

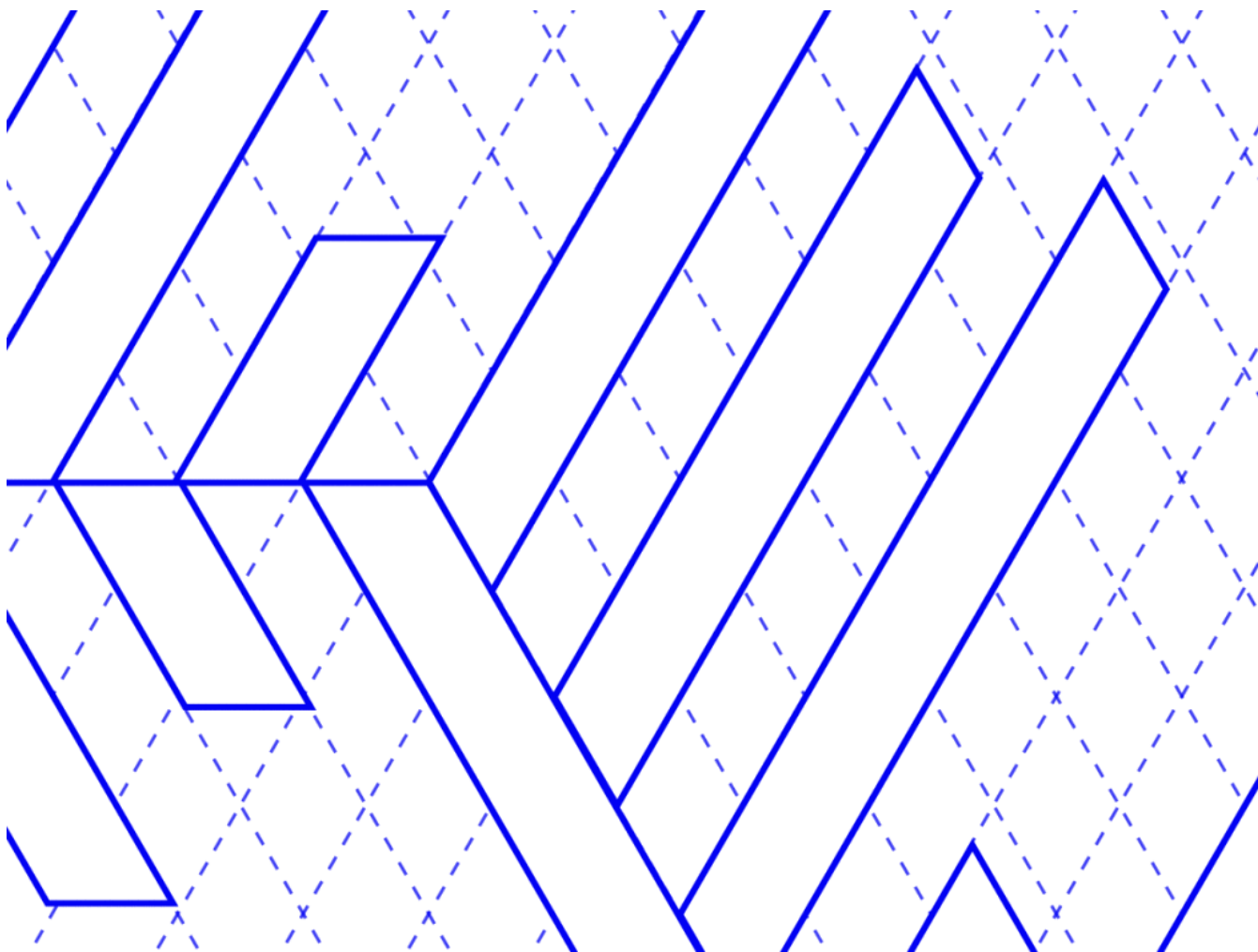
JAN. 12, 2016 AT 2:40 PM

How We're Forecasting The Primaries*

*And why we might be totally wrong.

By Nate Silver

Filed under 2016 Election



Forecasting primaries and caucuses is challenging, much more so than general elections. Polls shift rapidly and often prove to be fairly inaccurate, [even on the eve of the election](#). Non-polling factors, particularly [endorsements](#), can provide some additional guidance, but none of them is a magic bullet. And races with several viable candidates, like the one the Republicans are contesting this year, are especially hard to predict.

