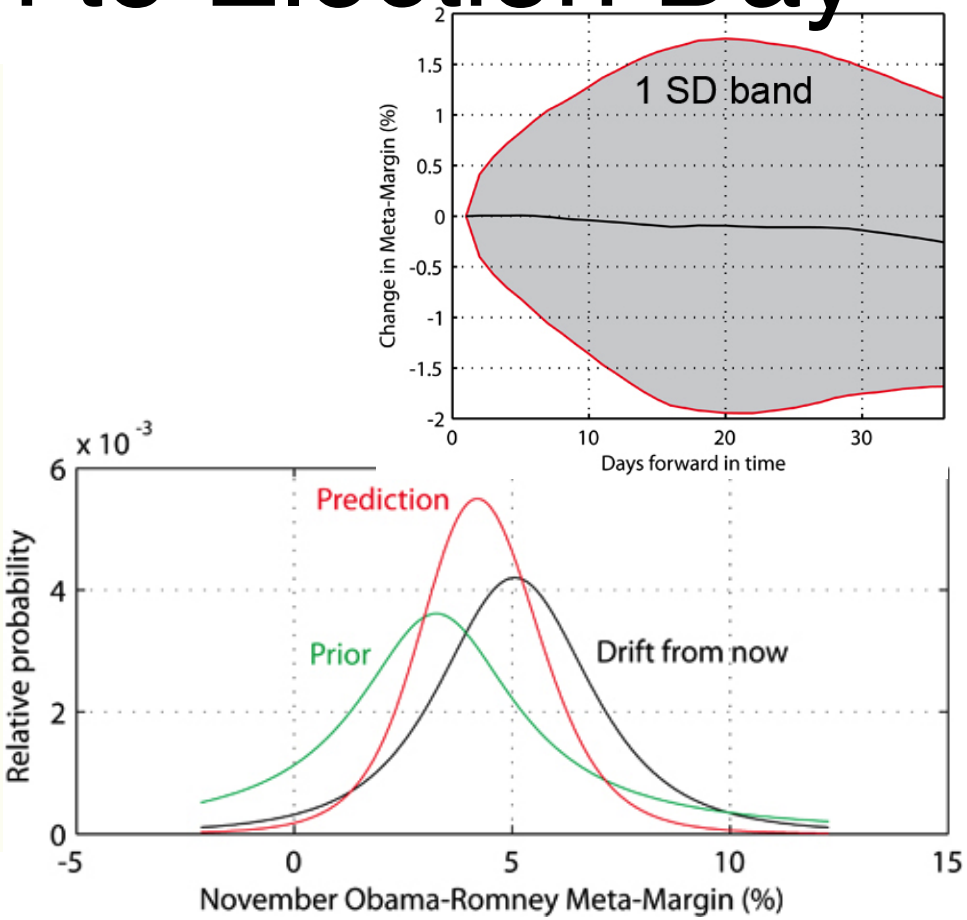
$$f_{X|Y=y}(x)$$

- Updating the prior with the knowledge you gained from *Y*



PEC – Projection to Election Day





Red (68% confidence interval): multiply prediction (based on today's snapshot and previous election variation) by a distribution for the drift observed throughout the election season

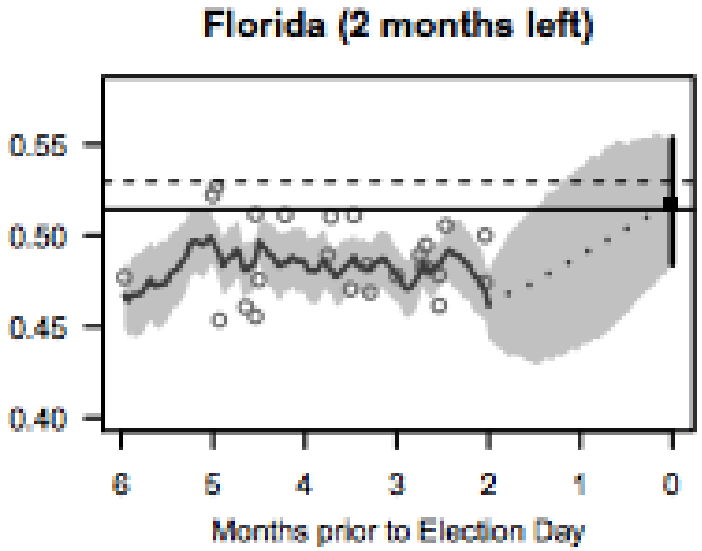
Yellow (95% confidence interval): enlarge today's 95% confidence interval by maximum drift. Contains contributions from drift and possible pollster error

