Forecasting the 2024 US General Election*

Analyzing the influence state on election outcome by using 538 Pollster Data with Bayesian hierarchical model, Predicting Harris Victory[need more details in result]

Tara Chakkithara Sakura Hu Luka Tosic

November 4, 2024

This paper forecasts the outcome of the 2024 U.S. general election. Using the poll of polls approach, a generalized linear model was developed with data from FiveThirtyEight [more spcific on method, and what is the date cutoff of the data] The model's prediction is based on the popular vote in the seven key swing states—Arizona, Nevada, Michigan, Georgia, Wisconsin, Pennsylvania, and North Carolina to estimate the overall winner. Results indicate that Harris is predicted to win, by [how much? need to be more spcific on the result]. This suggests that current polling trends may provide reliable insights into the likely election result, though future shifts in voter sentiment could impact the final outcome.

1 Introduction

The 2024 United States presidential election, the 60th quadrennial election, is scheduled for November 5, 2024. The stakes are high, as the outcome will shape domestic policies and impact international relations, with effects on trade, climate change, and security. In recent years, the domestic political landscape has changed significantly, marked by rising homelessness, income inequality, and social unrest. Alt-right movements increasingly target women's rights, minority rights, and immigration policies, leading to deep societal divisions. The current Democratic administration faces criticism for not addressing these pressing issues, with widespread protests against U.S. foreign involvement and the erosion of reproductive rights. Protests around police brutality and anti-immigration sentiments have also surfaced, reflecting the polarized nature of contemporary American society.

*Code and data are available at https://github.com/LukaTosic09/Forecasting-2024-Election

The primary candidates in this election, Democratic incumbent Kamala Harris and former Republican president Donald Trump, are expected to engage in a tight race, with polling indicating that the results could be very close. Swing states—those that do not consistently vote for one party—are vital in determining the overall outcome, as their electoral votes can sway the final result. Seven important swing states—Arizona, Nevada, Michigan, Georgia, Wisconsin, Pennsylvania, and North Carolina—are expected to be battlegrounds in this election, making their polling data essential for accurate forecasts.

This paper focuses on forecasting the election outcome by employing the "poll of polls" method, utilizing data from FiveThirtyEight, along with historical data from the 2016 Clinton vs. Trump election and the 2020 Trump vs. Biden election. A Bayesian updating and Monte Carlo Simulations is created to predict the winner of the election outcome based on the popular vote across these swing states[some more details for what was done?]. The primary estimand in this analysis is the probability distribution of Harris's win percentage, derived from current polling data and historical election results. The model predicted that Harris will win, and [add details to result]

Accurate forecasting of election outcomes is vital for several reasons. It enhances democratic engagement by informing voters about potential results, which encourages active participation in the electoral process. Furthermore, political campaigns can leverage these forecasts to tailor their strategies, optimize resource allocation, and concentrate efforts on pivotal swing states that may determine the election's outcome. Analyzing polling data also reveals trends in voter sentiment, providing valuable insights into the electorate's priorities and concerns that can inform policy discussions. Ultimately, the findings from this study can improve the accuracy of predictions in a political landscape characterized by rapid change and polarization, thereby enhancing the relevance and reliability of electoral analyses.

The remainder of this paper is structured as follows: ……

2 Data

2.1 Data Acquisition

Data for this paper is sourced from the Presidential General Election polls dataset, FiveThirtyEight (2024). This dataset includes approximately 16,000 observations of polls conducted from January 2023 to October 16, 2024. Each poll in the dataset has an associated pollscore, sample size, start date and end data, pct (percentage of votes), candidate name, and state. There are other features in the dataset such as time of publication, candidate sponsorship, etc. For this analysis, we will focus on the key veriables mentioned above.

2.2 Dataset Features

2.2.1 Pollscore

The pollscore feature serves as a metric for evaluating the quality of a poll. It is a mesaurement that tries to quantify the realiablity of a poll. It is calculated based on factors such as poll transparency, bias, and error, ABC News (2024). A more negative pollscore indicates higher quality, reflecting a greater confidence in the poll's accuracy. As seen in Figure 1, most polls collected by FiveThirtyEight have a pollscore under 0. In Figure 2, we see that lower poll scores are correlated with high transparency.

