The New York Times
FiveThirtyEight
Nate Silver's Political Calculus

DECEMBER 16, 2011, 12:26 PM

How Our Primary Forecasts Work

By NATE SILVER

//

Today we're officially launching the FiveThirtyEight polling-based forecasts of the upcoming presidential primaries and caucuses. This article will explain, in modestly but not excruciatingly technical detail, how they work.

First, however, let me explain what the basic objectives of the model are.

Suppose it's a week before a presidential primary. There are three major candidates in the race, who are polling at 35 percent, 30 percent and 25 percent, respectively. Given this information, what are the chances of each of the three candidates winning?

This is not a particularly deep question. But it's one that is of some interest to those of us who enjoy following the polls.

One hugely mistaken assumption would be to look at the margin of error associated with the poll. FiveThirtyEight has a database consisting of thousands of primary and caucus polls dating back to the 1970s. Each poll contains numbers for several candidates, so there are a total of about 17,000 observations. How often does a candidate's actual vote total fall within the theoretical margin of error?

The answer is, not very often. In theory, a candidate's actual vote total should fall outside the margin of error only 5 percent of the time. In reality, the candidate's vote total was outside the margin of error 65 percent of the time! Part of this is because the database includes some polls conducted months before the actual voting took place. But even if you restrict the analysis to polls conducted within the final week of the campaign, about 40 percent of the vote totals fell outside the margin of error — eight times more often than is supposed to happen if you could take the margin of error at face value.

Clearly, then, there is some value in developing some *realistic* and *empirically valid* notion of what the polls really tell us. That's all that we're really trying to accomplish. It's more an exercise in calibration than an exercise in prediction: we're trying to gauge how accurate our instruments (the polls) are in the real world. We're not interested (in this context) in coming to any grand theoretical proclamations about what makes candidates win in primaries, or in making a lot of assumptions about how the polls "should" behave. We're just trying to

determine how accurate the polls really are, and perhaps make a few reasonable adjustments around the margins that can reduce the uncertainty a little bit. It's as much about coming to an understanding about what we *don't* know as what we do know.

This turns out, nevertheless, to be a somewhat challenging question to answer. Therefore, we've chosen to narrow the focus in a couple of ways.

First, the analysis is concerned with what happens in the late stages of a primary campaign — specifically, in the 30 days or so before voters actually go to the polls. For this reason, we'll be releasing the forecasts a few states at a time, based on which states are next in line to vote. Longer than about a month before a primary or caucus, the polls can be so inaccurate — and so affected by what has happened in previous states — that there just isn't a lot of value in looking at them. The polls might be worth a passing glance, but you'll probably be better off thinking about the broader context of the nomination race. There isn't any kind of statistical magic that can make them much better.

Second, we're concerned with states in which there is a reasonably robust amount of polling, meaning that at least three polling organizations are active in the state. We will probably not be issuing forecasts for states in which there is just one stray poll here and there.

Third and finally, we are interested in what we can learn from the polls of the state, and the polls alone. There probably are a lot of factors you should consider in addition to the polls. In 2008, for instance, we had a lot of success by making forecasts based on the demographic factors in each state, without looking at the polls at all. (These forecasts were actually a bit better than those based on polling.) We will continue to consider these factors, especially if the Republican race stretches out for a long time. That just isn't our objective here, however.

The remainder of this post will describe how the forecasts are derived on a step-by-step basis and will get a bit technical; a more casual explanation can be found here.

Step 1. Determine the polling average. Since our interest is in states in which there are several different polls, the question naturally arises as to how to combine them. As is the case with our other forecasting products, polls are weighted based on their sample size and a pollster rating that accounts for the past accuracy of each polling firm as well as its methodological standards.

The polls are also weighted based on how recently the poll was conducted, a matter that deserves further discussion. Our analysis of thousands of past polls suggests that in primaries and caucuses, you'll be better served by paying a lot of attention to how recently the poll was conducted and by being aggressive about updating your assumptions in the face of new information. The chart below describes the average error in a primary poll based on how much lag time there was between the survey and the actual voting in a state. Polls conducted a day or two before the primary have historically been *much* more accurate than polls conducted a week or so beforehand, a reflection of the fact that voters often settle on their choice only at the last minute.

Therefore, the weights assigned to the polls, which are determined from a half-life formula, place a very heavy emphasis on recentness. A week-old poll will get only a fraction of the weight of a fresher poll, other factors being equal. The closer you get to the date of the primary or caucus, moreover, the more the premium on recentness increases, and the more the model will discount older polls.

