# Predicting a Harris Victory: A Bayesian Analysis of Polling Data and Voter Sentiment in the 2024 U.S. Presidential Election*

**Utilizing state-level polling data and probabilistic modeling to uncover regional support patterns and forecast electoral outcomes in a competitive race**

Chendong Fei Xinze Wu Claire Ma

November 4, 2024

This paper develops a model to forecast the outcome of the 2024 U.S. presidential election by analyzing aggregated polling data, or "poll-of-polls," sourced from FiveThirtyEight. Using a generalized linear model, we assess national trends alongside key battleground state polls to predict each candidate's likelihood of victory. The findings indicate a closely contested race, with specific demographic and regional factors creating narrow pathways to winning the presidency. This analysis highlights the value of aggregated polling data in understanding electoral dynamics and demonstrates the importance of statistical modeling in making informed predictions about major political events.

## 1 Introduction

The outcome of the U.S. presidential election has far-reaching implications, shaping both domestic policies and international relations. As the 2024 election approaches, voters and analysts turn to polls to understand the state of the race between Vice President Kamala Harris, the Democratic candidate, and former President Donald Trump, the Republican candidate. However, individual polls are often limited by their methodologies, timing, and sample demographics, leading to variations in predictions. To overcome these limitations, aggregating multiple polls—a technique known as "poll-of-polls"—provides a more stable and reliable indicator of public opinion. This paper applies a poll-of-polls approach, informed by methodologies from individual polls (Blumenthal 2014; Pasek 2015), to predict the outcome of the 2024 U.S. presidential election, focusing on data aggregated by FiveThirtyEight (**fivethirtyeight2024?**).

*Code and data are available at: https://github.com/ke3w/Prediction_US_presidential_election.git

The primary objective of this analysis is to forecast which candidate is likely to win the 2024 election based on aggregated national and battleground state polling data. By constructing a Bayesian generalized linear mixed model (GLMM), we aim to distill insights from the extensive polling data available, examining trends and key demographic indicators while accounting for inherent variability between pollsters and states.

The primary estimand in this analysis is the probability of each candidate winning the 2024 U.S. presidential election based on aggregated polling data. This probability is derived from a weighted average of poll results across national and battleground states, with adjustments for factors such as recent polling trends, sample sizes, pollster reliability, and state-specific electoral significance.

Our analysis reveals a highly competitive race, with key battleground states playing a pivotal role in determining the overall outcome. The model identifies specific regions and demographics that are likely to influence the election results, highlighting the polarized nature of the electorate. As of November 1, 2024, FiveThirtyEight's national polling average indicates a slight edge for Harris, who has 48.1% support compared to Trump's 46.7%. Despite this narrow national lead, the race in critical battleground states remains highly competitive. For instance, Pennsylvania is evenly split, with Harris holding marginal leads in states like Wisconsin and Michigan, while Trump shows slight advantages in Nevada, Georgia, and Arizona. These tight margins highlight the crucial role battleground states play in determining the election's outcome.

These findings underscore the importance of aggregated poll data in capturing the broader political landscape, offering insights that single polls may miss. By understanding the dynamics at play, this study contributes to a broader understanding of electoral processes and the predictive power of statistical models in forecasting complex political events.

This paper is organized as follows: Section 2 discusses the details of the dataset. **?@sec-Methodology** describes the Bayesian generalized linear mixed model used in the analysis. **?@sec-Results** presents the results, highlighting trends in polling, and **?@sec-Discussion** considers the implications of these results for future research on polling and public opinion.

# 2 Data

## 2.1 Overview

We use the statistical programming language R (R Core Team 2023) to analyze polling data from FiveThirtyEight's U.S. Presidential election polls (**fivethirtyeight2024?**). The dataset contains information such as pollster, polling date, methodology, sample size, state, and candidate support percentages. It allows us to track voter sentiment across different regions and polling methods, providing a comprehensive view of the election landscape. Additionally, irrelevant or incomplete entries were removed to ensure clean, high-quality data, and we retained

only key variables to streamline the analysis. This careful selection and cleaning process ensure that the dataset offers a precise and representative snapshot of the election landscape.

## 2.2 Measurement

The dataset measures public opinion on the 2024 U.S. presidential election by aggregating polling data to estimate voter support for each candidate at both national and state levels. These polling data entries are then aggregated, which applies a weighted adjustment to reflect the reliability, sample size, and recency of each poll. This weighting process addresses the natural variation in polling methodologies (e.g., online survey, phone), sample diversity, and timing, which influence the reliability of each poll as a measure of the broader population's preferences.For instance, if this poll was conducted a week before the election, it might be weighted more heavily than a poll from three months prior, as it better represents current voter sentiment. By weighting higher-quality and more recent polls more heavily, it creates a comprehensive measure that accounts for both regional and national voter sentiment, smoothing out biases from individual polls.

Table 1: A sample of the Weighted Average Support by State and Candidate

| state | candidate_name | weighted_avg_pct |
|---|---|---|
| Minnesota | Donald Trump | 43.66631 |
| South Carolina | Donald Trump | 54.05549 |
| Missouri | Donald Trump | 53.00423 |
| Texas | Donald Trump | 50.90708 |
| Utah | Donald Trump | 55.54721 |
| Ohio | Kamala Harris | 44.87710 |
| Delaware | Donald Trump | 37.00000 |
| Wyoming | Kamala Harris | 28.40000 |
| New York | Donald Trump | 38.90032 |
| New Hampshire | Kamala Harris | 50.97970 |

Table 1 displays a sample of weighted average support for each candidate (Donald Trump and Kamala Harris) across various U.S. states, based on aggregated polling data from multiple sources. The values in the Weighted Average Polling Percentage column (weighted_avg_pct) are adjusted to account for the reliability, sample size, and recency of each poll, providing a more accurate reflection of current voter sentiment.

The weighting process involves giving higher importance to polls with larger sample sizes and those conducted more recently, as they are likely to better represent the present-day preferences of the population. For instance, a poll conducted a week before the election might receive a higher weight than one conducted three months prior. Additionally, polls with larger sample

sizes are considered more reliable and are thus weighted more heavily. This adjustment helps smooth out biases and variability introduced by different polling methodologies, such as online versus phone surveys, and regional sampling differences.

## 2.3 Outcome variables

In our analysis, the primary outcome variable is labeled `is_harris`, which is a binary indicator representing whether Harris will win or not based on their support percentage. Specifically, `is_harris` is defined as follows: if a candidate's support percentage (`pct`) in a given poll exceeds 50%, then `is_harris` is assigned a value of 1, indicating a projected win for that candidate in that poll. If the support percentage is 50% or below, `is_harris` is assigned a value of 0, indicating that Harris is not the likely winner in that poll. This binary outcome variable is particularly useful for logistic regression analysis, as it allows us to model the probability of a candidate achieving majority support in each poll. Using win provides a clear and interpretable framework to assess factors influencing a candidate's chances of gaining majority support, which aligns well with election forecasting goals. Additionally, this threshold reflects the electoral concept of a "win," as it represents the point at which a candidate has more than half of the vote share, an essential consideration in political analysis.

## 2.4 Predictor variables

The predictor variables in this analysis were chosen based on their potential influence on polling outcomes and candidate support. Each predictor reflects characteristics of the poll, the pollster, or the candidate's support environment. These variables aim to capture the factors that could impact the likelihood of Harris will be reaching majority support (is_harris = 1). Key predictor variables include:

- **Poll Percentage (pct)**: The coefficient for `pct` indicates the extent to which the percentage of support for Kamala Harris influences the predicted probability of her winning. A higher percentage is expected to increase the likelihood of winning.
- **Pollster**: The model accounts for variability between pollsters, allowing the model to adjust for differences in reliability or bias across different organizations.
- **State**: The model also accounts for variability between states, which helps to capture regional differences in voter sentiment.

