

# Disentangling Bias and Variance in Election Polls

Houshmand Shirani-Mehr  
Stanford University

David Rothschild  
Microsoft Research

Sharad Goel  
Stanford University

Andrew Gelman  
Columbia University

## Abstract

It is well known among both researchers and practitioners that election polls suffer from a variety of sampling and non-sampling errors, often collectively referred to as *total survey error*. However, reported margins of error typically only capture sampling variability, and in particular, generally ignore errors in defining the target population (e.g., errors due to uncertainty in who will vote). Here we empirically analyze 4,221 polls for 608 state-level presidential, senatorial, and gubernatorial elections between 1998 and 2014, all of which were conducted during the final three weeks of the campaigns. Comparing to the actual election outcomes, we find that average survey error as measured by root mean squared error (RMSE) is approximately 3.5%, corresponding to a 95% confidence interval of  $\pm 7\%$ —twice the width of most reported intervals. Using hierarchical Bayesian latent variable models, we decompose survey error into election-level bias and variance terms. We find average absolute election-level bias is about 1.5%, indicating that polls for a given election often share a common component of error, likely in part because surveys, even when conducted by different polling organizations, rely on similar screening rules. We further find that average election-level variance is higher than what most reported margins of error would suggest. We conclude by offering recommendations for incorporating these results into polling practice.

# 1 Introduction

Election polling is arguably the most visible manifestation of statistics in everyday life, and random sampling is one of the great success stories of statistics, as is recounted in so many textbooks that give the story of the huge but uncontrolled Literary Digest poll being trounced by Gallup’s small, nimble random sample back in 1936. Sampling is part of our lives in many ways, with election polls having the highest profile, and also (with in the Literary Digest poll) providing a reality check on methods.

It has long been known that the margins of errors provided by survey organizations, and reported in the news, understate the total survey error. This is a hugely important topic in sampling but is difficult to address in general, for two reasons: first, we like to decompose error into bias and variance, but this can only be done with any precision if there are a large number of surveys (not merely a large sample in an individual survey) to do this assessment. Second, assessment of error requires a ground truth for comparison, which is typically not available (as the reason for conducting a sample survey in the first place is to estimate some population characteristic that you don’t already know).

In the present paper we decompose survey error in a large set of state-level pre-election polls. This dataset resolves both of the problems just addressed: First, the combination of multiple elections and many states gives us a large sample of polls; it is fortunate for this study that U.S. elections are frequently polls. Second, we can compare the polls to actual election results.

## 1.1 Background

Election polls typically survey a random sample of eligible or likely voters, and then generate population-level estimates by taking a weighted average of responses, where the weights are designed to correct for known differences between sample and population.<sup>1</sup> This

---

<sup>1</sup>One common technique for setting survey weights is raking, in which weights are defined so that the weighted distributions of various demographic features (e.g., age, sex, and race) of respondents in the sample

general analysis framework yields not only a point estimate of the election outcome, but also an estimate of the error in that prediction due to sample variance which accounts for the survey weights [Lohr, 2009]. In practice, weights in a sample tend to be close to equal, and so most major polling organizations simply report 95% margins of error identical to those from simple random sampling (SRS) without incorporating the effect of the weights, for example  $\pm 3.5\%$  for an election survey with 800 people.<sup>2</sup>

Though this approach to quantifying polling error is popular and convenient, it is well known by both researchers and practitioners that discrepancies between poll results and election outcomes is only partially attributable to sample variance [Ansolabehere and Belin, 1993]. As observed in the extensive literature on *total survey error* [Biemer, 2010, Groves and Lyberg, 2010], there are at least four additional types of error that are not reflected in the usually reported margins of error: frame, nonresponse, measurement, specification. Frame error occurs when there is a mismatch between the sampling frame and the 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, which pollsters recognize and attempt to correct for using likely voters screens, typically estimated from survey questions and with error. Nonresponse error occurs when missing values are systematically related to the response. For example, supporters of the trailing candidate may be less likely to respond to surveys [Gelman, Goel, Rivers, and Rothschild, 2016]. With nonresponse rates exceeding 90% for election surveys, this is a growing concern [Pew Research Center, 2016]. Measurement error arises when the survey instrument itself affects the response, for example due to order effects [McFarland, 1981] or question wording [Smith, 1987]. Finally, specification error occurs when a respon-

---

agree with the marginal distributions in the target population [Voss, Gelman, and King, 1995].

<sup>2</sup>For the 19 ABC, CBS, and Gallup surveys conducted during the 2012 election and deposited into Roper Center's iPoll, when weights in each survey were rescaled to have mean 1, the median respondent weight was 0.73, with an interquartile range of 0.45 to 1.28. For a sampling of 96 polls for 2012 Senate elections, only 19 reported margins of error higher than what one would compute using the SRS formula, and 14 of these exceptions were accounted for by YouGov, an internet poll that explicitly inflates variance to adjust for the sampling weights. Similarly, for a sampling of 36 state-level polls for the 2012 presidential election, only 9 reported higher-than-SRS margins of error.

dent’s interpretation of a question differs from what the surveyor intends to convey (e.g., due to language barriers). In addition to these four types of error common to nearly all surveys, election polls suffer from an additional complication: shifting attitudes. Whereas surveys typically seek to gauge what respondents will do would do on election day, they can only directly measure current beliefs.