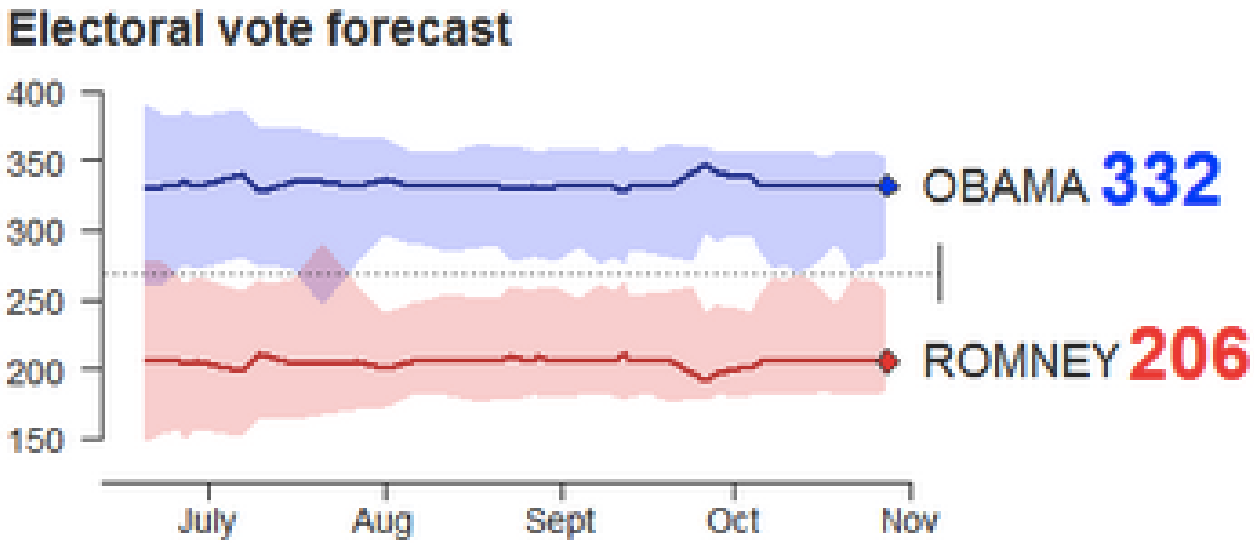
Past Performance

2008 Election:

- Obama 365 EV, McCain 173 EV
  - FiveThirtyEight predicted 384.5 EV for Obama
  - PEC predicted 352 EV for Obama
- States
  - FiveThirtyEight and PEC both predicted every state but Indiana correctly
- Senate
  - FiveThirtyEight predicted all 35 senate races correctly
  - PEC predicted 56 D, 2 I, 42 R (Franken won by ~250 votes to make it 57 D, 2 I, 41 R)

