USA pollster*

CHANGE SUBTITLE

Chris Yong Hong Sen

November 3, 2024

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

1 Introduction

Overview paragraph

Estimand paragraph

Results paragraph

Why it matters paragraph

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

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.

*Code and data are available at: https://github.com/Monoji77/USA-pollster.

2.3 Predictor Variables

Add graphs, tables and text. Use sub-sub-headings for each outcome variable or update the subheading to be singular.

Some of our data is of penguins (Figure 1), from Horst, Hill, and Gorman (2020).

```
# A tibble: 270 x 4
   poll_id pollster_id candidate_name winner
     <dbl>       <dbl> <chr>           <dbl>
 1   88766         383 Kamala Harris       1
 2   88766         383 Donald Trump        0
 3   88767         383 Kamala Harris       1
 4   88767         383 Donald Trump        0
 5   88764         770 Kamala Harris       1
 6   88764         770 Donald Trump        0
 7   88769        1554 Kamala Harris       0
 8   88769        1554 Donald Trump        1
 9   88739         770 Kamala Harris       1
10   88739         770 Donald Trump        0
# i 260 more rows
```

Talk more about it.

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

```
raw_pollster_data <- read_csv('../data/01-raw_data/raw_pollster_data.csv')
```

```
Rows: 15801 Columns: 52
-- Column specification --------------------------------------------------------
Delimiter: ","
chr (25): pollster, sponsors, display_name, pollster_rating_name, methodolog...
dbl (16): poll_id, pollster_id, pollster_rating_id, numeric_grade, pollscore...
num  (1): sponsor_ids
lgl (10): endorsed_candidate_id, endorsed_candidate_name, endorsed_candidate...

i Use `spec()` to retrieve the full column specification for this data.
i Specify the column types or set `show_col_types = FALSE` to quiet this message.
```