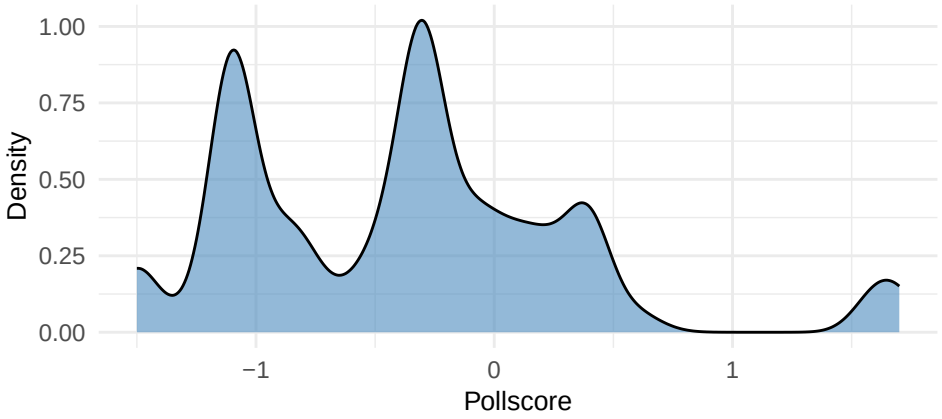


Figure 1: Distribution of Pollscores in the 2024 General US Election

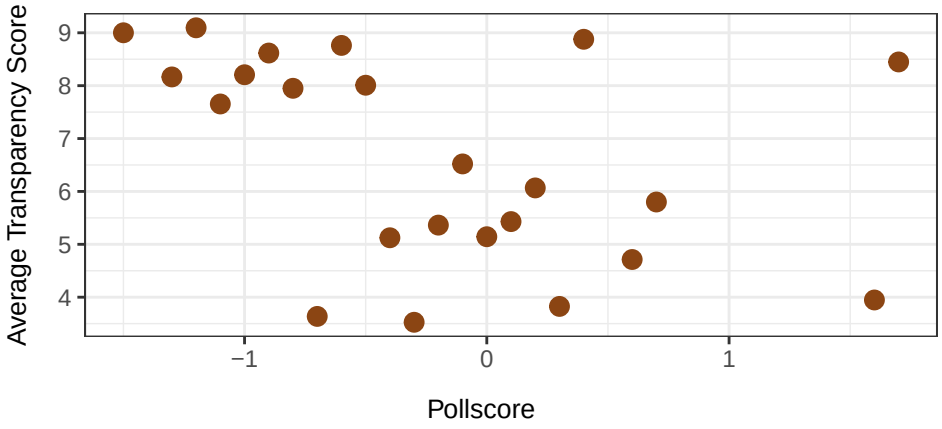
