My title*

My subtitle if needed

First author Yisu Hou

November 4, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

1 Introduction

Overview paragraph

note: For the literature review/why our thing matters section talk about how other predictions don't consider the biases in polls. Our thing is something new because we attempt to account for the bias in historical predictions for our 2024 prediction. (basically, for all models and predictions, I have an adjusted version that adjusts the support percentage using the difference between 2020 polls and real election outcomes, and a normal version that just uses the percentages from 2024 polls)

There's essentially 3 different levels of analysis: using multiple linear regression to look at the candidates' support over time on the national level, using logistic regression to predict the chance of winning on the national level, using state polls data to see who is ahead in electoral college votes.

Estimand paragraph

Results paragraph

Why it matters paragraph

Telegraphing paragraph: The remainder of this paper is structured as follows. Section 2….

*Code and data are available at: https://github.com/RohanAlexander/starter_folder.

2 Data

2.1 Overview

We use the statistical programming language R (R Core Team 2023)…. Our data (Toronto Shelter & Support Services 2024)…. Following Alexander (2023), we consider…

Overview text

2.2 Measurement

Some paragraphs about how we go from a phenomena in the world to an entry in the dataset.

2.3 Outcome variable

The outcome variable is the level of support for the two leading US presidential candidates, Vice President Kamala Harris and Former President Donald Trump, recorded as a percentage of likely voters. For the national-level analysis, the support percentage represents the percentage of polled individuals who responded in favor of a specific candidate to a question that asked about their preferred presidential candidate in a national-level poll. For the regional datasets, the support percentage represents the percentage of likely voters who preferred a specific candidate in a poll that targets a specific state or congressional district of the United States. Figure 1 shows the distribution of support percentages in the aggregated polling data from 2020 and 2024 presidential elections.

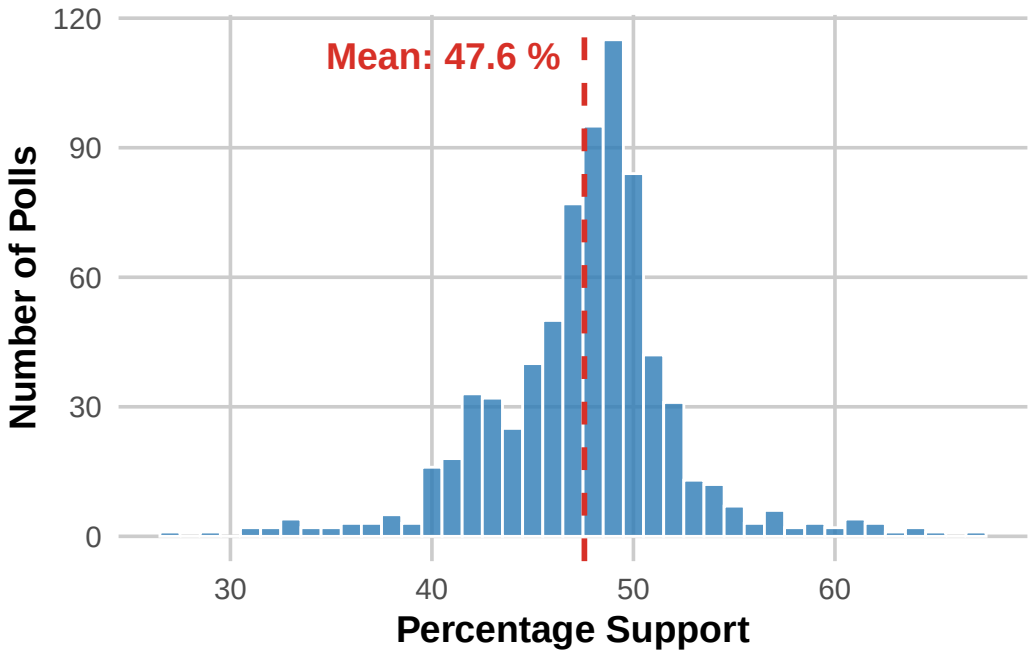

Figure 1: Percentage Support in US Presidential Polls

The aggregated polling data includes support percentages of different presidential candidates in polls with different FiveThirtyEight ratings [CITATION HERE https://abcnews.go.com/538/best-pollsters-america/story?id=105563951] across two elections. Please view [UPDATE APPENDIX] for summary statistics of sub-datasets.