Votamatic

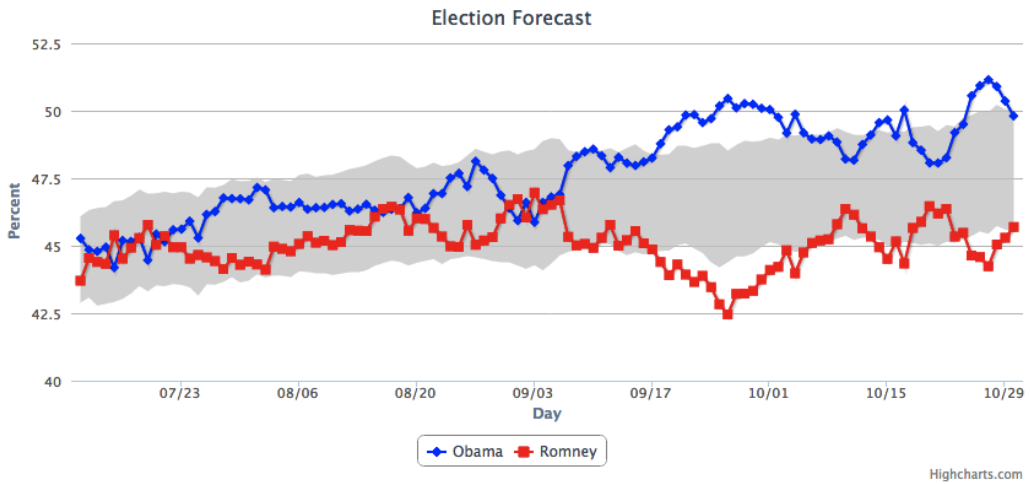


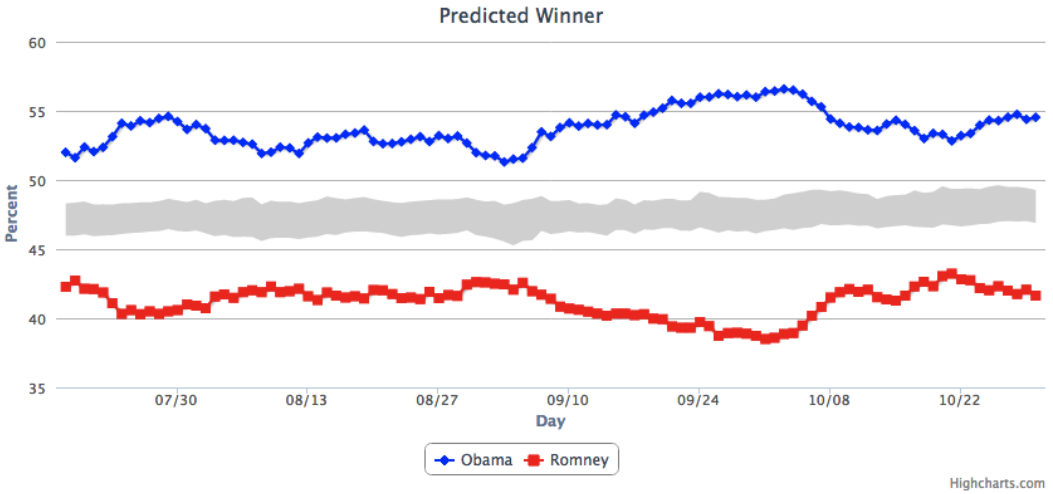


- Methodology:
  - Use an economic model to form a prior probability for the incumbent to win nationally on election day
  - The proportion of voters in a state favoring the incumbent consists of a unique state component, and a national component that follows national trends
  - Work backwards in time from prior and forward from state polls to get a posterior distribution for election day for each state
  - Use this posterior to do Monte Carlo simulation

[http://votamatic.org]

RAND Tracking Poll





(inside gray band, statistically indistinguishable)

- Methodology:
  - 3500 people are asked weekly:
    - What is the percent chance you will vote in the election?
    - What is the percent chance you will vote for ___?
    - What is the percent chance that ___ will win?
  - The same people are asked every week
  - The results are weighted for demographics
  - The first questions can be combined to give distributions
- Good for tracking changes over time, but there could be some bias in the initial sample or in the demographic weighting
- Does asking people multiple times create its own bias?