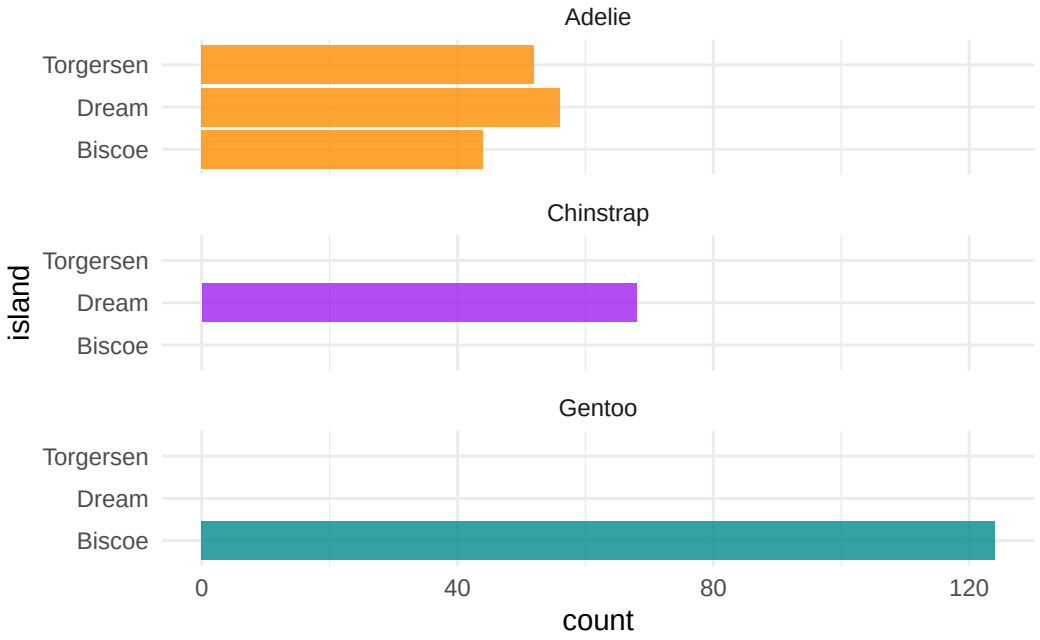


Figure 1: Bills of penguins

```
raw_pollster_data |>
  colnames()
```

```
 [1] "poll_id"                   "pollster_id"
 [3] "pollster"                  "sponsor_ids"
 [5] "sponsors"                  "display_name"
 [7] "pollster_rating_id"        "pollster_rating_name"
 [9] "numeric_grade"             "pollscore"
[11] "methodology"               "transparency_score"
[13] "state"                     "start_date"
[15] "end_date"                  "sponsor_candidate_id"
[17] "sponsor_candidate"         "sponsor_candidate_party"
[19] "endorsed_candidate_id"     "endorsed_candidate_name"
[21] "endorsed_candidate_party"  "question_id"
[23] "sample_size"               "population"
[25] "subpopulation"             "population_full"
[27] "tracking"                  "created_at"
[29] "notes"                     "url"
[31] "url_article"               "url_topline"
[33] "url_crosstab"              "source"
```

```
[35] "internal"                  "partisan"
[37] "race_id"                   "cycle"
[39] "office_type"               "seat_number"
[41] "seat_name"                 "election_date"
[43] "stage"                     "nationwide_batch"
[45] "ranked_choice_reallocated" "ranked_choice_round"
[47] "hypothetical"              "party"
[49] "answer"                    "candidate_id"
[51] "candidate_name"            "pct"
```

```
raw_pollster_data |>
  select(party, candidate_name) |>
  unique() |>
  arrange(party, candidate_name)
```

```
# A tibble: 70 x 2
   party candidate_name
   <chr> <chr>
 1 CON   Joel Skousen
 2 CON   Randall A. Terry
 3 DEM   Al Gore
 4 DEM   Amy Klobuchar
 5 DEM   Andy Beshear
 6 DEM   Bernie Sanders
 7 DEM   Cory A. Booker
 8 DEM   Elizabeth Ann Warren
 9 DEM   Gavin Newsom
10 DEM   Gretchen Whitmer
# i 60 more rows
```

```
raw_pollster_data |>
  select(party, candidate_name) |>
  unique() |>
  filter(grepl('Donald', candidate_name))
```

```
# A tibble: 3 x 2
  party candidate_name
  <chr> <chr>
1 REP   Donald Trump
2 IND   Donald Trump
3 REP   Donald Trump Jr.
```

```
# remove trailing time stamp
raw_pollster_data$created_at <- sapply(strsplit(raw_pollster_data$created_at, '\\s+'), '[', 1

time_var <- c('start_date', 'end_date', 'created_at')

# get day, month, year in character data type
raw_pollster_data$month <- sapply(strsplit(raw_pollster_data$created_at, '/'), '[', 1)
raw_pollster_data$day <- sapply(strsplit(raw_pollster_data$created_at, '/'), '[', 2)
raw_pollster_data$year <- sapply(strsplit(raw_pollster_data$created_at, '/'), '[', 3)

# observe how year is not cleaned
unique(raw_pollster_data$year)
```

```
[1] "24"   "2024" "23"   "2023" "22"   "2022" "21"   "2021"
```

```
# correct different formats for year
raw_pollster_data$year <- case_when(
  raw_pollster_data$year %in% c('24', '2024') ~ 2024,
  raw_pollster_data$year %in% c('23', '2023') ~ 2023,
  raw_pollster_data$year %in% c('22', '2022') ~ 2022,
  raw_pollster_data$year %in% c('21', '2021') ~ 2021,
  )

# check that year is cleaned
unique(raw_pollster_data$year)
```

```
[1] 2024 2023 2022 2021
```

```
raw_pollster_data <- raw_pollster_data |>
  mutate(day = ifelse(day < '10', paste0('0', day), day),
         month = ifelse(month < '10', paste0('0', month), month),
         created_at = as.Date(paste(year, month, day, sep='-')))

# get day, month, year in character data type
raw_pollster_data$month <- sapply(strsplit(raw_pollster_data$start_date, '/'), '[', 1)
raw_pollster_data$day <- sapply(strsplit(raw_pollster_data$start_date, '/'), '[', 2)
raw_pollster_data$year <- sapply(strsplit(raw_pollster_data$start_date, '/'), '[', 3)

# observe how year is not cleaned
unique(raw_pollster_data$year)
```

```
[1] "24"   "2024" "23"   "2023" "22"   "2022" "21"   "2021"
```

```
# correct different formats for year
raw_pollster_data$year <- case_when(
  raw_pollster_data$year %in% c('24', '2024') ~ 2024,
  raw_pollster_data$year %in% c('23', '2023') ~ 2023,
  raw_pollster_data$year %in% c('22', '2022') ~ 2022,
  raw_pollster_data$year %in% c('21', '2021') ~ 2021,
  )

# check that year is cleaned
unique(raw_pollster_data$year)
```

```
[1] 2024 2023 2022 2021
```

```
raw_pollster_data <- raw_pollster_data |>
  mutate(day = ifelse(day < '10', paste0('0', day), day),
         month = ifelse(month < '10', paste0('0', month), month),
         start_date = as.Date(paste(year, month, day, sep='-')))

