title: "Forecasting the 2024 U.S. Presidential Election"
subtitle: "Kamala Projected to Defeat Trump, 48.3% to 47.1% in the Popular Vote and 270 to 268 in the Electoral College Based on Poll of Polls and Bayesian Modeling"
author:
- Uma Sadhwani
- Arav Sri
thanks: "The code and data used to perform this presidential election forecast can be found at: [https://github.com/Aravsria/Term-Paper-2/tree/main].

date: today
date-format: long

abstract: "The U.S. Presidential election, set for Tuesday, November 5th features a tightly contested race between Vice President Kamala Harris and former President Donald Trump. In this paper, we employ a poll-of-polls methodology and apply a Bayesian model to forecast the likely winner of the popular vote and the outcomes. Using the results of our poll-of-polls analysis, we predict that Harris will win the popular vote, 48.3% to 47.1% and the electoral college 270 to 268. Our analysis shows that the race is extremely tight and former President Trump winning the popular vote, electoral college, or both is well within the margin of error. Our results show a statistical tie when we account for margin of error, bias, weaknesses, and limitations."
format: pdf
number-sections: true
bibliography: references.bib
---

## 1. Introduction

Predicting the outcome of the upcoming U.S. presidential election is a challenging but essential task for political analysts, pollsters and data scientists. By leveraging polling data, particularly through a “poll-of-polls” approach, we can capture a more comprehensive picture of voter intentions and potential election results.

A poll-of-polls approach aggregates data from various surveys, theoretically smoothing out biases or anomalies found in individual polls, and offering a more reliable forecast (Blumenthal, 2014; Pasek, 2015). This project aims to build a predictive model, either a linear or generalized linear model- to forecast the winner of the upcoming U.S. presidential election. This model will be developed and documented in an entirely reproducible manner to enhance transparency and reliability. Using data from publicly available polling sources, our analysis will also include an examination of the methodology of a specific pollster, providing insight into their sampling

strategy, recruitment approach, and the strengths and limitations of their methods. Additionally, we will design an idealized polling methodology, exploring how a budget of $100,000 could be allocated to maximize survey accuracy and data integrity.

## 2. Data

### 2.1 Overview

The dataset was downloaded on October 21, 2024, and only includes polling data available up to that date; any polls released afterward are not considered in this analysis. The presidential polling dataset from FiveThirtyEight covers both national and state-level polls for all 50 states, as well as the congressional districts in Maine and Nebraska [@fivethirtyeight]. These polls were conducted by a range of pollsters, including YouGov, Siena/NYT, CES/YouGov, Marquette Law School, The Washington Post, and McCourtney Institute/YouGov. Our analysis focuses on national data and state-level data for all states expected to play a critical role in determining the election outcome.

To prepare the dataset for analysis, we filtered it to include only high-quality polls conducted on or after July 21, 2024, and focused on likely voters. The data was cleaned using the tidyverse and janitor packages in R to enhance quality and relevance. We standardized column names for clarity and selected key columns: pollster, pollscore, sample_size, candidate_name, and pct (indicating candidate support percentage). Rows with missing values in these essential columns were removed to ensure completeness.

To ensure numerical integrity, we converted pct to numeric and limited its range between 0 and 100, representing valid percentages of support. We also converted pollscore to numeric and sample_size to integer types. We then renamed pct as candidate_support and pollscore as pollster_score for greater clarity. Any remaining rows with missing values were dropped to produce a clean dataset, which was then saved as analysis_data.csv for further analysis.

### 2.2 Measurement

Our primary objective is to gauge public opinion and translate it into a forecast for the 2024 U.S. presidential election. Surveys are widely used to measure public sentiment. During election cycles, pollsters conduct surveys to assess public opinion and candidate preferences. These surveys aim to represent the U.S. electorate by sampling likely or registered voters and gathering information on demographics, partisan affiliation, candidate preferences, and issue stances. Each survey response reflects an individual's voting preference, which pollsters aggregate and adjust to reflect the broader population. This process includes weighting by state and demographic factors (e.g., age, education, race, gender) and accounting for the likelihood of voting [@wfu]. These adjustments convert raw opinions into projected support percentages for each candidate, enabling predictions of potential election outcomes.