In contrast to errors due to sample variance, it is difficult—and perhaps impossible—to build a useful and general statistical theory for the remaining components of total survey error. Moreover, even empirically measuring total survey error can be difficult, as it involves comparing the results of repeated surveys to a ground truth obtained, for example, via a census. For these reasons, it is not surprising that many survey organizations continue to use estimates of error based on theoretical sampling variation, simply acknowledging the limitations of the approach. Indeed, Gallup [2007] explicitly states that their methodology assumes “other sources of error, such as nonresponse, by some members of the targeted sample are equal,” and further notes that “other errors that can affect survey validity include measurement error associated with the questionnaire, such as translation issues and coverage error, where a part or parts of the target population...have a zero probability of being selected for the survey.”

## 1.2 Our study

Here we empirically and systematically study error in election polling, taking advantage of the fact that multiple polls are typically conducted for each election, and that the election outcome can be taken to be the ground truth. We investigate 4,221 polls for 608 state-level presidential, senatorial, and gubernatorial elections between 1998 and 2014, all of which were conducted in the final three weeks of the election campaigns. By focusing on the final weeks of the campaigns, we seek to minimize the impact of errors due to changing attitudes in the electorate, and hence to isolate the effects of the remaining components of survey error.

We find that the average difference between poll results and election outcomes (as mea-

sured by RMSE) is 3.5%, corresponding to a 95% confidence interval of  $\pm 7\%$ , twice the width of most reported intervals. To decompose this survey error into election-level bias and variance terms, we apply hierarchical Bayesian latent variable models [Gelman and Hill, 2007]. We find that average absolute election-level bias is about 1.5%, indicating that polls for a given election often share a common component of error. This result is likely driven in part by the fact that most polls, even when conducted by different polling organizations, rely on similar likely voter models, and thus surprises in election day turnout can have comparable effects on all the polls. Moreover, these correlated frame errors extend to the various elections—presidential, senatorial, and gubernatorial—across the state. Past political commentators have suggested polling organizations “herd”—intentionally manipulating survey results to match those of previously reported polls—which should in turn decrease election-level poll variance.<sup>3</sup> We find, however, that average election-level standard deviation is about 2.5%, well above the 2% implied by most reported margins of errors, which suggests the variance-reduction effects of any herding are smaller than the variance-inducing differences between surveys.

## 2 Data description

Our primary analysis is based on 4,221 polls completed during the final three weeks of 608 state-level presidential, senatorial, and gubernatorial elections between 1998 and 2014. Polls are typically conducted over the course of several days, and following convention, we throughout associate the “date” of the poll with the last date during which it was in the field. We do not include House elections in our analysis since polling is only available for a small and non-representative subset of such races.

To construct this dataset, we started with the 4,154 state-level polls for elections in 1998–2013 that were collected and made available by FiveThirtyEight, all of which were completed during the final three weeks of the campaigns. We augment these polls with

---

<sup>3</sup>See <http://fivethirtyeight.com/features/heres-proof-some-pollsters-are-putting-a-thumb-on-the-scale>.

the 67 corresponding ones for 2014 posted on Pollster.com, where for consistency with the FiveThirtyEight data, we consider only those completed in the last three weeks of the campaigns. In total, we end up with 1,646 polls for 241 senatorial elections, 1,496 polls for 179 state-level presidential elections, and 1,079 polls for 188 gubernatorial elections.

In addition to our primary dataset described above, we also consider 7,040 polls completed during the last 100 days of 314 state-level presidential, senatorial, and gubernatorial elections between 2004 and 2012. All polls for this secondary dataset were obtained from Pollster.com and RealClearPolitics.com. Whereas this complementary set of polls covers only the more recent elections, it has the advantage of containing polls conducted earlier in the campaign cycle.

### 3 Estimating total survey error

For each poll in our primary dataset (polls conducted during the final three weeks of the campaign), we estimate total survey error by computing the difference between: (1) support for the Republican candidate in the poll; and (2) the final vote share for that candidate on election day. As is standard in the literature, we consider *two-party* poll and vote share: we divide support for the Republican candidate by total support for the Republican and Democratic candidates, excluding undecideds and supporters of any third-party candidates.

Figure 1 shows the distribution of these differences, where positive values on the  $x$ -axis indicate the Republican candidate received more support in the poll than in the election. For comparison, the dotted line shows the theoretical distribution of polling errors assuming simple random sampling (SRS) (i.e., the distribution of errors if each poll were a simple random sample centered at the true election outcome). The plot highlights two points. First, for all three political offices, polling errors are approximately centered at zero. Thus, at least across all the elections and years that we consider, polls are not systematically biased toward either party. Indeed, it would be surprising if we had found systematic error, since