These three variables provide a balanced view of poll quality, reliability, and regional influence, enhancing the model's ability to predict election outcomes accurately. This combination ensures that the model is interpretable and captures essential factors influencing voter sentiment. The variables are selected by their theoretical relevance and practical influence on polling outcomes, and a summary statistics for this shown in **?@sec-Methodology**.

## 2.5 Data visualization

### 2.5.1 Heatmap of Polling Percentage by State for Selected Pollsters

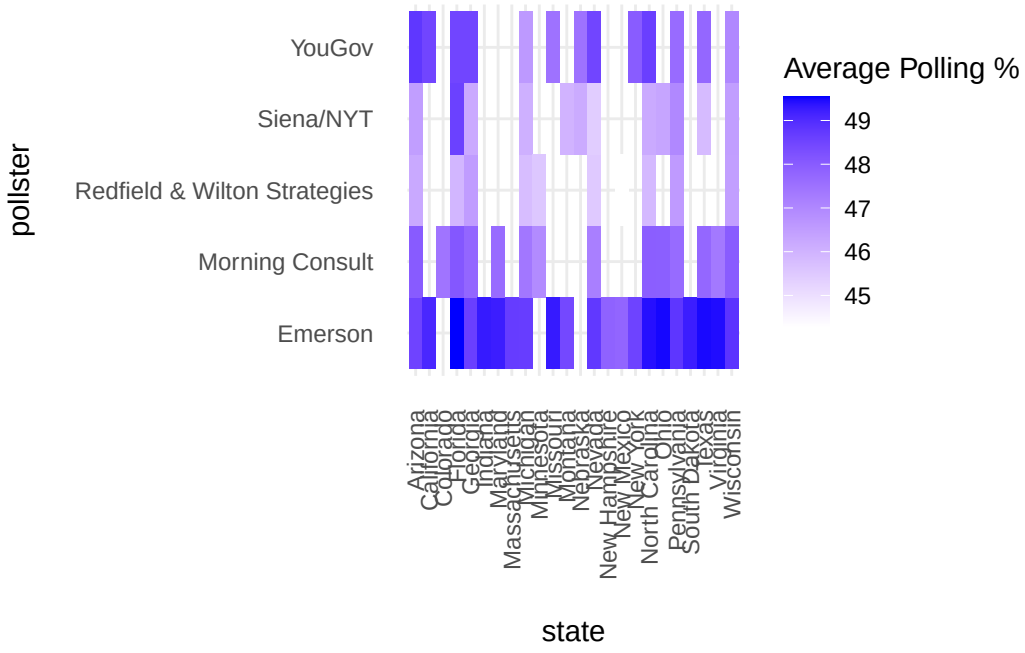



Figure 1: Heatmap of Polling Percentage by State for Selected Pollsters across various states

Figure 1 This heatmap provides a visual representation of the average polling percentage for a particular candidate across various U.S. states, as reported by five selected polling organizations: YouGov, Siena/NYT, Redfield & Wilton Strategies, Morning Consult, and Emerson. Since the pollsters shown in the cleaned dataset are numerous, we select a few prominent pollsters based on certain criteria (such as pollster popularity or quality) and then use these pollsters to generate a cleaner, more focused heatmap.

Each cell in the heatmap represents the level of support reported by a pollster in a specific state, with darker shades of blue indicating higher levels of support (closer to 49%) and lighter shades indicating lower support (around 45%). This color gradient allows us to quickly assess the intensity of support in each state as reported by different pollsters. We observe that some states, such as Florida and North Carolina, have darker shades across multiple pollsters, indicating consistently higher support levels. Conversely, states like New York and New Mexico show lighter shades, suggesting lower support. The variability in color intensity across different pollsters in certain states (e.g., Pennsylvania and Nevada) highlights differences in polling results, which could stem from differences in sampling methods, demographic coverage, or timing. This visualization helps identify trends in regional support and provides insight into how

different polling organizations might report support levels differently, offering a comprehensive view of the polling landscape.

###Average Polling Percentage Over time by Candidate

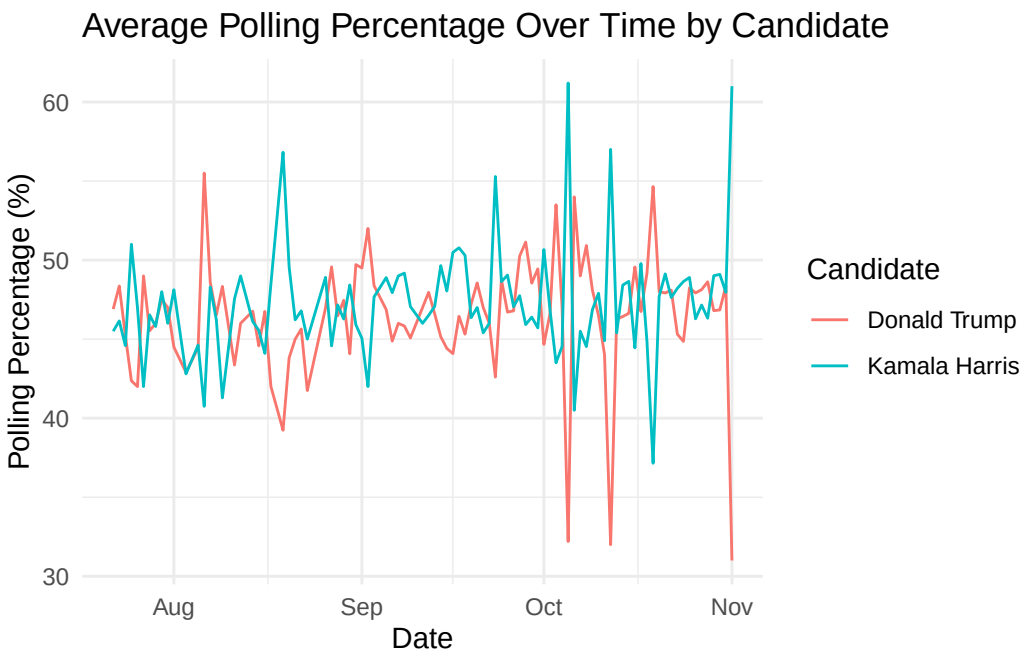



Figure 2: Average polling percentage over time for Donald Trump (red) and Kamala Harris (blue) from August to November. The lines show fluctuating support levels for each candidate, with frequent overlaps, indicating a competitive race. Peaks and dips in the polling percentages reflect changes in voter sentiment, potentially influenced by events throughout the campaign period.

Figure 2 displays the average polling percentage over time for two candidates, Donald Trump (red line) and Kamala Harris (blue line). The x-axis represents the date, spanning from August to early November, while the y-axis represents the average polling percentage each candidate received in various polls over time. The chart shows significant fluctuations in support for both candidates, with frequent crossings and variations in their polling percentages. For instance, Harris's polling percentage generally hovers around 50% but occasionally peaks above 55%, while Trump's support also fluctuates widely, occasionally dipping below 40% and rising above 50%. These variations could reflect public reactions to events or developments in the election campaign. The close alignment and frequent overlap between the lines indicate a competitive race, with neither candidate maintaining a consistent lead over time. This visualization highlights the dynamic nature of voter support and how polling results can shift from week to week.