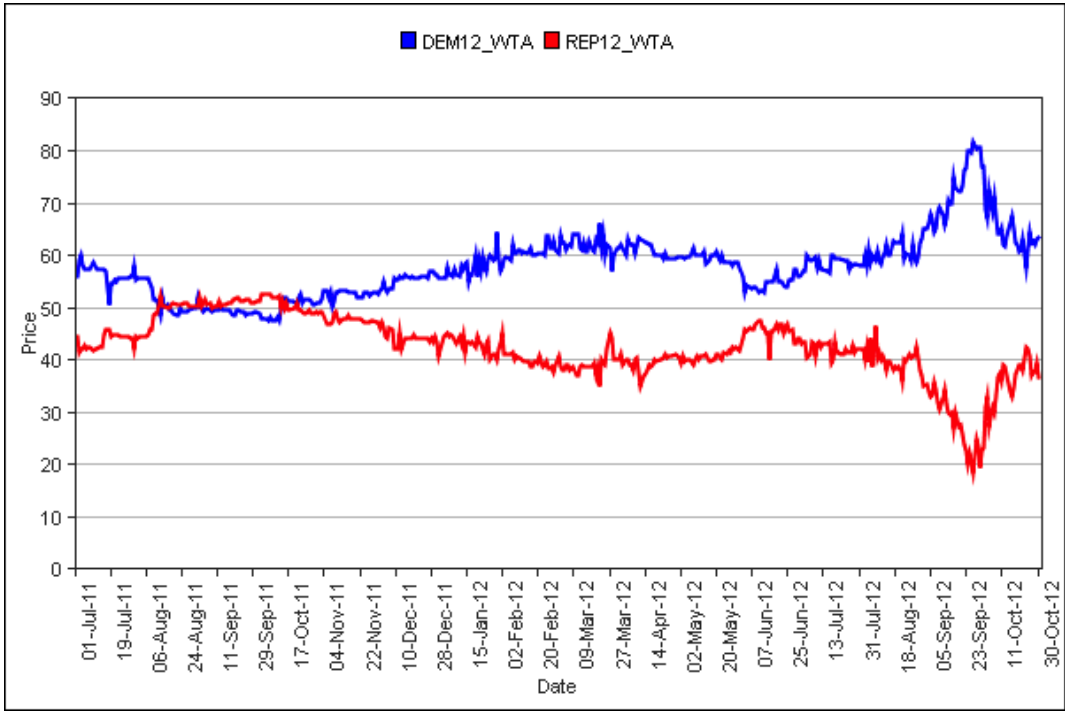
[https://mmicdata.rand.org/alp/?page=election]

Markets

You buy a contract on election results:

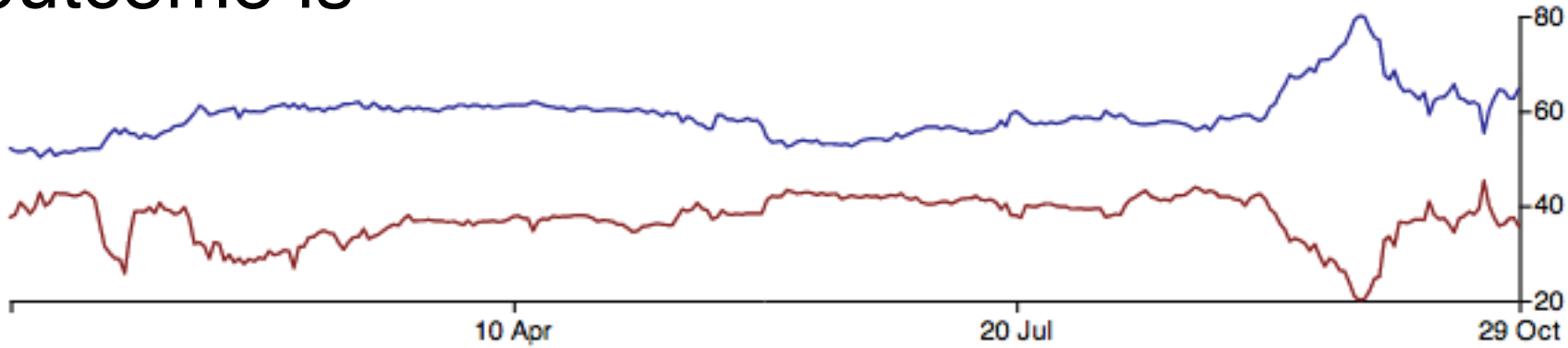
- Candidate wins, contract is worth $$
- Candidate loses, contract is worth 0