### Difference between poll results and election outcomes

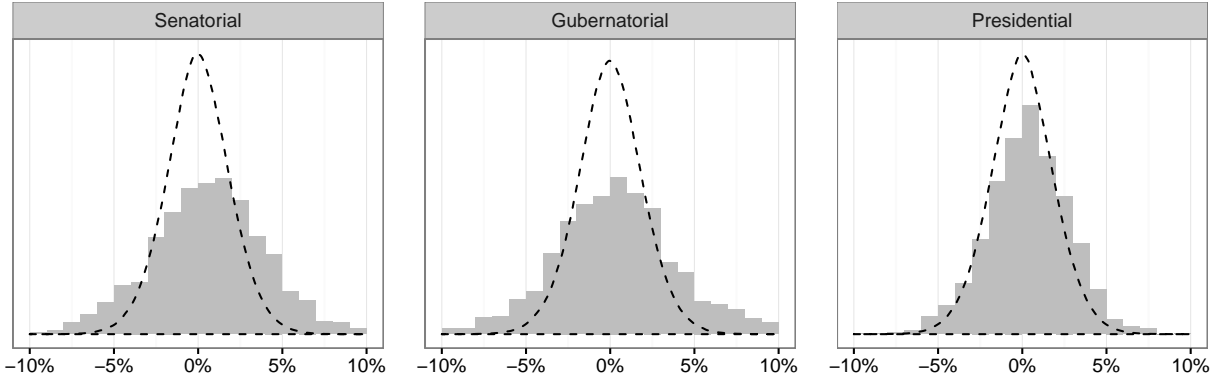


Figure 1: *The distribution of polling errors (Republican share of two-party support in the poll, minus Republican share of the two-party vote in the election) for state-level presidential, senatorial, and gubernatorial election polls between 1998 and 2014. Values indicate the difference between the two-party vote share of the Republican candidate predicted by the poll and the corresponding election outcome. Positive values thus indicate the Republican candidate received more support in the poll than in the election. For comparison, the dashed lines shows the theoretical distribution of polling errors assuming each poll is generated via simple random sampling.*

pollsters would have been highly motivated to notice and corrected for any such bias.

Second, the polls exhibit substantially larger errors than one would expect from simple random sampling. For example, it is not uncommon for senatorial and gubernatorial polls to miss the election outcome by more than 5 percentage points, an event that would rarely occur if respondents were simple random draws from the electorate. Adding quantitative detail to this visually apparent observation, Table 1 lists the root mean square error (RMSE) of the polls, as well as the expected RMSE under SRS.<sup>4</sup> Elections for all three offices have error larger than what one would expect from SRS. The senatorial and gubernatorial polls, in particular, have substantially larger RMSE (3.7% and 3.9%, respectively) than SRS (1.9%). In contrast, the RMSE for state-level presidential polls is 2.5%, much more in line with what one would expect from SRS. Importantly, because reported margins of error are typically

---

<sup>4</sup>For each poll  $i \in \{1, \dots, N\}$ , let  $y_i$  denote the two-party support for the Republican candidate, and let  $v_i$  denote the final two-party vote share of the Republican candidate in the corresponding election. Then RMSE is  $\sqrt{\frac{1}{N} \sum_{i=1}^N (y_i - v_i)^2}$ .

Senatorial	Gubernatorial	Presidential	SRS
3.7%	3.9%	2.5%	1.9%

Table 1: *Estimates of average poll error (RMSE) in state-level presidential, senatorial, and gubernatorial races. For comparison, expected error from SRS is also given. Because reported margins of error are typically derived from theoretical SRS error rates, the traditional margins of error are too conservative.*

derived from theoretical SRS error rates, the traditional margins of error are too conservative. Namely, SRS-based 95% confidence intervals cover the actual outcome for only 71% of senatorial polls, 72% of gubernatorial polls, and 87% of presidential polls. It is not immediately clear why presidential polls fare better, but one possibility is that turnout in such elections is easier to predict and so these polls suffer less from frame error.

We have thus far focused on polls conducted in the three weeks prior to election day, in an attempt to minimize the effects of error due to changing attitudes in the electorate. To examine the robustness of this assumption, we now turn to our secondary polling dataset and, in Figure 2, plot average poll error as a function of the number of days to the election. Due to the relatively small number of polls conducted on any given day, we include in each point in the plot all the polls completed in a seven-day window centered at the focal date (i.e., polls completed within three days before or after that day). As expected, polls early in the campaign season indeed exhibit more error than those taken near election day. Average error, however, appears to stabilize in the final weeks, with little difference in RMSE one month before the election versus one week before the election. Thus, the polling errors that we see during the final weeks of the campaigns are likely not driven by changing attitudes, but rather result from a combination of frame and nonresponse error. Measurement and specification error also likely play a role, though election polls are arguably less susceptible to such forms of error.



