US Presidential Poll Errors, 2016 and 2020

Ian Bogley

11/26/2020

Foreword

Thanks to my professor Jonathan Davis for his constant work on providing me and the rest of my classmates here at the University of Oregon with the support we need to produce work like this. You can find his information here:

https://sites.google.com/site/jonathanmvdavis/ (https://sites.google.com/site/jonathanmvdavis/)

Also feel free to take a look at my own website or social media:

https://ibogley.github.io/website/ (https://ibogley.github.io/website/)

https://www.linkedin.com/in/ian-bogley-857088196/ (https://www.linkedin.com/in/ian-bogley-857088196/)

The original files I used can be found here:

https://github.com/ibogley/Code-Portfolio/tree/master/blog (https://github.com/ibogley/Code-Portfolio/tree/master/blog)

Setup

Before we start, these will be the packages we need:

```
library(pacman)
p_load(rvest,httr,tidyverse,janitor,lubridate,ggpubr,usmap,ggrepel)
```

Also note that for presentation purposes, the code used will be provided at the end for replication.

Introduction

The presidential election of 2016, pitting Hillary Clinton against Donald Trump, was a major turning point in American history. One of the storylines that came out was the errors in General Election polling, which consistently underestimated Trump's performance in key states.

Take for instance some states that were traditionally considered reliably democratic such as Michigan, Wisconsin, and Pennsylvania.

Michigan: According to archived fivethirtyeight analysis based off of national and statewide polling, Clinton was consistently projected to have a 3-4% advantage over Trump in terms of final vote share. In Wisconsin, Clinton was purported to have an advantage of up to 5%. Pennsylvania also showed signs of a democratic voteshare 3% higher than Trumps.

However, Trump was able to win all three of these "Blue Wall" states, even if by razer thin margins. These numbers gave Clinton an extremely good chance of winning.

The goal of polling is to give us an accurate representation of vote shares on election day, so multiple instances of polling errors in Trump's direction point to some sort of systematic error in the prediction.

Now consider the 2020 presidential election: Donald Trump against Joe Biden. While the democratic candidate had a very clear polling lead throughout the election season, there again seems to be an issue with underestimating the performance of Trump.

Let's take a quick look at the same "Blue wall" states, and their election results in 2020. Bear in mind that with this cycle occuring during the COVID-19 pandemic, vote counts are taking longer due to the large amount of mail-in-ballots. This means that way may have to wait for official results a while yet. However, with 99% of precincts reporting, we can begin to think about the final vote shares, and how different they looked from the polling this cycle.

As of November 26th (2020), with 99% of the total vote counted, Michigan was polling at a difference of almost 10% which was instead won by Biden with only a 3% difference. In Wisconsin, an 8% margin for Biden evaporated into a democratic win by less than 1%. Pennsylvania went from a Biden lead of 5% to a final win by less than 2%.

Analysts and pollsters will likely struggle for decades to come over why these predictions were so far off, and more so why they were all failing to account for some amount of Trump's support. One question that can be asked immediately is the following, **was the accuracy of the polls different from 2016?**

We will be using data from fivethirtyeight for polling data regarding both the 2016 and 2020 election cycles, while the final vote shares will be scrapped from wikipedia (which are originally provided by the Associated Press).

We will interpret **final poll average percentages as estimates of the final voting shares in each state**, and we will only consider the ratio between Trump and Biden.

The data

To this end, let's begin by loading in our polling data. Please note that we will only be using polls that were completed in September of the election year or later, and also that we will only be considering the ratio between Trump and his democratic opponent. We also will only be considering statewide results and polling. Polling data is sourced from fivethirtyeight

```
polls_2016 %>% head(5)
```

```
## # A tibble: 5 x 5
##   state        cycle dem_poll rep_poll trump_poll_ratio
##   <chr>        <int>    <dbl>    <dbl>            <dbl>
## 1 Virginia      2016       48       42            0.467
## 2 Florida       2016       48       45            0.484
## 3 Pennsylvania  2016       47       44            0.484
## 4 Florida       2016       42       46            0.523
## 5 California    2016       54       28            0.341
```