So market price should give some indication of how likely people think an outcome is

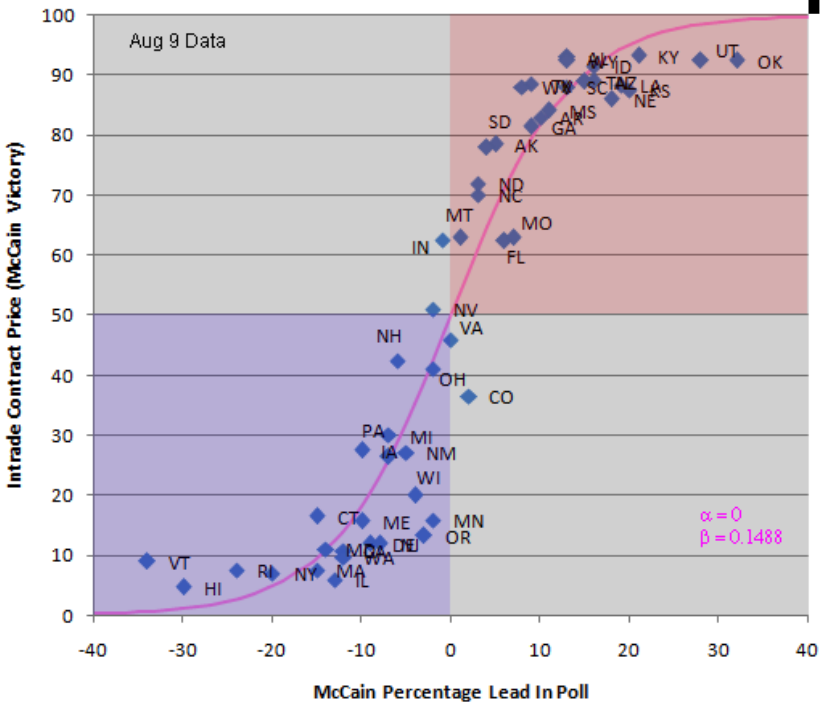


Iowa Electronic Market [http://iemweb.biz.uiowa.edu/graphs/graph_Pres12_WTA.cfm]

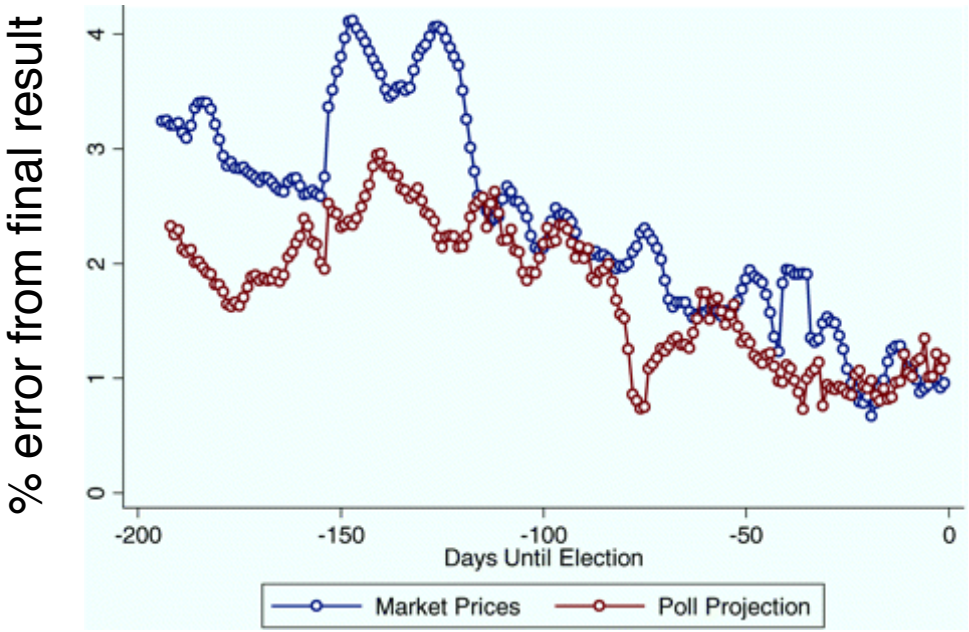
InTrade [http://www.intrade.com/v4/misc/scoreboard/]