As displayed by Figure 1, the distribution of support percentages is close to a normal distribution, with most data points between 45 and 52 percent.

2.4 Predictor variables

2.4.1 End Date

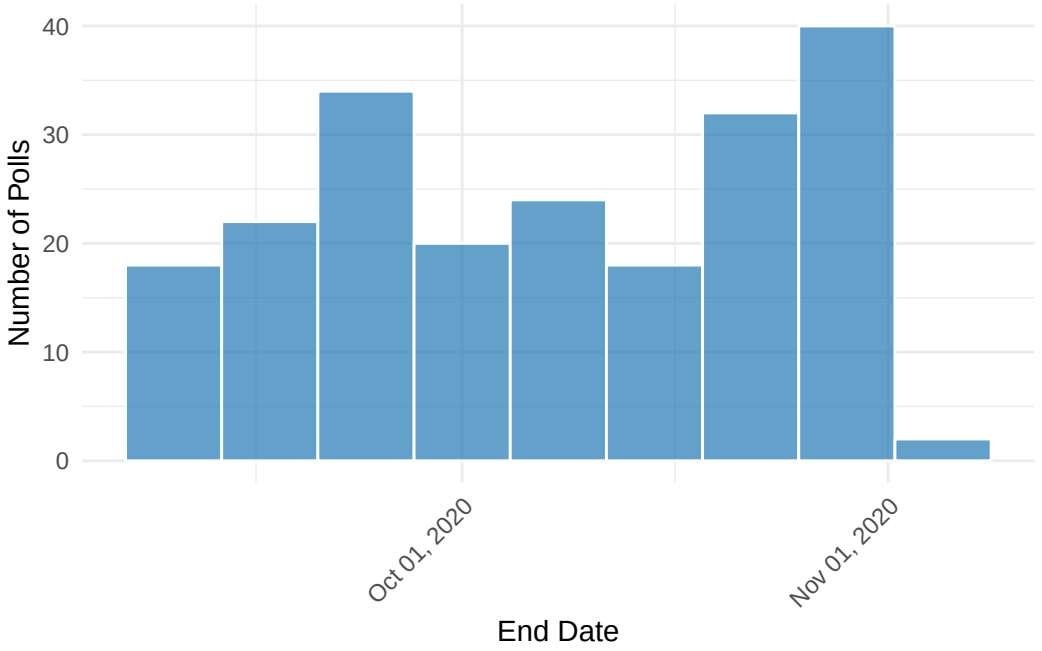

Figure 2: Distribution of Poll End Date, 2020 Cycle

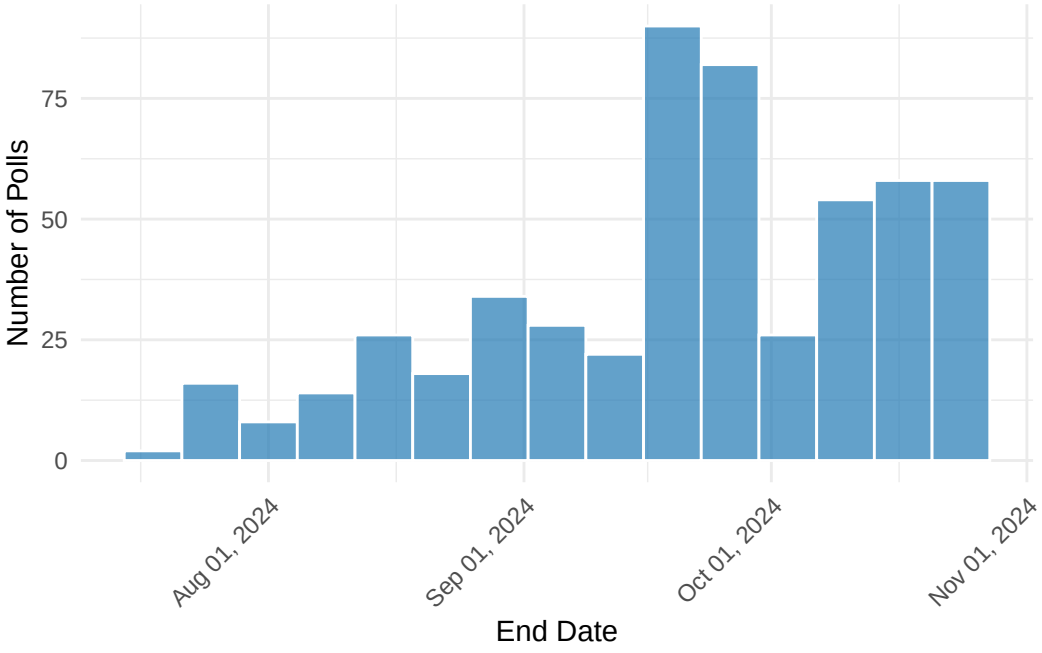

Figure 3: Distribution of Poll End Date, 2024 Cycle

2.4.2 Sponsored

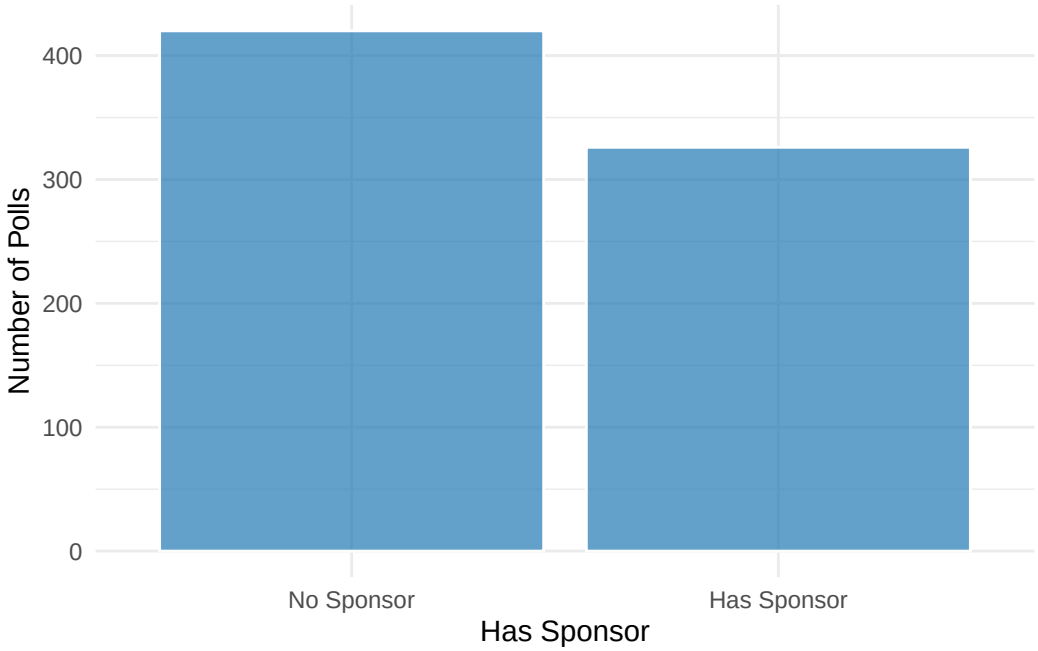

Figure 4: Number of Polls with and without Sponsors

2.4.3 Transparency Score

