Forecasting US Presidency Election Results for 2024*

Analyzing Trends in Trump's Support Leading Up to the 2024 U.S. Presidential Election

Ruizi Liu Yuechen Zhang Bruce Zhang

November 4, 2024

This paper forecasts the 2024 U.S. Presidential election with a specific focus on Donald Trump's support, utilizing polling data from FiveThirtyEight. By applying linear models, we analyze data collected from reputable pollsters such as YouGov and Siena/NYT. Our results indicate that Trump's support has remained stable, ranging from 40% to 50% in both national and state polls between August and October 2024. Notably, higher-quality pollsters provided more consistent and reliable predictions. This research is significant because it offers insights into voter stability and the effectiveness of polling methodologies, highlighting broader implications for understanding electoral behavior and forecasting political outcomes. However, limitations such as the assumption of linearity and the exclusion of other candidates underscore the need for cautious interpretation and continued monitoring to refine predictions as the election nears.

1 Introduction

Predicting the outcome of the United States presidential election is a critical and multifaceted task, with polling data as a key indicator of voter sentiment and election trends. The aggregation of data from multiple pollsters enhances the reliability of these analyses, as it provides a broader and more representative view of voter preferences. However, the accuracy of such predictions is complicated by the diverse methods used by different polling organizations, variations in sampling strategies, and the timing of surveys. Despite abundant polling data, these inherent uncertainties make accurate election forecasting a persistent challenge.

*Code and data are available at: https://github.com/RIRI0527/2024US_election_forecast.

This paper is structured to first introduce the importance of reliable election forecasting and the factors influencing polling data accuracy. We then detail our methodology, focusing on constructing linear models using FiveThirtyEight's presidential election poll data, which includes contributions from various pollsters such as YouGov and RMG Research. Our analysis aims to explore how key variables—such as geographic distribution, polling methods, and non-response bias—impact the stability of voter support. Additionally, we will investigate the influence of a high-profile pollster's methods on the projected support for candidates, with a specific focus on Donald Trump.

The estimand in our study is the percentage of voter support for Donald Trump at the state and national levels, as predicted by polling data over time. By estimating this quantity, we aim to understand the dynamics of voter sentiment and forecast how stable or variable Trump's support might be in the lead-up to the 2024 election.

The importance of this work lies in its potential to improve the understanding of electoral dynamics, inform campaign strategies, and highlight the limitations of existing polling practices. Our findings indicate that Trump's support has generally remained stable and even increased in key states over the analyzed period. We observed consistent polling trends among the selected pollsters, underscoring the value of analyzing high-quality, aggregated polling data for election forecasting.

2 Data

2.1 Overview

We utilized the statistical programming language R (R Core Team 2023) to process and analyze polling data sourced from FiveThirtyEight's comprehensive database on the 2024 U.S. Presidential election (FiveThirtyEight 2024). This dataset Table 1 includes polling results from organizations such as YouGov and Siena/NYT, collected using diverse methodologies like online surveys and phone interviews. The data captures critical details, including the pollster, methodology, sample size, geographic location, and polling dates. By cleaning and organizing the data, we are able to examine how public opinion shifts over time and across different regions, facilitating robust analysis of candidate support trends.

Table 1: Sample of 2024 US election forecast

Percentage	Start Date	End Date	State	Methodology	Pollster	Poll Score
48	10/8/24	2024-10-11	National	Online Panel	YouGov	-1.1
47	10/7/24	2024-10-10	Pennsylvania	Live Phone	Siena/NYT	-1.5
45	10/7/24	2024-10-10	Pennsylvania	Live Phone	Siena/NYT	-1.5
47	10/7/24	2024-10-10	Pennsylvania	Live Phone	Siena/NYT	-1.5

Percentage	Start Date	End Date	State	Methodology	Pollster	Poll Score
45	10/7/24	2024-10-10	Pennsylvania	Live Phone	Siena/NYT	-1.5

2.2 Measurement

The data represents real-world voter sentiment collected from various polls conducted by different organizations. Each poll surveys a representative sample of voters, asking them about their candidate preferences and using methods such as live phone interviews or online panels. Information such as the percentage of respondents supporting a specific candidate, the poll's location, and the date range of the survey is recorded in a structured format. For example, a YouGov poll conducted in Michigan in early October provides data on voter support for each candidate.