Nonetheless, we've developed a pair of relatively simple statistical models that we hope can shed some light on the upcoming [primaries](#). These models don't claim to have a lot of precision, sometimes showing very wide potential ranges of results for each candidate. (For a variety of reasons, the ranges are especially wide for the Iowa Republican caucuses. Ted Cruz's most likely range spans all the way from about 14 percent to 41

percent of the vote, for example. There's more on the forecast in Iowa and New Hampshire [here](#).)

But we do think our models do a reasonable job of laying odds and accounting for the uncertainty involved in the races. How safe is a 15-percentage-point lead in the polls with three weeks to go? How likely is a candidate in fourth place to jump to first? How should you weigh endorsements against polls? Our forecasts can answer questions like these.

ADVERTISEMENT



As I mentioned, we're running two separate (although related) forecasts this year that we call **polls-only** and **polls-plus**:

- The **polls-only** model is based only on polls from a particular state;¹ for example, only polls of New Hampshire are used in the New Hampshire forecast.
- The **polls-plus** model is based on state polls, national polls and endorsements. (National polls are used in a [slightly unusual way](#); they're a contrarian indicator. More about that later.) The polls-plus model also seeks to account for how the projected results in Iowa could affect the results in New Hampshire and how the results in those states could affect the results in subsequent contests.

In theory, the polls-plus model should be more accurate than the polls-only model, but it's a pretty small difference; in our backtesting, polls-plus was more accurate at predicting a candidate's actual result 57 percent of the time, while polls-only was more accurate 43 percent of the time. That's something, but there are plenty of times when the polls-only model will give the more accurate answer. Therefore, we think the models are more useful when looked at together.

You'll also notice that in some states — Nevada is one example — we list a weighted polling average for each candidate, but not a polls-only or polls-plus forecast. This is a compromise of sorts. The polls-only and polls-plus models are [trained](#) on past elections when there was quite a bit of polling data in the final 60 days of the campaign. So if polling data is sparse in a state this year, or if voting is still a long way away there, we won't run a forecast. But we may list a polling average, which you can think of as a FiveThirtyEight version of the polling averages published at RealClearPolitics and Huffington Post Pollster.

Maybe you're the type of reader who's interested in the fine print? What follows is a more detailed, step-by-step guide to how the models make their forecasts.

Step 1: Calculate polling averages

We start by calculating a weighted polling average for each candidate in each state.² The weights reflect the quality of each survey as determined by [FiveThirtyEight's pollster ratings](#), which grades polls based on their past accuracy and methodological standards. The poll weights also adjust for a poll's sample size and how recently it was conducted. All polls are included in the weighted average unless they were internal polls released by a candidate or a candidate's super PAC or if we have good reason to suspect that the poll faked its data or committed other gross ethical violations.

This process of weighting polls is highly similar to the one FiveThirtyEight uses for its [general election forecasts](#). An important difference, however, is that public opinion shifts much more quickly in the primaries, so recency is at more of a premium when calculating a polling average. Thus, a poll of middling quality that's hot off the presses will sometimes receive more weight than a top-quality one that's a week old. (We wish it weren't that way, but our research is pretty emphatic on the value of preferring newer polling data.)³

HOUSE EFFECTS

The models do have a defense mechanism against potential outlier polls, however. Namely, polls are adjusted for [house effects](#), which is a tendency for a pollster to consistently show different results for a candidate than the average of other polls.⁴ If a certain pollster consistently has Hillary Clinton polling 3 percentage points higher than other polls conducted in the same states at about the same times, for example, the model will subtract a fraction of that 3-point house effect from Clinton's numbers whenever that pollster issues a new poll for her.⁵ The house effects adjustment is designed such that higher-quality polls (as rated by our pollster ratings) have more say in calibrating

the polling averages. Thus, it partially corrects for low-quality polls that might “flood the zone” with frequent releases of questionable data.

Step 2: Polls-only forecast

What's the difference between the weighted polling average in a state, as described in Step 1, and the polls-only forecast? In fact, the differences are pretty minor.

First, undecided voters are allocated to the candidates in the polls-only forecast. The allocation is a combination of proportional⁶ and equal.

Second, the polls-only forecast incorporates an estimate of the probability that a candidate will drop out before the primary. (This is described in Step 3.)

Third, the polls-only forecast accounts for uncertainty by calculating a range of possible outcomes and uses this range to estimate the candidate's chance of winning the primary. (This is described in Step 4.)

Step 3: Polls-plus forecast

In designing the polls-plus forecast, we considered an array of possible predictors, including: endorsements, state and national fundraising totals, favorability ratings, ideology ratings and national polls. Just about all of these have some positive correlation with primary and caucus outcomes: Candidates with higher favorability ratings are more likely to see their ballot-test numbers go up than down, for example. And candidates who are good ideological “fits” for their states overperform their polls more often than not.