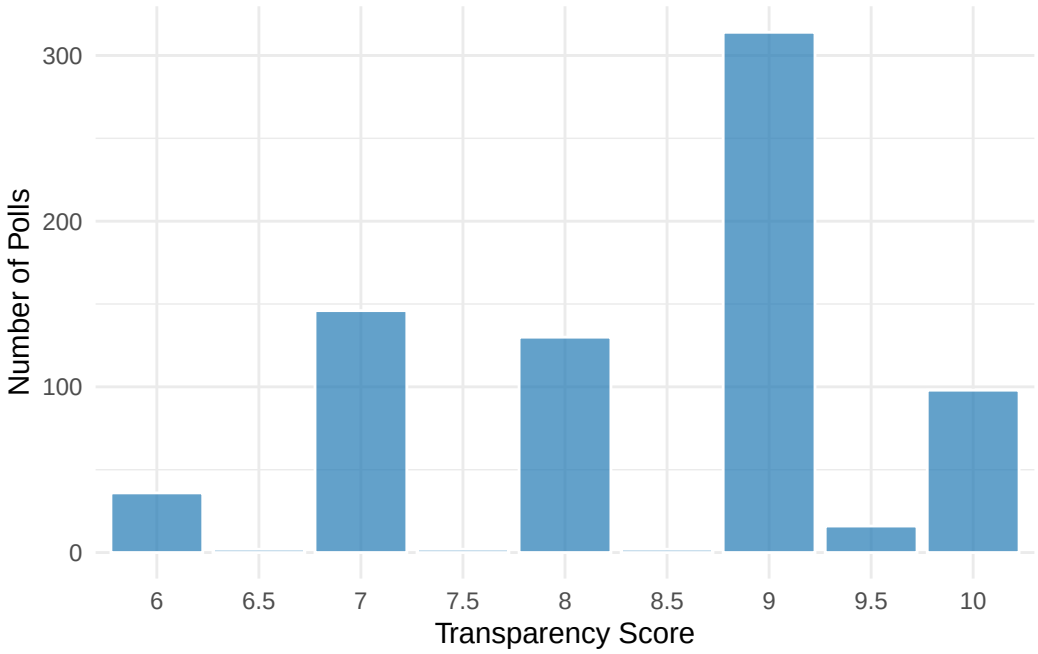

Figure 5: Distribution of Transparency Scores

2.4.4 Sample Size

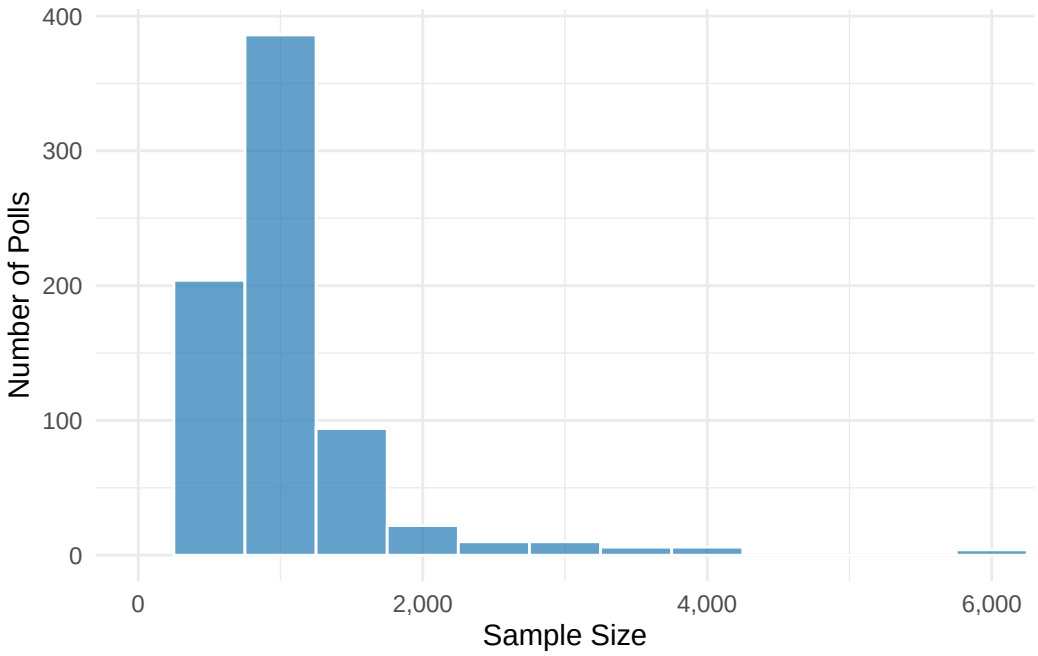

Figure 6: Distribution of Sample Sizes

3 Model

The goal of our modelling strategy is twofold. Firstly,...

Here we briefly describe the Bayesian analysis model used to investigate... Background details and diagnostics are included in Appendix B.

3.1 Model set-up

Define y_i as the number of seconds that the plane remained aloft. Then β_i is the wing width and γ_i is the wing length, both measured in millimeters.

$$y_i|\mu_i, \sigma \sim \text{Normal}(\mu_i, \sigma) \tag{1}$$
$$\mu_i = \alpha + \beta_i + \gamma_i \tag{2}$$
$$\alpha \sim \text{Normal}(0, 2.5) \tag{3}$$
$$\beta \sim \text{Normal}(0, 2.5) \tag{4}$$
$$\gamma \sim \text{Normal}(0, 2.5) \tag{5}$$
$$\sigma \sim \text{Exponential}(1) \tag{6}$$

We run the model in R (R Core Team 2023) using the `rstanarm` package of Goodrich et al. (2022). We use the default priors from `rstanarm`.

