

Special Section: Statistics for Democratic Processes

Predicting Presidential and Other Multistage Election Outcomes Using State-Level Pre-Election Polls

William F. CHRISTENSEN and Lindsay W. FLORENCE

Although much of the media attention during presidential election years focuses on polls tracking popular support for the major candidates, the complicated role played by the Electoral College in this multistage election process must be accounted for in order to address the issue of winning the presidency. State-level pre-election polls are used in a manner that allows the structure of multistage election processes to be addressed directly. We consider frequentist and Bayesian approaches for predicting election outcomes and discuss ways to incorporate such analyses in a course project suitable for undergraduates or graduate students studying statistics. Using state-level pre-election polling data, we consider the U.S. presidential election of 2004 and we also apply this approach to predict the control of the U.S. Senate in 2006. This class exercise has proved to be a useful “capstone project” which requires students to address a complicated problem by synthesizing multiple sources of available data and applying a combination of statistical methods. Using simulation-based approaches for addressing the multistage nature of presidential elections and control-of-Congress processes can be valuable and instructive for students of statistics and political science, and can be beneficial to the media in providing consumers with political news.

KEY WORDS: Capstone project; Electoral College; News media; Politics; Simulation; U.S. president.

1. INTRODUCTION

The problem of understanding and predicting election outcomes has long been part of political science research. Several authors have used regression models for nationwide polling data to forecast the outcome of the popular vote (Campbell and Wink 1990; Campbell 1996) while others use state-level polls to predict the outcome of the election in each state (Campbell 1992;

William Christensen is Associate Professor, Department of Statistics, Brigham Young University, Provo, UT 84602 (E-mail: william@stat.byu.edu). Lindsay Florence is a Master’s Student, Department of Statistics, Brigham Young University, Provo, UT 84602 (E-mail: lindsay.florence@gmail.com). The authors thank Scott Grimshaw, Shane Reese, the associate editor, and the referees for insightful comments that improved the article.

Cohen 1998; Holbrook and DeSart 1999). However, analyses of popular opinion trends do not address the issue of principal interest—that is, a prediction of who will actually win the presidency. This is not always the same as winning the national popular vote (as Al Gore did in the 2000 U.S. presidential election). Park, Gelman, and Bafumi (2004) employed a multilevel logistic regression model to generate estimates of state-level vote shares. Their model employs national opinion data and state-level demographic covariates to obtain estimates in a manner that is related to the small-area estimation problem. Instead of using national-level polls and covariates to derive information about each state, we use state-level polls directly.

Although national popular opinion during U.S. presidential races is most commonly measured and discussed in the media, the U.S. presidential election is based on the Electoral College, in which each state has a number of electors equal to the number of its U.S. senators plus the number of its U.S. representatives. Additionally, the District of Columbia acts as a “state” with a number of electors proportional to its population, but not exceeding the number of electors assigned to any of the states. The people in each state vote for the state-level electors who then vote for a presidential candidate, with most states using a winner-take-all policy for casting votes in the Electoral College. Thus, although much of the media attention during election years focuses on polls tracking popular support for the major candidates, the complicated role played by the Electoral College in this multistage election process must be accounted for in order to address the issue of winning the presidency. The methods discussed herein apply equally well to other multistage electoral processes, such as the control of the U.S. Senate and U.S. House of Representatives.

The principal facets of the analyses considered in this article were originally formulated as a nonparametric statistical methods class project during the month before the 2004 presidential election. The easy accessibility of rich data combined with the inherent interest in elections proved this exercise to be one of the students’ favorite components of the course. Moreover, it is sometimes difficult to provide opportunities for students to address complex problems in a real-world environment. This class exercise has proved to be a useful “capstone” project which requires students to address a complicated problem by synthesizing available data with a combination of existing statistical

tools. The exercise can be simple enough to be implemented by advanced undergraduate statistics students, but can also be adapted for M.S. level statistics students.

Because our discussion focuses predominantly on the use of the 2004 state-level pre-election polls, throughout the article we refer to the 2004 candidates (Bush and Kerry), but the methods apply to any presidential or other multistage election process. Our method addresses each stage of the election process by first using state-level polls to calculate the probability of Bush winning each state, and then simulating a large number of elections in order to approximate the sampling distribution for “Bush electoral votes.” We can then estimate the probability that Bush receives more than 269 electoral votes. This simulation-based approach also allows the students to explore somewhat complicated questions such as:

- What is the probability of an Electoral College tie (and the controversy that would almost surely accompany it)?
- What is the probability that less than 20 electoral votes will separate the two candidates (increasing the likelihood of a contested election)?
- What is the probability that Bush wins the election if Kerry wins Florida?
- What is the probability that Bush wins the election if Bush wins at least two of the three large “battleground” states (Florida, Ohio, and Pennsylvania)?

Section 2 discusses pre-election poll data and how to access it. Section 3 discusses the simulation of elections using both frequentist and Bayesian approaches. Section 4 presents results based on pre-election poll data associated with the 2004 U.S. presidential election. We also illustrate how our simulation-based approach can be applied to other multistage electoral processes by evaluating the probability that the Republican Party would retain control of the U.S. Senate in the 2006 election. Section 5 discusses the potential value of this exercise in the classroom, describes statistics courses where state pre-election poll data could be useful in teaching statistical concepts, and makes recommendations for implementation in classrooms and the media.