In the end, however, we opted for a relatively simple three-variable model, rather than a “kitchen sink” approach. The variables are state polls, endorsements and national polls. The model also considers how the projected results in Iowa might affect New Hampshire and how the results in those states might affect subsequent states. I've already described the process by which state polls are used, so I'll focus on the other factors now.

ENDORSEMENTS

As described in the book “[The Party Decides](#),” party elites play an important role in the nomination process, and their endorsements are historically a leading indicator of success in the primaries. The media often puts a lot of emphasis on these endorsements during the early, “[invisible primary](#)” phase of the campaign but forgets about them once voting is underway. Our research suggests, however, that endorsements have historically remained a leading indicator of popular support throughout the nomination process.⁷

Even when party elites have failed to come to a consensus on a candidate before Iowa

and New Hampshire — as seems to be the case on the Republican side this year — they've rallied behind a choice after the first few states voted, perhaps greasing the skids for his or her eventual victory. For instance, party leaders rallied behind John McCain in 2008, John Kerry in 2004 and Michael Dukakis in 1988 after their success in the early states, and nomination processes that once looked divisive turned out to be fairly smooth.

We measure endorsements by [endorsement points](#), which assign a candidate 10 points for each endorsement by a current governor from his or her party, 5 for each endorsement by a U.S. senator, and one for each endorsement by a member of the U.S. House of Representatives. More recent endorsements are weighted more heavily. So, for example, while Jeb Bush narrowly leads Marco Rubio in overall endorsement points among Republicans, Rubio gets more credit in the model because he's received many more endorsements than Bush recently.

NATIONAL POLLS

As my colleague Harry Enten [discovered](#), national polls have some predictive power in helping forecast the outcome in individual states — but not in the way you might expect. Instead, they have *negative* predictive power. For example, if you take two candidates polling at 15 percent in New Hampshire, the one polling at 10 percent in national polls is more likely to finish higher in the Granite State than the one polling at 20 percent in national polls.

How to explain this? There are a few plausible explanations, but the most intuitive one is as follows. The gap between state and national polls is a good proxy for how well-suited the candidate is to a particular state. Maybe a candidate with strong evangelical roots is polling well in Iowa, for example, or a candidate who has spent months building a great ground game in New Hampshire is doing well there. Once these advantages begin to show up in the state polls, they tend to expand over time. Momentum of various kinds — a candidate who is beating expectations in a state will usually get favorable press coverage for it and may double-down on the resources he's investing there — may contribute to the process.

Does that mean a candidate will be hurt in our model if his standing rises in national polls? Not exactly. When new national polling data comes out, our model waits until there's fresh data from the state to figure out what to make of it. If a candidate gains in both state and national polls, he'll be helped in the model. But if a candidate gains in the national polls and his state polls don't improve at all, that can be a bearish indicator.⁸

Also, note that as the election approaches in each state, the polls-only and polls-plus forecasts will tend to converge. The model is set up such that the weight given to endorsements is set to zero by Election Day in each state, while the weight given to national polls is reduced.

ACCOUNTING FOR IOWA AND NEW HAMPSHIRE

The polls-plus model also includes an adjustment for the projected results in Iowa and New Hampshire. The process takes place in two phases:

- Until Iowa votes, 20 percent of a candidate's New Hampshire forecast is made up of his Iowa forecast. And until New Hampshire votes, 20 percent of his forecast in states subsequent to New Hampshire is made up of his New Hampshire forecast.
- Once Iowa and New Hampshire vote, the projected results are replaced by the actual results of the voting in those states. This effect fades to zero once new polling from the subsequent states comes in and we know the actual effect on the polls instead of having to guess at it.⁹

Overall, this is a relatively conservative way to account for the results in Iowa and New Hampshire, which often prove to be extremely influential on the rest of the race. However, their influence can be hard to predict. Often, [whether a candidate beats his polls](#) is as important as his absolute finish in determining the media-driven momentum he gets out of these states.

WILL A CANDIDATE DROP OUT?