3.1.1 Model justification

We expect a positive relationship between the size of the wings and time spent aloft. In particular...

We can use maths by including latex between dollar signs, for instance θ.

4 Results

Our results are summarized in **?@tbl-modelresults**.

5 Discussion

5.1 First discussion point

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

5.2 Second discussion point

Please don't use these as sub-heading labels - change them to be what your point actually is.

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

.1 Pollster methodology overview and evaluation

The Quinnipiac University Poll conducts independent polling in swing states - to analyze their methodology, we look specifically into their October 2024 Pennsylvania polls.

Their target population is likely voters aged 18 and older in Pennsylvania. To reach this population, they use likely voters aged 18 and older with phone numbers (both landline and cell) as their sampling frame, i.e. the frame of possible subjects that they sample observations from. Quinnipiac University employed Random Digit Dialing (RDD) to generate their sample of 2,186 respondents. This dual-frame approach reflects modern communication patterns, with 1,644 cell phone and 542 landline completions. However, using phones as a sampling frame means they cannot reach voters without phone access, introducing potential coverage bias.

Their sampling approach uses stratification by Census division according to area code, meaning that they divide Pennsylvania into geographic regions before using RDD to sample within each region. This strategy ensures even geographic representation but adds complexity to the sampling process. For each selected number, they attempt contact at least three times before marking it as non-responsive. For landline calls, they ask to speak with the household member who has the next birthday, a simple but effective randomization technique. Afterwards, a series of screening questions confirm that the subject is indeed a likely voter, after which the subject's responses are then formally taken as part of the sample.

After collecting responses, Quinnipiac adjusts their data through post-stratification weighting. In this weighting, they compare their sample's demographic composition to known population benchmarks from the Census (like age, gender, education, and race distributions in Pennsylvania) and adjust the weight given to each response to match these benchmarks. For example, if their sample has too few young voters compared to Census data, responses from young voters would be weighted more heavily. While this helps correct for sampling imbalances, it can increase the variance in their estimates if the weights vary substantially.

The survey administration addresses measurement issues through its design. Live interviewers conduct all interviews, enabling question clarification and generating higher response rates compared to automated systems. However, live interviewers may introduce social desirability bias, where respondents might modify their answers to appear more socially acceptable. This becomes particularly relevant in political polling, where respondents might hesitate to express unpopular political views.

Several types of bias affect the poll's results. Self-selection bias occurs because certain types of people (typically those more politically engaged or with stronger views) are more likely to agree to participate in the survey. Non-response bias arises when people who respond differ systematically from those who don't - for instance, busier people might be less likely to answer calls, potentially underrepresenting certain occupational groups. Coverage bias means some groups (like those without phones) have no chance of being included in the sample.

Ethics-wise, the survey does a good job of informing individuals of the details of the survey, such as its purpose and how participants' data is to be used. Personally-identifiable data (i.e. name, phone number, etc.) is not collected during the survey other than to ensure that repeat numbers are not drawn, and the usage of live interviewers to conduct the survey ensures that concerns about the ethics of the survey can be voiced and answered on-the-spot. Combined with the lack of financial incentive for participants, which suggests that all participants participate of their complete free will and intention, this is an ethically sound methodology for conducting a political survey.

Finally, the poll also faces common challenges in political polls such as this one - the five-day field period (October 24-28) may miss opinion changes close to election day, and while weighting adjustments help correct for demographic imbalances, they may increase variance in the estimates if some groups need to be weighted heavily to match population benchmarks, for example.

Overall, Quinnipiac's methodology represents a balanced approach to managing practical constraints and statistical rigor in modern political polling - while some common biases are still likely to skew the poll results off the true support levels for Harris, for example, the poll uses methods such as post-stratification weighting to tradeoff biases at the cost of model variance. Using live interviewers unavoidably introduces social desirability bias, however, and significantly increases poll costs per quota. Modifying the methodology to remove this aspect of the survey would potentially reduce bias and allow for larger samples to be taken, in turn opening up possibilities for cross-validation and the such, which then reduces the effect of increased model variance on the final results.