# get day, month, year in character data type
raw_pollster_data$month <- sapply(strsplit(raw_pollster_data$end_date, '/'), '[', 1)
raw_pollster_data$day <- sapply(strsplit(raw_pollster_data$end_date, '/'), '[', 2)
raw_pollster_data$year <- sapply(strsplit(raw_pollster_data$end_date, '/'), '[', 3)

# observe how year is not cleaned
unique(raw_pollster_data$year)
```

```
[1] "24"   "2024" "23"   "2023" "22"   "2022" "21"   "2021"
```

```
# correct different formats for year
raw_pollster_data$year <- case_when(
  raw_pollster_data$year %in% c('24', '2024') ~ 2024,
  raw_pollster_data$year %in% c('23', '2023') ~ 2023,
  raw_pollster_data$year %in% c('22', '2022') ~ 2022,
  raw_pollster_data$year %in% c('21', '2021') ~ 2021,
  )

# check that year is cleaned
unique(raw_pollster_data$year)
```

```
[1] 2024 2023 2022 2021
```

```
raw_pollster_data <- raw_pollster_data |>
  mutate(day = ifelse(day < '10', paste0('0', day), day),
         month = ifelse(month < '10', paste0('0', month), month),
         end_date = as.Date(paste(year, month, day, sep='-')))

raw_pollster_data |>
  select(start_date, end_date, created_at) |>
  summarise(min(start_date))
```

```
# A tibble: 1 x 1
  `min(start_date)`
  <date>
1 2021-04-03
```

```
sum(is.na(raw_pollster_data$partisan))
```

```
[1] 14423
```

```
raw_pollster_data |>
  filter(!is.na(state))
```

```
# A tibble: 8,264 x 55
   poll_id pollster_id pollster     sponsor_ids sponsors            display_name
     <dbl>       <dbl> <chr>              <dbl> <chr>               <chr>
 1   88766         383 PPP                  618 Northwest Progress~ Public Poli~
 2   88766         383 PPP                  618 Northwest Progress~ Public Poli~
 3   88767         383 PPP                  618 Northwest Progress~ Public Poli~
 4   88767         383 PPP                  618 Northwest Progress~ Public Poli~
 5   88769        1554 RMG Research        2178 Napolitan News Ser~ RMG Research
 6   88769        1554 RMG Research        2178 Napolitan News Ser~ RMG Research
 7   88771        1741 ActiVote              NA <NA>                ActiVote
 8   88771        1741 ActiVote              NA <NA>                ActiVote
 9   88770        1741 ActiVote              NA <NA>                ActiVote
10   88770        1741 ActiVote              NA <NA>                ActiVote
# i 8,254 more rows
# i 49 more variables: pollster_rating_id <dbl>, pollster_rating_name <chr>,
#   numeric_grade <dbl>, pollscore <dbl>, methodology <chr>,
#   transparency_score <dbl>, state <chr>, start_date <date>, end_date <date>,
```

```
#   sponsor_candidate_id <dbl>, sponsor_candidate <chr>,
#   sponsor_candidate_party <chr>, endorsed_candidate_id <lgl>,
#   endorsed_candidate_name <lgl>, endorsed_candidate_party <lgl>, ...
```

```
raw_pollster_data |>
  select(office_type) |>
  unique()
```

```
# A tibble: 1 x 1
  office_type
  <chr>
1 U.S. President
```

```
raw_pollster_data |>
  summarise(min(pollscore, na.rm=T), max(pollscore, na.rm=T))
```

```
# A tibble: 1 x 2
  `min(pollscore, na.rm = T)` `max(pollscore, na.rm = T)`
                        <dbl>                       <dbl>
1                        -1.5                         1.7
```

```
raw_pollster_data |>
  summarise(min(transparency_score, na.rm=T), max(transparency_score, na.rm=T))
```

```
# A tibble: 1 x 2
  `min(transparency_score, na.rm = T)` `max(transparency_score, na.rm = T)`
                                 <dbl>                                <dbl>
1                                    0                                   10
```

```
raw_pollster_data |>
  summarise(min(numeric_grade, na.rm=T), max(numeric_grade, na.rm=T))
```

```
# A tibble: 1 x 2
  `min(numeric_grade, na.rm = T)` `max(numeric_grade, na.rm = T)`
                            <dbl>                           <dbl>
1                             0.5                               3
```

```
raw_pollster_data |>
  summarise(min(start_date, na.rm=T), max(start_date, na.rm=T),
            min(end_date, na.rm=T), max(end_date, na.rm=T))
```

```
# A tibble: 1 x 4
  `min(start_date, na.rm = T)` max(start_date, na.rm = ~1 min(end_date, na.rm ~2
  <date>                       <date>                     <date>
1 2021-04-03                   2024-10-16                 2021-04-07
# i abbreviated names: 1: `max(start_date, na.rm = T)`,
#   2: `min(end_date, na.rm = T)`
# i 1 more variable: `max(end_date, na.rm = T)` <date>
```