To ensure data quality, we cleaned the dataset by selecting polls with a numerical grade of 3 or higher, as rated by FiveThirtyEight FiveThirtyEight (2024). We also addressed missing values and standardized variables to maintain consistency and reliability in our analysis.

2.3 Outcome variables

The primary outcome variable in our analysis is **pct**, which denotes the percentage of respondents in a poll who support a particular candidate. This variable is crucial for assessing each candidate's level of public support over time and in different regions.

2.4 Predictor Variables

We utilized several predictor variables to explain variations in the polling outcomes:

- **pollster**: Identifies the organization conducting the poll (e.g., YouGov, Siena/NYT). Differences in methodologies between pollsters can affect the results, making this variable important for our analysis.
- **state**: Indicates the geographic region or national focus of the poll, capturing variations in voter behavior across different states.
- **end_date**: The date when the polling concluded, allowing us to analyze how voter support evolves over time.
- **methodology**: Describes the data collection method used (e.g., online survey, live phone interview). This variable helps assess how different approaches may impact polling results.

These predictor variables are essential for modeling polling percentages and forecasting trends, providing insights into how factors like geography, polling organization, and methodology influence public opinion.

3 Methodology

3.1 Data Collection and Preprocessing

We gathered polling data for the 2024 U.S. Presidential Election from FiveThirtyEight and processed it using R (R Core Team 2023). The dataset contains polling results across various demographics and time periods, with contributions from reputable pollsters such as YouGov and Siena/NYT. To address missing values, we applied imputation techniques, and generated dummy variables for categorical predictors like region and pollster. Additionally, post-stratification weighting was implemented to ensure our sample accurately represents the U.S. population, minimizing sampling bias.We chose the pollster based on active selection of the pollster score. We classified all pollsters with a pollscore of more than 3 as high quality and worthy of including in our analysis.

3.2 Model Selection

We tested a range of regression models to predict candidate vote percentages, with predictors including polling date, candidate, and pollster. After evaluating the performance of these models, we selected simple linear regression for its interpretability and effectiveness in identifying trends in polling data. We opted not to use multiple regression models due to low R-squared values that suggested limited explanatory power.

3.3 Simple Linear Regression

We first employed a simple linear regression model to estimate Trump's support percentage based on polling end dates. The model follows this equation:

$$pct = \beta_0 + \beta_1 \ end \ date$$

.

3.4 Assumptions and Limitations

We assume that there is a linear relationship between the independent variables (e.g., pollster, state, sample size, poll date) and the outcome (Trump's polling percentage) in a linear model. This means that we assumed that each factor had a constant effect on Trump's poll numbers, which may contradict the reality. In addition, the complex nature of polling data may not be fully represented by a relatively simple linear model. The model is unable capture potential interactions between variables such as state, methodology, and time. At the same time, poll results for other candidates could have indirect effects on Trump's support, which were not accounted for in this analysis.

3.5 Software and Tools

We conducted our analysis by using R (R Core Team 2023) and utilized a variety of packages, including tidyverse (Wickham et al. 2019), here (Müller 2020), ggplot2 (Wickham et al. 2020), janitor (Firke 2021), rstanarm (Goodrich et al. 2020), knitr (Xie 2020), and kableExtra (Wang 2021), for data cleaning, visualization, and regression modeling.