.2 Idealised methodology

With a $100,000 budget, our approach focuses on producing accurate state-level estimates in key battleground states, which would then inform our national forecast. We prioritize Pennsylvania, Michigan, Wisconsin, Georgia, Arizona, and Nevada, allocating resources proportionally based on each state's electoral importance and expected margin of victory.

Our sampling strategy employs both probability and non-probability methods. Probability sampling (60% of budget, i.e. $60k) means every member of our target population has a known, non-zero chance of being selected - this allows us to calculate proper margins of error and make statistical inferences about the population. For this, we use dual-frame random digit dialing (RDD) for phone surveys and address-based sampling (ABS) for mail-to-web recruitment. RDD involves generating random phone numbers within active area codes, while ABS uses the U.S. Postal Service's delivery database as a sampling frame. The ABS approach helps reach households without reliable phone access. We stratify our sample by geography, demographics, and previous voting patterns to ensure representation across key subgroups - meaning we divide the population into these subgroups and sample from each independently.

For non-probability sampling (40% of budget, i.e. $40k), where respondents' selection probabilities are unknown and not everyone has a chance of being selected, we recruit through multiple online panel vendors and use targeted social media advertising to reach traditionally underrepresented groups. While this approach introduces potential selection bias because participants self-select into the sample, it helps reach younger voters who are less responsive to traditional survey methods. We implement quota sampling within these non-probability samples to match key demographic targets - for example, stopping collection from certain demographic groups once their quota is filled.

Respondent recruitment uses multiple contact methods - mail, email, text, and phone - with attempts made at different times and days to maximize response rates. We offer a $10 gift card incentive for completed surveys and provide both English and Spanish language options. This mixed-mode contact strategy helps reduce non-response bias by providing multiple ways to participate.

Data validation is crucial for maintaining quality. We cross-reference responses with voter files where available - meaning we check if respondents' self-reported registration status matches official records. We screen for duplicate responses using IP addresses and phone numbers, and implement attention checks within the survey (questions with known correct answers to ensure respondents are reading carefully). Speed checks identify rushed responses that might indicate low-quality data by flagging completions that fall below a minimum reasonable completion time, while consistency checks across related questions help identify potentially fraudulent responses by looking for logical contradictions in answers.

Our weighting approach uses post-stratification to known population benchmarks - this means we adjust the weight given to each response so that our sample matches known population characteristics. For example, if our sample has 30% college graduates but the population has 40%, we would give more weight to responses from college graduates. We include demographics (age, race, education, gender), geographic location, past voting behavior, and party registration in our weighting scheme. We produce daily estimates using a 7-day rolling average, which helps smooth out daily fluctuations while remaining responsive to real changes in voter preferences.

Finally, how ethics are handled is a crucial part of any survey methodology. In this idealised methodology, consent will be asked for at the beginning of the survey, full disclosure of how information is used will be given beforehand and no self-identifiable information will be recorded (so no names, phone numbers, etc.). The $10 incentive is enough to hopefully make it worthwile for participants' time, but also not ideally not significant enough of an incentive to make individuals suppress otherwise deal-breaking concerns with the survey purely for the sake of the incentive.

The survey instrument itself focuses on six key areas: screening questions to identify likely voters, voting intentions (including direct questions about Trump vs. Harris preferences), political preferences, demographics, issue priorities, and media consumption patterns. We've implemented this survey design in Google Forms, which can be found here: https://forms.gle/pk7vDiMHwEGLMK849

This methodology balances statistical rigor with practical constraints, while acknowledging and attempting to address the key challenges in modern political polling: declining response rates, coverage bias, and the increasing difficulty of reaching a representative sample of likely voters.

.3 Idealised survey

The survey, made using Google Forms, is linked here: https://forms.gle/pk7vDiMHwEGLMK849 Note that the questions are identical for both the phone and online surveys bar q6. A copy of the survey that is identical to the one implemented in the Google Forms above is presented below: Thank you for participating in this survey about the 2024 U.S. Presidential Election. This survey is part of a research project at the University of Toronto studying voting intentions and political attitudes.