Root mean square poll error over time

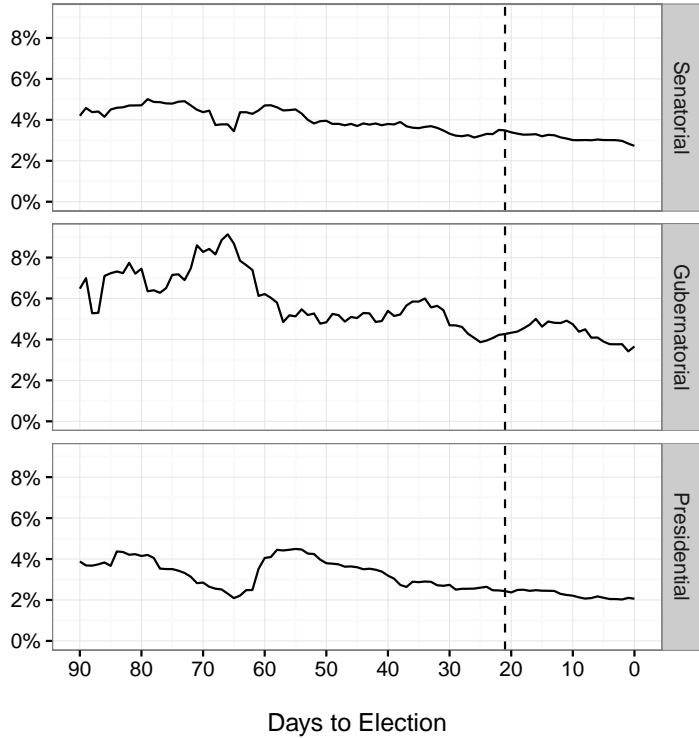


Figure 2: *Poll error, as measured by RMSE, over the course of elections. The RMSE on each day  $x$  indicates the average error for polls completed in a seven-day window centered at  $x$ . The dashed vertical line at the three-week mark shows that poll error is relatively stable during the final stretches of the campaigns, suggesting that the discrepancies we see between poll results and election outcomes are by and large not due to shifting attitudes in the electorate.*

## 4 Estimating election-level bias and variance

In principle, Figure 1 is consistent with two distinct possibilities. On one hand, election polls may typically be unbiased but have large variance; on the other hand, polls may generally have non-zero bias, but in aggregate these biases cancel to yield the depicted distribution. To check, we next decompose the observed poll error into election-level bias and variance components. The bias term captures systematic errors shared by all polls in the election (e.g., due to shared frame errors), while the variance term captures traditional sampling variation as well as variation due to differing survey methodologies across polls and polling organizations.

### Difference between poll averages and election outcomes

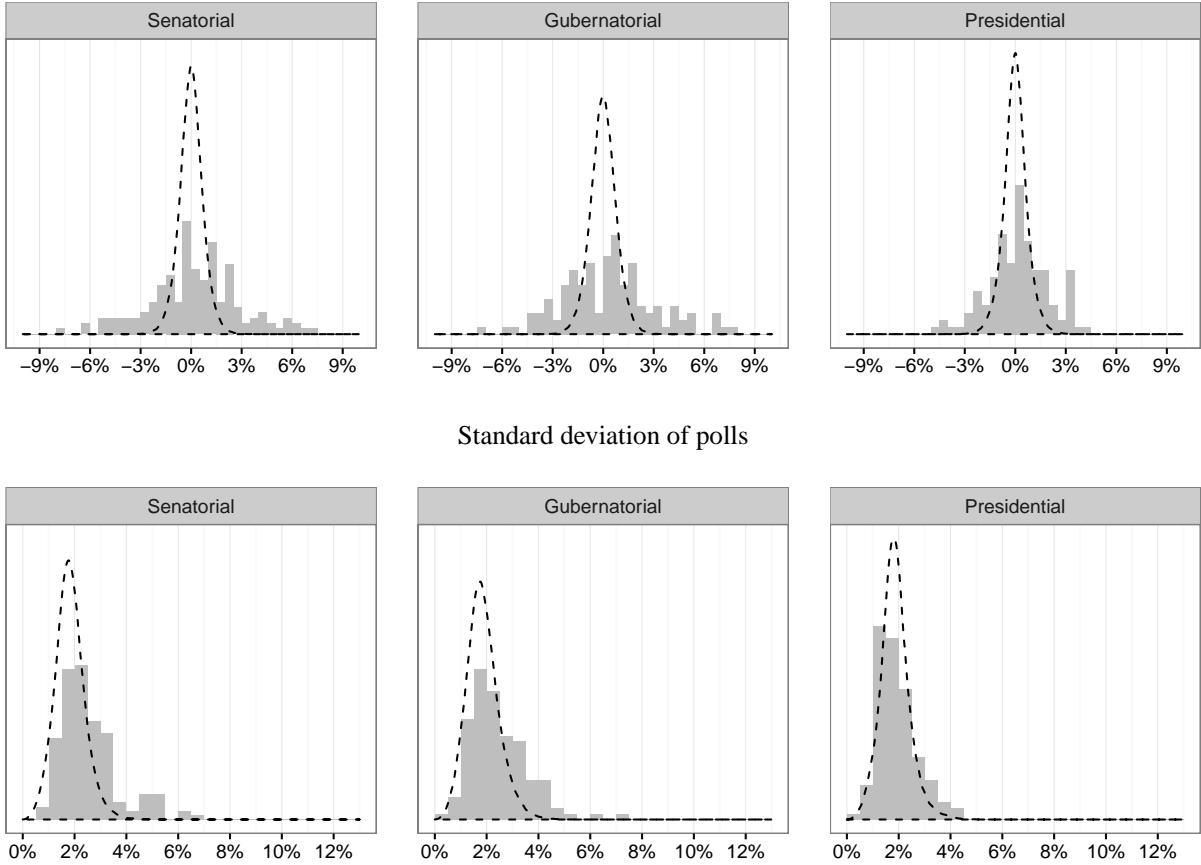


Figure 3: *Simple estimates of election-level bias and variance, obtained by taking the difference between the average of polls in an election and the election outcome (top plot), and the standard deviation of polls in an election (bottom plot). In the top plot, positive values indicate the Republican candidate received more support in the polls than in the election. For comparison, the dashed lines shows the theoretical distribution of results if polls were generated via SRS.*