Markets



[http://election.princeton.edu/2008/11/03/inefficiencies-in-intrade/]



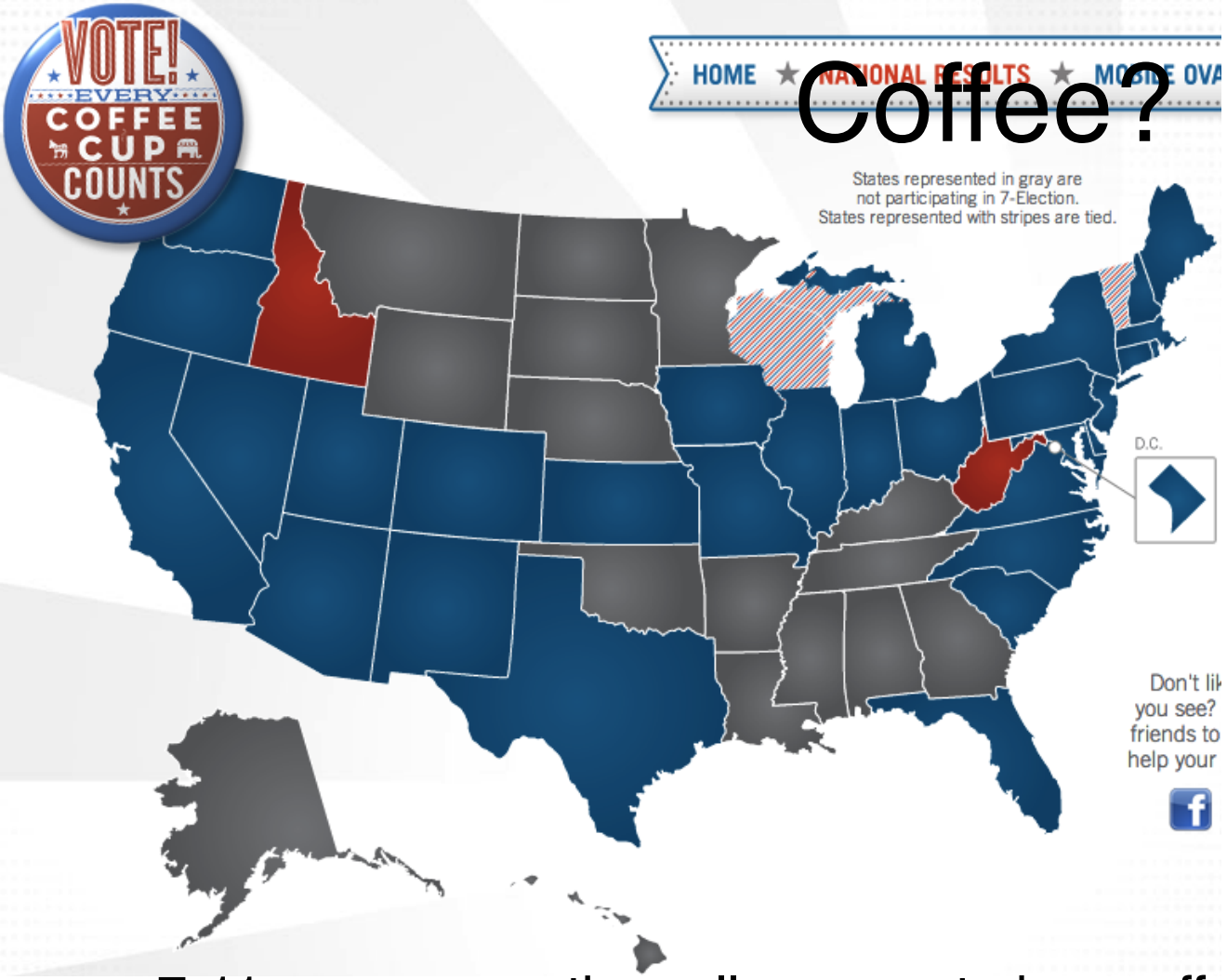
[R. Erikson & C. Wlezien, Public Opin Q 2008 72(2)]