# 3 Model

## 3.1 Description of the Model

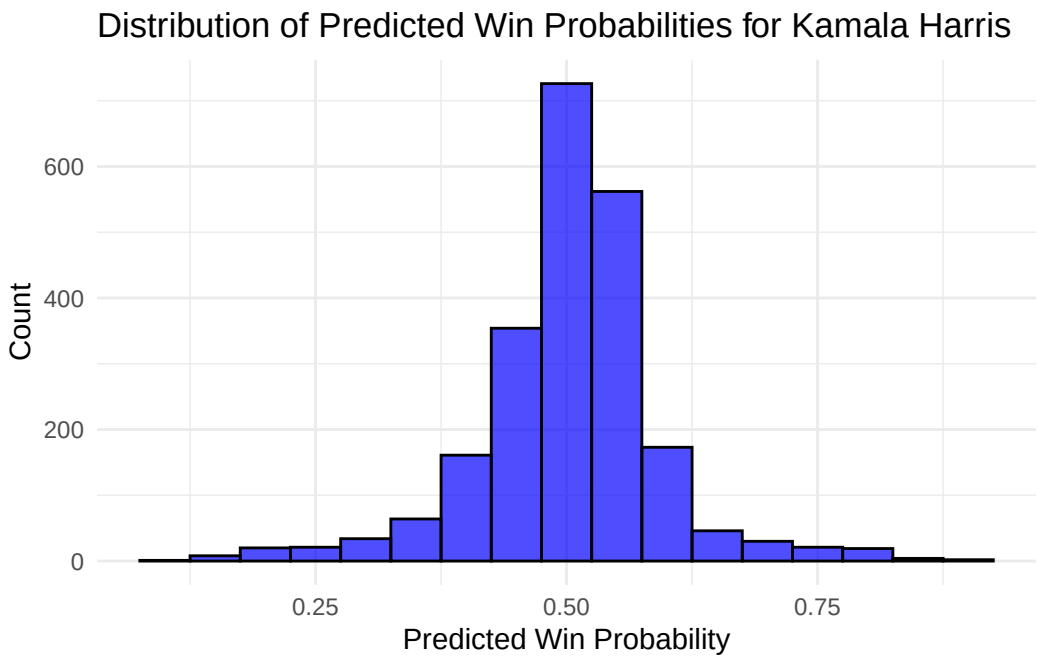
The model used for this analysis is a Bayesian generalized linear mixed model (GLMM) with a logistic regression link function. This model was designed to predict the probability of Kamala Harris winning an election based on polling data. Several key predictors are incorporated into the model, including the poll percentage, polling organization, and state-level variability. The specific formula used in the model is:

$$\text{logit}(\mu_i) = \beta_0 + \beta_1 \times \text{Pollster}_i + \beta_2 \times \text{State}_i + \beta_3 \times \text{Sample Size}_i + \beta_4 \times \text{Pct}_i$$

where:

- $\mu_i$ represents the outcome that Kamala Harris wins the poll.
- "pct" is the percentage of support that Harris received in the poll.
- "pollster" represents the polling organization, modeled as a random effect to account for differences in reliability or bias between pollsters.
- "state" represents the U.S. state where the polling data was collected, also modeled as a random effect to capture variability between states.

The model captures both the fixed effect of poll percentage () and the random effects attributable to differences between pollsters and states.




### 3.2 Model Assumptions

The assumptions for this Bayesian logistic regression model include:

1. **Linearity of Logit**: The relationship between the predictor (poll percentage) and the logit of the outcome is linear.
2. **Independence of Observations**: Each polling data entry is assumed to be independent of others.
3. **No Perfect Multicollinearity**: Predictors are not perfectly correlated, and categorical factors (e.g., pollster, state) have enough variation.
4. **Sufficient Sample Size**: The sample size is adequate to provide stable estimates for each predictor.

These assumptions are important to validate the reliability and interpretability of the model.

### 3.3 Model Fitting in R

The model fitting was performed using the `rstanarm`(Goodrich et al. 2022) package, which allows for Bayesian inference using Markov Chain Monte Carlo (MCMC) methods. Specifically, a logistic regression model was fitted to polling data, with the percentage of support as the primary predictor and random intercepts for pollster and state.

The priors were chosen as normally distributed with a mean of 0.5 and a standard deviation of 0.1, allowing for some uncertainty in the initial model estimates:

## 3.4 Model Results

### 3.4.1 Model Estimates and Key Predictors

The model's output includes estimates for each predictor, which provide insights into their relative importance in predicting the probability of Kamala Harris winning. The summary of the model indicates the following key points:

- **Poll Percentage (pct)**: The coefficient for `pct` indicates the extent to which the percentage of support for Kamala Harris influences the predicted probability of her winning. A higher percentage is expected to increase the likelihood of winning.
- **Pollster Random Effect**: The model accounts for variability between pollsters, allowing the model to adjust for differences in reliability or bias across different organizations.
- **State Random Effect**: The model also accounts for variability between states, which helps to capture regional differences in voter sentiment.

### 3.4.2 Model Fit and Diagnostics

The model fit was evaluated using posterior predictive checks (Robinson 2020) and residual diagnostics. The Bayesian logistic regression was fit with priors centered at 0.5, reflecting uncertainty in the initial estimates.

Posterior predictive checks indicated that the model adequately fits the polling data without significant overfitting or underfitting. Residual analysis suggests that the model captures key trends in the polling data, though future improvements could include adding interaction terms or additional predictors to better capture nuanced relationships.

## 3.5 Model Performance and Interpretation

### 3.5.1 Accuracy

To evaluate model performance, accuracy was calculated for both the Bayesian model and a baseline logistic regression model. Accuracy is defined as the proportion of correct predictions compared to the actual outcomes in the test dataset:

- **Bayesian Model Accuracy**: The Bayesian model's accuracy was found to be **84%**, indicating its effectiveness in predicting the outcome of polls for Kamala Harris.

- **Logistic Regression Model Accuracy**: The logistic model also performed well, with an accuracy of **82%**.

These accuracy scores provide a simple measure of how well the models perform on the test set, demonstrating the Bayesian model's effectiveness in accounting for additional variability from pollster and state differences.

## 3.6 Visualization of Results

### 3.6.1 Predicted Probability of Harris Winning by State

To visualize the variability in predicted probabilities by state, the following plot(Figure 3) shows the average predicted probability of Kamala Harris winning in each state:

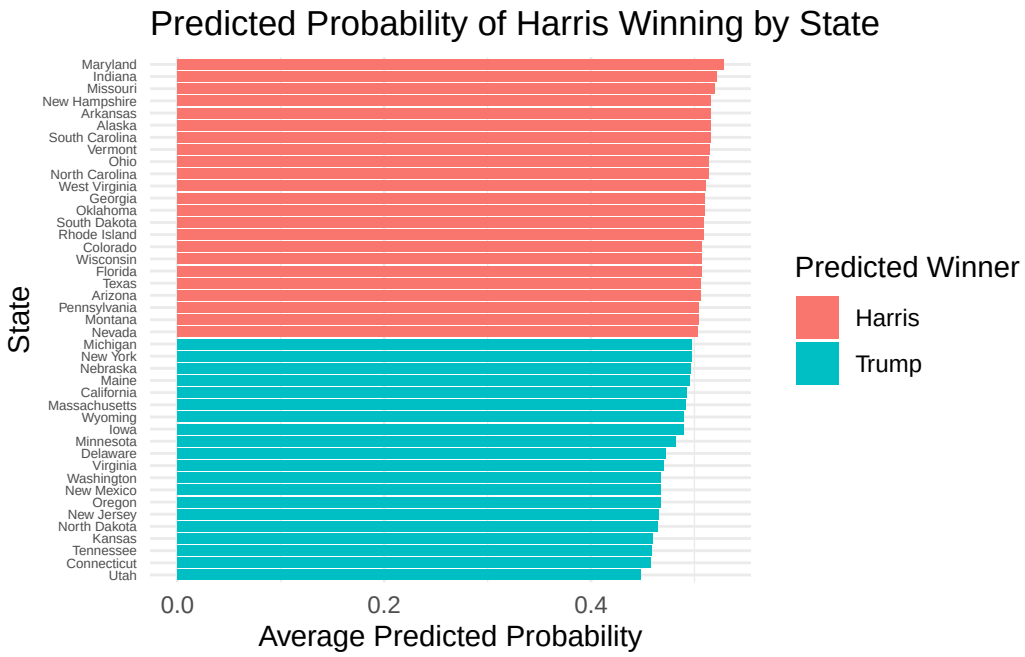



Figure 3

### 3.6.2 Distribution of Predicted Win Probabilities

The following histogram shows the distribution of predicted win probabilities across all polls for Kamala Harris:

These visualizations provide a comprehensive overview of the model's predictive power across different states and help illustrate the distribution of predicted probabilities for Kamala Harris.