## 4.1 Simple sample estimates

To estimate election-level bias and variance, we start by imagining poll results in a given election are independent draws from an unknown, election-specific *poll distribution*. This poll distribution reflects both the usual sampling variation, as well as uncertainty arising from nonresponse, frame, and other sources of polling error. With this setup, a simple and intuitive estimate of the election-level poll bias is the difference between the average of the

	Senatorial	Gubernatorial	Presidential	SRS
Average error (RMSE)	3.7%	3.9%	2.5%	1.9%
Average absolute bias	2.1%	2.3%	1.4%	0%
Average standard deviation	2.6%	2.6%	1.9%	1.9%

Table 2: *Simple estimates of election-level RMSE, bias, and variance. In particular, election-level bias is estimated by taking the difference between the average of polls in a election and the election outcome. For comparison, we include the corresponding terms for SRS.*

poll results in that election and the election outcome itself. Similarly, we can estimate the variance of the election-specific poll distribution via the sample variance of the observed poll results. As before, we consider two-party support for the Republican candidate throughout our analysis.

Returning to our primary dataset of polls completed within the final three weeks of the campaigns, we compute election-level bias and variance for the 397 races for which we have at least four polls. Figure 3 shows the resulting distribution of estimated election-level poll bias and variance. The dashed lines show what one would expect from SRS, if the polls were simple random samples centered at the true election outcome. We note that even though SRS is unbiased, the bias of SRS as *estimated* by this method will never be identically zero, as indicated by the dashed lines. The figure clearly shows that election-level polling bias—particularly for senatorial and gubernatorial races—is often substantial, at times in excess of 5%. The election-level standard deviation of polls is likewise larger than one would expect from SRS. As summarized in Table 2, senatorial and gubernatorial races have average absolute bias of more than 2%, and presidential races have average absolute bias of 1.4%. The bias term, which is not reflected in traditional margins of error, is as big as the theoretical sampling variation from simple random sampling.

## 4.2 A Bayesian approach

Our analysis above indicates that election-level bias is a key component of polling error. However, given the relatively small number of polls in each election, simple poll averages

yield imprecise estimates of election-level bias. In particular, since the observed election-level average of poll results is itself a noisy estimate of the true (unknown) mean of the election-specific poll distribution, the method above will overestimate the average absolute election-level poll bias. As a stark illustration of this fact, the estimated average absolute bias of SRS is 0.5%, significantly larger than the true value of zero.

To address this issue and accurately estimate election-level bias and variance, we fit hierarchical Bayesian latent variable models [Gelman and Hill, 2007]. The *latent variables* here refer to parameterizations of election-level bias and variance, and the hierarchical Bayesian framework allows us to make reasonable inferences even for races with relatively small numbers of polls. Whereas the above, simple approach conflates noise in the estimate of bias with actual bias in the underlying poll distribution, this more nuanced technique overcomes that shortcoming.

For each poll  $i$  in election  $r[i]$ , let  $y_i$  denote the two-party support for the Republican candidate (as measured by the poll), where the poll has  $n_i$  respondents and was conducted  $t_i$  months before the election (since we restrict to the last three weeks of the campaign, we have  $0 \leq t_i < 1$ ). Let  $v_i$  denote the final two-party vote share for the Republican candidate. Then we assume the poll outcome  $y_i$  is a random draw from a normal distribution parameterized as follows:

$$y_i \sim \text{Normal} \left( v_i + \alpha_{r[i]} + t_i \beta_{r[i]}, \sqrt{\frac{v_i(1-v_i)}{n_i} + \tau_{r[i]}} \right). \quad (1)$$

Here,  $\alpha_{r[i]} + t_i \beta_{r[i]}$  is the election-specific bias (positive values indicate the poll is likely to overestimate support for the Republican candidate), where we allow the bias to change linearly over time. In reality, bias is not perfectly linear in time, but given the relative stability of late-season polls (Figure 2), this seems like a natural and reasonable functional form to assume. The possibility of election-specific excess variance (relative to SRS) in poll results is captured by the  $\tau_{r[i]}$  term. Estimating excess variance is statistically and computationally tricky, and there are many possible ways to model it; for simplicity, we use

an additive term, and note that our final results are largely robust to the exact specification.

To help deal with the relatively limited number of polls in each election, we further assume the parameters for election-level bias ( $\alpha$  and  $\beta$ ) and variance ( $\tau$ ) are themselves drawn from normal distributions, leading to a hierarchical model structure:

$$\alpha_j \sim \text{Normal}(\mu_\alpha, \sigma_\alpha)$$

$$\beta_j \sim \text{Normal}(\mu_\beta, \sigma_\beta)$$

$$\tau_j \sim \text{Half-Normal}(0, \sigma_\tau)$$

Finally, weakly informative priors are assigned to the hyper-parameters  $\mu_\alpha$ ,  $\sigma_\alpha$ ,  $\mu_\beta$ ,  $\sigma_\beta$  and  $\sigma_\tau$ . The hierarchical priors have the effect of pulling the parameter estimates of bias and variance in any given election toward the average over all elections, where the magnitude of the effect is related to the number of polls in the race and the overall distribution of these terms across all races. Thus, even for races with few polls, one can obtain accurate estimates of bias and variance by statistically grounding off of the estimates inferred for other races.