Now let's webscrape the final vote shares for each state from wikipedia, which is originally sourced from the Associated Press. Until the final vote results are posted on wikipedia, we will use a google search state by state to get the results we don't have yet. The vote counts are still provided by the Associated Press, meaning that they will be close to the official vote counts when all is said and done. Also helping is that 99%-100% of most vote shares have been counted so far (11/26/2020), allowing for close estimates to our final actual vote shares.

```
results_2016 %>% head(5)
```

```
## # A tibble: 5 x 7
##   state      cycle trump_vote_ratio dem_votes rep_votes dem_pct rep_pct
##   <chr>      <dbl>            <dbl>     <int>     <int>   <dbl>   <dbl>
## 1 Alabama     2016            0.644    729547   1318255   0.344   0.621
## 2 Alaska      2016            0.584    116454    163387   0.366   0.513
## 3 Arizona     2016            0.519   1161167   1252401   0.451   0.487
## 4 Arkansas    2016            0.643    380494    684872   0.336   0.606
## 5 California  2016            0.339   8753788   4483810   0.617   0.316
```

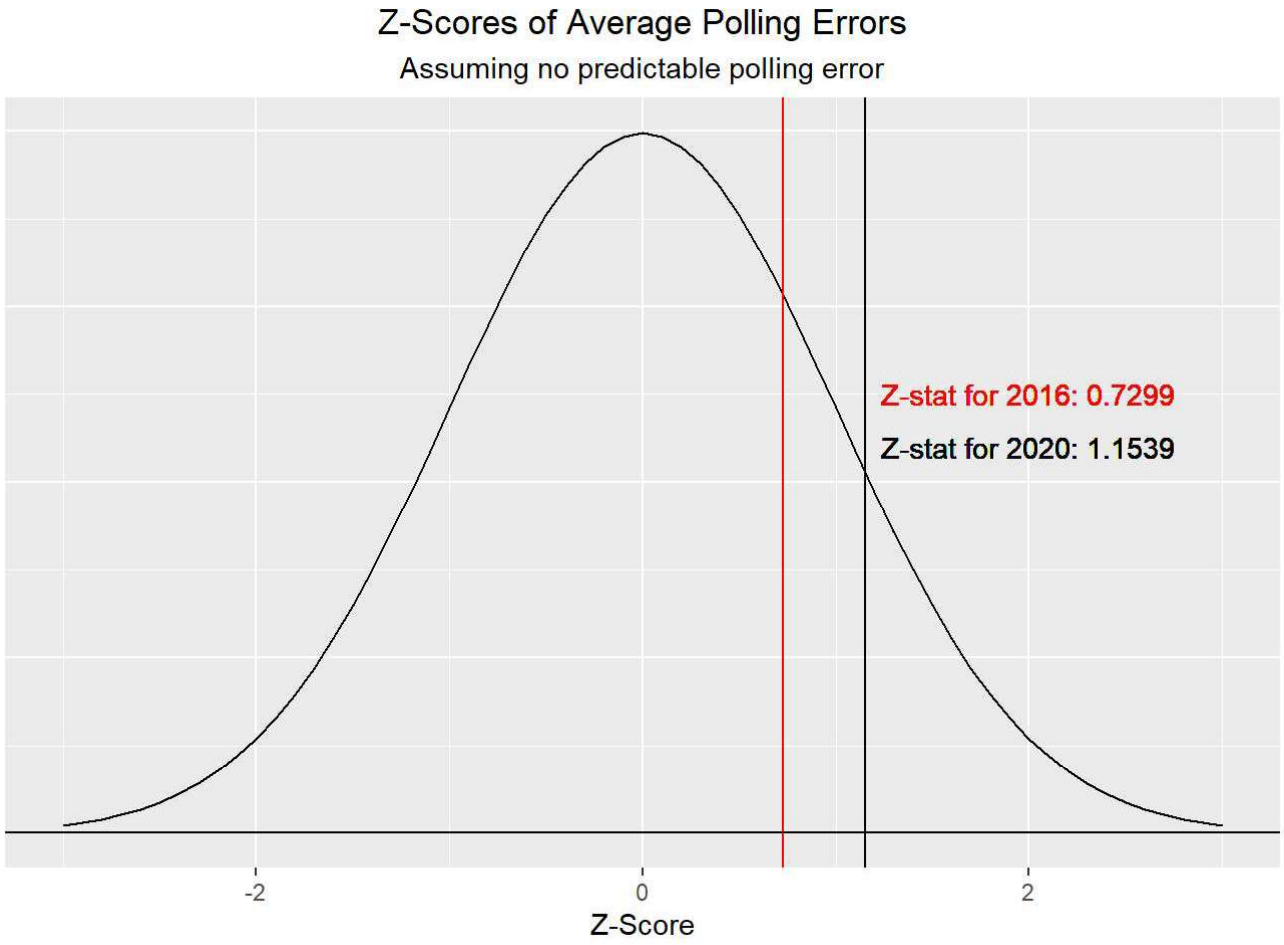
Now we can aggregate our polling data into average poll ratios by state, and combine with the results. Note that we will treat the polling error as the differene between the polled ratio $Trump/(Trump + Biden)$ vs the voteshares using the same ratio.

```
final_2016 %>% head(5)
```

```
## # A tibble: 5 x 11
##   state cycle poll_error trump_vote_ratio trump_poll_ratio dem_votes rep_votes
##   <chr> <dbl>      <dbl>            <dbl>            <dbl>     <int>     <int>
## 1 Virg~  2016    0.00507            0.472            0.467   1981473   1769443
## 2 Flor~  2016    0.0223             0.506            0.484   4504975   4617886
## 3 Penn~  2016    0.0202             0.504            0.484   2926441   2970733
## 4 Flor~  2016   -0.0165             0.506            0.523   4504975   4617886
## 5 Cali~  2016   -0.00275            0.339            0.341   8753788   4483810
## # ... with 4 more variables: dem_pct <dbl>, rep_pct <dbl>, dem_poll <dbl>,
## #   rep_poll <dbl>
```