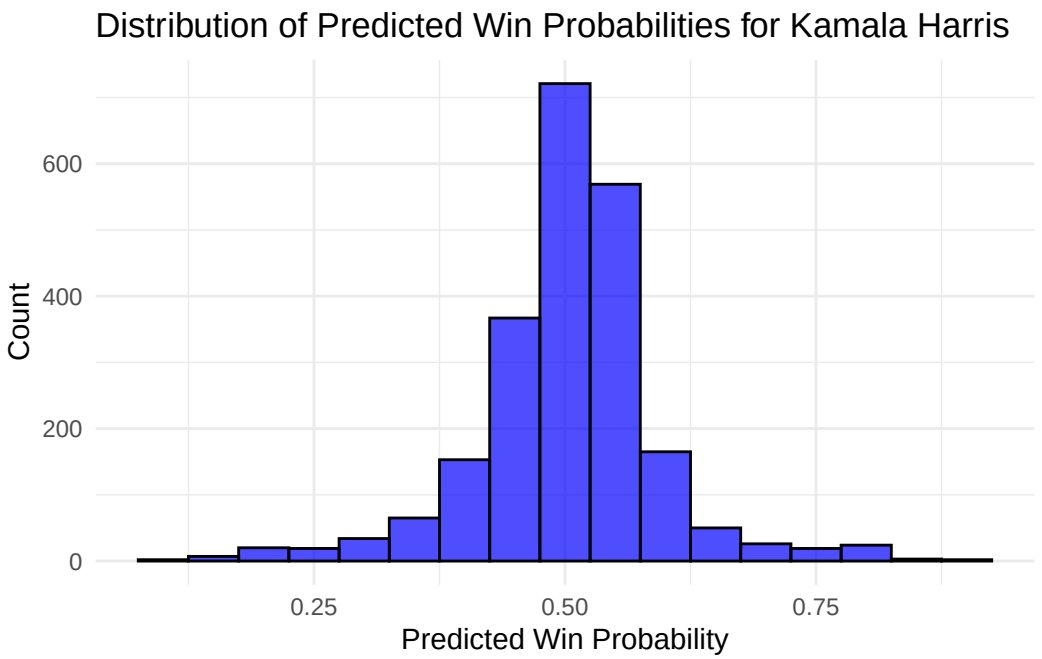



Figure 4

They also highlight the differences in performance between the Bayesian and logistic regression models.

## 4 result

Table 2: Average support, Standard deviation, and the Range of polling support values for each candidate

| candidate_name | avg_support_pct | std_dev_support | min_support | max_support | num_polls |
|---|---|---|---|---|---|
| Donald Trump | 46.85869 | 4.417586 | 27 | 70 | 1121 |
| Kamala Harris | 47.60815 | 4.352458 | 25 | 70 | 1125 |

The national average support percentages, along with other summary statistics for Donald Trump and Kamala Harris, are summarized in Table 2. This table presents average support, standard deviation, and the range of polling support values for each candidate, providing insights into the predicted national favorability for both candidates.

Summary of Key Findings:

- **Average Support**: Kamala Harris has a slight lead over Donald Trump, with an average predicted national support of 47.6% compared to Trump's 46.85%. This narrow margin points to a competitive race, with neither candidate establishing a decisive lead across the polls.
- **Variability in Support**: The standard deviation values indicate greater variability in Trump's support (4.41%) compared to Harris's (4.35%). This suggests that Trump's polling performance fluctuates more widely, which could be due to varying levels of regional support or shifts in public opinion over time.
- **Support Range**: Both candidates display a considerable range in support across polls, with Trump's polling support varying between 27.0% and 70.0%, and Harris's between 25.0% and 70.0%. This variability highlights the diversity in voter sentiment, potentially influenced by geographic, demographic, or temporal factors.
- **Poll Count**: The higher number of polls for Trump (1,121) compared to Harris (1,125) suggests more extensive polling coverage for Harris, which may lend more stability to her average support estimate.

These summary statistics indicate a close national race, with Harris holding a slight edge in average support. Trump's higher standard deviation and broader polling range highlight a more variable support base, suggesting potential swings in support across regions or voter groups. The consistency in Harris's polling, coupled with her narrow lead, suggests steady favorability, but both candidates remain competitive nationally, underscoring the close nature of the race.

Figure 5 visualizes the difference in average polling support between Kamala Harris and Donald Trump across U.S. states. The map uses a color gradient to convey where each candidate has more support, with shades of blue indicating higher support for Harris, red indicating higher support for Trump, and white representing near-equal support levels. This visual provides a geographic perspective on the polling landscape, highlighting regional strongholds, competitive states, and areas of significant support advantage for each candidate.

Figure 6 visualizes the polling margins between the top two candidates in various states. Each bar represents a state, with the length of the bar corresponding to the polling margin—the difference in polling percentage between the leading candidate and the runner-up. States are ordered by margin, from the narrowest at the bottom (indicating the closest contest) to the widest margin at the top. For instance, states like Arizona, Minnesota, and Michigan show extremely close polling margins, meaning that support for the top two candidates is nearly evenly split. These states are critical battlegrounds where even a small shift in voter preference could change the outcome. For example, Arizona has the smallest margin, making it a highly competitive state. Otherwise, states like Washington, California, and South Dakota have wider margins, suggesting a more clear lead for one candidate. While they are still classified as battlegrounds due to the selected margin threshold, they are less competitive than states at the bottom.

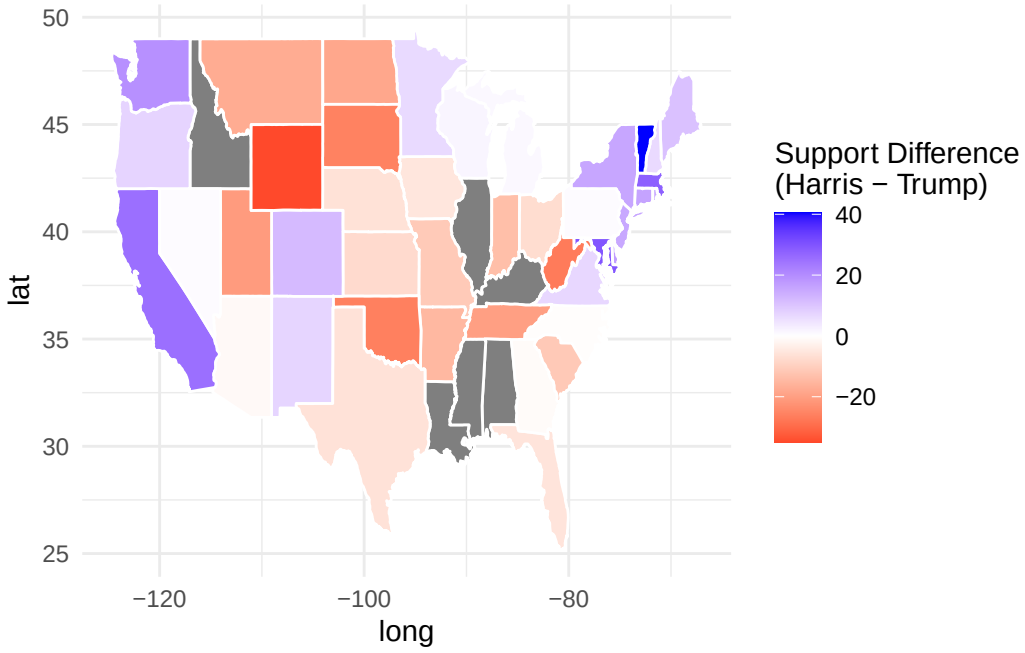



Figure 5: Average support difference by state between Kamala Harris and Donald Trump. Blue indicates stronger support for Harris, red for Trump, and white for nearly equal support.