Talk way more about it.

2.4 Predictor variables

Add graphs, tables and text.

Use sub-sub-headings for each outcome variable and feel free to combine a few into one if they go together naturally.

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 C.

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 Table 1.

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.

Table 1: Explanatory models of flight time based on wing width and wing length

	First model
(Intercept)	1.12
	(1.70)
length	0.01
	(0.01)
width	−0.01
	(0.02)
Num.Obs.	19
R2	0.320
R2 Adj.	0.019
Log.Lik.	−18.128
ELPD	−21.6
ELPD s.e.	2.1
LOOIC	43.2
LOOIC s.e.	4.3
WAIC	42.7
RMSE	0.60

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Patriot Polling: Wisconsin Presidential Analysis Data Collection

A.1 Methodology: Overview

From 12 to 14 October 2024, Patriot Polling has conducted phone surveys in the state of Wisconsin of the United States of America, which borders Minnesota, Iowa, Michegan, and Illinois. Their target population are voters in Wisconsin. For sample of landlines which are characterised by households, an equal number of phone numbers are randomly selected across every landline block in Wisconsin using Random Digit Dialing (RDD). As for the sample of mobile numbers, which are characterised by an individual with access to a digital mobile device, the sample was purchased from a consumer contact number database. For this poll, a total of 803 respondents has completed the survey either from contact through the landline or their personal contact number. (Ruggieri 2024)

Ruggieri (2024) conducted polling in such a way where a randomly contacted respondent would hear pre-recorded voice messages and users are able to interact with the automated phone system. This system is called Interactive voice response (IVR) which can handle outbound calls more systematically since the same questions would be asked sequentially in the survey. There were 2 questions provided by Patriot Polling that were asked to respondents. Both questions begin with 'How will you vote for president?'. This is followed by the candidates name as it appears in the voting ballot. There was no indication whether this was all the questions that were asked but there were tables provided by the pollster showing the breakdown by sex, education, preferred political affiliation and income. Therefore, it would be safe to assume that there were questions related to the socio-demographic characteristics of the respondents even though we don't know the specific questions asked.

A.2 Methodology: Population

The targeted population of this poll was not explicitly stated. While the pollster did conduct phone surveys, they did not mention any inclusion or exclusion criteria. Furthermore, the use of IVR during phone surveys and lack of information provided regarding the questions asked as discussed in Section A.1, it will be difficult to guarantee that the polls are targeting all eligible voters in Wisconsin. Given that nature of the polling article focuses on Wisconsin, it would be safe to assume that the targeted population would be eligible voters above 18 years old in Wisconsin.

A.3 Methodology: Sampling Frame

Since phone surveys were conducted on both landlines and smartphones, the sampling frame of this poll would be eligible voters above 18 years old in Wisconsin with a working landline or possess a phone with a working contact number.

A.4 Strengths

This automated method of phone surveys is a lot more convenient and efficient than in person type of surveys since it removes the need for a surveyor to visit respondents face to face and also removes the need for a surveyor to contact multiple respondents over the phone. This decreases labour effort and cost.

Furthermore, with real-time feedback from phone surveys, the list of questions can be easily and quickly adjusted to ensure the ethical integrity of the survey being conducted. According to Ruggieri (2024), this survey took place in 2 days which could have taken a lot longer if other in-person type of surveying methods were used in this research.