Our dataset, sourced from @fivethirtyeight, is a compilation of presidential polls conducted by various pollsters that were conducted during the 2024 presidential election cycle. Each entry represents the percentage of respondents to a unique poll supporting Vice President Harris (after July 21, 2024), President Joe Biden (before July 21, 2024), former President Trump, and third-party candidates. Unique polls are identified by a poll id, and each entry includes includes detailed information about the poll, such as the conducting pollster, target pollster, its population, sample size, and the methodology that was used and information about its quality and accuracy, including its numeric grade, pollscore, and transparency score [@fivethirtyeight].

The process of translating individual opinions into structured entries in our dataset involves three key steps, allowing us to analyze trends and forecast election outcomes:

- Survey: selected voters respond to a survey.

- Adjustment: survey responses are aggregated and weighted to estimate support for each candidate.

- Reporting: the results from the adjustment step are recorded as dataset entries, which serve as snapshots of public opinion over time.

**2.3 Outcome and predictor variables**

We will use end_date (the date a poll was completed), state, pollster, and pollscore to predict support for Vice President Harris and former President Trump at both the national level and the state level for each state. The tables and visualizations below illustrate potential relationships between these predictor variables and support for Vice President Harris and former President Trump.

[INSERT GRAPH from Arav]

```
# Load necessary libraries
library(ggplot2)
library(dplyr)

# Load the data
data <- read.csv("data/02-analysis_data/analysis_data.csv")

# Convert end_date to Date format (assuming end_date is in "YYYY-MM-DD" format)
data$end_date <- as.Date(data$end_date)

# Filter data for relevant states (if needed) or use all states
```

```
# For this example, I'm assuming we want to look at multiple states
# To focus on battleground states or a subset, uncomment and adjust the code below
# battleground_states <- c("Arizona", "Georgia", "Nevada", "North Carolina", "Wisconsin",
"Michigan", "Pennsylvania")
# data <- data %>% filter(state %in% battleground_states)

# Plotting support over time by state and pollster
ggplot(data, aes(x = end_date, y = pollscore, color = pollster, group = pollster)) +
  geom_line(aes(linetype = Candidate), size = 1) + # Line graph for trends
  geom_point(aes(shape = Candidate), size = 2) + # Points for individual poll data
  facet_wrap(~ state) + # Separate panels for each state
  labs(title = "Polling Support Trends for Vice President Harris and Former President Trump",
       subtitle = "By Date, State, and Pollster",
       x = "Poll End Date",
       y = "Support Score",
       color = "Pollster",
       linetype = "Candidate",
       shape = "Candidate") +
  theme_minimal() +
  theme(legend.position = "bottom")
```

**2.3.1 Divide in support for Harris vs Trump by state**

The level of support for Vice President Harris and former President Trump varies across states; in some states, Vice President Harris has a higher-than-average national support level, while in others, it falls below the national average. The outcome of the 2024 presidential election is expected to hinge on seven key battleground states: Arizona, Georgia, Nevada, North Carolina, Wisconsin, Michigan, and Pennsylvania, as well as Nebraska's second congressional district (270 to Win, 2024). Notably, Maine and Nebraska allocate their electoral votes uniquely, awarding one vote to the winner in each congressional district and an additional two votes to the overall statewide popular vote winner (270 to Win, 2024).