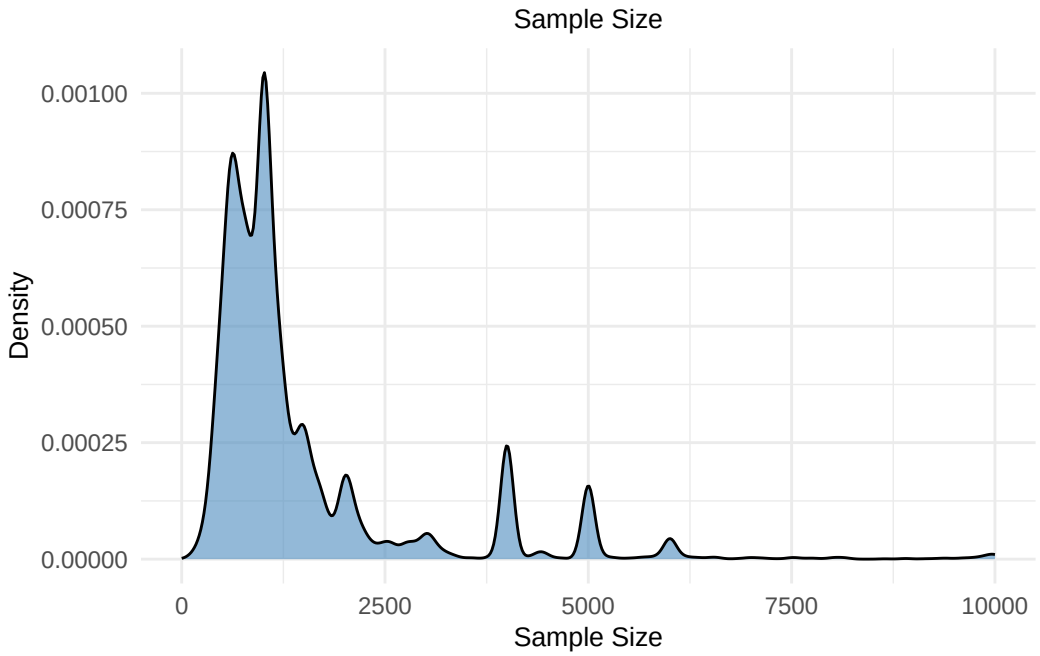
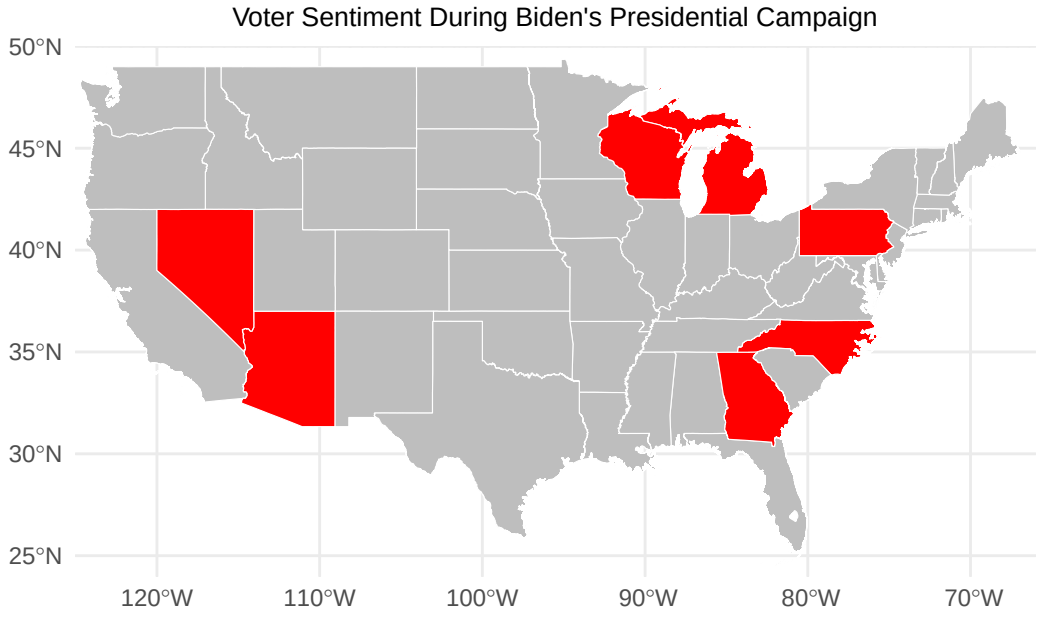
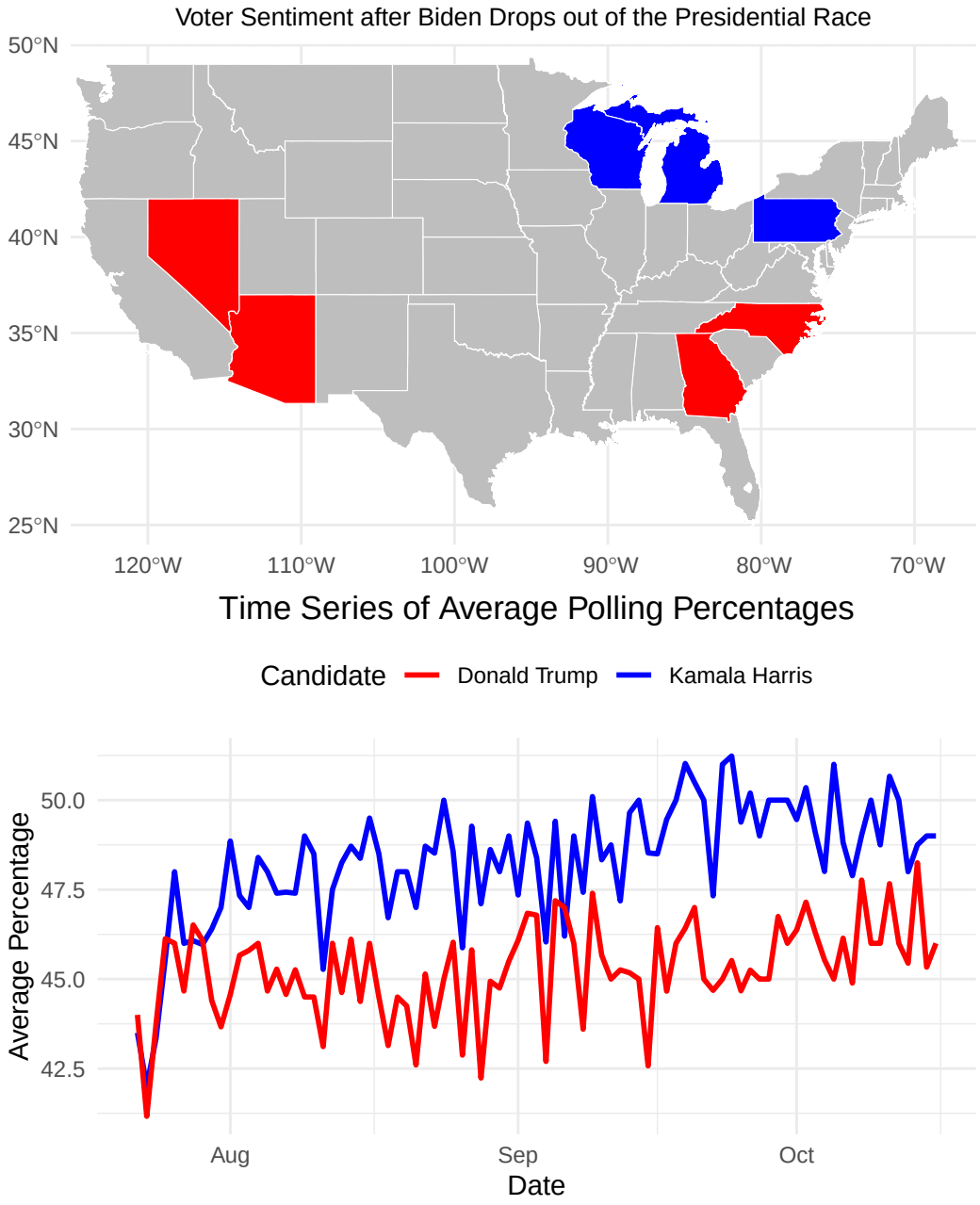