4 Results

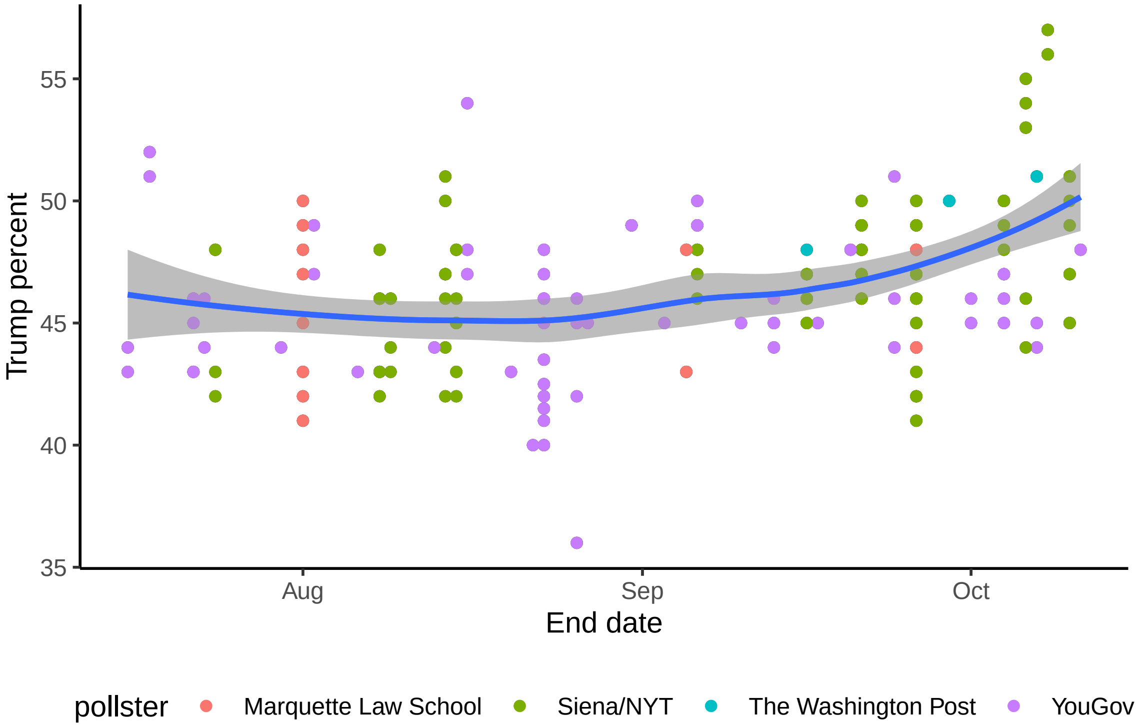

Figure 1: Simple Linear Regression Model of Trump's Support Percentage Relative to End Date

Figure 1 shows the percentage support that Donald Trump has by based on different pollsters relative to the end date of the polls during the year of 2024. The distribution of the data points shows that all pollsters besides The Washington Post had a significant number of samples that spanned from early August to late October. The Washington Post only began sampling in late September with a smaller sample size. The linear regression model shows a clear upward trend in the percentage of support Trump is getting as time passes. The specific pollsters did not show clear differences in the percentage of Trump supporters that diverged from the overall trend. However, it can be visually observed that the support data obtained from YouGov was relatively flatter and did not show a clear upward trend. On the other hand, The Washington Post obtained data that showed support percentages that were generally higher than the model predictions.