This model is fit separately for senatorial, presidential and gubernatorial elections. Posterior distributions for the parameters are obtained via Hamiltonian Monte Carlo [Hoffman and Gelman, 2014] as implemented in Stan, an open-source modeling language for full Bayesian statistical inference. To check performance, we first fit the model on synthetic data generated via SRS, preserving the empirically observed election outcomes, the number of polls in each election, and the size of each poll. On this synthetic dataset, the model estimates average absolute bias is 0.2% and average excess standard deviation is 0.1%, in line with the theoretically correct answer of 0 for both quantities. In particular, the model-estimated average absolute bias, while not perfect, is considerably better than what we obtained via the simple method of Section 4.1.

The results from the fitted model are summarized in Table 3. (The full distribution of election-level estimates is provided in the Appendix.) Consistent with our previous analysis,

	Senatorial	Gubernatorial	Presidential	SRS
Average error (RMSE)	3.7%	3.9%	2.5%	1.9%
Average absolute bias	1.8%	2.0%	1.0%	0%
Average absolute bias on election day	1.6%	1.9%	1.0%	0%
Average standard deviation	2.6%	2.6%	2.2%	1.9%

Table 3: *Model-based decomposition of election-level error into bias and variance, both components of which are higher than would be expected from sampling alone.*

elections for all three offices exhibit substantial average absolute bias, approximately 2% for senatorial and gubernatorial elections and 1% for presidential elections. As expected, bias as estimated by the Bayesian model is somewhat smaller (and ostensibly more accurate) than what we obtained from the simple sample averages. As before, however, we still have that the bias term is about as big as the theoretical sampling variation from SRS. The third line in Table 3 shows estimated average absolute bias on the day of the election (i.e., as given by the  $\alpha_j$  terms). The slight decrease in election day bias suggests that at least part of the error in poll results comes from public sentiment changing over the course of the campaign. Given that we have intentionally focused on polls conducted during the final three weeks of the election cycle to mitigate the effect of such movements, it is not surprising that the decrease in bias is relatively small.

Why do polls exhibit non-negligible election-level bias? We offer two possibilities. First, as discussed above, polls in a given election often have similar sampling frames. Telephone surveys, regardless of the organization that conducts them, will miss those who do not have a telephone. Relatedly, projections about who will vote—often based on standard likely voter screens—do not vary much from poll to poll, and as a consequence, election day surprises (e.g., an unexpectedly high number of minorities or young people turning out to vote) affect all polls similarly. Second, since polls often apply similar methods to correct for nonresponse, errors in these methods can again affect all polls in a systematic way. For example, it has recently been shown that supporters of the trailing candidate are less likely to respond to polls, even after adjusting for demographics [Gelman et al., 2016]. Since most polling

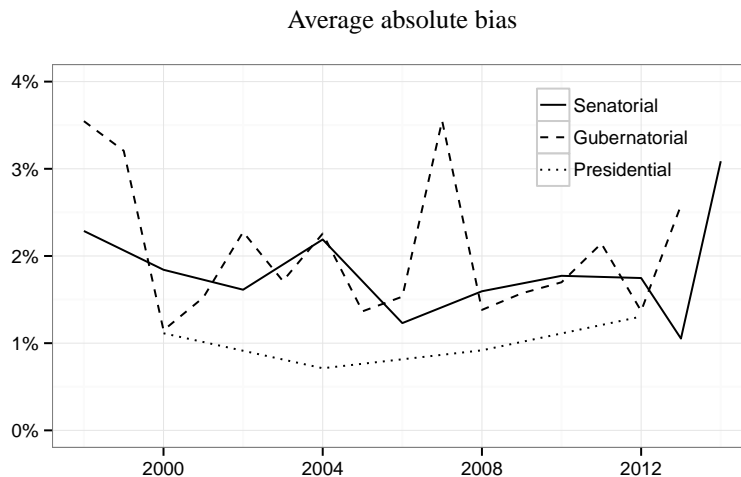


Figure 4: *Model-based estimates of average absolute bias show no consistent trends over time.*

organizations do not correct for such partisan selection bias, their polls are all likely to be systematically skewed.

Figure 4 shows how the average absolute election-level bias changes from one election cycle to the next. While there is noticeable year-to-year variation, there does not appear to be any consistent trend over time. Given that survey response rates have plummeted during this period (from an average of 36% in 1998 to 9% in 2012), it is perhaps surprising that we do not see an accompanying rise in poll bias [Pew Research Center, 2012]. Nevertheless, the consistency over time provides further evidence that the effects we observe are real and persistent.