2. ELECTION POLL DATA

Historically, one of the main challenges associated with forecasting election outcomes has been the lack of state-level pre-election poll data (Cohen 1998), but opinion polls are now easily accessible on the Internet. For example, in 2004, state-level poll data for all 50 states and the District of Columbia were available from several Web pages such as the *LA Times* Web site (where most of the data for these analyses were obtained). Using Internet resources is convenient for students and makes it possible for them to follow the election without requiring the instructor to personally update datasets. Although pre-election polling data are inevitably flawed, they can still provide much insight about national and regional trends, and the nontrivial biases inherent in such data provide a vehicle for class discussion

about data validity and the associated validity of statistical inference.

Researchers have noted that presidential pre-election polling data may not be useful until at least early September after the two parties’ national conventions (Campbell and Wink 1990; Campbell 1996). For the analyses of the 2004 presidential election discussed here, we recorded state-level opinion poll updates 12 different times beginning on October 12, 2004, and ending on November 2, the day before the election. During the 22-day window in which we recorded poll results, some states had no new updates while others had as many as ten. When we began collecting data on October 12, 2004, we assumed the polls for each state were taken on that day even though some polls may have been older. Multiday polls were treated as if the data were gathered on the day the poll was reported. All datasets used in this manuscript, along with documentation about the Web sites used to gather data, can be found at the Web site <http://statistics.byu.edu/faculty/wfc/electionpollproject>.

State polls are not updated on a consistent basis. This raises a concern about how to incorporate the incoming data. The following questions can be addressed by students: Should one use only the latest poll one has for each state? Should one combine all previous polls, weighting each respondent participating in each poll equally? Should one combine poll results but down-weight poll results that are older? We consider each of these approaches in the next section and compare the probabilities arising from each.

3. SIMULATING ELECTIONS

3.1 Frequentist Methods

Predicting election day voting outcomes based on early pre-election polling is a very complicated problem because such a prediction would require a consideration of opinion trends, future campaign spending, and historical voter behavior. Further, the actual election day results will be affected by many unpredictable factors arising in the final days of the campaign, including world events and candidate mistakes. Consequently, we focus on the simpler problem of estimating the probability that Bush would win the election if it were held on the day of the recent poll.

To estimate the probability of Bush winning the election, we begin by quantifying the evidence that Bush is more popular than Kerry in each state. For this discussion, we refer to 51 “states” because the District of Columbia also votes in the Electoral College. We further note that Maine and Nebraska could be further partitioned into several “substates” because these allow congressional districts to cast votes for different candidates. For simplicity (and because a within-state partitioning of electoral votes has not yet occurred), we treat each state as winner-take-all. A more thorough analysis could be undertaken which accounts for this nonstandard behavior in Maine and Nebraska, but this would require congressional-district-level polls which are not as readily available. Alternatively, the same state-level polls can be used to simulate outcomes in each congressional district separately. However, applied to the 2004 presidential

polling data, such an approach yields daily estimates of $\Pr\{\text{Bush win}\}$ that differ by less than 0.25% in our analyses.

State-level pre-election poll results were converted to a trinomial vector consisting of the percent favoring Bush, the percent favoring Kerry, and the percent undecided/favoring other. Because there was no viable third candidate in October 2004, we began by quantifying the evidence that Bush was more popular than Kerry. From a frequentist perspective, one can summarize this evidence in state i using the following asymptotically normal test statistic based on the multinomial distribution:

$$z_i = \frac{p_{\text{Bush}|i} - p_{\text{Kerry}|i}}{\sqrt{\frac{(p_{\text{Bush}|i} + p_{\text{Kerry}|i} - (p_{\text{Bush}|i} - p_{\text{Kerry}|i})^2)}{n_i}}}, \quad (1)$$

where $i = 1, \dots, 51$ represents the states, and n_i is the sample size for the poll. Note that although this statistic accounts for those who are undecided and favoring other candidates, we are implicitly assuming that these voters will not impact the outcome of the race. This approach for summarizing evidence would be invalid if, for example, undecided voters are dramatically more likely to side with a particular candidate or if a third party candidate has an opportunity to win.

One of the first issues confronting students (particularly those who have not been exposed to Bayesian reasoning) is the challenge of converting the p value calculated from (1) into a quantity that represents the probability of Bush winning state i , which we denote θ_i . That is, if we denote our observed data for state i using \mathbf{x}_i (a vector of ones and zeros corresponding to respondents favoring Bush and Kerry, respectively), can one use

$$\begin{aligned} \hat{\theta}_i &= 1 - \Pr\{\text{data more favorable to Bush than } \mathbf{x}_i \mid \text{race is tied}\} \\ &= \Phi(z_i) \end{aligned}$$

as a reasonable measure of

$$\theta_i = \Pr\{\text{Bush wins} \mid \mathbf{x}_i\}?$$

The relationship between p values and Bayesian evidence was addressed directly by Casella and Berger (1987), Berger and Sellke (1987), and the discussants' comments that followed these articles. The comment of Morris (1987) is particularly useful in the context of this problem because the author considered an election scenario that is closely related to the problem at hand, and the argument is short and reasonably digestible by advanced undergraduates majoring in statistics. Under reasonable prior specifications for the proportion of voters supporting a candidate, Morris (1987) illustrated that one minus the frequentist p value is too small to represent the Bayesian posterior probability, but that the two quantities become equal as $n_i \rightarrow \infty$. As a part of the project in our class, students are asked to summarize the Morris (1987) argument in their reports when justifying the use of one minus the frequentist p value as an appropriate measure of the probability of Bush winning the state.