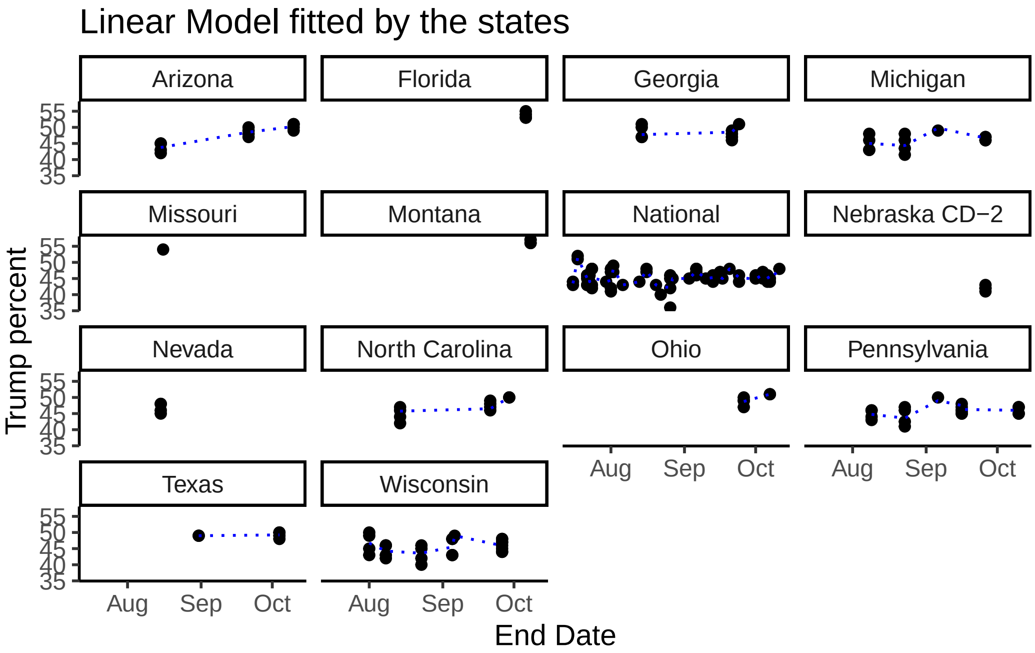

Figure 2: Simple Linear Regression Model of Trump's Support Percentage by State Relative to End Date

Figure 2 breaks down the support by state. The model shows a high variety of supporting trends. It is also seen that a few states have very limited number of data points that are insufficient to conclude a trend. In Arizona, there was a clear upward trend in the support percentage for Donald Trump. In most other states, the support for Trump was around 40% to 50%, but there was no clear upward or downward trend. Instead, Trump's support was relatively consistent over time.

Figure 3 separates the data by pollster and looks at the trends obtained per pollster. This was briefly seen in Figure 1 but can be more clearly observed in Figure 3. Based on this figure, it is clear that the data obtained by YouGov did not display a clear upward trend in Trump's support percentage. Instead, the number fluctuated largely over time but maintained at a consistent average of around 45%. All other three pollsters, including The Washington Post, Siena/NYT, and Marquette Law School obtained datapoints that displayed an upward trend in Trump's support percentage over time.