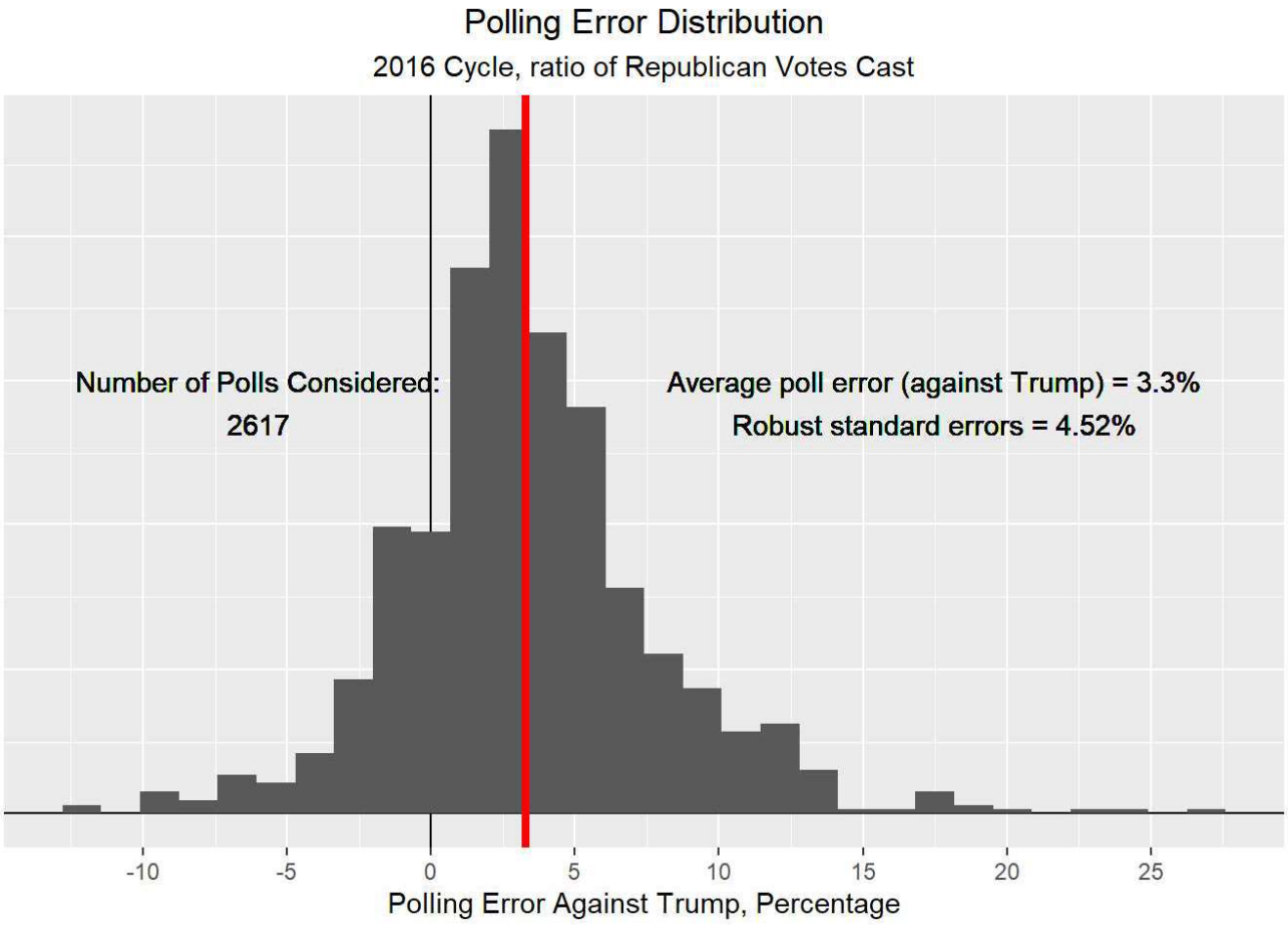
Now let's try plotting the distributions of each cycle's polling errors on a distribution assuming there isn't a predictable polling error:

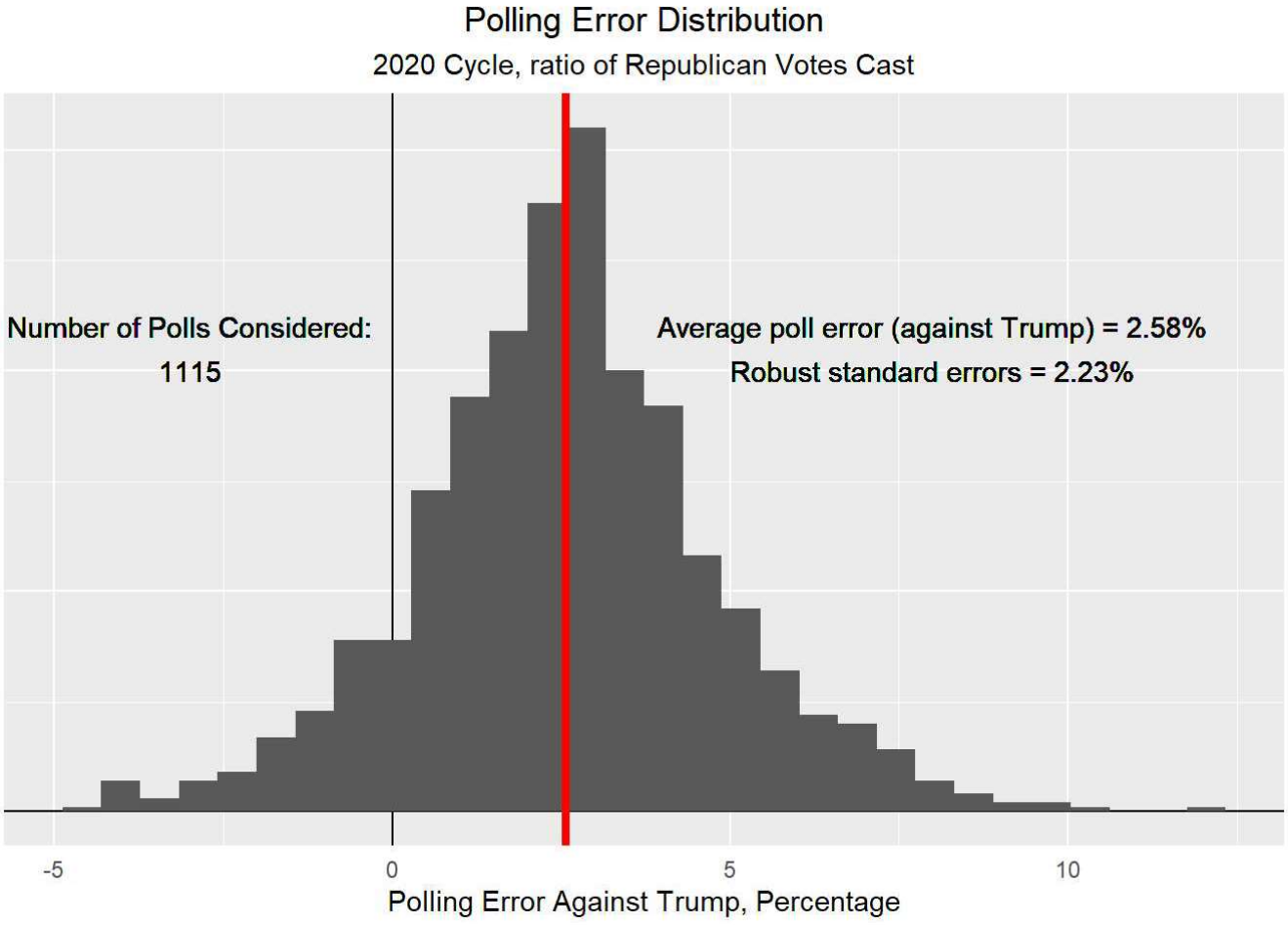
$H_0 : pollerror = 0$ $H_A : pollerror > 0$