We then consider a possible election outcome by drawing 51 Bernoulli responses using the estimated probabilities of Bush

winning each state. Each Bernoulli response is multiplied by the number of electoral votes allocated to each state and the resulting values are summed to obtain the electoral votes for Bush. Following this procedure, we simulate 100,000 elections as a means for approximating the sampling distribution for "Bush electoral votes." We can then calculate the proportion of elections where Bush receives at least 269 electoral votes, and use this proportion as an estimate of

$$\theta_{\text{EC}} = \Pr\{\text{Bush wins electoral college} \mid \mathbf{x}_1, \dots, \mathbf{x}_{51}\}.$$

Even though Bush would not have automatically won the election in the case of a 269-to-269 tie, we considered a tie in the Electoral College to be a victory for Bush because Republicans in the 2004 U.S. House of Representatives controlled a majority of their respective state delegations and looked very likely to maintain a majority of state delegations in the 2004 election. Thus, Bush most certainly would have won the election if the decision were made by an early 2005 House vote consisting of one vote per state.

Because pre-election polls are continually being updated during the weeks prior to the election, one must consider the way in which new and old information should be assimilated when calculating the state-level probabilities. We consider three different data assimilation methods when calculating these probabilities and compare the difference in outcomes. The methods considered are:

1. *Latest Poll.* Consider only the most recent poll for each state.
2. *Combined Polls.* Combine all previous polls up the present time and treat it as a single sample, weighting only by sample size.
3. *Weighted Polls.* Combine all previous polls but adjust the sample size according to a weight function depending on the day the poll is taken. The weight function we use is $w(t) = 1 - \frac{t}{26}$, where t is the number of days since the poll was carried out. This weight is then multiplied by the sample size. As an example, when conducting an analysis on November 2, a poll reported on October 20 would have only half the weight of a poll reported on November 2. Brown and Chappell (1999) also used a weight function that changes according to how close the poll is to the election. Because we use data from only the last 22 days of the 2004 presidential campaign, we consider only one weight function in this analysis. The effect of the weight function is considered in the analysis of the 2006 Senate election in Section 4.2.

3.2 Bayesian Methods

For each state, there exists readily available information about past voting behavior, current party affiliation, satisfaction with the incumbent, or other demographic information. Consequently, a Bayesian hierarchical approach is appealing, particularly since the justification for the frequentist approach to obtain

state-level probabilities requires that we appeal to an asymptotic result. For simplicity, we treat each response obtained from a pre-election poll as if it were a Bernoulli outcome (i.e., a vote for Bush or a vote for Kerry). As with the frequentist approach, we are implicitly assuming that undecided voters will eventually favor and vote for the candidates in the same proportions as the voters in general. Other researchers have addressed this issue by either dividing the percentage of undecided voters proportionally or equally between the two major candidates (Campbell and Wink 1990; Campbell 1996). Our approach allows us to address the issues using a binomial distribution to describe voter preference in pre-election polls, but more sophisticated methods using the multinomial distribution could also be used.

A great deal of effort can be invested in creating sophisticated prior distributions for π_i , the proportion voting for Bush in state i . Insights about constructing the prior can be obtained from the literature. Campbell (1992) used the general state support for a party as a regression model parameter for predicting vote share, and Holbrook and DeSart (1999) used the mean percentage of votes for the two major parties from the previous two elections. One might also consider vote shares for each party as a function of primary election turnout, primary election results, presidential approval ratings, or recent local election outcomes. In order to provide a simple and accessible approach to prior distribution construction for undergraduates with little or no knowledge of political science, we use the state-level proportions voting for the 2000 Republican candidate (Bush) and the 2000 Democratic candidate (Gore) to construct beta prior distributions for π_i . Although more simplistic than other possible approaches, this approach conformed with the conventional wisdom that little would change from the polarizing 2000 election. Additionally, results from the 2000 presidential election seemed a more promising a priori predictor than 2004 party affiliation and allowed students a concrete starting point for constructing reasonable priors.

In state i , let y_i be the number of the n_i polled voters favoring Bush and let π_i be the actual proportion of voters favoring Bush, where y_i and n_i are obtained using one of the data-assimilation methods described in Section 3.1. Then the likelihood for y_i is

$$f(y_i|\pi_i) \sim \text{Bin}(n_i, \pi_i),$$

and the (conjugate) prior distribution for π_i is

$$f(\pi_i) \sim \text{Beta}(m\pi_{i,\text{Bush2000}}, m\pi_{i,\text{Gore2000}}),$$

where $\pi_{i,\text{Bush2000}}$ and $\pi_{i,\text{Gore2000}}$ are the proportions voting for Bush and Gore in 2000 and m represents a prior “weight.” Our use of $m = 300$ implies that we give the 2000 election the same weight in the analysis as a newly reported pre-election poll with 300 respondents (approximately half the influence of a typical 600-person opinion poll). When using the weighted polls data-assimilation method discussed in Section 3.1, $m = 300$ gives the prior distribution about the same influence as a two-week-old poll with 600 respondents. In our analyses we found that if we restrict the data to the latest poll, we could observe small but perceptible differences in outcomes when varying m in the range of 1 to 500. However, when using either the combined

polls or weighted polls assimilation methods, varying m in the range of 1 to 500 had relatively little effect because these methods’ accumulated sample sizes quickly became large enough to swamp out the impact of the 2000 election results.