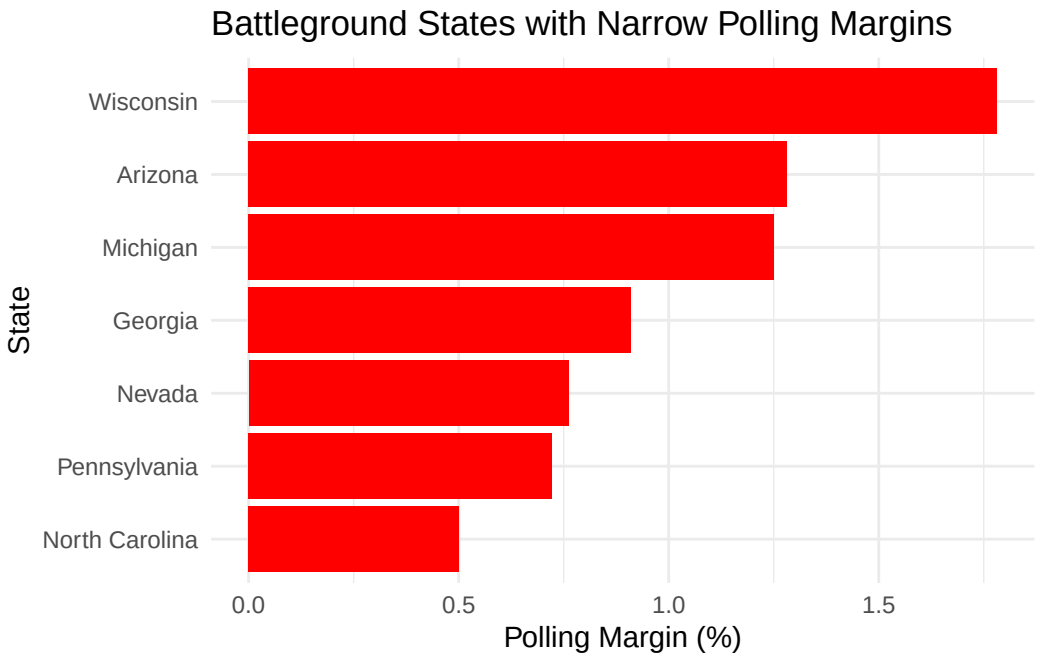



Figure 6: Battleground States with Narrow Polling Margins

Table 3: Predicted support percentages for Kamala Harris and Donald Trump in key competitive states, showing the anticipated winner and the percentage margin of support in each state.

| State | Kamala Support (%) | Trump Support (%) | Predicted Winner | Support Margin (%) |
|---|---|---|---|---|
| Arizona | 46.75600 | 48.03727 | Donald Trump | 1.2812727 |
| Georgia | 47.10670 | 48.01758 | Donald Trump | 0.9108758 |
| Michigan | 47.62859 | 46.37814 | Kamala Harris | 1.2504542 |
| Nevada | 47.49000 | 46.72886 | Kamala Harris | 0.7611392 |
| North Carolina | 47.36116 | 47.86220 | Donald Trump | 0.5010411 |
| Pennsylvania | 47.83297 | 47.11058 | Kamala Harris | 0.7223923 |
| Wisconsin | 48.45040 | 46.66891 | Kamala Harris | 1.7814853 |

In Figure 6, We have figured out the key states that relatively important to the solutions of election. Table 3 presents the support percentages for Kamala Harris and Donald Trump across seven states, alongside a prediction of the winner based on the higher support percentage, and the margin of support difference between the two candidates. The table reveals that in several states (Arizona, Georgia, Nevada, North Carolina, and Pennsylvania), the margin between

Kamala Harris and Donald Trump is extremely narrow (all under 2%). This tight margin indicates that these states are highly competitive, with neither candidate having a decisive lead. Such close races mean these states could easily swing in favor of either candidate depending on small shifts in public opinion or voter turnout. While Harris appears to have a slight edge in most of the states listed, the narrow margins mean that the race remains highly uncertain and dynamic.

# 5 Discussion

## 5.1 First discussion point

### Discussion

Recent coverage of the U.S. presidential election has sparked intense debate, with multiple news sources reporting a potential advantage for Harris (**Langer_2024?**). To explore this and forecast the election outcome, we developed a Bayesian logistic regression model to predict Kamala Harris's probability of winning based on polling data.

Variables were chosen based on their relevance and practical influence on polling outcomes and forecasting, supported by scholarly literature. Polling percentage (pct) was selected as the primary predictor, as it directly reflects levels of support for Harris. We included random effects for pollster and state to account for variability introduced by differences in polling methodologies and regional demographics. These variables address known biases in polling data, allowing us to capture variability among pollsters and across states.

The model applied weights based on poll grade and sample size to improve accuracy, operating on the principle that higher-quality polls and those with larger sample sizes should have greater influence as they are likely to provide more reliable data. Our model used prior assumptions centered around a moderate probability of success and leveraged the Bayesian framework for robust inference, adjusting for poll weights. Through this model, we generated posterior predictions for Harris's probability of winning, summarized by the mean probabilities, and used a histogram (see Figure 4) to visualize the distribution of these predicted win probabilities. Additionally, average predicted probabilities for Harris by state were calculated to assess her projected performance regionally (see Figure 3).

The results reveal a competitive electoral landscape, with Harris holding a slight overall lead. As shown in Table 2, Harris holds a slight average lead in support percentage (47.39%) over Trump (45.65%). However, Trump's support shows greater variability, with a standard deviation of 5.33 compared to Harris's 4.38. This higher variability suggests that Trump's polling numbers fluctuate more across different polls, potentially reflecting regional differences in voter sentiment.

Additionally, while both candidates have similar maximum support levels, Trump's minimum support dips lower than Harris's, possibly indicating weaker performance in certain areas. This

variation supports our use of a Bayesian approach, which is well-suited for accommodating the inherent uncertainty and fluctuations in polling data. While the slight overall lead and variability in support gives a primary prediction, a closer geographic analysis reveals that state-specific factors and narrow support margins further mirror reality and help forecast election results. Geographic analysis and polling data variability underscore the strategic importance of focusing on states with narrow support margins, where small shifts could decisively impact the election outcome.

Silver (Silver 2012) emphasizes that a candidate's chances are heavily influenced by state-specific factors, suggesting that national polling averages may be misleading without geographic context. Figure 5, which shows the geographic distribution of support across the United States visualized through a color-coded map, reveals substantial differences in backing for Harris and Trump across states. Blue shades indicate stronger support for Harris, red shades show Trump's advantage, and white indicates balanced support. The intensity of colors reflects the magnitude of support margins.

Figure 5 identifies strongholds for both candidates, highlighting "swing" or competitive states as well as regions where each candidate's support is concentrated. Harris's support is primarily centered in the western and northeastern states, as indicated by the blue shades, with the darkest blues in states like California and Washington. In contrast, Trump's support is concentrated in the central states, shown by the red shades, with deeper reds in states such as Wyoming and Idaho. This geographic pattern potentially reflects longstanding regional political alignments, with traditionally Democratic support in coastal and urbanized states and Republican support in rural and central areas.