In addition to bias, Table 3 also shows the average election-level standard deviation. Though the standard deviation of presidential elections (2.2%) is not much larger than for SRS (1.9%), both senatorial and gubernatorial elections have standard deviations approximately 0.7 percentage points more than SRS, a large value relative to the magnitude of typically reported errors. As with the observed bias, it is difficult to isolate the specific cause for the excess variation, but we can again speculate about possible mechanisms. Since different polling organizations often use different survey methodologies—such as survey mode (telephone vs. Internet), and question wording and ordering—measurement error

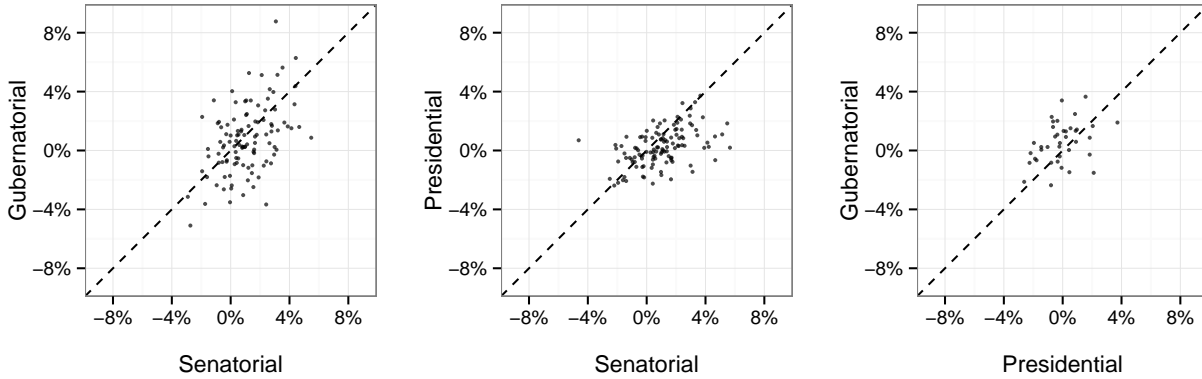


Figure 5: *Comparison of election-level polling bias in various pairs of state-level elections. Each point indicates the estimated bias in two different elections in the same state in the same year. The plots show modest correlations, suggesting a mix of frame and nonresponse errors.*

likely contributes to poll-to-poll variation. Election-level variation is also likely in part due to differences in the precise timing of the polls, and idiosyncratic differences in likely voter screens.

Finally, Figure 5 shows the relationship between election-level bias in elections for different offices within a state. Each point corresponds to a state, and the panels plot estimated bias for the two elections indicated on the axes. Overall, we find moderate correlation in bias for elections within the state: 0.54 for gubernatorial vs. senatorial, 0.49 for presidential vs. senatorial, and 0.28 for gubernatorial vs. presidential. Such correlation again likely comes from a combination of frame and nonresponse errors. For example, since party-line voting is relatively common, an unusually high turnout of Democrats on election day could affect the accuracy of polling in multiple races. This correlated bias in turn leads to correlated errors, and illustrates the importance of treating polling results as correlated rather than independent samples of public sentiment.



## 5 Discussion

Particularly in polls for senatorial and gubernatorial elections, we find substantial election-level bias and excess variance. At the very least, this observation suggests that care should be taken when using poll results to assess a candidate’s reported lead in a competitive race. Moreover, in light of the correlated polling errors that we find, close poll results should give one pause not only for predicting the outcome of a single election, but also for predicting the collective outcome of related races. To mitigate the recognized uncertainty in any single poll, it has become increasingly common to turn to aggregated poll results, whose nominal variance is often temptingly small. While aggregating results is generally sensible, it is particularly important in this case to remember that shared election-level poll bias persists unchanged, even when averaging over a large number of surveys.

Taking a step further, our analysis offers a starting point for polling organizations to quantify the errors left unmeasured by traditional margins of errors. Instead of simply stating that these commonly reported metrics miss significant sources of error, which is the status quo, these organizations could—and we feel should—start quantifying and reporting the gap between theory and practice. Indeed, empirical election-level bias and variance could be directly incorporated into reported margins of error. Though it is hard to estimate these quantities for any particular election, historical averages could be used as proxies.

Large election-level bias does not afflict all estimated quantities equally. For example, it is common to track movements in sentiment over time, where the precise absolute level of support is not as important as the change in support. A stakeholder may primarily be interested in whether a candidate is on an up or downswing rather than his or her exact standing. In this case, the bias terms—if they are constant over time—cancel, and traditional methods may adequately capture poll error.

Given the considerable influence election polls have on campaign strategy, media narratives, and popular opinion, it is important to not only have accurate estimates of candidate support, but also accurate accounting of the error in those estimates. Looking forward,

we hope our analysis and methodological approach provide a framework for understanding, interpreting, and reporting errors in election polling.