Estimated completion time: 8-10 minutes

Your responses will be kept confidential and used only for research purposes. Email information, and any other information that may personally identify you, is not gathered. You may skip any questions you prefer not to answer, though complete responses are most helpful for our research.

For questions or concerns about this survey, please contact: andrew.goh@mail.utoronto.ca

SCREENING SECTION: Q1. Are you 18 years of age or older?

Yes No [END SURVEY]

Q2. Are you a U.S. citizen?

Yes No [END SURVEY]

Q3. Are you registered to vote at your current address?

Yes No Not sure [If No or Not sure: Do you plan to register before the November 2024 election?]

VOTING INTENTION: Q4. How likely are you to vote in the 2024 presidential election?

Definitely will vote Probably will vote Might or might not vote Probably will not vote Definitely will not vote

Q5. If the 2024 presidential election were held today, and the candidates were Kamala Harris (Democrat) and Donald Trump (Republican), who would you vote for?

Kamala Harris Donald Trump Another candidate (please specify) Would not vote Not sure

ATTENTION CHECK: Q6. To ensure you're reading carefully, please select "Somewhat disagree" for this question: "I enjoy following political news."

Strongly agree Somewhat agree Somewhat disagree Strongly disagree No opinion

POLITICAL PREFERENCES: Q7. Generally speaking, you consider yourself a:

Democrat Republican Independent Something else (please specify)

Q8. How would you rate the current state of the U.S. economy?

Very poor Poor Fair Good Excellent

ISSUE PRIORITIES: Q9. Which ONE of the following issues is most important to you when deciding how to vote?

Economy and jobs Immigration Healthcare Climate change Crime and public safety Education National security Abortion rights Gun policy Something else (please specify)

For each of the following issues, please indicate whether you think Kamala Harris or Donald Trump would do a better job handling it:

Q10. Economy and jobs:

Kamala Harris would do better Donald Trump would do better No difference Not sure

Q11. Human rights and freedom of speech: [Same options] Q12. Abortion: [Same options] Q13. Healthcare: [Same options] Q14. Immigration: [Same options] Q15. National security: [Same options]

MEDIA CONSUMPTION: Q16. Where do you most often get your news about politics? (Select all that apply)

Network TV news (ABC, CBS, NBC) Cable TV news (CNN, Fox News, MSNBC) Local TV news Radio Print newspapers News websites Social media Friends and family Other (please specify)

Q17. How many hours per day do you typically spend following news about politics?

Less than 1 hour 1-2 hours 2-4 hours More than 4 hours

DEMOGRAPHICS: Q18. What is your age?

18-24 25-34 35-44 45-54 55-64 65 or older

Q19. What is your gender?

Male Female Non-binary/Other Prefer not to say

Q20. What is your race/ethnicity? (Select all that apply)

White Black or African American Hispanic or Latino Asian Native American Other (please specify) Prefer not to say

Q21. What is the highest level of education you have completed?

Less than high school High school graduate Some college Associate's degree Bachelor's degree Graduate degree Prefer not to say

Q22. What was your total household income before taxes in 2023?

Under $25,000 $25,000-$49,999 $50,000-$74,999 $75,000-$99,999 $100,000-$149,999 $150,000 or more Prefer not to say

CONSISTENCY CHECK: Q23. Looking ahead to November 2024, if Kamala Harris is the Democratic nominee and Donald Trump is the Republican nominee, how do you think you will vote?

Kamala Harris Donald Trump Another candidate (please specify) Would not vote Not sure

[END OF SURVEY] Thank you for completing this survey about the 2024 U.S. Presidential Election. Your responses will help us better understand voter preferences and political attitudes across the country. If you have any questions about this research or would like to be informed about the results, please contact andrew.goh@mail.utoronto.ca. Your time and participation is greatly appreciated.

A Additional data details

B Model details

B.1 Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows…

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows…

B.2 Diagnostics

?@fig-stanareyouokay-1 is a trace plot. It shows… This suggests…

?@fig-stanareyouokay-2 is a Rhat plot. It shows… This suggests…

References

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

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "rstanarm: Bayesian applied regression modeling via Stan." https://mc-stan.org/rstanarm/.

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

Toronto Shelter & Support Services. 2024. *Deaths of Shelter Residents*. https://open.toronto.ca/dataset/deaths-of-shelter-residents/.