As it turns out, the average polling error in 2020 is less likely given no predictable polling error. This seems to further add to the narrative that the polls are consistently under-predicting Trump's performance. With p-values of .2327 for 2016 and .1243 for 2020, neither effect gives enough evidence under traditional confidence intervals (1%,5%,10%). However, it seems that **the 2020 average polling error was even less likely than that in 2016 given no predictable effect underestimating Trump's performance.**

Now lets look at the distribution of the polling errors:

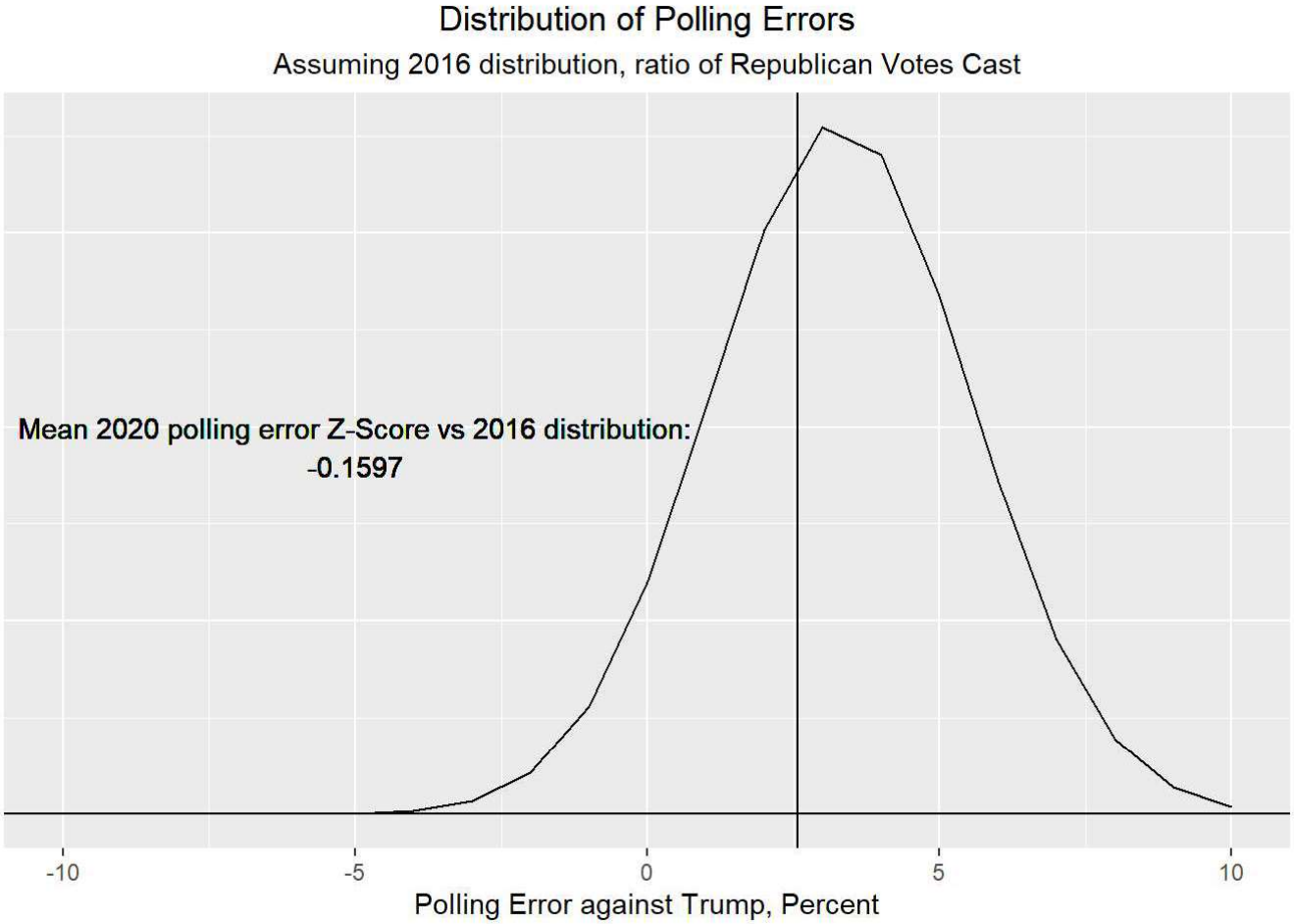

Polling Error Distribution
2016 Cycle, ratio of Republican Votes Cast
Number of Polls Considered:
2617
Average poll error (against Trump) = 3.3%
Robust standard errors = 4.52%
-10
-5
0
5
10
15
20
25
Polling Error Against Trump, Percentage


Polling Error Distribution
2020 Cycle, ratio of Republican Votes Cast
Number of Polls Considered:
1115
Average poll error (against Trump) = 2.58%
Robust standard errors = 2.23%