Particularly, the polling company used a repertoire of RDD sampling and stratified sampling on landlines to ensure that every landline block (strata) has equal numbers of respondents to be included in the final sample. This helps to ensure diversity from the sample by including different geographical landline blocks, reducing sampling bias. The purchasing of mobile phones from phone number databases abates sampling bias in areas where a particular demographic - younger people who are fully dependant on mobile phones and not landlines - would otherwise be underrepresented if only RDD sampling was conducted on landlines.

A.5 Limitations and weakness

Firstly, there was a lack of information regarding how RDD was performed for the landline sample, how they handled non-working landlines, as well as the total sample size attributed from landline and mobile phone surveys.

The handling of no-response entities were also not explicitly mentioned by Patriot Polling. This could translate to respondents who were chosen but did not pick up their mobile phones/landlines as it was a busy period in the day or due to potential heightened awareness of phone based scams. This means that the final sample obtained might not be representative of the Wisconsin voter population.

Also, the initial phrasing to the second question can very easily confuse respondents. Although the initial phrasing is followed by the senate candidate's name, the initial phrasing 'How will you vote for president' still uses the wrong word 'president' instead of 'senate'. This does not reflect a high quality survey, which could lead to decreased trust from the respondent to

Patriot Polling, in turn potentially resulting in a premature end to the surveys, voiding the initial response.

Overall, the biggest weakness in the methodology would be the purchasing of phone numbers from a consumer database as it introduces selection bias. Individuals from middle to higher socio-economic status would have greater purchasing power than their lower income counterparts, which means individuals from lower socio-economic status may not have phones. In addition, elderlies within Wisconsin may not necessarily have mobile phones due to the lack of skills to use such technologies. Also, individuals would've opted for do-not-call registries, opted out of marketing databases or simply have prepaid phones which do not link to their personal information. This could mean an under representation of individuals from the said groups.

A.6 Simulation exploring weaknesses in Patriot's Polling's methodology

To illustrate the biggest concern raised in Section A.5, we will be simulating a Patriot's Polling's methodology by using US Census data to obtain random our pool of samples that are representive of the population of Wisconsin with information about their age, race and county. Since the raw data provided by IPUM (Ruggles et al. 2024) encoded states and counties, we combined the raw data with a supplementary excel file provided by IPUM to obtain the actual county name. We will be using the variable $CISMRTPHN$ from the census data which represents whether the particular respondent of the census data has a smartphone. This variable will be used to mirror a Patriot's Polling methodology of obtaining personal contact number by purchasing from an online database. We used this approach as only Wisconsin respondents with smartphones can possibly be included in a smartphone contact number database.

For the sake of discussion, we will be obtaining 2 samples. **Sample A will include respondents above the voting age of 18 in Wisconsin regardless of whether they possess a smartphone while sample B will only include respondents above the voting age of 18 in Wisconsin who have smartphones.** Sample B will mimic purchasing phone numbers from a database.

Table 2 shows us the proportion of white, black and asian races from sample A while Table 3 shows us that from sample B. We notice that black individuals from Wisconsin are underrepresented (8.7% vs 9.4%). According to Johnson (2018), black poverty rates in Wisconsin are 2.5 times higher than the overall Wisconsin poverty rate. This supports the finding that if we were to include only individuals with smartphones, certain racial groups might be underrepresented. Likewise, we are able to observe Asians being overrepresented (4.4% vs 3.8%). Ricketts and Kent (2024) exerts that for various educational attainment levels, Asians are earning more than individuals who belong to white, black or hispanic households. It is unsurprising that Asians are being overrepresented as higher wealth among the Asian population in Wisconsin would translate to higher purchasing power to purchase smartphones. Asians would naturally be more likely to be included in a sample where only smartphone users are considered, by

extension, Asians would be more likely to be included in a sample obtained from a consumer database containing user contact numbers.

Table 2: proportion of races from non-smartphone-discriminating sample

race	count	proportion
Asian	38	3.8%
Black	94	9.4%
White	768	76.8%
others	100	10%

Table 3: proportion of races from smartphone-only sample

race	count	proportion_chosen
Asian	44	4.4%
Black	87	8.7%
White	759	75.9%
others	110	11%

This becomes an issue because underrepresentation of black communities and overrepresentation of asian communities in Wisconsin would not reflect the actual voting preferences of a particular racial demographic, resulting in sampling bias, potentially misleading people who interpret a Patriot Polling's poll on Wisconsin or even their written articles.