Standard calculations can be used to show that the binomial likelihood and the beta prior yields the posterior distribution

$$f(\pi_i|y_i) \sim \text{Beta}\left(y_i + m\pi_{i,\text{Bush2000}}, n_i - y_i + m\pi_{i,\text{Gore2000}}\right).$$

Using the Bayesian framework, we simulate an election by first computing $\hat{\theta}_i$ from each state’s posterior distribution for π_i . That is, $\hat{\theta}_i = \int_{0.5}^1 f(\pi_i|y_i)d\pi_i$ is the Bayesian estimate of the probability that Bush wins state i . As with the frequentist approach, once we have obtained estimates of the state-level probabilities ($\hat{\theta}_i$), we then simulate a large number of elections and use the relative frequency of Bush victories (≥ 269 electoral votes) to estimate θ_{EC} .

4. RESULTS

4.1 U.S. Presidential Election 2004

Because the results of the Electoral College simulations depend upon a complex combination of state-level probabilities (θ_i), we begin our analysis by illustrating the differences in the frequentist and Bayesian approaches for calculating θ_i . Figure 1 illustrates the value of θ_{Florida} when using the weighted polls data-assimilation method in combination with the frequentist and Bayesian ($m = 300$) approaches. The probability was somewhat volatile during the final week due to polls on October 27 and November 2 that indicated higher rates of support for Kerry, but the important concept illustrated in Figure 1 relates to the relationship between the frequentist and Bayesian estimates over time. Early in the data-collection process, the prior for θ_{Florida} (centered at 0.5) has greater impact than it does later in the process when the accumulated pre-election polling data swamp out the effect of the prior. As noted by Morris (1987), the frequentist estimates of θ_i are more extreme than the Bayesian estimates. This plot also confirms the argument of Morris (1987) that (when using some reasonable assumptions about prior distributions) an increasing sample size yields a frequentist estimate of θ_i that converges to the Bayesian estimate.

The behavior of Electoral College outcomes is volatile because small changes in pre-election polling data from important swing states can have a dramatic impact on the Electoral College, particularly in a close election year such as 2004. Further, the comparison of Bayesian and frequentist approaches is more complicated than the election example of Morris (1987) because our situation involves a multistage election process with a different prior distribution for each state’s π_i . Thus, when the θ_i values from the state-level analyses are incorporated into an analysis of possible Electoral College outcomes, the volatility of the Electoral College process yields results that are more complex. Figure 2 illustrates the number of electoral votes received by Bush in 100,000 simulated elections when using the Bayesian approach described in Section 3.2. State-level probabilities are based on the weighted polls data assimilation method

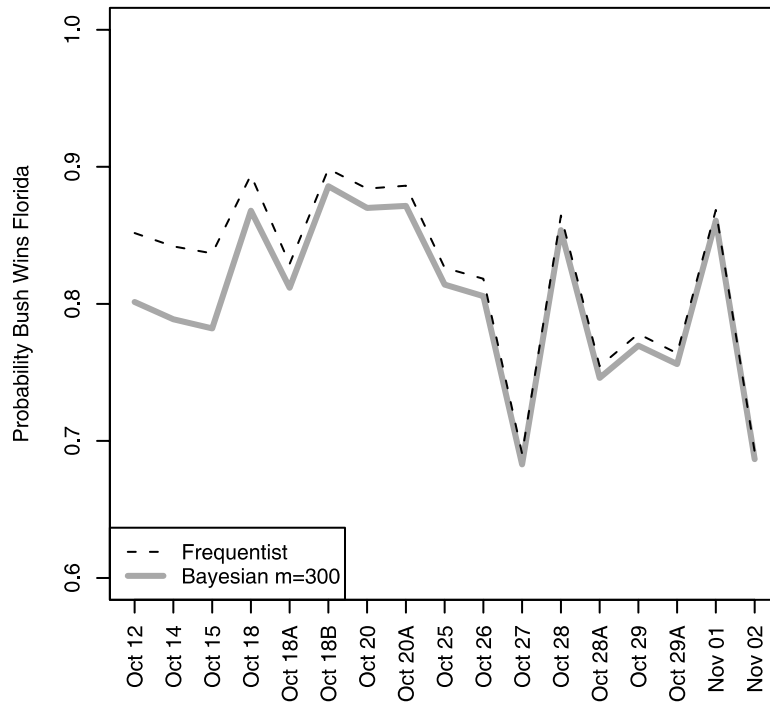


Figure 1. Daily estimates of θ_{Florida} when using the weighted polls data-assimilation method and each of the frequentist and Bayesian ($m = 300$) approaches. Dates such as “Oct 18A” exist because some states’ polls were updated more than once on some days.

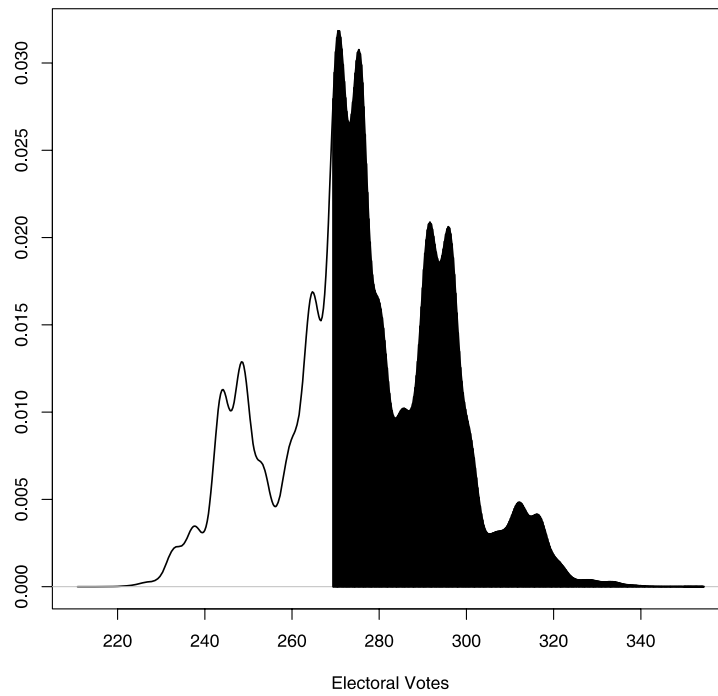


Figure 2. Approximate sampling distribution for “Bush electoral votes” based on 100,000 simulated elections when using the Bayesian approach described in Section 3.2. State-level probabilities (θ_i) are calculated using the weighted polls data-assimilation method on November 2, 2004. The shaded area to the right of 269 electoral votes represents the probability that Bush wins.