Figure 2: Average Transparency Scores Across Varying Pollscores

2.2.2 Sample Size







3 Data Overview

Our data is the Project 538 [cite] Presidential General Election polls dataset (current cycle) from October 17th. The data contains roughly 16000 observations of polls taken from January 2023 at the earliest to October 16th at the latest. Each poll has been given an associated

pollscore, numeric grade and transparency score, and the methodology used to conduct it is also given. We also know the duration of the poll by start and end date, as well as the time of publication and the number of particpants. The candidate sponsoring each poll is given as well as the candidate endorsed by each poll.

(can be deleted) Each poll has an associated pollster, and each pollster is assigned a numeric grade (0.5 to 3.0) and pollscore (-1 to 1) corresponding to the quality and reliability of the pollster and a transparency score (0 to 10) corresponding to the level of transparency the pollster provides in regards to their methodology. Furthermore, each poll has an associated methodology. [graphics to include?, pollscore vs numeric_grade and numeric_grade vs transparency score can show strong linear relationship, providing justification to use one]

Important to our modeling are the variables answer and pct, where answer indicates what candidate the poll predicts will receive the associated percentage (pct) of votes. Thus each poll is captured at least twice in the dataset, once for each polled candidate. The associated unique poll_id is a variable that then captures multiple observations of the same poll. Another important variable is Hypothetical, which notes whether a poll is between candidates that ran against each other in reality, or in a hypothetical scenario. Another important variable for our modeling is state, which indicates which state, if any, the poll was focused. [pct by state for both candidates can show that swing states are less clear]

4 Data Cleaning

To clean this data, we first filtered out all hypothetical polls. Then, we filtered out any polls not focused on swing states. Next, we looked only at polls answering a percentage of Trump as he appeared in the last two presidential races, and we removed all data from from before Biden dropped out to get results for Trump vs Harris.

Next, we wanted to focus only on high quality polls, so we filtered out all polls with a pollscore greater than zero.

Talk more about it.

And also planes (**?@fig-planes**). (You can change the height and width, but don't worry about doing that until you have finished every other aspect of the paper - Quarto will try to make it look nice and the defaults usually work well once you have enough text.)

Talk way more about it.

5 Model

5.1 Model Justification

The unique structure of the electoral college voting system in the United States makes it unlikely for any candidate other than Donald Trump or Kamala Harris to be competitive in the 2024 presidential election. Given Trump's historical performance in previous campaigns, particularly the 2016 election, it is more insightful to model his election outcomes. To achieve this, we have decided to create a Bayesian Hierarchical Model.

Our goal is to model the probability distribution of the proportion of electoral college seats that Donald Trump may secure, utilizing R Core Team (2023). The 2024 election closely resembles the 2016 election, which Donald Trump won, and he is facing a female opponent again. To capture our initial beliefs about the election being competitive, with Trump holding a slight advantage, we have selected a weak prior of Beta(4,3). This prior distribution is characterized by a mode that aligns closely with Trump's previous percentage of electoral seats allocated in 2016. The Beta distribution is particularly effective in Bayesian modeling due to its ease of updating with new data.

To refine our prior and incorporate more recent data, we will focus on the election outcomes from key swing states. The swing states for the 2024 general election are: Nevada, Arizona, Wisconsin, Michigan, Pennsylvania, North Carolina, and Georgia. Model validation and additional details are found in Appendix I.