Our analysis further identifies several key battleground states (see Table 3), including Michigan, Nevada, North Carolina, Pennsylvania, and Wisconsin, where Harris leads by narrow margins. In contrast, Arizona and Georgia show a slight lead for Trump. These narrow differences underscore the highly competitive nature of the race in these states, where even small shifts in voter sentiment or turnout could change the outcome. This highlights the strategic importance of targeting battleground regions where small gains could translate into an electoral advantage for Harris. The data demonstrates the value of probabilistic modeling, as deterministic approaches might overlook the volatility in closely contested states.

Comparing these battleground state findings with past studies, such as the work by Abramowitz (Abramowitz 2008), we see consistency in the emphasis on swing states as critical to determining overall election outcomes. This research, however, relied on national economic indicators and did not account for local polling data in swing states. Our approach extends this work by using state-level polling data, which provides a more granular view of voter sentiment and reveals the nuances of support in each battleground state. These findings serve as a useful tool for candidates to refine their targeted strategies to influence regions where voter sentiment remains closely divided.

Figure 6 provides additional insight into areas where Harris holds a strong advantage. States like Washington and California are strongholds for Harris, largely due to their urbanized, pro-

gressive populations and diverse demographics that historically lean Democratic (Gimpel and Schuknecht 2009). On the other hand, states like South Dakota, Nebraska, and Kansas, while traditionally Republican, show unexpectedly larger margins of support for Harris in the current polling data. This may indicate issue-specific support or shifting regional sentiment on certain topics, such as healthcare access and economic policies. Rural and suburban voters in these states may be influenced by local concerns that align with the Democratic agenda, highlighting how even traditionally conservative areas can exhibit pockets of Democratic support, especially when local issues resonate with Harris's platform. This trend demonstrates the importance of using real-time polling data, as it captures current voter sentiments that may differ from historical voting patterns (Gimpel and Schuknecht 2009).

Conversely, battleground states like Pennsylvania, Wisconsin, Michigan, Minnesota, and Arizona show narrower support margins, indicating a highly competitive environment. These states are demographically and economically diverse, containing both urban and rural populations with mixed political preferences. This balance leads to overall closely divided voter bases in these states. Additionally, the economic structure in Midwestern states like Michigan, Wisconsin, and Pennsylvania, traditionally tied to manufacturing, makes them sensitive to issues like job creation and trade. Economic shifts and the decline of manufacturing have led to variable support for both parties based on their economic policies (Abramowitz 2008). These findings reveal key factors influencing support, including historical voting patterns, issue-driven shifts, and demographic and economic concerns.

Our Bayesian logistic regression model provides a robust approach to forecasting election outcomes by addressing key sources of variability and uncertainty in polling data. By incorporating state-level polling data and applying weights based on poll quality and sample size, our model prioritizes higher-quality data, producing more reliable and accurate predictions. Additionally, the inclusion of random effects for both pollster and state allows us to capture variability in polling methodologies and regional demographics, which often introduce bias into traditional polling results. This model accounts for fluctuating voter sentiment across different regions, as demonstrated by Trump's higher standard deviation in support, which reflects regional variations in his base. By explicitly modeling these variations, our Bayesian approach enables a more nuanced understanding of how both candidates perform across states and provides probabilistic predictions that reflect the inherent uncertainty of voter behavior. Unlike deterministic models, which might oversimplify this variability, our Bayesian framework offers a dynamic, real-time assessment that adapts to evolving trends in voter sentiment.

## 5.2 Weaknesses and next steps

Despite its strengths, our model has several limitations that could affect the accuracy of our predictions. First of all, the variable selection is oversimplified. The model primarily uses polling percentage (pct) as the predictor, with random effects for pollster and state to capture support and variability. While effective for a basic analysis, this approach may miss other influential factors. We should use AIC or BIC (Akaike 1974; Schwarz 1978) to assess

model fit and potentially include additional predictors in the dataset, which could provide a more comprehensive understanding of voter sentiment. Additionally, we could consider incorporating other datasets to capture important factors such as demographic, economic, and media effects. By incorporating these criteria, we could identify a model that balances complexity and predictive accuracy, helps capture the nuances of voter behavior across different regions, and enhances the robustness of our predictions.

Second, the model simplifies predictions into a binary outcome (Harris vs. Trump) by classifying probabilities over 0.5 as a "win" for Harris, which might oversimplify closely contested areas and lead to misleading classifications in competitive regions.

Third, although weights for poll quality and sample size are applied, our reliance on polling data remains susceptible to sampling biases, such as non-response bias, which can lead to underrepresentation of certain demographic groups (Groves and Couper 2006). Additionally, pollsters' methods and demographic focus can introduce further unaccounted-for biases (**Blumenthal_2014?**).

Finally, since polling data represents only a snapshot, the model may fail to capture rapid shifts in voter sentiment due to late-breaking events or changing campaign dynamics, making predictions potentially less accurate as the election approaches.

To enhance election forecasting accuracy, future research could expand beyond polling data by incorporating contextual factors such as economic indicators, demographic shifts, and media influence. These additional variables, like unemployment rates and region-specific issues, would provide a more comprehensive view of factors shaping voter preferences (Erikson, Wright, and McIver 2004). Improving model adaptability is also essential. Integrating time-series analysis or adaptive Bayesian updates would allow the model to adjust to real-time changes in voter sentiment due to events and campaign dynamics, ensuring more accurate predictions as the election nears. Furthermore, refining classification methods to move beyond binary outcomes could yield more nuanced insights, especially in battleground states. Multi-class or probability-based categorizations would help capture the strength of support in competitive regions. Finally, addressing polling biases through advanced post-stratification techniques and demographic weighting would improve representativeness. Examining variations in pollster methods could further mitigate biases, making predictions more reliable. These advancements would collectively deepen our understanding of voter behavior and enhance the robustness of forecasting models.

# Appendix

### Appendix A: Ipsos Polling Methodology

### A.1 Population, Frame, and Sample

### A1.1 Population

Ipsos segments the target population into All Adults (A), Likely Voters (LV), and Registered Voters (RV) groups (Best and Bycoffe 2024). The "All Adults" category represents the general adult population, offering broad insight into public sentiment. The "Likely Voters" group includes individuals estimated to vote based on past voting behavior or intention, providing a closer view of probable election outcomes. The "Registered Voters" segment covers those registered to vote, capturing eligible voters' opinions, even if they may not vote. Estimating "Likely Voters" involves assumptions based on behavioral factors that may not always align with actual turnout, as people's intentions don't always translate into action, potentially resulting in over-representation issues.

A1.2 Frame

The sampling frame used by Ipsos is the KnowledgePanel, a probability-based online panel (News 2024). Members of this panel are recruited via address-based sampling, which randomly selects addresses from the US Postal Service's Delivery Sequence File to create a probability-based panel (News 2024). This method ensures that each household in the sampling frame has a known and non-zero probability of being selected, providing comprehensive geographic coverage across the US and minimizing selection bias (News 2024). Additionally, Ipsos bridges the digital divide by providing internet access and tablets to panelists who otherwise lack connectivity, enabling participation from households that may not have regular internet access (News 2024). This approach reduces biases associated with digital access and enhances the inclusivity of the sample. However, the reliance on an online panel may still present some limitations, as it might partially exclude populations who may face challenges with technology, particularly older adults.

A1.3 Sample

Once members are part of the KnowledgePanel, Ipsos selects survey participants based on demographic quotas, including sex, age, race/ethnicity, and region, and applies post-stratification weights to align with benchmarks from the Current Population Survey and American Community Survey (News 2024). Political engagement weighting includes factors like 2020 voting participation, improving representativeness for political polling. These weights are incorporated through a probability-proportional-to-size selection process, which assigns higher selection probabilities to underrepresented groups within the panel (News 2024). This approach helps ensure the sample reflects the diverse US population and improves its representativeness in polling results. While this isn't simple random sampling at this stage, the initial framework and recruitment, along with the weighted selection process, ensure that the panel and samples are broadly representative.