This will lead to some very aggressive forecasts, with the projections often changing considerably based on the outcome of a single survey. Sometimes this means the model will chase down a specious trend, but it will home in on the right answer slightly more often than not over the long run.

The model also makes a special provision for Iowa and New Hampshire, which can have an especially large influence on the race. A New Hampshire poll conducted one day after the Iowa caucuses will receive considerably more weight than one conducted just one day before Iowa. In practice, this means that the first New Hampshire polls conducted after Iowa will all but eliminate those conducted just before it.

Step 2. Allocate undecided voters. Forecasts can also be improved by assigning undecided voters to the candidates. The specific method used represents a compromise between dividing these votes evenly among the candidates and dividing them proportionately, based on what would have produced the best results on the historical data.

Step 3. Determine "momentum." In the spirit of making rather aggressive forecasts, the model also considers the trajectory of each candidate's polling. It determines this by comparing a candidate's polling average to the average that would be obtained if less of a premium were placed on recentness (specifically, if the half-life associated with each poll were doubled). It then assigns a bonus or penalty to each candidate based on the magnitude of the difference, as well as how robust the polling data is; a trend observed across several polls is much more likely to be meaningful than one based on just one or two of them.

Step 4. Project vote for each candidate. The model then projects the vote for each candidate by combining the polling average (Step 2) with the momentum factor (Step 3).

Step 5. Estimate uncertainty in forecast. This is only half the battle, however; it is just as important to estimate how much error there is in the forecast. It is predictable when the forecasts are relatively more reliable . Specifically, the forecast error increases based on five fairly intuitive factors:

1. The forecast error is larger the further away you are from the actual voting date of the primary or caucus.

2. The forecast error is larger the fewer the number of reliable polls.

3. The forecast error is larger the more undecided voters there are. Also, when a candidate drops out of the race, their voters are considered undecided and forecast error increases until they settle upon a new candidate.

4. The forecast error is higher the earlier you are in the nomination process. Polls in Iowa and New Hampshire fairly often go astray. By late in the race, however, voter preferences will have hardened, and pollsters will have a better sense for who will turn out and who won't.

5. The forecast error is higher the more inconsistency there is the candidate's polling (such as can be measured by the standard deviation of a candidate's polls). A candidate who polls at 30 percent in some surveys but 10 percent in others has more upside potential and may be more threatening to a front-runner than one who polls at 20 percent across the board.

Based upon these factors, the model estimates an error parameter for each candidate: how much it expects its prediction to miss by, on average. It then applies this error estimate in a couple of ways.

Step 6a. Determine win probability estimates. The model first estimates the chances that each candidate will win the caucus or primary based on his or her projected vote share (Step 4) and the projected vote share of the top two candidates. (Polling at 25 percent means one thing if it puts you in first place in a multicandidate race and quite another if you trail another candidate 75-25). Because these initial estimates (which rely upon logistic regression) do not necessarily add up to 100 percent, the model uses an iterative process to calibrate them more reliably.

However, the win probability estimates are discounted based on the expected uncertainty in each candidate's forecast (Step 5). Having a 10-point lead several months before the primary in a state where just one or two polling firms have been active is much less robust than having a 10-point lead across multiple recent surveys on election eve. The model accounts for these properties. But in some cases, the uncertainty is so great as to swamp everything else. This is why, for example we do not (yet) list win probability estimates for states like Florida and South Carolina as they will be greatly affected by the voting in Iowa and New Hampshire; you might as well just pick a name out of a hat.

Step 6b. Develop confidence intervals. In addition to the win probability estimates for each candidate, the model also provides an estimate of the range of possible outcomes. Specifically, it provides an estimate of the 5th and 95th percentile of possible outcomes for each candidate such as is determined by quantile regression. A candidate's actual vote total should theoretically fall within this range 90 percent of the time.

The range is often quite wide because the error in primary polls is often quite large: Hillary Rodham Clinton's upset of Barack Obama in the 2008 Democratic primary in New Hampshire is one example, but hardly the only one. The error range can also be somewhat asymmetric, especially for the candidates lower in the polls. For instance, Jon M. Huntsman Jr. is projected to get 6 percent of the vote in Iowa, but his confidence interval runs from 0 to 15 percent. Again, this is a reflection of how these polls have behaved historically. A

candidate polling at 6 percent will occasionally beat his numbers by 8 points and finish with 14 percent of the vote. But he can't underachieve them by 8 points and finish at negative 2.

The important thing to keep in mind is that the considerable degree of uncertainty in the model is simply a reflection of the considerable degree of uncertainty in primary polling. We are not claiming to have found any miracles. The idea, instead, is to give you an honest and thoughtful perspective on the way the polls have behaved in the real world.