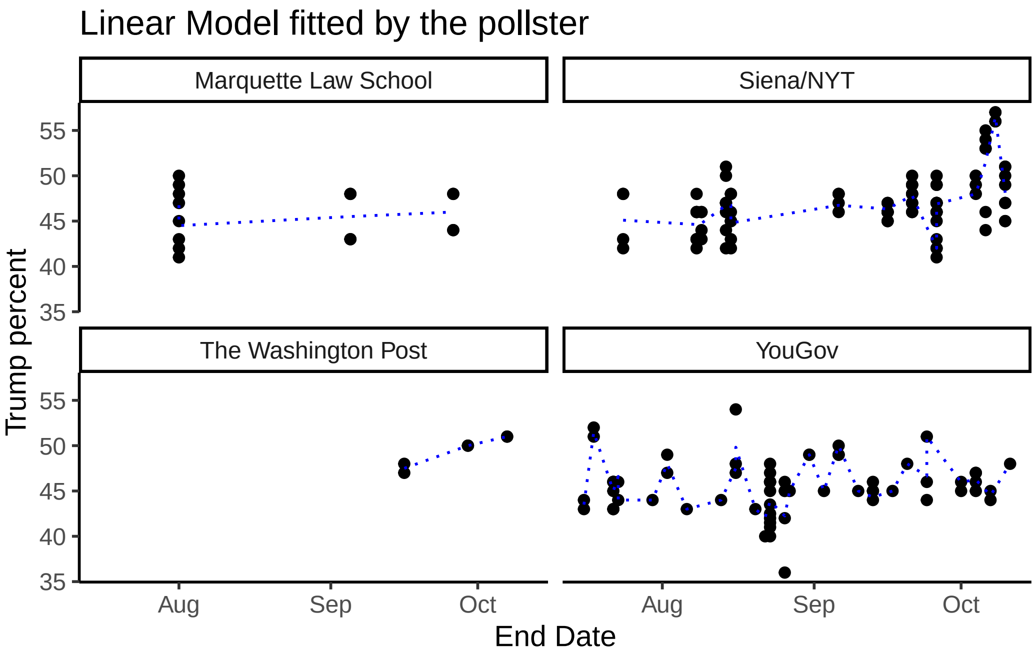

Figure 3: Simple Linear Regression Model of Trump's Support Percentage by Pollster Relative to End Date

5 Discussion

5.1 Overview

In this study, we aimed to predict the outcome of the 2024 U.S. Presidential election by analyzing polling data, with a particular focus on Donald Trump's support across different states and pollsters. Our analysis utilized polling data sourced from FiveThirtyEight, with variables such as candidate support percentages, polling methodologies, and the geographical distribution of respondents. The data was analyzed using both linear regression models and Bayesian methods, and our results indicate several key trends.

5.2 Key Findings

The data suggests that Trump's support has remained relatively stable, with polling percentages generally falling between 40% and 50% from August to October 2024. These trends were consistent across both national and state-level polls, though slight variations were observed depending on the pollster and the region in question. Pollsters with higher ratings, as determined by their numeric grade, generally showed more consistent results in predicting Trump's support levels, supporting the hypothesis that higher-quality polling organizations provide more reliable insights into voter behavior.

Additionally, the time factor (i.e., the start and end dates of polls) appeared to have a notable influence on the predictions. This aligns with the understanding that public opinion is dynamic and can shift throughout the election cycle based on external events, candidate campaigns, and media coverage. However, the state-specific analysis showed a smaller degree of fluctuation, with several key states consistently reporting stable support levels for Trump.

5.3 Implications of the Results

Based on what was found in the results, Trump's support has been consistent and on the rise. This shows that Trump has been successful with his campaigns and many public events may have been in his favor. The linear regression model suggests that the support for Trump is either increasing or consistent across different states and different pollsters. This indicates that the public opinion on his suitability for presidency is high, especially considering that his support is consistent at around 45% instead of at a low percentage.

5.4 Weaknesses and Next Steps

One of the main weaknesses of this study is the reliance on polling data alone. Polls are prone to various forms of bias, including non-response bias and sampling errors. Additionally, the changing nature of campaigns, external events, and voter turnout can lead to discrepancies between polling predictions and actual election outcomes. Another limitation is the focus on national-level polling; while state-level analysis was conducted, more granular regional data might provide deeper insights into key battleground areas.

Future research should explore the inclusion of additional variables such as media coverage, campaign spending, and real-time voter engagement metrics to complement the polling data. Furthermore, expanding the analysis to include dynamic models that account for time-varying effects could provide a more accurate forecast of electoral outcomes.

6 Conclusion

This paper examines polling data for the 2024 U.S. Presidential election, with a focus on trends in candidate support and the factors affecting polling outcomes. Using data compiled by FiveThirtyEight from sources such as YouGov, Siena/NYT, Marquette Law School, and The Washington Post, we modeled polling percentages over various time periods, pollsters, and states. The analysis was done specifically for the support of Donald Trump. To account for variations in the data and estimate uncertainty in the predictions, a linear regression model was employed.