Both the polls-only and polls-plus forecasts attempt to project the likelihood that the candidate will drop out before the state votes. What factors predict when candidates drop out? The most important considerations in the model, based on an analysis of when candidates have dropped out in the past, are as follows:

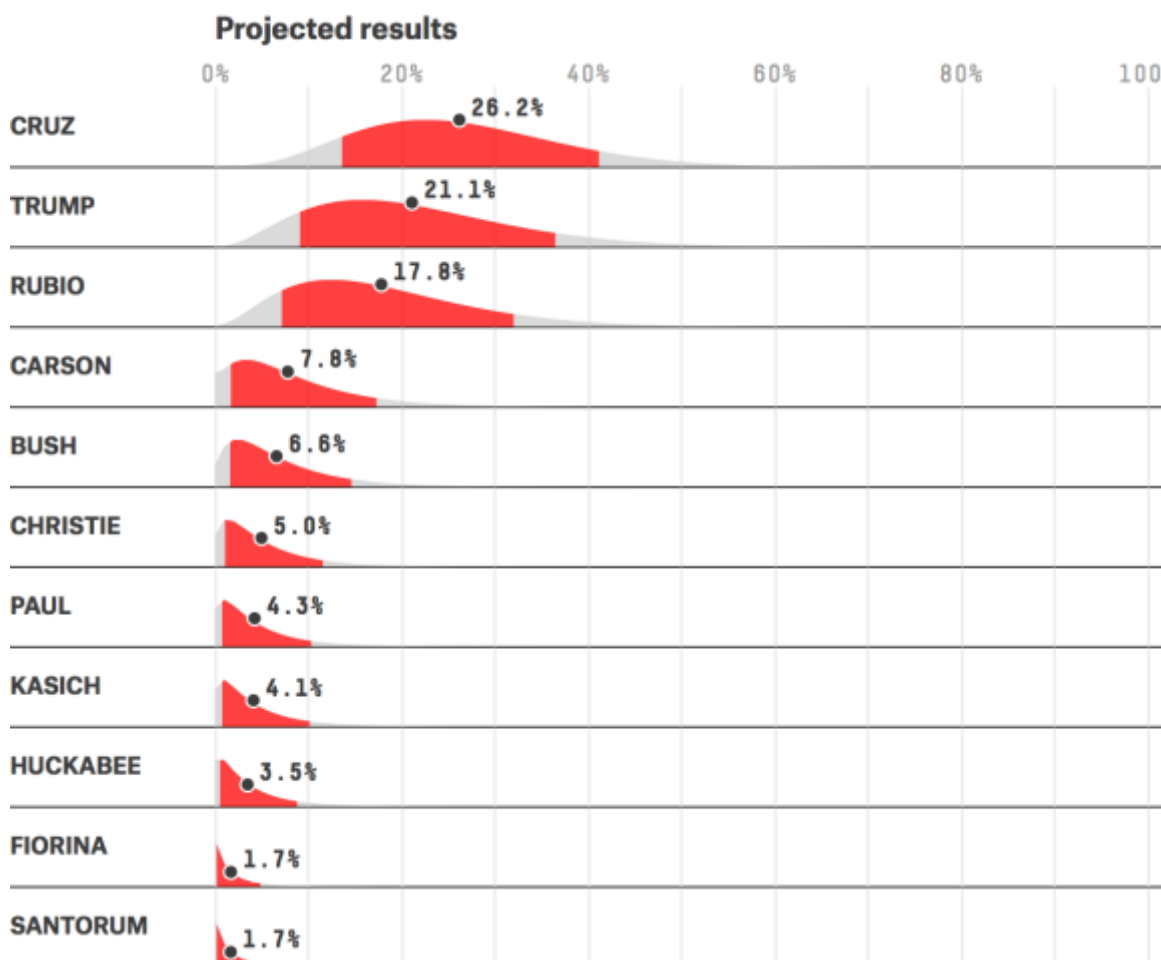
- Candidates rarely drop out immediately before a major primary or caucus. If they've made it to within a couple of days of voting, they'll usually play out the string, even if their position looks hopeless.
- The field usually winnows significantly after Iowa and New Hampshire.
- A candidate's decision to drop out is influenced both by his absolute standing and his trajectory in the polls in upcoming primaries and caucuses. Candidates with negative momentum in the polls are at risk of dropping out, while those with favorable momentum rarely drop out.

- National polls and endorsements have only minor effects on a candidate's likelihood of dropping out. Instead, a candidate's decision is influenced mostly by his position in the next couple of states scheduled to vote.¹⁰

Why go through all this trouble to calculate a candidate's dropout probability? It's important from a technical standpoint because it can make a candidate's range of possible outcomes **irregularly shaped**. A candidate polling at 20 percent in some state might ordinarily have a **confidence interval** that runs between 14 percent and 27 percent of the vote, for instance. But depending on other factors, there may be some chance he'll drop out, in which case he could get close to zero percent of the vote instead.