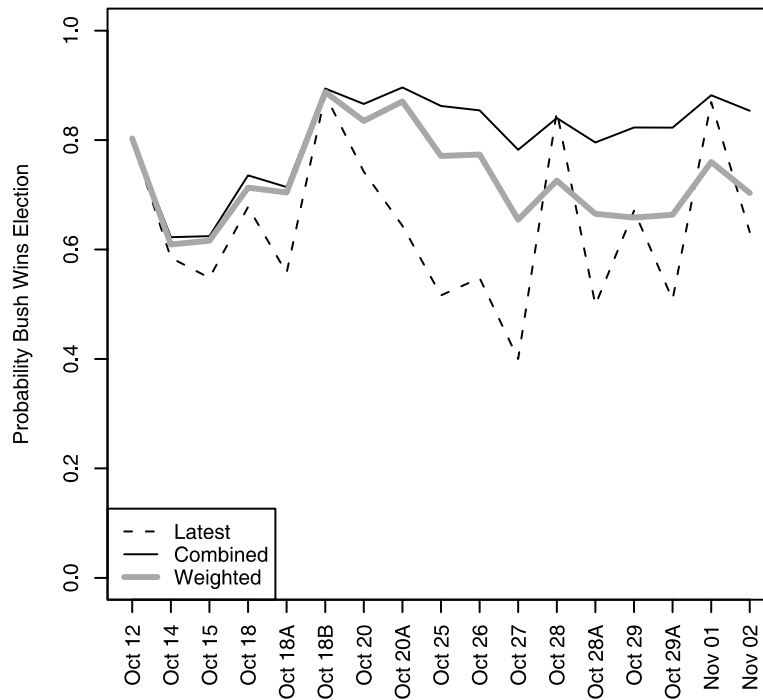


Figure 3. Bayesian estimates of θ_{EC} (probability of a Bush win) using each of the three different data assimilation methods discussed in Section 3.1: latest poll (dashed line), combined polls (solid black line), and weighted polls (solid gray line). Estimated probabilities are based on 100,000 simulated elections.

on November 2, 2004. On this day, the estimate of θ_{EC} is 70.3%. The number of electoral votes actually won by Bush is 286.

Figure 3 compares the Bayesian estimates of θ_{EC} for each of the three data-assimilation methods, calculated on each of the days where new state-level polls became available. Because of the smaller associated sample sizes, the latest poll assimilation method yields much more volatile estimates than the other two methods.

Because of the closeness of the 2000 election and the complexities of the presidential election process, there was much interest in the possibility of an Electoral College tie and the almost certain controversy that would ensue. Roughly two months before the 2004 election, Sracic and Ritchey (2004) attracted attention in the media with their analysis considering the probability of a tie ($\Pr\{\text{tie}\}$). Their approach considered all possible ways of assigning 17 designated “swing states” to Bush or Kerry (with the other 34 states considered to be a certain win for one candidate or the other). They found that 1.5% of the possible outcomes yielded an Electoral College tie. Less than a week before the election, a recalculation assuming only seven swing states yielded an estimate of $\Pr\{\text{tie}\}$ equal to 3.1% (Curl 2004). Note that their approach resembles ours except that they effectually assign each state a θ_i of either 0, 0.5, or 1.

In contrast to Sracic and Ritchey (2004), our approach provides a simple way to calculate the probability of a tie in a manner that incorporates available polling data, considers all 51 “states,” and (when using the Bayesian formulation) uses additional prior knowledge about voter behavior. Figure 4 shows the Bayesian estimate of $\Pr\{\text{tie}\}$ using the weighted polls assimilation method. During the 22-day window in which we gathered polling data, estimates of $\Pr\{\text{tie}\}$ were in the range of 1.9% to

4.7%, with a mean of 3.2%. The substantial day-to-day variability is an unavoidable consequence of the structure of the Electoral College system. For example, note that in Figure 2, the probability of a tie is related to the height of the probability mass function at 269 electoral votes. Because the probability mass function is very steep in the neighborhood of 269, a change in θ_i for even one small, closely divided state will dramatically change the probability of a tie. To compare with the analysis of Sracic and Ritchey performed one week before the election (Curl 2004), our estimates of $\Pr\{\text{tie}\}$ for October 26–28 were in the range of 3.6% to 4.7%. On the day before the election (November 2), the Bayesian estimate of $\Pr\{\text{tie}\}$ was 3.1%.

4.2 Control of the U.S. Senate 2006

To illustrate the applicability of this analysis approach to any multistage election process, we also consider the control of the U.S. Senate for the 2006 election. Although the election of each U.S. senator depends only on voters within each of the 50 states, the determination of which party controls the U.S. Senate is in fact a multistage election process. After each even-year election, the party in the majority (with independents caucusing with the major party of their choice) assumes “control” of the Senate. Because the party in control has clear advantages in promoting its agenda, the question of controlling the Senate and House is of significant interest during even-numbered years.