5.2 Model Set Up

Let X be the random variable: proportion of electoral college seats Donald Trump will win.

$$X \sim \text{Beta}(4, 3)$$

For each swing state j, let $Y_{j,1}, Y_{j,2}, Y_{j,3}, ..., Y_{j,n_j}$ be independent and identically distributed (i.i.d) random variables, where $1 \leq j \leq 7$ and $Y_{j,i}$ represents the percentage of votes for Donald Trump in state j. F_j is some unknown probability distribution of voting percentages of state j.

$$Y_{j,i} \sim F_j(x) \quad \text{for } j = 1, 2, \dots, 7 \text{ and } i = 1, 2, \dots, n_j$$

We can use Kernel Density Estimation to estimate F_j.

$$\widehat{F}_j(x) = \frac{1}{n} \sum_{i=1}^{n_j} K_h(x - Y_{j,i})$$

Where K_h is the gaussian kernel with bandwidth $h = n_j - \frac{1}{5} n_j^{-\frac{1}{5}} \frac{n_j - 1}{5}$. Let $\hat{Z}_1, \hat{Z}_2, ..., \hat{Z}_7$ be simulated values drawn from $\hat{F}_1, \hat{F}_2, ..., \hat{F}_7$ respectively. Let M_j be the election outcome of the state (win or lose). M_j can be modeled using a binomial distribution where n_j is the number of polls.

$$M_j \sim \text{Binomial}(n_j, \hat{Z}_j)$$

We can now simulate an election and create an estimator for X, the random variable representing the proportion of electoral seats Donald Trump will hold after the election. We start by allocating Trump all the seats of prodominatly red states which are 219 seats. We then add additional seats for each swing state before diving by the total number of seats to estimate the proportion of seats Trump will win. Here, s_j is the number of electoral college seats in state j.

$$\hat{X} = \frac{(s_1)\hat{M}_1 + (s_2)\hat{M}_2 + ... + (s_7)\hat{M}_7 + 219}{538}$$

Using monte carlo simulation we can generate 10,000 estimates $\hat{X}$. We apply a learning rate of 0.01 to avoid overfitting and update our prior.

$$X \sim Beta(4 + W, 3 + L)$$

Where W is the number of estimates greater than 0.5 multiplied by the learning rate of 0.01, and L is the number of estimates less than or equal to 0.5 multiplied by the same learning rate of 0.01.

5.3 Model Assumptions & Limitations

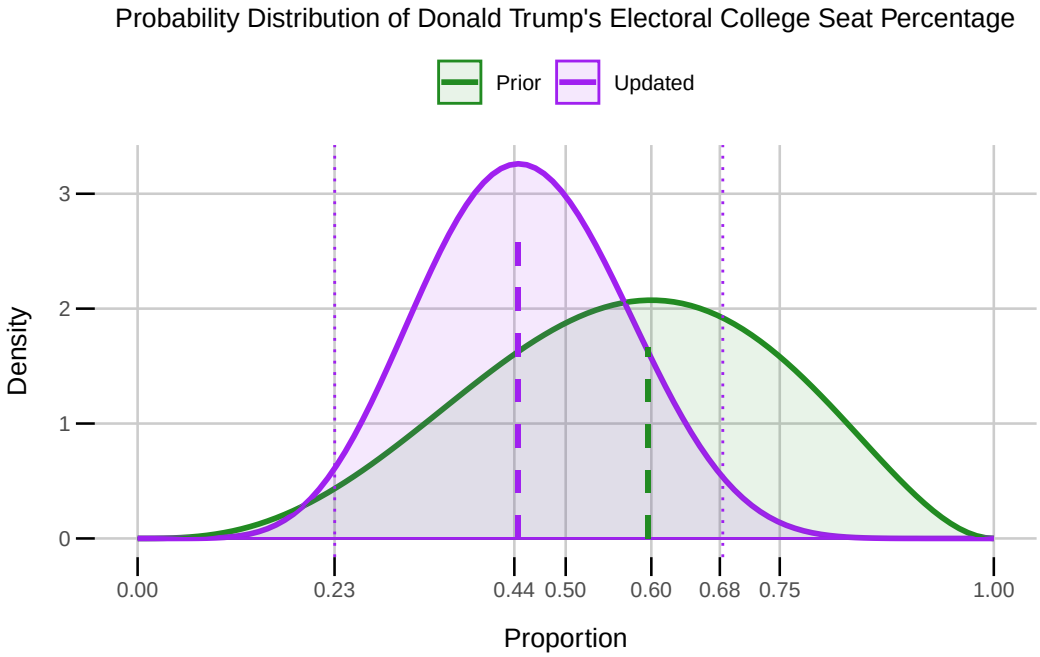
Our model relies on polling data collected after Joe Biden, the former primary candidate for the Democratic Party, withdrew from the race. While this data is recent, it does not account for any shifts in voter sentiment that may have occurred between October 19, 2024, and November 5, 2024.

In today's highly polarized political climate, there is an increased likelihood that voters may not respond honestly in polls. This lack of transparency can lead to significant data distortions, ultimately compromising the accuracy of predictions.

Additionally, the model's simplicity poses a limitation, as it does not consider various critical factors that can influence electoral outcomes. Economic conditions, social dynamics, and demographic shifts are all significant elements that can impact voter behavior but are not incorporated into the model.

This model designed specifically for the context of United States elections. It is built on assumptions inherent to the electoral college system, such as the presence of two dominant candidates and a winner-takes-all allocation of electoral seats. As a result, the insights derived from this model may not be easily generalizable to elections in different countries or political contexts.