## References

- Stephen Ansolabehere and Thomas R. Belin. Poll faulting. *Chance*, 6, 1993.
- Paul P. Biemer. Total survey error: Design, implementation, and evaluation. *Public Opinion Quarterly*, 74(5):817–848, 2010. ISSN 0033-362X.
- Gallup. Gallup world poll research design. <http://media.gallup.com/WorldPoll/PDF/WPResearchDesign091007bleeds.pdf>, 2007. Accessed: 2016-04-07.
- Andrew Gelman and Jennifer Hill. *Data Analysis Using Regression and Multi-level/Hierarchical models*. Cambridge University Press, 2007.
- Andrew Gelman, Sharad Goel, Douglas Rivers, and David Rothschild. The mythical swing voter. *Quarterly Journal of Political Science*, 2016.
- Robert M. Groves and Lars Lyberg. Total survey error: Past, present, and future. *Public Opinion Quarterly*, 74(5):849–879, 2010. ISSN 0033-362X.
- Matthew D. Hoffman and Andrew Gelman. The no-U-turn sampler: Adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15 (Apr):1593–1623, 2014.
- Sharon Lohr. *Sampling: Design and Analysis*. Nelson Education, 2009.
- Sam G. McFarland. Effects of question order on survey responses. *Public Opinion Quarterly*, 45(2):208–215, 1981.

Pew Research Center. Assessing the representativeness of public opinion surveys. <http://www.people-press.org/2012/05/15/assessing-the-representativeness-of-public-opinion-surveys>, 2012. Accessed: 2016-04-07.

Pew Research Center. Our survey methodology in detail. <http://www.people-press.org/methodology/our-survey-methodology-in-detail>, 2016. Accessed: 2016-04-07.

Tom W. Smith. That which we call welfare by any other name would smell sweeter: An analysis of the impact of question wording on response patterns. *Public Opinion Quarterly*, 51(1):75–83, 1987.

D. Stephen Voss, Andrew Gelman, and Gary King. Pre-election survey methodology: Details from nine polling organizations, 1988 and 1992. *Public Opinion Quarterly*, 59:98–132, 1995.

# A Appendix

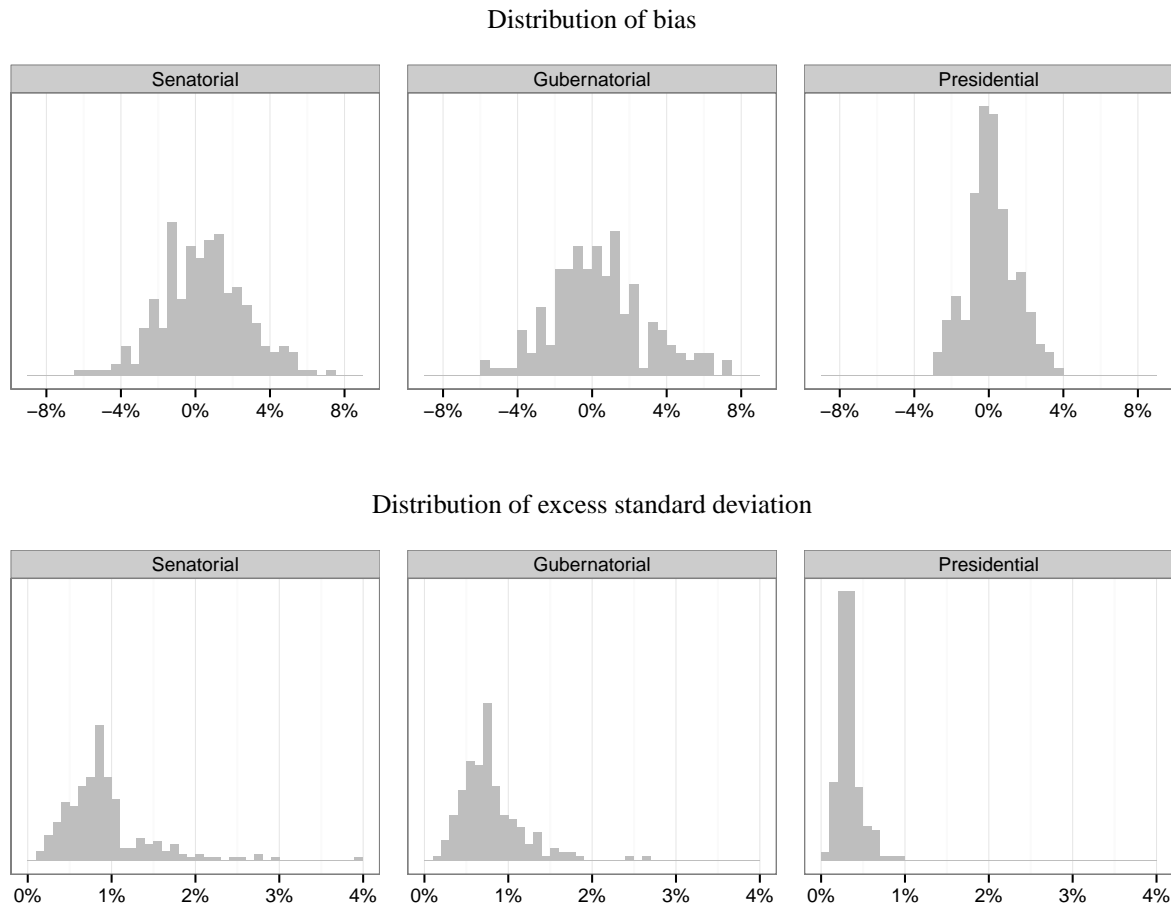


Figure 6: *Distribution of model-estimated election-level bias and excess standard deviation. In the top plot, positive values indicate the Republican candidate received more support in the polls than in the election.*