Prior to the 2006 election, the Senate was composed of 55 Republicans, 44 Democrats, and 1 Independent (who could effectively be grouped with the Democratic caucus). In the 2006 election, the Republicans were defending 15 Senate seats and the Democrats were defending 18, with 40 Republicans and 27

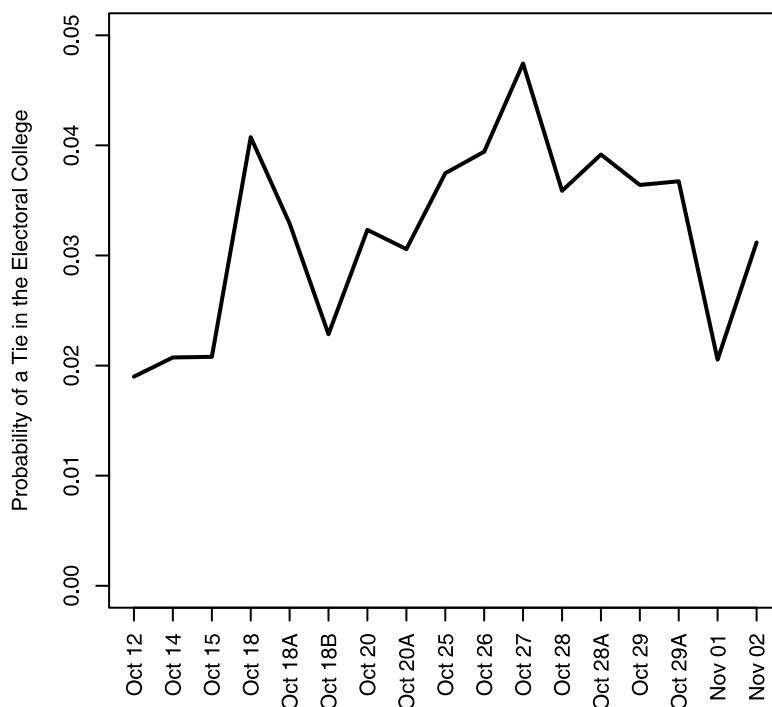


Figure 4. Bayesian estimates of the probability of a tie in the Electoral College. State-level probabilities (θ_i) are calculated using the weighted polls data-assimilation method. Estimated probabilities are based on 100,000 simulated elections.

Democrats holding seats that will not be contested until 2008 or 2010. We did not include Connecticut in the simulation because each of the viable candidates would caucus with the Democratic party. Indiana was not included because the incumbent was running virtually unopposed. Using state-level pre-election polls about Senate races, we wish to construct a daily-updated distribution for the number of seats Republicans would hold if the election were held that day. For most states, the polling data used in the analyses were gathered on August 1, 2006, or later. The exceptions were Massachusetts, Mississippi, and Wyoming where the only polls available before the last week of the campaign were conducted prior to August 1. We began analyses on August 31, after obtaining at least one poll for every state.

We use an approach similar to that described in Section 3.2 and construct beta priors for each of the 31 Senate races of interest. As described in our discussion of prior distributions for the 2004 U.S. presidential race, sophisticated approaches can be developed for constructing priors for each Senate race from a host of demographic and historical measurements. However, in the interest of providing an easily comprehensible approach for undergraduates studying statistics, we construct priors for 2006 Senate vote shares by using two different sources of historical voter behavior. First, the state-level vote shares for Bush and Kerry from the 2004 presidential election are used as measures of party popularity. Second, if one of the candidates in the 2006 election is an incumbent elected in 2000, we also incorporate the vote shares for the two parties from the 2000 Senate election. Because the conventional wisdom during the autumn of 2006 is that past voter behavior may be less useful during the current political climate of dissatisfaction with elected officials, we use the relatively small prior weights of $m_{2004} = 50$ for the

2004 presidential vote shares and $m_{2000} = 50$ for the 2000 senatorial vote shares. If neither candidate in a state's Senate race is an incumbent elected in 2000, then $m_{2000} = 0$.

For state i , let y_i be the number of the n_i polled voters favoring the Republican Senate candidate and let π_i be the actual proportion of voters favoring the Republican, where y_i and n_i are obtained using the weighted polls data-assimilation method described in Section 3.1. Because the state-level pre-election polling data were gathered over a much longer time period than in the 2004 presidential election example, we considered two different weight functions—one that gives the estimator a long memory of the past polls and one that gives a short memory. The long memory weight function used here is

$$w(t) = \begin{cases} 1 - \frac{t}{70}, & t \leq 56 \\ 0.2, & t > 56 \end{cases}, \quad (2)$$

where t is the number of days since the poll was carried out. The short memory weight function is defined similarly by

$$w(t) = \begin{cases} 1 - \frac{t}{14}, & t \leq 13 \\ 0.05, & t > 13 \end{cases}. \quad (3)$$

Following the argument in Section 3.2, the posterior distribution for the probability of a Republican win in state i is

$$f(\pi_i | y_i) \sim \text{Beta} \left(y_i + m_{2004}\pi_{i,\text{Bush}2004} + m_{2000}\pi_{i,\text{Rep}2000}, \right. \\ \left. n_i - y_i + m_{2004}\pi_{i,\text{Kerry}2004} + m_{2000}\pi_{i,\text{Dem}2000} \right).$$

A total of 100,000 elections were simulated and Figure 5 gives daily Bayesian estimates of $\text{Pr}\{\text{at least 50 Republicans}\}$