6 Results



The updated probability distribution for Donald Trump's electoral college seat percentage is modeled as Beta(7.67,9.33). This distribution has a mode of approximately 0.44, indicating that the most likely outcome is that Trump would secure around 44% of the electoral college seats. This suggests that Kamala Harris is favored to win the overall election.

However, it's crucial to emphasize that we cannot confidently assert that Kamala Harris will win with 95% certainty. The shape of the probability distribution is similar to that of a coin flip, reflecting significant uncertainty in electoral outcomes. While Harris may have a higher likelihood of winning, there remains a considerable chance for Trump to secure a majority of the seats.

To quantify Trump's chances of victory, we evaluated the cumulative distribution function (CDF) of the Beta distribution at the 50% threshold, which corresponds to the majority of electoral college seats. By subtracting this CDF value from 1, we calculated the probability of Trump winning the election. The result of this analysis indicates that Trump has a 35% probability of securing victory.

7 Discussion

7.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

7.2 Second discussion point

7.3 Third discussion point

7.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix 1

The pollster that is chosen to be analyzed is The New York Times/Siena College. The population targeted by this poll consists of the entire U.S. electorate. The frame for the poll is derived from the L2 voter file, which is a nonpartisan database, and further supplemented by matched cellular telephone numbers from Marketing Systems Group (The New York Times, 2024). From this frame, respondents are randomly selected from a national list of registered voters to participated in telephone interviews in English and Spanish (The New York Times, 2024). To enhance its regional and demographic representativeness, the national poll includes separate samples for Florida (622 voters) and Texas (617 voters) and applies oversampling to ensure adequate responses from key demographic groups, specifically Black voters and Hispanic voters (The New York Times, 2024). The poll's oversample includes 589 Black voters and 902 Hispanic voters, with subcategories for individuals identifying as a single or multiple races or ethnicities (The New York Times, 2024). By adjusting the weights of these subgroups in the final analysis, the poll aims to produce results reflective of the national electorate (The New York Times, 2024).

The sample for this poll was recruited through a stratified sampling method, which considered factors such as state, party affiliation, race, gender, and age (The New York Times, 2024). Each stratum was adjusted to address differential coverage and expected response rates, with the initial selection weights reflecting the reciprocal of a stratum's telephone coverage and modeled response rate based on historical nonresponse data from prior polls (The New York Times, 2024). The poll's stratification process included additional adjustments by state to accommodate varying levels of telephone number productivity (The New York Times, 2024). The sampling approach allows for representation of diverse demographics and achieves a more balanced sample, but it introduces trade-offs. While oversampling minority groups improves representation and allows for deeper analysis, the increased sampling error for these smaller subgroups complicates the interpretation of subgroup findings relative to the broader electorate (The New York Times, 2024).

Nonresponse is managed through stratified sampling adjustments, where mean response rates for each stratum help inform weighting to minimize potential bias from nonresponse (The New York Times, 2024). Additionally, the survey package in R is used to apply a series of weights, which adjusts the data based on various demographic characteristics and compensates for nonresponse at both the sample and question level (The New York Times, 2024). For instance, demographic adjustments are made based on age, party affiliation, race, and educational attainment (The New York Times, 2024). Moreover, among respondents with multiple contact numbers, the poll selects the number with the highest predicted response rate, further mitigating nonresponse bias and increasing the likelihood of respondent participation (The New York Times, 2024).

The questionnaire used in The New York Times/Siena College poll has several strengths aimed at enhancing accuracy, representativeness, and reliability. One advantage lies in the survey's duration, with calls limited to less than 15 minutes to maintain respondent engagement (The

New York Times, 2024). This approach helps reduce the likelihood of incomplete responses due to participant fatigue, thereby supporting data integrity. By incorporating a stratified sampling approach that oversamples in battleground states, the questionnaire captures opinions from critical regions likely to influence the presidential election outcome (The New York Times, 2024). This focus aligns with the Electoral College system, making the data more actionable for understanding the dynamics within these decisive states. Moreover, the reliance on voter registration files, which contain detailed demographic and geographic data for 200 million Americans, supports robust sample balancing and demographic matching. The voter file's information is used to ensure proportional representation across party lines, regions, race, age, and other critical attributes. This allows the questionnaire to better represent national diversity and enhances accuracy through weighting adjustments for underrepresented demographics, such as those without a college degree (The New York Times, 2024). The use of telephone interviews, despite declining response rates, remains one of the few methods to quickly access a randomized sample of voters. Alternative methods like email or mail recruitment introduce their own issues, such as attracting only highly interested individuals, potentially biasing the sample toward the politically active rather than the average voter (The New York Times, 2024). Furthermore, the poll's careful management of sampling error, with a +- 2.4-point margin for the national sample and +- 5 points for state polls, ensures a strong level of accuracy while acknowledging the inherent uncertainty in survey results.