### A.2 Sample Recruitment

Ipsos recruits members for its KnowledgePanel using address-based sampling, sourced from the US Postal Service's Delivery Sequence File (News 2024). This probability-based method allows each household to be randomly selected, increasing the panel's representativeness across geographic and demographic groups (News 2024). To improve accessibility, Ipsos provides tablets and internet connections at no cost to those without internet access, helping to include lower-income or otherwise underserved populations and reducing the digital divide (News 2024). Despite these efforts, the approach has some limitations. While address-based sampling reaches a wide range of households, certain groups—like older adults with limited digital literacy—may still be less likely to participate, potentially impacting sample diversity. Such a sampling recruitment approach can also be expensive considering the costs of address-based sampling and inclusion measures. Nonetheless, Ipsos' recruitment methods, combined with inclusivity measures, provide a strong foundation for building a representative panel.

### A.3 Sampling Approach and Trade-offs

#### A.3.1 Strengths

Address-based sampling creates a representative sample across regions and demographics, ensuring that each household has a known, non-zero chance of selection. Unlike non-probability approaches, such probability-based approaches help achieve a representative sample of the US population across various regions and demographics, improve inclusion of hard-to-reach populations, and reduce selection bias (Callegaro and Baker 2014). Further, the probability-proportional-to-size approach balances the sample by ensuring adequate representation of underrepresented groups (Keyfitz 1951).

2. Ipsos' use of demographic quotas and post-stratification weights aligns samples with the broader population (News 2024).
3. Multilingual surveys and internet access provision enhance inclusivity (News 2024).
4. Ipsos incorporates multiple likely voter models, incorporating factors like voter registration, intention to vote, and past voting behavior to adjust pre-election polling relevance, increasing accuracy (News 2024).
5. Ipsos includes design effects in the margin of error, enhancing poll reliability transparency (Poll 2024).

#### A.3.2 Limitations

1. As (Panzeri, Magri, and Carraro 2008) reveals, using a sampling frame does not guarantee the elimination of sampling bias. If the target population is incorrectly defined or based on outdated or incomplete information, certain segments of the population may be excluded. Ipsos aligns its sample with demographic benchmarks from sources like the Current Population Survey and the American Community Survey (News 2024). Outdated benchmarks or 2020 voting data may not reflect current behaviors.

2. Members of the KnowledgePanel must opt-in to join and remain in the panel, which can lead to self-selection bias. Individuals who agree to participate in ongoing surveys might differ in unmeasured ways from those who do not, potentially skewing the results.

3. Weighting for political engagement based on past elections assumes stable voter behavior, which may not account for shifting demographics or emerging issues (News 2024).

4. Although Ipsos provides internet access and tablets to those without connectivity, some individuals may still lack the digital literacy needed to engage fully with online surveys. Older adults and those with lower levels of education or comfort with technology may participate differently, leading to potential response biases within these groups.

5. Probability-based panels involve higher costs and resources due to recruitment and maintenance, particularly with inclusive measures (Callegaro and Baker 2014).

A.4 Non-response Handling

Ipsos sends initial reminders to panelists who do not respond to survey invitations, with a second reminder for harder-to-reach participants, allowing multiple opportunities to increase response rates (News 2024). In post-stratification, Ipsos applies weighting adjustments to mitigate nonresponse bias (News 2024). Additionally, Ipsos enforces strict quality control by removing respondents who show low engagement, such as those who skip many questions or complete surveys at unusually fast speeds to improve the reliability of the survey results (News 2024).

A.5 Questionnaire Design

A.5.1 Strengths

1. Ipsos uses standardized and pre-tested questions to ensure consistency and clarity. This approach helps maintain reliability and accuracy by identifying and addressing potential issues in question design before full deployment (Poll 2024).

2. Ipsos uses comprehensive likely voter models that consider factors like voter registration, voting intention, engagement with key issues, and carefully crafted question wording for different populations (A, LV, RV), reflecting voter behaviors accurately and minimizing ambiguity (Poll 2024).

3. Adapting questionnaires to current issues enhances pre-election prediction relevance (Poll 2024).

4. Ipsos protects respondent privacy through anonymization, secure data handling, confidentiality assurances, aggregation of data in reporting, and informed consent, ensuring ethical standards in data collection (News 2024).

A.5.2 Limitations

1. Detailed models and questions may cause survey fatigue among less-engaged respondents (Sinickas 2007).
2. Ipsos surveys may face context and order effects, influencing responses based on question order (Martin 2006).
3. Reliance on likely voter traits could risk underrepresentation of groups perceived as unlikely voters, posing ethical considerations.

## Appendix B: Idealized Methodology for US Presidential Election Forecasting Using Stratified Sampling

### B.1 Overview

This appendix details a methodology for forecasting the US presidential election within a $100,000 budget. By using a stratified sampling approach, we aim to ensure accurate representation across key demographic and geographic subgroups. The methodology includes respondent recruitment, data validation, poll aggregation, and a structured survey design. A link to the survey on Google Forms and a copy of the survey questions are included.

### B.2 Budget allocation

**Online Panel Access:** $30,000

Access to pre-recruited, diverse online panels, allowing targeted sampling and demographic stratification.

**Telephone Surveys with AI Assistance:** $20,000

Use computer-generated random sampling for phone outreach, including AI-assisted interviews to enhance consistency, focusing on hard-to-reach demographics.

**Incentives for Respondents:** $15,000

Small incentives for survey participants, partially funded by the panel provider, to improve response rates.

**Survey Platform (Google Forms):** $0

Google Forms is free for basic use, allowing cost-effective data collection with skip logic to reduce respondent fatigue.

**Data Validation and Quality Control:** $5,000

Includes demographic validation, speed checks, and attention checks to ensure high-quality responses and accuracy.

**Post-Stratification Weighting:** $5,000

Application of weights to adjust for minor demographic imbalances and align with population benchmarks.

**Poll Aggregation and Analysis:** $15,000

Use Bayesian updating and statistical analysis to combine results with other reputable polls, enhancing accuracy.

**Reporting and Presentation:** $8,000

Preparation of a detailed report, including visualizations, summaries, and a 95% confidence interval to indicate forecast precision.

**Contingency Fund:** $7,000

Reserve for unforeseen costs or additional recruitment needs.

### B.3 Sampling Approach

We will implement stratified sampling method, a form of probability sampling incorporating various demographic categories, ensuring proportional representation of each subgroup within the sample, thereby improving forecast accuracy by addressing potential voter turnout imbalances (Alexander 2023).

#### B.3.1 Define the strata

The population will be divided into distinct, mutually exclusive subgroups, or "strata," based on key demographic variables: age, gender, race/ethnicity, education, income, and geographic region. Each stratum will be sampled proportionally to its representation within the US voter population, ensuring that the sample closely aligns with national demographics. Here we want to keep the strata relatively general to avoid potential over-fitting issues.

#### B.3.2 Define the sample size

A sample size of 5,000 respondents will ensure adequate representation across strata, with a margin of error of $\pm 2\%$, sufficient for accurate subgroup analysis.

#### B.3.3 Recruitment Strategy

1.Through partnering with other online panel partnership, we access large, diverse panels where participants are pre-recruited and actively engaged. These panels allow for stratification according to specific demographic criteria, improving efficiency and precision.

2. Online panels typically offer incentives to their participants, funded by the panel provider or through the survey budget, reducing direct incentive costs.