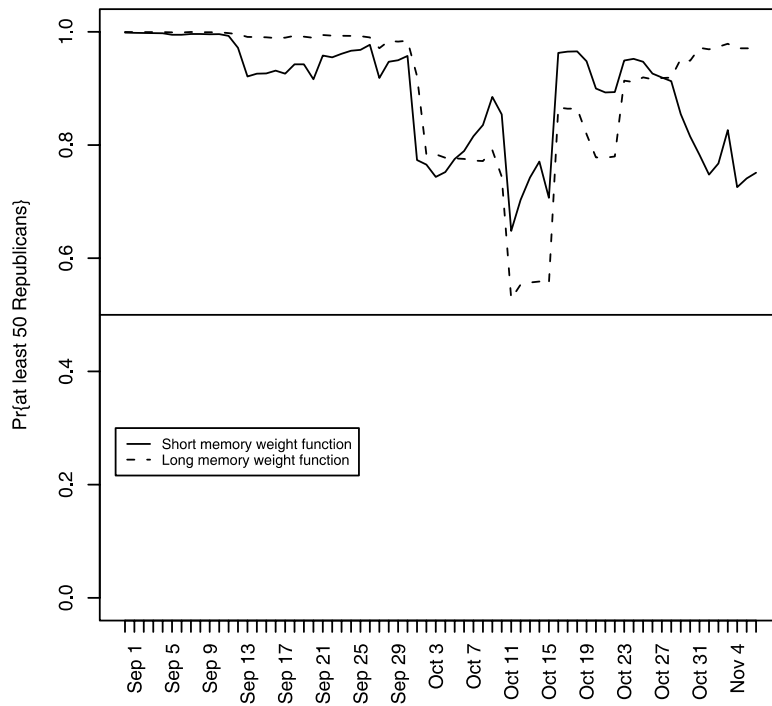


Figure 5. Daily Bayesian estimates of $\Pr\{\text{at least 50 Republicans}\}$. State-level probabilities (θ_i) are calculated using the weighted polls data-assimilation method with both the long-memory weight function (dashed line) and the short-memory weight function (solid line). Estimated probabilities are based on 100,000 simulated elections.

when using each of the weights in (2) and (3). Estimates are updated daily from August 31, 2006, to November 6, 2006. In the actual November 7, 2006 election the Republican party retained only 49 seats—one seat short of that necessary to maintain control. Notwithstanding, on November 6 the estimate for $\Pr\{\text{at least 50 Republicans}\}$ was equal to 97% for the long-memory analysis and 75% for the short-memory analysis. These high probabilities are due in great part to the Virginia Senate race where the Democratic challenger (Jim Webb) had a final week surge in the polls and beat the Republican incumbent (George Allen) by a margin of 0.4%. For both the long- and short-memory analyses, Virginia is the only state where $\Pr\{\text{Republican win}\}$ does not correctly predict the eventual winner, with $\Pr\{\text{Allen win}\}$ equal to 96% for the long-memory analysis and 65% for the short-memory analysis. The final week improvements in the polling numbers for the Democratic challenger in both Virginia and (to a lesser degree) Missouri were the primary reasons for the difference between the final estimates of $\Pr\{\text{at least 50 Republicans}\}$ in the long- and short-memory analyses. Figure 6 illustrates the estimated distribution for the total number of Republicans in the Senate as calculated using the short-memory analysis on November 6. Although both the short- and long-memory analyses indicate that the most likely number of Republicans is 50, the Democratic wins in each of the three races decided by less than 2.3% (Virginia, Montana, and Missouri) resulted in a total of only 49 Republicans in the new Senate.

5. DISCUSSION AND CONCLUSION

In this article, we have discussed a “capstone project” which challenges students to synthesize multiple sources of data and use multiple statistical concepts and methods when addressing a complex problem with no clear answer. The body of widely available pre-election poll data is a rich resource for class projects and discussion. Most college students have at least a moderate degree of interest in politics, particularly during presidential election years. Because statistics students often have some familiarity with the issues of the day and the temporally evolving nature of public opinion, they are better able to appreciate the complex relationship between data and subject matter. A comprehension of this complexity is difficult to obtain from most textbook problems or “black box” data analysis projects, yet this “big picture” perspective is increasingly important in modern settings where statisticians must work at the interface of statistical science and other disciplines. The project discussed herein is used to synthesize several important statistical concepts including: the relationship of frequentist p values and Bayesian posterior probabilities, the effects of pooling data when estimating a temporally evolving parameter, the impact of the prior distribution as the sample size increases, and the approximation of sampling distributions via simulation.

Projects based on pre-election poll data can be useful in a wide variety of statistics courses. In a nonparametric methods course, several facets of the project have been tailored to emphasize the utility of distribution-free methods. For example, questions are posed about the regional nature of voting behaviors in the United States that are most appropriately addressed using Fisher’s exact test and permutation tests. In a Bayesian