-5
0
5
10
Polling Error Against Trump, Percentage

While both the mean and standard errors for polling errors are smaller in 2020 than in 2016, notice that a 0% polling error is now more than 1 standard deviation from the mean. So while we have improved our accuracy for statewide polls, they still fail to account for an effect similarly seen in 2016. In fact, **the 2020 cycle describes a smaller effect, but seems to more distinctly specify it.**

Let's think about this question now: do the polling errors in both cycles seem to be the same effect?

$H_0 : avgpollerror_{2020} = avgpollerror_{2016}$ $H_A : avgpollerror_{2020} < avgpollerror_{2016}$



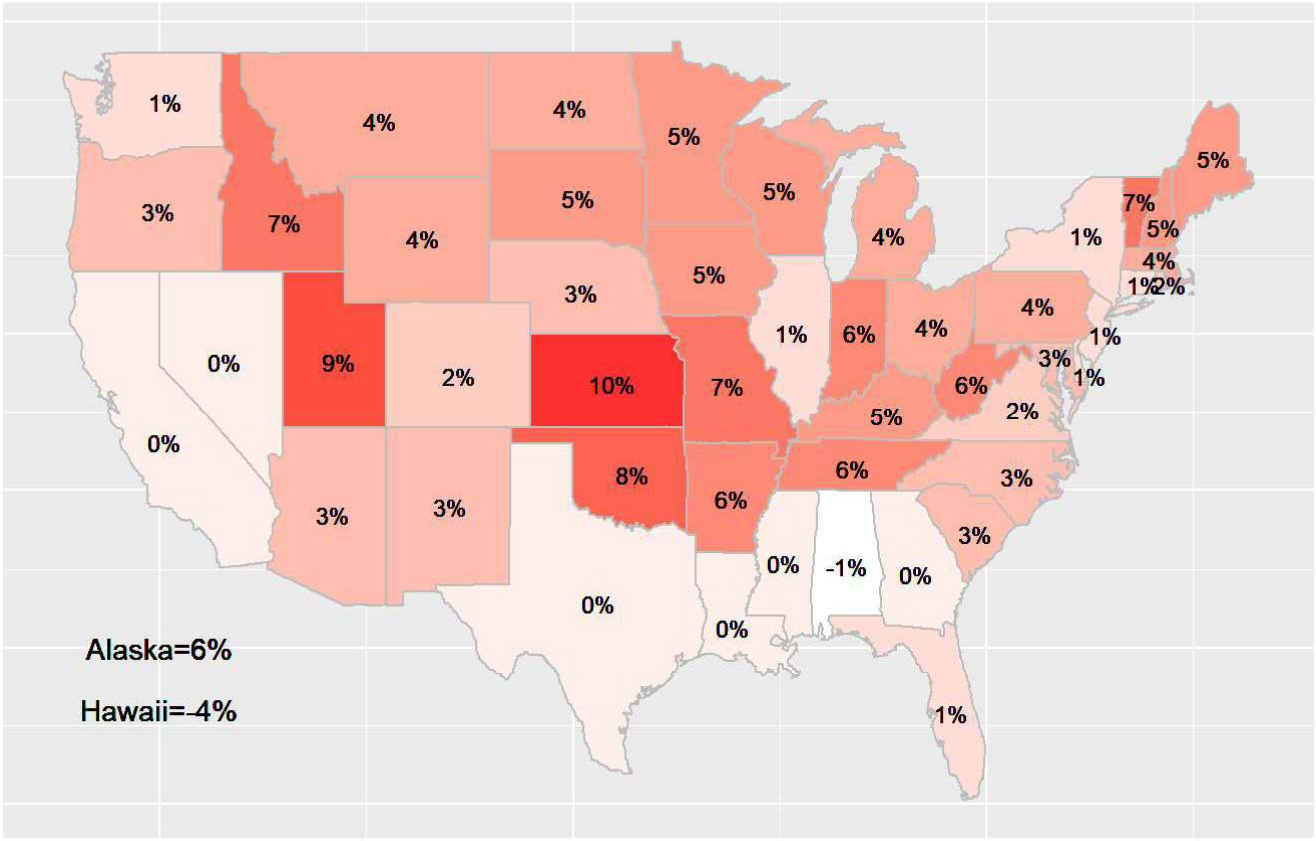
Let's assume a single tailed approach, so as to test if the 2020 polls were more accurate (with an effect closer to 0). The p-value here of getting the 2020 error given 2016's predictable effect is 0.8745. This seems to imply that **the error being picked up in 2016 is very similar to that observed in the 2020 cycle.**