The results showed that the support for Donald Trump was stable and increasing over time from early August to late October. The state-specific trends all show the support for Donald

Trump at around 40-50% support rate, which is high considering that there are 5 candidates currently in the running as of October of 2024. The poll-specific trends out of the pollsters that were graded at 3 or higher by FiveThirtyEight showed similar results where Trump's support was around 40-50%. Overall, this indicates stable support for Trump that's higher and still on the increasing trend. With the current situation, it is likely that Trump prevails in the presidency election. However, it is important to continue to monitor each candidates' campaigns and potential turn of events in support rates to provide the most accurate forecast possible.

A Appendix A – Idealized Methodology and Survey

A.1 Overview

YouGov is an internationally recognized polling organization known for its expertise in political forecasting. It is particularly noted for its use of the "poll-of-polls" approach, which aggregates multiple surveys to create a more accurate picture of public opinion. Relying heavily on online panel polling, YouGov provides timely, cost-effective data, particularly useful in the context of U.S. presidential elections. Below is an examination of the key components of YouGov's polling techniques.

A.2 Sampling

YouGov's polling targets U.S. likely voters, identified based on eligibility and past voting behavior, voter registration status, and intent to vote in the upcoming election. The firm uses a vast online panel, consisting of millions of U.S. residents recruited through various digital channels. This panel is strategically segmented and weighted to match the U.S. population's key demographic characteristics such as age, gender, race, education, and geographic region. Participants for YouGov's polls are recruited through a variety of methods, including targeted online advertisements, partnerships with affiliate websites, and incentive-based surveys where respondents earn rewards for participation. While these methods help build a large and diverse panel, there is an inherent risk of self-selection bias, as individuals who voluntarily join the panel may not be fully representative of the broader population. YouGov uses a stratified sampling method, establishing quotas for demographic categories like age, race, and region, and applies post-stratification weighting to ensure the sample more closely reflects the overall population.

A.3 Advantages and Disadvantages

An advantage of YouGov is cost and speed. Online panels are less expensive and faster to implement than traditional methods like phone surveys. Another advantage is targeting efficiency. They target specific groups, such as likely voters, that can be reached more easily and efficiently. A disadvantage is the non-probability sampling method used. This is because not all individuals have an equal chance of being selected, which introduces potential bias. Another disadvantage is internet accessibility. Having the surveys online leads to the issue where individuals without internet access, particularly those from older or lower-income demographics, may be underrepresented, which could affect the results.

A.3.1 Survey Design

YouGov's surveys are designed to be clear and consistent, often pre-tested on smaller groups to ensure accuracy. The questions are regularly updated to stay relevant to current political events. However, the use of online surveys limits the depth of questions, often favoring simpler, multiple-choice formats. This simplicity can lead to superficial responses, especially if participants experience survey fatigue due to frequent polling requests. Additionally, individuals without internet access are excluded, which may reduce the representativeness of certain demographic groups.

A.4 Dealing with Non-response

To address non-response and improve data accuracy, YouGov employs several strategies. They begin by offering rewards to incentivize participation, which encourages more people to complete surveys. For longer surveys, they follow up with reminders to prompt respondents to finish, further increasing completion rates. To account for any remaining gaps in representation, YouGov applies weighting adjustments, such as assigning greater weight to younger voters if they are underrepresented in the sample. Despite these efforts, some non-response bias may persist, as individuals who choose not to participate may differ systematically from those who respond, particularly in terms of political engagement.

A.5 Conclusion

YouGov's use of online panels offers significant advantages in terms of cost, speed, and flexibility, making it a valuable tool for political forecasting. While it effectively uses stratification and weighting to mitigate some biases, challenges such as self-selection, non-response bias, and the exclusion of individuals without internet access persist. Despite these limitations, YouGov remains a key player in the field of modern political polling, providing valuable insights for political analysis and forecasting.

B Appendix B – Idealized Methodology and Survey

B.1 Overview

This appendix explains an idealized methodology for the data collection phase to predict the US presidential election results. Current pollster methodologies mainly include panels, live phone calls, interactive voice responses (IVR), and text-to-web. Many characteristics of a poll can determine the quality of the poll, including but not limited to timing, suitability of methodology for the target, scope of distribution, and many more (Change Research (2022)).