Next, Table 4 shows us the the percentage difference in individuals chosen categorised by age groups from sample A and B. We are immediately alerted to a 20% decrease in the number of elderlies chosen from sampling method B compared to sampling method A. As of 2024, elderlies above the age of 65 are either from the baby boomer generation or silent generation, who are known to experience difficulty when using smartphones. Pew Research Center (2024) shows us that 97% of Americans between 18 to 49 have smartphones, 89% for that of age groups 50 to 64, and a lower 76% of the American elderlies above 65 year old who own smartphones. Therefore, by proceeding with sample B (parallel to purchasing phone numbers from consumer database) for which the inclusion critera selects individuals who own smartphones, we might be discounting the votes of elderly population, resulting in selection bias.

Table 4: percentage difference in individuals chosen from sample A and B by age groups

age bin	sample A	sample B	perc diff
above 65	238	191	-20%
18-30	191	206	+8%
30-65	571	603	+6%

B Idealised methodology for presidential election forecasting

B.1 Context of presidential election

US General elections runs every 4 years. The current state of the US electoral system is a two-party system where the republican and democratic parties dominate the political space (U.S. Embassy & Consulate in the Kingdom of Denmark, n.d.). During general elections, eligible voters in America are actually voting for a group of people called electors within their state. Elector candidates from 38 out of 50 states are usually pledged to support a specific political party. While there had been 'faithless electors' who cast a vote to the opposing party, they account for lesser than 1% of the elector population from the past 58 presidential elections. 'Faithless electors' had never affected the outcome of the elections (Otis 2024). If a elector candidate receives majority of votes by voters in their state, they gain the privilege of all electoral votes allocated for that state. All electors make up the electoral college. The electoral college would then casts their votes to their preferred presidential nominee, for which the majority would determine who becomes president.

B.2 Split electoral votes in Nebraska and Maine

In the states of Nebraska and Maine, individuals are voting not at the state level, but at the congressional district level. Nebraska has 3 districts while Maine has 2. One elector is chosen for each congressional district, and an additional 2 electors is given for the statewide majority. Unlike the other states which operate on a winner-takes-all, this difference in voting allows for greater variation in the outcome of the general elections (Encyclopædia Britannica, Inc 2024).

B.3 Idealised methodology: Pilot

The purpose of a pilot testing is to ensure a well-designed election forecasting survey. Due to the nature of the general election characterised by the votes of the electoral college, random sampling has to be performed across all 50 states of America. $5,000 should go towards a pilot test obtaining 200 respondents per state. This pilot test would target battleground states such as Wisconsin as well as Maine and Nebraska to test specific issues around a split vote between the two presidential candidate. Begin by using stratified sampling that divides each state population into strata by various socio demographic factors (e.g. gender, highest educational attainment, race, income), then conduct systematic random sampling within each strata. An optional $5 digital or mailed voucher can be used to incentivise respondents to complete the survey.

Respondents should be recruited quickly using by random digit dialing (RDD) to reduce selection bias. Firstly, obtain the list of possible area codes corresponding to the first 6 digits

of a contact number within a state. Then randomly sample the remaining 4 digits. Coupled with the use of an automated interactive voice response system, this would increase efficiency while taking note of non-response from either premature end to phone surveys or non-working landlines/phonelines. This also decreases response bias as unlisted contact numbers would be included in the sampling frame as well. Theoretically, a larger than expected sample has to be recruited in order to account for contact numbers that are not in use. Also, it is important to distinguish landline RDD sampling and cellphone RDD sampling as they represent different communties within the population as mentioned in Section A.4 and Section A.5

As for the contents of the phone survey, begin by providing detailed, succint information. We would introduce our company name, the purpose of the phone survey, and how only their polling preference is recorded and how their contact information would be used (but not saved) for the optional voucher redeeming purposes. If the phone survey is conducted in the first congressional district of Maine, explicitly mention that this survey is used to obtain polling preferences among voters within the first congressional district of Maine, and there are no political affiliations. It is paramount to obtain consent from the respondent. Therefore, have an option to let a respondent opt out of the survey without any form of discrimination. If a respondent chooses to continue the survey, we can ask them for their highest educational background, race, gender and/or other socio demographic information with the option to indicate non-response for any particular question. If a respondent successfully completes the phone survey, record the information into a secure, password protected excel sheet for further data analysis. It is also important to honor the incentives should the respondent complete the survey by using only the phone number. At the end, thank the respondent for their coorperative participation.