On the other hand, do note the findings in our histograms earlier, with the **2020 cycle having a smaller average polling error and smaller standard errors**.

To finish off, let's try to track down where these errors might be coming from, and whether the sources of errors are different this time around:

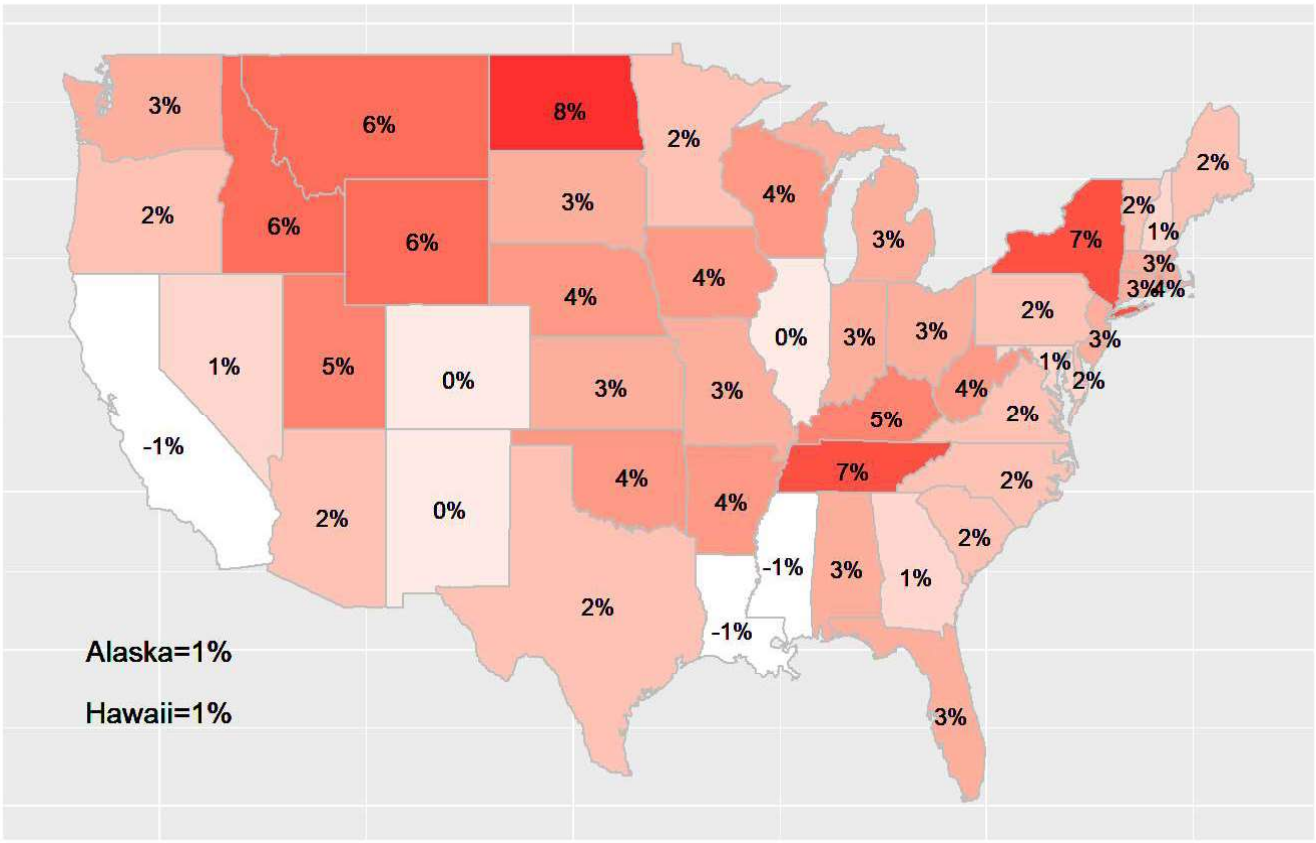
2016 Average Polling Errors Against Trump