The idealized methodology described below considers these aspects in the context of predicting election results and shapes the poll in a way that best suits this purpose.

B.2 Sample Frame and Sample

An idealized methodology, given a large budget of $100K, would aim to recruit participants that are representative of the entire nation. A way of recruiting participants that is most likely to gain responses that is most representative of the nation is stratified sampling by state. Eligible voters will be stratified based on the state they would be voting from. A weighted random sample would be conducted, meaning that a specific number of participants will be randomly selected from each state based on the population size of the state relative to the country. The optimal sample size for a total number of voters of more than 160M (based on previous numbers of total voters) is around 2500 people (Statista Research Department (2024), Morgan (1983)). We will aim for 3000 people to account for potential incomplete responses, no responses, or other issues with specific respondents. Based on the populations of each state, a specific number out of the 3000 total respondents will be allocated to the state. The respondents will be selected at random and contacted by email and in person. The respondents will be given a survey, which will be described in detail in the next few paragraphs. The responses will be gathered and analyzed to generate a forecast for the election results of the current election cycle.

B.3 Polling Methodology

The ideal polling method to reach the amount of people we aim to reach would be an online survey-style questionnaire. This method would allow us to reach a large amount of people in a short amount of time and have quantifiable data that's easy to analyze. This can also allow us to allocate most of the budget to sample selection to ensure that the sample is representative.

The ideal survey will be a set of questions regarding the political positions, preferences for each candidate, tendency to vote a specific candidate, and potentially other questions. Each question will be worded in a way that is easy to understand. Each question will be asked in a neutral way to ensure that ideas aren't primed for the respondents. The order of the questions is also important. It is difficult to balance primacy, recency, and moderacy effects for each participant because only one can be first and one can be last. A way to control for this is to scramble to order of questions on parties and candidates within stratified samples. For example, for two participants living in California, one would be asked about their thoughts on the Democratic party first while the other would be asked about the Republican party first. Similar order manipulation can be done for the candidates. Another aspect that can be added to the surveys is demographic information. Although this may not be directly related to forecasting the election, it could offer interesting statistical patterns that can be analyzed for other purposes in the future. These questions can include “what is your ethnicity?”, “what is your self-identified gender identity?”, and many more.

B.4 Survey

Based on the description of an idealized methodology and survey, a survey has been constructed to abide to these conditions. This has been attached in the link and sample figures below. https://forms.gle/WpFwHWs5YtPzGRTm9

References

Change Research. 2022. "Change Research 2022 Accuracy Report Methodology." https://changeresearch.com/wp-content/uploads/2023/01/Change-Research-2022-Accuracy-Report-Methodology.pdf.

Firke, Sam. 2021. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

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

Goodrich, Ben, Jonah Gabry, Imad Ali, Sam Brilleman, Jessica Lee, Lauren Kennedy, Jiqiang Guo, et al. 2020. *Rstanarm: Bayesian Applied Regression Modeling via Stan*. https://mc-stan.org/rstanarm/.

Morgan, Peter B. 1983. "Search and Optimal Sample Sizes." *The Review of Economic Studies* 50 (4): 659–75. https://doi.org/10.2307/2297768.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

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

Statista Research Department. 2024. "Number of Registered Voters in the United States." 2024. https://www.statista.com/statistics/273743/number-of-registered-voters-in-the-united-states/.

Wang, Haozhu. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. *Tidyverse: Easily Install and Load the Tidyverse*. https://CRAN.R-project.org/package=tidyverse.

Wickham, Hadley, Winston Chang, Lionel Henry, Thomas Lin Pedersen, Kohske Takahashi, Claus Wilke, Kara Woo, Hiroaki Yutani, and Dewey Dunnington. 2020. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. https://CRAN.R-project.org/package=ggplot2.

Xie, Yihui. 2020. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.