Despite these advantages, the questionnaire also faces several limitations. As response rates have decreased in recent years, the representativeness of telephone surveys has likely suffered, as individuals are more reluctant to answer phone calls from unfamiliar numbers. This trend has increased the costs of conducting such surveys, as more calls are needed to reach a proportional sample, and some demographic groups may remain underrepresented despite weighting adjustments (The New York Times, 2024). The reliance on telephone interviews also risks excluding younger voters and others who rely primarily on digital communication, potentially skewing the data if not properly addressed through demographic balancing and weighting. Additionally, while the questionnaire's margin of error provides a statistical measure of confidence, this margin can be significant, particularly in close races where even a two-point margin could change interpretations and outcomes. This issue highlights the limitations of polling as a "blunt instrument" where minor fluctuations in results can feel disproportionately impactful in a close election context (The New York Times, 2024).

A Appendix 2 {Idealized Survey}

We are provided a budget of $100'000 to forecast the 2024 presidential election. First we must define our key terms. We are looking to forecast the winner of the US general election, so our key parameter of interest for survey participants is the candidate they would support and/or vote for. There are other secondary parameters to discuss later. Our target population is registered voters in the US as they would be the ones informing our parameter of interest, with our sampling frame being registered voters that we can reach with our survey. The sample

then is the registered voters who end up taking the survey. We must also define a sample size that is both realistically achievable, and adequately large to make a meaningful prediction of the election winner. A sample size in the range of 2500-6000 is within the realm of possibility and would be sufficient.

B 2.1 Sampling

The first task is to develop a sample from a sampling frame through some sample method(s). We would make use of our large budget in order to go about a probabilistic sampling strategy, specifically a stratified sampling approach with additional simple random sampling within the given strata. We would stratify along state, age, education, sex, income, and self-identified political affiliation. The motivation behind these strata is that we know these factors create bias; note that democrats are more likely to answer surveys [], as are those with higher educational attainment []. Additionally, certain battleground states as they are referred to like Georgia or Pennsylvania are important to gather data on as they are considered the most important states in deciding the election []; another reason is that the US population is not split evenly among states, and even in a large poll we would expect only a small, and less reliable amount of responses from certain states with smaller populations.

The specifics of the stratification along the various strata will depend on facts about distribution of population, educational attainment, age, sex, and more among registered voters. Once such statistics are collected likely through census data, we can specify the quantities we hope to see in our sample. ie if we have a sample of size n, and a population of registered voters in a given state of x, we hope to see in our data (x*n)/y where y is the number of registered voters in th US.

C 2.2 Survey Method

The survey method we will employ is an online panel of questions in the form of a Google form. These forms are secure and anonymous, and provide a simple way to store the results we receive. Additionally, the format is low cost to produce and takes no additional labour to maintain unless issues arise, in contrast with a live phone that must be either directly manned or monitored. Another advantage is the transparency provided, users can read and reread instructions, and return to questions they are unsure of easily. Keeping the form short, up to 15 minutes can help stave off disengagement from possible participants.

The maintaining of the survey would take some paid labour. The survey would run for 1 to 2 weeks over the course of 3 months prior to the election as a way to collect data that captures shifting trends in public opinion. Doing this would require us to note when the survey was submitted.

D 2.3 Recruitment

Distributing the survey will require multiple things. We must choose how we will distribute the survey. To keep people who are not registered voters from answering, there must be a way for participants to verify that they are registered voters. We propose they register with an email to receive the survey, and they would only receive the survey if their information matches a registered voter in the US, this system would have to be automated and not save sensitive information in order to maintain anonymity. Of course, we'd need to somehow access information on registered voters in the US. Such a system would of course take a portion of our budget to implement

In order to incentivize participation, we could allocate part of the budget to providing an incentive of monetary value for participants. Something like a small giftcard or voucher that is guaranteed upon survey submission.

To reach potential participants, a campaign to advertise the survey would be launched. We could pay businesses to send a promotional email to users of their service, particularly of services used by demographics we are trying to account for in our strata. A powerful tool for advertising this survey would be social media, especially since in the modern day, social media are used by a very wide range of people from the most partisan Republicans and Democrats to the median individual. We can advertise physically as well, more so in the states that we are hoping to account for per our stratification.

E 2.4 Data Validation

Once the data is collected, it would have to be parsed and cleaned. We would want to know how many questions were unanswered, and to produce basic summary statistics on things like the demographics and state of participants in order to gain an initial understanding of our data. Doing this while the survey is running can tell us if we need to push it further in certain parts of the nation or to certain groups with the hope of amassing more responses in those categories.

As survey cycles continue, we would clean the data to remove non-sensical or uninterpretable responses as well as automated or repeated responses.