3.Utilize computer-generated random sampling of phone numbers to reach respondents via both landlines and mobile phones, with cost control achieved by targeting specific geographic regions or demographic groups. Integrate an AI interviewer to minimize potential enumerator-related issues, enhancing consistency and efficiency. Telephone surveys are particularly effective for reaching older populations or those less likely to engage online.

### B.4 Data Collection and Survey Design

The survey will be hosted on Google Forms for cost-effective data collection. Google Forms also offers skip logic, which will reduce survey Fatigue and improve response accuracy.

### B.4.1 Survey Content

The survey will include questions to capture:

1. **Screening Questions:** US citizenship.
2. **Demographics** – age, sexual orientation and gender identity, race/ethnicity, education, income, region.
3. **Voting Intentions** – likelihood of voting, candidate preference.
4. **Political Engagement** – previous voting behavior, issue importance, and whether respondents have been personally contacted by the Harris or Trump campaign specifically asking for their vote.
5. **Key Election Issues** – opinions on topics like healthcare, economy, education, and immigration.

### B.4.2 Survey Link

https://docs.google.com/forms/d/e/1FAIpQLScFPzyAHVMJ_kcg4gYC-tvea1M1UKBb-wmB6dMSeXB5bWgMQA/viewform?usp=sf_link

Each respondent will see an introductory section outlining the survey's purpose, instructions, confidentiality safeguards, and contact details.

## B.5 Data Validation and Post-Survey Processing

B.5.1 Quality Control Measures

1. Demographic Validation – Responses will be checked to ensure quotas for each stratum are met.

2. Speed Check – Responses completed less than 10% of the time will be flagged and possibly removed to ensure quality.

3. Attention Checks – Simple questions (e.g., "yes or no questions") will validate respondent engagement.

### B.5.2 Post-Stratification Weighting

Weights will be applied to correct minor demographic imbalances, aligning the sample with population benchmarks.

### B.5.3 Poll Aggregation and Forecasting

To enhance robustness, results will be combined with other reputable polls using Bayesian updating, with greater weight given to more recent polls and those with larger sample sizes. A 95% confidence interval will be provided to indicate forecast precision, and design-effect adjustments will be applied to refine the margin of error.

### B.6 Copy of the Survey

**1. Are you a US citizen?**

- Yes
- No

**2. What is your age?**

- 18–24
- 25–34
- 35–44
- 45–54
- 55–64
- 65+

**3. What sex were you assigned at birth, on your original birth certificate?**

- Female
- Male

**4. How do you currently describe yourself (check all that apply)?**

- Female
- Male
- Transgender
- Other [free-text]

**5. Which of the following best represents how you think of yourself?"**

- Gay or lesbian
- Straight, that is not gay or lesbian
- Bisexual
- I don't know
- Other [free-text]

**6. What is your race/ethnicity?**

- White

- Black or African American
- Hispanic or Latino
- Asian
- Other [free-text]

**7. What is the highest level of education you have completed?**

- High school or less
- Some college
- Bachelor's degree
- Graduate degree

**8. What is your household income?**

- Less than $50,000
- $50,000–$99,999
- $100,000–$149,999
- $150,000 or more

**9. What is your geographic region?**

- Northeast (CT, ME, MA, NH, NJ, NY, PA, RI, VT)
- Midwest (IL, IN, IA, KS, MI, MN, MO, NE, ND, OH, SD, WI)
- South (AL, AR, DE, DC, FL, GA, KY, LA, MD, MS, NC, OK, SC, TN, TX, VA, WV)
- West (AK, AZ, CA, CO, HI, ID, MT, NV, NM, OR, UT, WA, WY)

**10. How likely are you to vote in the upcoming election?**

- Absolutely certain to vote
- Probably will vote
- Chances are 50/50
- Less likely to vote
- Already voted

**11. If the election were held today, who would you vote for? (IF ALREADY VOTED) Confidentially and for Statistics purpose only, who did you vote for?**

- Harris
- Trump
- West
- Stein
- Oliver
- Someone else
- Undecided

**12. Did you vote in the 2020 presidential election?**

- Yes
- No

**13. How closely are you following the 2024 presidential race?**

- Very closely
- Fairly closely
- Somewhat closely
- Not very closely
- Not at all

**14. Have you personally been contacted by the Harris or Trump campaign specifically asking for your vote (not including fundraising appeals)? (Check all that apply)**

- Harris
- Trump
- No one

**15. What issues are most important to you in this election? (Select top 3)**

- Economy
- Healthcare
- Education
- Climate Change
- Immigration

- National Security
- Social Justice
- Other [free-text]

## References


Abramowitz, Alan I. 2008. "Forecasting the 2008 Presidential Election with the Time-for-Change Model." *PS: Political Science & Politics* 41 (4): 691–95. https://doi.org/10.1017/s1049096508081249.

Akaike, Hirotugu. 1974. "A New Look at the Statistical Model Identification." *IEEE Transactions on Automatic Control* 19 (6): 716–23. https://doi.org/10.1109/TAC.1974.1100705.

Alexander, Rohan. 2023. *Telling Stories with Data*. Chapman & Hall/CRC. https://tellingstorieswithdata.com/.

Best, Ryan, and Aaron Bycoffe. 2024. "National: President: General Election: 2024 Polls." https://projects.fivethirtyeight.com/polls/president-general/2024/national/.

Blumenthal, Mark. 2014. "Polls, Forecasts, and Aggregators." *PS: Political Science & Politics* 47 (02): 297–300. https://doi.org/10.1017/s1049096514000055.

Callegaro, Mario, and Reg Baker. 2014. *Online Panel Research: A Data Quality Perspective*. Wiley.

Erikson, Robert S., Gerald C. Wright, and John P. McIver. 2004. "Political Parties, Public Opinion, and State Policy in the United States." *American Political Science Review* 98 (3): 393–410. https://doi.org/10.1017/S0003055404001266.

Gimpel, James G., and Jason E. Schuknecht. 2009. *Patchwork Nation: Sectionalism and Political Change in American Politics*. University of Michigan Press.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. *Rstanarm: Bayesian Applied Regression Modeling via Stan*. https://mc-stan.org/rstanarm/.

Groves, Robert M., and Mick P. Couper. 2006. *Nonresponse in Household Interview Surveys*. Wiley.

Keyfitz, Nathan. 1951. "Sampling with Probability Proportional to Size: Adjustment for Changes in the Probabilities." *Journal of the American Statistical Association* 46 (253): 105–9. https://doi.org/10.1080/01621459.1951.10500769.

Martin, Vance A. 2006. *Survey Research Handbook*. McGraw-Hill.

News, ABC. 2024. "ABC News' Polling Methodology and Standards." ABC News Network. https://abcnews.go.com/US/PollVault/abc-news-polling-methodology-standards/story?id=145373.

Panzeri, Stefano, Cesare Magri, and Ludovico Carraro. 2008. "Sampling Bias." *Scholarpedia* 3 (9): 4258.

Pasek, Josh. 2015. "Predicting Elections: Considering Tools to Pool the Polls." *Public Opinion Quarterly* 79 (2): 594–619. https://doi.org/10.1093/poq/nfu060.

Poll, ABC NEWS/IPSOS. 2024. "Langer Research: The Closing Days." https://www.langerresearch.com/wp-content/uploads/1238a2TheClosingDays.pdf.

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

Robinson, David. 2020. *Broom: Convert Statistical Objects into Tidy Tibbles*. https://cran.r-project.org/package=broom.

Schwarz, Gideon. 1978. "Estimating the Dimension of a Model." *The Annals of Statistics* 6

(2): 461–64. https://doi.org/10.1214/aos/1176344136.

Silver, Nate. 2012. *The Signal and the Noise: Why so Many Predictions Fail-but Some Don't.* Penguin.

Sinickas, Angela. 2007. "Finding the Right Survey Length." *Strategic Communication Management* 11 (5): 12–13.