Table 1. Presents the polling averages for Harris and Trump both nationally and at the state level, highlighting a narrow lead for Vice President Harris in the popular vote and highly competitive races in the seven swing states—Arizona, Georgia, Nevada, North Carolina, Michigan, Pennsylvania, and Wisconsin—based on data available as of October 21, 2024.

```
# Load necessary libraries
library(tidyverse)
library(janitor)
```

```
# Load the data
polls_data <-
read_csv("/home/rstudio/polling_data/polling_data/data/01-raw_data/president_polls.csv")

# Filter for Harris and Trump data only
filtered_data <- polls_data %>%
  filter(candidate_name %in% c("Kamala Harris", "Donald Trump"))

# Create a table with state, Harris %, and Trump %
table_data <- filtered_data %>%
  select(state, candidate_name, pct) %>%
  group_by(state, candidate_name) %>%
  summarise(avg_pct = mean(pct, na.rm = TRUE)) %>%
  pivot_wider(names_from = candidate_name, values_from = avg_pct) %>%
  clean_names() %>%
  rename("state" = "state",
         "harris %" = "kamala_harris",
         "trump %" = "donald_trump")

# Display the table
print(table_data)
```

According to our FiveThirtyEight (2024) presidential polling statistics, Vice President Harris is just 50.5% ahead of former President Trump in the popular vote. Trump is leading in Arizona, Georgia, and North Carolina, while Harris is leading in Nevada, Michigan, Wisconsin, and Nebraska's second congressional district. These seven battleground states are displaying close contests. Both candidates are tied in Pennsylvania. Although these states are not expected to have a significant impact on the election outcome (270 to Win 2024), the dataset also includes polling data from states that are predicted to favor Republicans (such as Florida, Missouri, Montana, Nebraska, Ohio, and Texas) and those that are likely to lean Democratic (such as Minnesota, New Hampshire, and Virginia). Trump and Harris' polling averages differ slightly among the battleground states; in four of them, Trump's polling average is higher than the national average, while Harris's is lower in five.

Polling has indicated that she and Trump are in a close race since Vice President Harris emerged as the presumed Democratic contender and President Biden halted his reelection campaign. Appendix C.2 provides polling averages for the six months before to election day, starting prior to Biden's departure on July 21, 2024. While Figure 2 shows state-level polling for both candidates in the seven battleground states and Nebraska's second congressional district, Figure 1 summarizes the national polling averages for Harris and Trump from July 21.

Shortly after being officially announced as the Democratic contender in late July, Vice President Harris first outperformed former President Trump in popular vote surveys. But in the middle of August, her lead started to wane, and the race has been neck and neck ever since.

The battleground states and Nebraska's second congressional district have seen different trends in Harris and Trump's state-level polling. Trump has held a slim advantage in Arizona since early September, and he has continuously outperformed Harris in Georgia since she entered the race. In Michigan, the two contenders were deadlocked until early October, when Harris gained a little advantage. Although there is little information available for Nebraska's second congressional district, since late September, Harris has continuously held a larger advantage there than either candidate has in any other battleground state. Prior to Trump's slight victory in October, Harris and Trump were almost even in North Carolina. At the moment, Harris has maintained a slim advantage in Wisconsin since August, while the two are essentially even in Pennsylvania.

[INSERT GRAPH HERE FOR HARRIS VS TRUMP % vs september, nov, dec]

**Figure 1.** Illustrates Harris's lead over Trump in the national popular vote polling, with lines representing the moving average of polls and individual high-quality polls shown as points. The color shading of the points indicates which candidate led in each poll.

The shifts in Harris and Trump's polling averages over time vary significantly across the seven battleground states and Nebraska's second congressional district. Since early September, Trump has maintained a slight lead over Harris in Arizona and has consistently led in Georgia since she entered the race. Harris and Trump were neck-and-neck in Michigan until early October, when Harris gained a slight edge. While polling data for Nebraska's second congressional district is limited, Harris has shown a relatively stronger lead there than in any other battleground state since late September. In North Carolina, Harris and Trump were almost tied until Trump took a slight lead in October, with the two candidates now nearly tied in Pennsylvania. Harris has also held a slight lead in Wisconsin since August.