In order to expand the sampling frame to include voters who may not prefer to do a phone surveys, we can conduct a Google forms survey as well which is free of charge. The format of the form should mimic that of the RDD random sampling method. Refer to this google form for an example of the survey questions aksed.

After obtaining all pilot samples, validate the pilot data by checking for any patterns or outliers. In particular, we have to pay special attention to ensure that the samples received are balanced and representive of the voting population in that particular state by cross referencing to census data available from IPUMS US census data (Ruggles et al. 2024). If the data is not representive of a particular population, it might be necessary to conduct in person surveys in regions of the state where certain demographics are underrepresented.
These forms of recruiting respondents should mimic the actual sampling conducted on a larger scale after a successful pilot.

B.4 Idealised methodology: Large scale sampling

By refining our phone surveys and google form surveys, we can expand the surveys to all states in America. For this we budget out $80,000 to obtain between 2000 to 5000 samples across

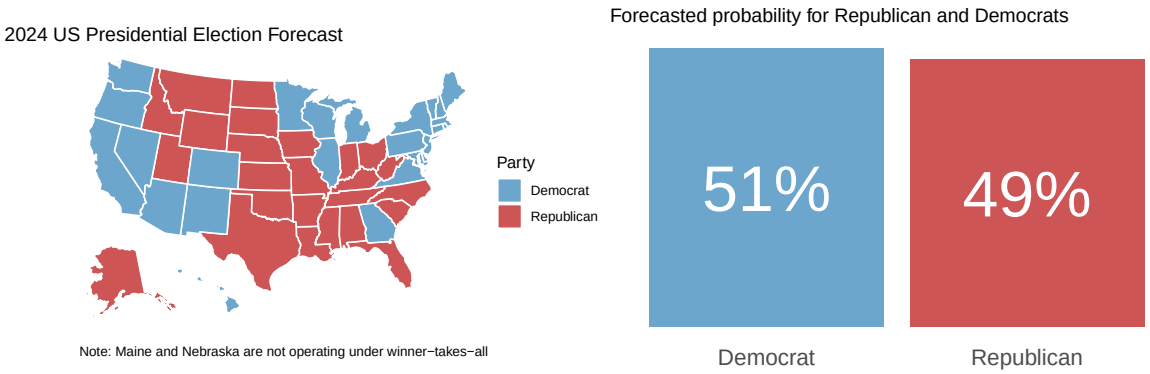
each state, depending on the state's population. For the states of Nebraska and Maine, we will be performing both analysis for individual congressional districts and statewide majority votes to determine the number of estimated electoral votes as described in Section B.2. Particularly for all battleground states as seen from the past 2020 elections like Arizona, Florida and Maine's second congressional district, we will be oversampling to ensure that the preferences of sampled voters are better representive of the overall voter population in their particular state or congressional district, leading to reduced sampling error and more precise forecasting of the presidential elections.

If the pilot discovers regions where certain demographics are underrepresented (e.g. poorer communities without digital communication devices nor landlines), surveyors would have to be employed in these regions to obtain samples directly.

Ultimately, the algorithm to forecast the presidential winner involves obtaining the majority vote based on samples obtained within each state (and districts in Nebraska and Maine). Then based on the number of electors assigned to that region, we will sum up the number of electors affiliated to the presidential candidates. This sum divided by the 538 total electors would be the predicted probability that a presidential candidate would win in the upcoming election. $10,000 will be spent on the cost of surveyors. The remaining $5,000 from our budget will be used as contigency funds.

B.5 Presentation of polling results

We can map out the results to give a visual representation of how the votes mapped out to each state. Figure 2a provides a quick overview to an estimated presidential candidate preference in each state in America from the obtained samples. The colour distinction makes it quickly identifiable the majority within each state. Figure 2b provides the final predicted probability of how likely a party (presidential candidate) would be elected president. The following map utilises data obtained from Cook Political Report (Wasserman et al., n.d.). Tigris library (Walker 2024) coupled with tidyverse library (Wickham et al. 2019) was used to generate the US map shown in Figure 2a. Tidyverse library was also used to generate the bar chart shown in Figure 2b.



