The New York Times

SURVEY SAYS?

When You Hear the Margin of Error Is Plus or Minus 3 Percent, Think 7 Instead

By David Rothschild and Sharad Goel

Oct. 5, 2016

As anyone who follows election polling can tell you, when you survey 1,000 people, the margin of error is plus or minus three percentage points. This roughly means that 95 percent of the time, the survey estimate should be within three percentage points of the true answer.

If 54 percent of people support Hillary Clinton, the survey estimate might be as high as 57 percent or as low as 51 percent, but it is unlikely to be 49 percent. This truism of modern polling, heralded as one of the great success stories of statistics, is included in textbooks and taught in college classes, including our own.

But the real-world margin of error of election polls is not three percentage points. It is about twice as big.

In a new paper with Andrew Gelman and Houshmand Shirani-Mehr, we examined 4,221 late-campaign polls — every public poll we could find — for 608 state-level presidential, Senate and governor's races between 1998 and 2014. Comparing those polls' results with actual electoral results, we find the historical margin of error is plus or minus six to seven percentage points. (Yes, that's an error range of 12 to 14 points, not the typically reported 6 or 7.)

The implication? Even if you see a poll in early November that has Donald J. Trump up by three points or Mrs. Clinton up by five, you should still not be so sure who is going to win the election.

What explains this big gap between the stated and observed error rates? All the polls we analyzed were conducted during the final three weeks of the campaign, by which time any changes in aggregate public opinion were very small, so the disparity is not simply driven by changes in sentiment as an election progresses.

Let's start with the stated margin of error, which captures sampling variation: error that occurs because surveys are based on only a subset of the full population of likely voters. Even if this sample of respondents is selected randomly from the full population, it is not a perfect representation of attitudes in the full population. That's where the usual 3 percent error rate comes from.

But the stated margin of error misses other important forms of error. Frame error occurs when there is a mismatch between the people who are possibly included in the poll (the sampling frame) and the true target population.

For example, for phone-based surveys, people without phones would never be included in any sample. Of particular import for election surveys, the sampling frame includes many adults who are not likely to vote. Pollsters try to correct for this by using likely-voter screens — typically asking respondents if they will vote — but this screen itself can introduce error that can at times be larger than the bias it was intended to correct.

And then there is nonresponse error, when the likelihood of responding to a survey is systematically related to how one would have answered the survey. For example, as another one of our papers shows, supporters of the trailing candidate are less likely to respond to surveys, biasing the result in favor of the more popular politician.

A similar effect probably explains part of Mrs. Clinton's recent dip in the polls, as Democrats became less enthusiastic about answering surveys when she appeared to be struggling. With nonresponse rates exceeding 90 percent for election surveys, this is a growing concern.

Finally, there is error in the analysis phase. In one example, as Nate Cohn showed in an Upshot article, four pollsters arrived at different estimates even when starting from the same raw polling data.

Other errors, which we believe are less important in United States election surveys, include effects of survey wording and interviewer bias.

All these nonsampling errors show up in two ways. First, polls within a race vary from one another slightly more than one would expect from classical textbook explanations. Second, and most markedly, polls tend to systematically overestimate or underestimate the true answer. Thus, even when many polls are averaged together, the result is biased in favor of one candidate or the other.

The accompanying graph shows, for each race, how much the average of all polls in that race differs from the final election outcome. If sampling variation were the only source of error, the polling average would be quite close to the election outcome. Standard theory says the average of this many polls should be within about half a percentage point of the true answer, and that this difference shrinks to zero as more polls are conducted.

But, in reality, we find that the polling average can easily be two points off from the final vote share. For any given race, the polls are consistently too high for one of the candidates and too low for the other.

One explanation for this nonsampling error, which persists even after averaging many polls, is that most survey organizations make similar assumptions about who will vote, and so Election Day surprises (such as an unexpectedly high turnout of minority or young voters) affect all polls similarly. Likewise, if there is nonresponse bias (for instance, Trump or Clinton supporters being more or less likely to participate in a survey), this too will happen with all polls together.

We can measure this error after the fact, but during the campaign it's hard to identify which candidate has the inflated lead. We do not observe any systematic bias in favor of Democratic or Republican candidates. This makes sense. If we saw this, pollsters would correct for it.

Over all, neither party has consistently outperformed the polls, but within a given year the errors are correlated between states. Even if a candidate appears to be trailing slightly in all the swing states, he or she could conceivably end up winning them all.

This November, we would not be at all surprised to see Mrs. Clinton or Mr. Trump beat the state-by-state polling averages by about two percentage points. We just don't know which one would do it.

David Rothschild is an economist at Microsoft Research. Follow him on Twitter at @DavMicRot or at PredictWise.com. Sharad Goel is an assistant professor at Stanford University in the School of Engineering.

A version of this article appears in print on Oct. 5, 2016, on Page A18 of the New York edition with the headline: If a Poll's Margin of Error Is Plus or Minus 3 Points, Think 7