The next part of validation is applying stratification weights to ensure our data conforms to the stratification criteria we outlined above. In other words, apply post-stratification weights to align survey responses with Census-based demographics, using our previously outlined strata as variables. Use iterative proportional fitting (raking) to adjust for sample imbalances.

F 2.5 Poll Aggregation and Forecasting

We would aggregate weekly poll results, adjusting for recent responses with higher weights []. Next we would employ Bayesian updating as in our paper to forecast trends based on historical election and polling data, which accounts for typical shifts and volatility in key regions.

We would also have to account for a margin of error and integrate relevant data (e.g., voter registration numbers, early voting patterns) into final our models to refine predictions, particularly in the aforementioned battleground states.

G 2.6 Survey Example

A complete example survey can be found here: https://forms.gle/tkFYQruLWuwCMLAm9

The questions provided are: Are you currently a registered voter? What is the current state you are registered to vote in? How would you describe your political leanings? Indicate your age in years Indicate your sex What is your highest level of educational attainment? What was your 2023 taxable income amount? (can prefer not to say) If the general election were held tomorrow, what party would you vote for? How Likely are you to vote in this election? If you voted in the 2020 general election, who did you vote for? How well do you feel your most preferred candidate's policies address your concerns and most important issues as a voter What are your most top 3 most important issues when deciding who to vote for? How confident are you in the outcome of this 2024 election?

H Additional Data Details

I Model Details

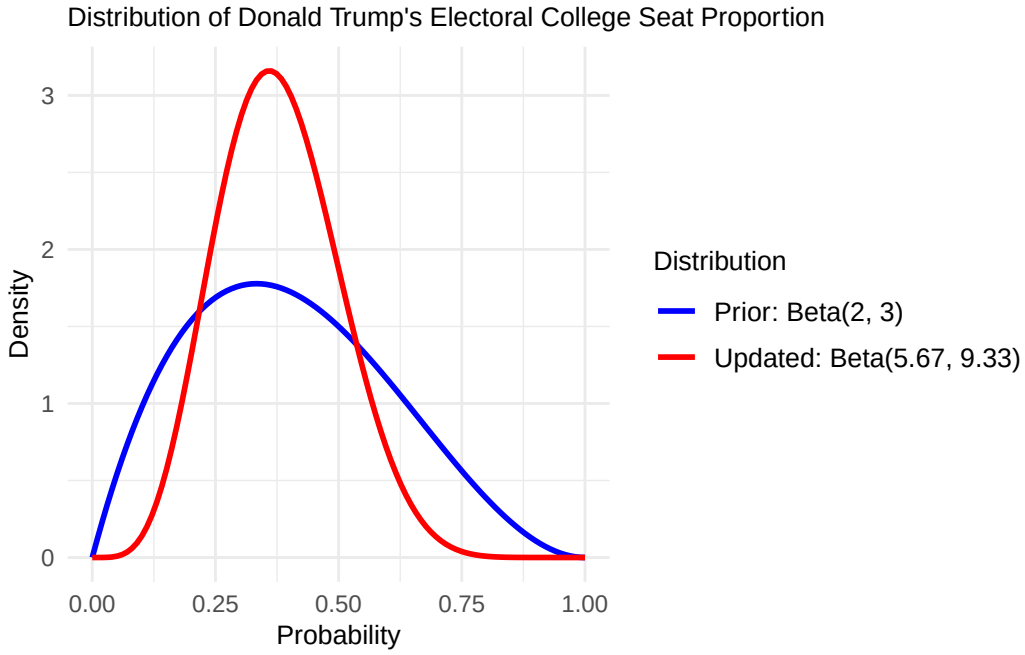
I.1 Alternate Models

In our analysis, we considered two additional models: a similar Bayesian hierarchical model with a different prior and a logistic regression model.

We opted not to select logistic regression for several key reasons. First, generalized linear models (GLMs) provide point estimates or binary outcomes, whereas our objective was to capture a full probability distribution of the election outcomes, in addition to identifying a winner. Second, we observed that the percentage of votes (PCT) did not exhibit a linear relationship with any of its potential predictors. While this issue could potentially be addressed using splines or polynomial regression, we believed it would be better to use another modeling

approach since we wanted to incorporate prior data as well. Due to this, we decided to adopt a Bayesian framework.

The first Bayesian hierarchical model we considered employed a prior of Beta(2, 3) to model the proportion of seats that Donald Trump would win. This model's mode closely reflects his his 2020 election results. However, we decided against this prior because we felt that the 2016 election more accurately represented the current circumstances. Notably, even with this model, we arrived at similar conclusions: Kamala Harris was still projected to win.



I.2 Posterior Predictive Check

References

ABC News. 2024. "How 538's Pollster Ratings Work." https://abcnews.go.com/538/538s-pollster-ratings-work/story?id=105398138.

FiveThirtyEight. 2024. "2024 National Presidential Polls." https://projects.fivethirtyeight.com/polls/president-general/2024/national/.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.