(a) Map forecast

(b) Overall forecast

Figure 2: Forecast of US presidential elections

C Model details

C.1 Posterior predictive checks

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

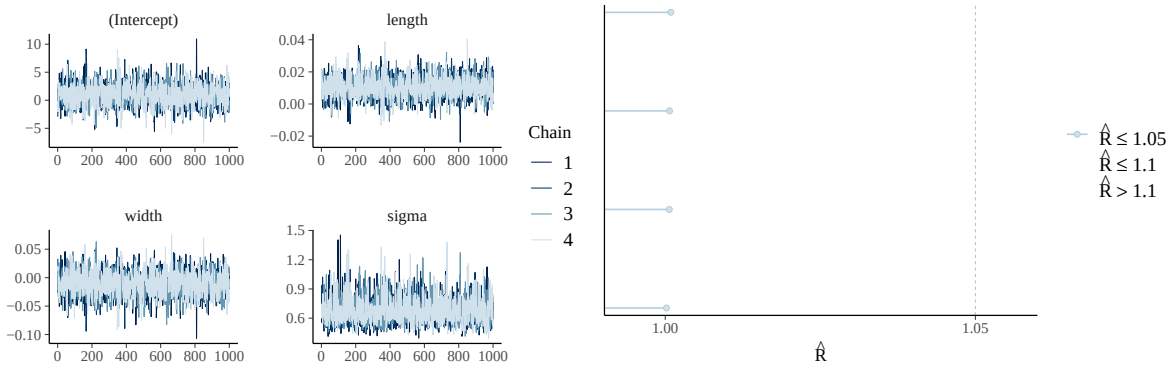
In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows…

Examining how the model fits, and is affected
by, the data

C.2 Diagnostics

Figure 3a is a trace plot. It shows… This suggests…

Figure 3b is a Rhat plot. It shows… This suggests…



(a) Trace plot

(b) Rhat plot

Figure 3: Checking the convergence of the MCMC algorithm

References

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

Encyclopædia Britannica, Inc. 2024. *Electoral College*. https://www.britannica.com/video/demystified-how-does-electoral-college-work/-250292.

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

Horst, Allison Marie, Alison Presmanes Hill, and Kristen B Gorman. 2020. *palmerpenguins: Palmer Archipelago (Antarctica) penguin data*. https://doi.org/10.5281/zenodo.3960218.

Johnson, Deborah. 2018. *Study Finds Wisconsin's African American Poverty Rate Three to Four Times Higher Than White Poverty Rate*. University Of Wisconsin-Madison. https://news.wisc.edu/study-finds-wisconsins-african-american-poverty-rate-three-to-four-times-higher-than-white-poverty-rate/.

Otis, Deb. 2024. *Do Faithless Electors Change Presidential Election Results?* https://fairvote.org/do-faithless-electors-change-presidential-election-results/.

Pew Research Center. 2024. *Mobile Fact Sheet*. https://www.pewresearch.org/internet/fact-sheet/mobile/?tabItem=5b319c90-7363-4881-8e6f-f98925683a2f.

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

Ricketts, Lowell R., and Ana Hernández Kent. 2024. *Wealth and Its Distribution: A Look at Asian American Households in 2022*. Federal Reserve Bank of St. Louis. https://www.stlouisfed.org/on-the-economy/2024/aug/wealth-distribution-look-asian-american-households-2022#:~:text=Indeed%2C%20Asian%20household%20wealth%20is,several%20factors%2C%20such%20as%20education.

Ruggieri, Lucca. 2024. *Patriot Polling Data Collection Methodology*. https://patriotpolling.com/our-polls/f/trump-and-baldwin-hold-narrow-leads-in-wisconsin.

Ruggles, Steven, Sarah Flood, Matthew Sobek, Daniel Backman, Annie Chen, Grace Cooper, Stephanie Richards, Renae Rogers, and Megan Schouweiler. 2024. IPUMS USA: Version 15.0 [dataset]. Minneapolis, MN: IPUMS, 2024. https://doi.org/https://doi.org/10.18128/D010.V15.0.

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

U.S. Embassy & Consulate in the Kingdom of Denmark. n.d. *Presidential Elections and the American Political System*. https://dk.usembassy.gov/usa-i-skolen/presidential-elections-and-the-american-political-system/#:~:text=That%20means%20that%20two%20parties,Party%20and%20Natural%20Law%20Party.

Walker, Kyle. 2024. *Tigris: Load Census TIGER/Line Shapefiles*. https://CRAN.R-project.org/package=tigris.

Wasserman, David, Sophie Andrews, Leo Saenger, Lev Cohen, Ally Flinn, and Griff Tatarsky. n.d. Cook Political Report. https://www.cookpolitical.com/vote-tracker/2020/electoral-college.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.