Step 4: Determine probability distributions and estimate chance of winning

In our interactive, you'll see a bunch of funky-looking curves like the ones below for each candidate; they represent the model's estimate of the possible distribution of his vote share. The red part of the curve represents a candidate's 80 percent **confidence interval**. If the model is calibrated correctly, then he should finish within this range 80 percent of the time, above it 10 percent of the time, and below it 10 percent of the time.



But how are these curves calculated? First, the model recognizes that the uncertainty is higher under certain circumstances. In particular, it considers the following:

- The uncertainty is higher the further removed you are from Election Day.
- The uncertainty is higher when a candidate has a larger projected share of the vote. (There's not much uncertainty about where a candidate polling at 1 percent is likely to finish, in other words.)
- The uncertainty is higher in caucuses than in primaries.
- The uncertainty is higher when you have less polling data. (The benefit of additional polling data is greatly diminishing, though, perhaps because of [herding](#).)
- The uncertainty is higher earlier in the primary calendar than later on (perhaps because pollsters learn how to correct their mistakes from earlier in the process).
- The uncertainty is higher when there are more candidates in the race.
- The uncertainty is higher when there's a wider gap between the polls-only and polls-plus forecasts. In other words, candidates who have some favorable indicators and some unfavorable ones face more uncertainty in their forecast than those for whom everything is seemingly in alignment. (This means that an unusual candidate like Donald Trump tends to have especially uncertain forecasts, for example.)

Note that almost all these factors align to create a highly uncertain outcome in the first couple of Republican contests; there is an unprecedented number of candidates remaining, and endorsements, state polls and national polls are [not in terribly strong alignment](#).

You'll also notice that the probability distributions are asymmetrical; typically, they have a longer right tail than left tail. (If you're wondering, they're calculated using a [probit transformation](#).) In fact, this is a mathematical necessity for candidates with a small projected share of the vote. A candidate currently polling at 5 percent has some chance of gaining 10 points and finishing at 15 percent instead but no chance of losing 10 points and finishing with -5 percent. Accounting for a candidate's probability of dropping out can add to the asymmetry.

GOING FROM PROBABILITY DISTRIBUTIONS TO WIN PROBABILITIES

Given how we've calculated these ranges, we could theoretically just draw a random outcome from each candidate's distribution and award the primary to whichever candidate finishes with the highest number. Except, it's not *quite* that simple because

every percent of the vote a candidate gains must come from some other candidate; if Bernie Sanders finishes toward the high end of his range, that means Clinton has probably finished toward the lower end of hers, for instance. The models use a fairly simple technique to estimate these effects and then run 10,000 simulations to estimate each candidate's chances of winning.

There's also an awful lot our models don't consider. For instance, they don't do much to consider the interrelationships in the vote between different types of candidates. For example, if John Kasich gains a vote in New Hampshire, it's probably more likely to come from a similar candidate like Bush than a dissimilar candidate like Ben Carson. Still, we hope they can form a reasonable benchmark for following the upcoming primaries, even if they're almost certain to get a few races wrong.

Footnotes

1. Well, that's *almost* true. Technically, the house effects adjustment used in the polls-only model incorporates polls from all states, along with national polls.
2. We also apply this process to national polls to calculate a national polling average for each candidate.
3. Furthermore, in the model, the premium on recency increases the closer you get to the date of the primary. With 30 days to go, you shouldn't necessarily sweat every tick in the polls — but with three days left, it becomes extremely important.
4. We've previously included house effects adjustments in our general election forecasts, but they're a new feature in the primary model, having not been included in our version [four years ago](#).
5. The exact magnitude of the adjustment depends on how many polls that pollster has conducted and how consistent the effect has been.
6. A proportional allocation would mean that candidates get a share of the undecided vote that is equal to their share of the committed vote. Thus, a candidate polling at 20 percent would get twice as many undecided voters as one polling at 10 percent.
7. If anything, in fact, their importance seems to increase the deeper the party is into its nomination process, as party leaders seek to achieve party unity and avoid a contested convention.
8. Note that you shouldn't attempt to apply this "trick" in states where there isn't a lot of recent polling data. Other things held equal, if a candidate has gained or lost a bunch of ground in national polls, you should expect the same to be true in upcoming states. It's only when you have enough state polling to prove that the gap is widening that this heuristic proves useful.
9. If, for example, the model is expecting a 5-percentage-point bounce in New Hampshire for Ted Cruz because of his performance in Iowa but a number of New Hampshire polls show that he's

gotten no bounce at all, the model will quickly cease to give him credit for his Iowa performance.

10. The polls-only model doesn't use national polls or endorsements in predicting a candidate's likelihood of dropping out. Instead, it looks at a candidate's polling only in the state the forecast pertains to.