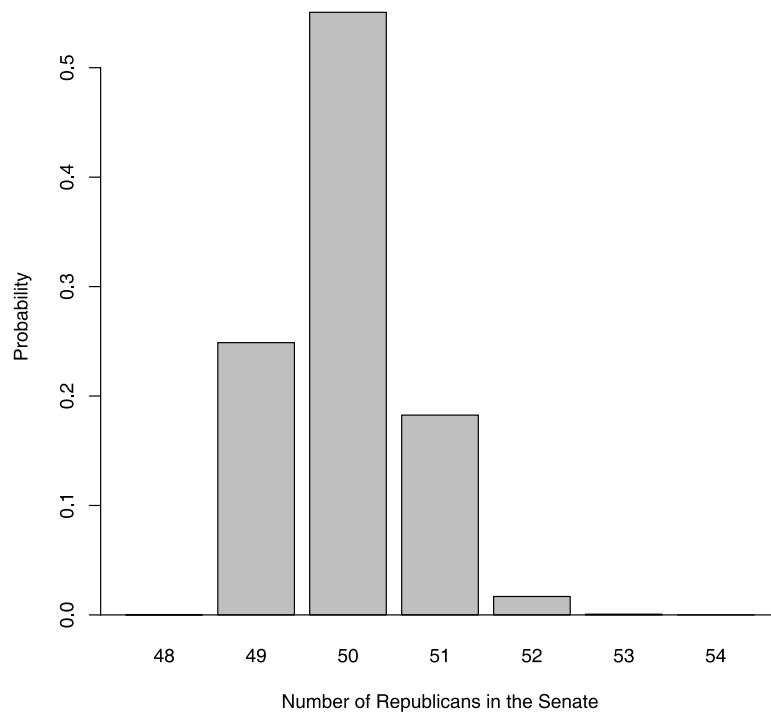


Figure 6. Approximate sampling distribution for the number of Republican senators from the 2006 election based on 100,000 simulated elections. State-level probabilities (θ_i) are calculated using the short-memory weighted polls data-assimilation method on November 6, 2006.

methods course, various models for voter behavior can be explored, and Markov chain Monte Carlo methods can be applied to nonconjugate analysis settings. Survey sampling courses can focus on data-collection methods and data validity, and time series courses can more carefully address the temporal nature of the polling data in estimation and trend forecasting. Thus, by choosing to address various facets of the data or model in a more thorough manner, this undergraduate capstone project can be adapted for M.S. level statistics courses.

Students who completed this project in a nonparametric methods course said they enjoyed applying the concepts learned in the classroom to a real-life situation of interest. Some even said that they plan to carry out the analyses in future elections just for fun. For many undergraduate students, simulation is a new concept, and some students expressed that the programming required for the computer simulations helped them solidify concepts. Although the assignment requires only limited computer programming knowledge, some of our undergraduate students (including nonmajors) have had little programming experience in statistical packages such as R. Consequently, during the course we provide some basic instruction on R and we place students in groups of two for this project, ensuring that less-experienced programmers work with someone more experienced. An additional challenge in implementing the project relates to the students' familiarity with the Bayesian paradigm. Whether the students use the frequentist approximation of a Bayesian posterior probability or the fully Bayesian approach, many students will need a brief introduction to the Bayesian paradigm before proceeding to the analyses. We have done this during the lecture in which the project is introduced and discussed.

Finally, on the topic of predicting multistage election out-

comes such as the U.S. presidential election, we conclude that the traditional daily tracking of nationwide popular support for candidates or parties can easily be replaced by a more instructive analysis such as that presented in Section 4. The popular support for candidates and parties (on a "generic ballot") is only of peripheral interest to most consumers of political news. Our approach can be used to answer questions such as "What is the probability that the Democratic presidential candidate will win the presidency?" or "What is the probability of an Electoral College tie?" or "What is the probability that the Republicans retain control of the Senate?" The analysis of such questions has traditionally been the sole domain of political analysts. However, with the wide availability of polling data, such questions can also be addressed quantitatively. Using simulation-based approaches for addressing the multistage nature of presidential elections and control-of-congress processes can be valuable and instructive for students of statistics and political science, and can be beneficial to the media in providing consumers with pertinent coverage of the election process.

[Received October 2006. Revised August 2007.]

REFERENCES

- Berger, J. O., and Sellke, T. (1987), "Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence," *Journal of the American Statistical Association*, 82, 112–122.
- Brown, L. B., and Chappell, H. W. Jr. (1999), "Forecasting Presidential Elections using History and Polls," *International Journal of Forecasting*, 15, 127–135.
- Campbell, J. E. (1992), "Forecasting the Presidential Vote in the States," *American Journal of Political Science*, 36, 386–407.
- (1996), "Polls and Votes: The Trial Heat Presidential Election Forecasting Model, Certainty, and Political Campaigns," *American Politics Quar-*

- terly, 24, 408–433.
- Campbell, J. E., and Wink, K. A. (1990), “Trial-Heat Forecasts of the Presidential Vote,” *American Politics Quarterly*, 18, 251–269.
- Casella, G., and Berger, R. L. (1987), “Reconciling Bayesian and Frequentist Evidence in the One-Sided Testing Problem,” *Journal of the American Statistical Association*, 82, 106–111.
- Cohen, J. E. (1998), “State-Level Public Opinion Polls as Predictors of Presidential Election Results: The 1996 Race,” *American Politics Quarterly*, 26, 139–159.
- Curl, J. (2004), “Electoral College Fiasco Looks More Likely,” *The (DC) Washington Times*, Oct. 28.
- Holbrook, T. M., and DeSart, J. A. (1999), “Using State Polls to Forecast Presidential Election Outcomes in the American States,” *International Journal of Forecasting*, 15, 137–142.
- Morris, C. N. (1987), “Testing a Point Null Hypothesis: The Irreconcilability of P Values and Evidence: Comment,” *Journal of the American Statistical Association*, 82, 131–133.
- Park, D. K., Gelman, A., and Bafumi, J. (2004), “Bayesian Multilevel Estimation with Poststratification: State-Level Estimates from National Polls,” *Political Analysis*, 12, 375–385.
- Sricac, P., and Ritchey, N. P. (2004), “All Tied Up in Presidential What-Ifs,” *The Washington Post*, August 22, 2004.