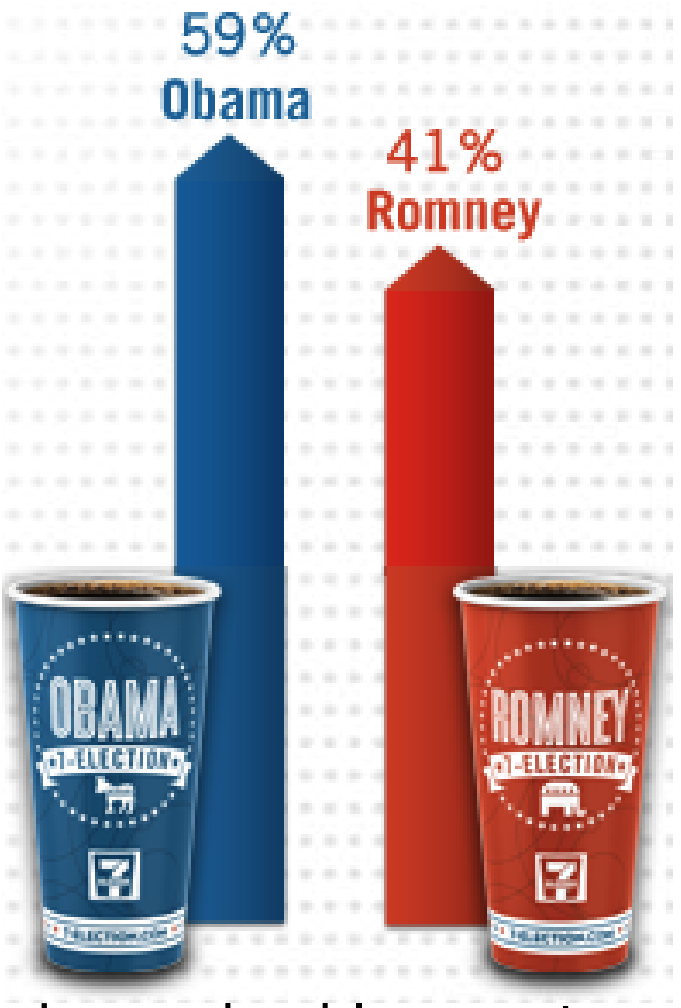
- Markets are underconfident even for landslide polling
- The claim is that gamblers take into account factors that polls don't like GOTV efforts and voter fraud
- But markets don't seem any better than polls:
  - Not much liquidity in the markets (can be manipulated)
  - Are the markets predicting, or following the polls?

Masks?


THE OFFICIAL PRESIDENTIAL MASK POLL 2012 SM
100% ACCURATE SINCE 2000!
SEAL OF THE BUYCOSTUMES.COM 2012 PRESIDENTIAL MASK POLL
Like
Send
396 people like this. Sign Up to see what your friends like.
1 mask = 1 vote. This poll can be bought! Vote for your favorites by buying one, ten, even 100 masks today!
51%
DEMOCRATS
49%
REPUBLICANS
Barack Obama Mask
$19.99
Reg $24.99
Read 2 Reviews
FROM: SPIRITHALLOWEEN.COM
New For 2012
Mitt Romney Mask
$19.99
Reg $21.99

Coffee?





- 7-11, as a promotion, allows you to buy coffee in a red or blue cup to cast a vote for a Presidential candidate
- Past History:
  - 2000 – Bush 21%, Gore 20%
  - 2004 – Bush 51%, Kerry 49%
  - 2008 – Obama 52%, McCain 46%

[http://www.7-eleven.com/7-Election/NationalResults.aspx]