Considering the ratio of Democratic vs Republican Votes Only



2020 Average Polling Errors Against Trump

Considering the ratio of Democratic vs Republican Votes Only



Note at the time of this writing (11/26/2020), New York is still only reporting at 85%, and it is estimated that most of the ballots yet to be counted will be democratically leaning. This means that in the coming weeks the polling error in New York may drop as well.

It appears that the polling errors did decrease between 2016 and 2020. Note especially the drops in the midwest. Where many had a polling error of 5~6% in 2016, 2020 saw in between 3~4% for most of the region.

Consider Wisconsin and Pennsylvania, each of which were swing states in the 2020 election, and were part of the blue wall before 2016. Both states saw their polling errors decrease by 2~3%.

Conclusion

Both cycles describe a consistent polling error underestimating Trump. Note that **even though 2020 saw a smaller average effect, it seems to even solidify the idea that there is some systematic effect that is failing to be accounted for.** In fact, the errors seem so similar that I would argue that it could be the same effect in both cycles. The source of this error, however, seems to still be a mystery. Additionally, the polls seemed to be as accurate as those in the past, meaning that these errors may be unavoidable due to the complications with trying to get an accurate sample of the American voterbase. (Nate Silver, 2020)

Consider the following:

Individuals will either respond to polls with the candidate they believe they will vote for at the time, or poll undecided. If they poll one way or the other, they may still change their minds. However, taking polls at different times during the election cycle is meant to account for the changing opinions of the general populous, so it seems that the issue may be occurring in how we predict undecided voters to act. Perhaps undecided voters were more likely to vote for Trump, or perhaps it was that undecided voters were more apprehensive about giving their true political leaning to the pollsters. Another possible issue that has been discussed at length is the problem of non-response, with “Response rates to polls in the low single digits” (Nate Silver, 2020)

However, I would recommend that you checkout fivethirtyeight for further research and reading. All the data I have used here comes from them originally, with only the final vote totals being provided by the Associated Press.

Sources

Silver, Nate. “The Polls Are All Right.” FiveThirtyEight, FiveThirtyEight, 30 May 2018, fivethirtyeight.com/features/the-polls-are-all-right/.

Silver, Nate. “The Polls Weren’t Great. But That’s Pretty Normal.” FiveThirtyEight, FiveThirtyEight, 11 Nov. 2020, fivethirtyeight.com/features/the-polls-werent-great-but-thats-pretty-normal/.

Polling data was extracted from a csv file that can be found here:

https://www.kaggle.com/fivethirtyeight/2016-election-polls (https://www.kaggle.com/fivethirtyeight/2016-election-polls)

https://projects.fivethirtyeight.com/polls/president-general/ (https://projects.fivethirtyeight.com/polls/president-general/)

Appendix 2: Reproducable Code

I will also include the csv files in a github repo associated with this project for your use.

The final vote shares were scrapped from wikipedia, with the Associated Press being the original source.

At the time of this writing, the official results have neither been certified nor put on wikipedia. As such, I wrote code to fill in the gaps with vote shares as of when the code is run. All the vote counts come from Associated press, which is the final source for both wikipedia articles. The temporary figures will also be from the Associated press.