### 2.3.2 Divide in support for Harris vs Trump by pollsters and pollscores

Different polling organizations may present results that lean toward one candidate due to factors such as methodology, how respondents are selected, sample size, and handling of non-responses. Within our dataset, poll averages for Harris and Trump are generally close across various pollsters, though minor discrepancies exist. Figure 3 illustrates the national polling averages for Harris and Trump, showing variation by pollster. Siena/NYT polls reflected a near tie between

Harris and Trump until mid-September when Harris gained a slight edge, though the two have been almost evenly matched again since mid-October. YouGov polls have consistently shown a slight lead for Harris since mid-August, while Marquette Law School polls indicate a gradual decline in Harris's lead since she became the Democratic nominee. Polling outcomes for Harris and Trump can be influenced by the specific polling organization conducting the survey, which is why we "pool the polls," or combine results from multiple polls, to minimize individual pollster bias . Our dataset contains fewer polls from CES/YouGov and McCourtney Institute/YouGov, but both have shown a small lead for Harris.

The "pollscore" metric is an indicator of a pollster's accuracy, with positive scores reflecting accuracy above a theoretical baseline and negative scores indicating higher-than-average precision. Our dataset includes only high-quality polls, each graded with a score of 3, which signifies a negative pollscore, transparency, and strong accuracy. Within our dataset, Siena/NYT is the top-performing pollster, demonstrating high pollscore reliability and close polling averages for Harris and Trump .

**Appendix 1: Sampling Approach and Pollster Methodology**

The Times/Siena late October 2024 battleground poll uses a sampling technique known as response-rate-adjusted stratified sampling. This probability-based approach divides the voter population in the seven battleground states into specific categories or strata. These strata are based on various characteristics, including legislative district, political affiliation, ethnicity, gender, marital status, household size, voting history, age, and home ownership statusomly selecting individuals within each stratum, the poll aims to achieve a representative sample of likely voters across these states. Adjustments are made for different selection probabilities within each stratum, considering the likelihood of each respondent participating in the 2024 election. These adjustments help align the sample with the projected composition of the electorate in 2024.

Additreater weight is assigned to groups that were underrepresented in the sample to reduce the impact of non-response bias. This approach addresses historical biases observed in the 2016 and 2020 elections, where pollsters were found to have underestimated support for Trump due to non-response bias . The final sampl poll included 7,879 likely voters from the seven battleground states: Arizona, Nevada, Georgia, North Carolina, Wisconsin, Michigan, and Pennsylvania. Specifically, there were 1,025 from Arizona, 1,010 from Nevada, 1,004 from Georgia, 1,010 from North Carolina, 1,305 from Wisconsin, 998 from Michigan, and 1,527 from Pennsylvania .

Furthermore, the sample nsin was supplemented with additional unregistered voters, acknowledging the state's record of substantial same-day voter registration.

The advantage of the Times/Siena poll lies in its concentrated focus on key battleground states and the use of stratified sampling. By honing in on seven critical states instead of the national popular vote, the poll can make predictions about the election outcome based on state-specific data. Stratified sampling also helps mitigate issues of non-response bias and the under-representation of certain groups, such as Trump supporters. This is achieved by weighting each stratum to reflect the overall population distribution and adjusting for the likelihood of responses from typically underrepresented groups.

The Times/Siena polls, however, do not use patterns as a weighting criterion, which has been controversial. While this approach avoids the potential pitfalls of past election recall bias, it could overlook the stabilizing effects of historical voting patterns. Some research has shown that recalling previous voting behavior could make polls more accurate, particularly given the observed patterns since 2004 where recalling vote history has influenced poll reliability and support estimates.

**[insert citation]**

```
@manual{timessienapolls,
  title = {Cross-Tabs: Late October 2024 Times/Siena Poll of the Likely Electorate},
  author = {{The New York Times}},
  publisher = {The New York Times},
  year = {2024},
  url = {https://www.nytimes.com/interactive/2024/10/26/us/elections/times-siena-nyc-poll-registered-voter-crosstabs.html},
}
```