W203 Lab 1 Assignment

Section 2: Tuesday 4PM PST

Ziling Huang, Dhyani Parekh, Amber Rashid, Alice Ye

10/26/2020

Question #1: Do US voters have more respect for the police or for journalists?

Introduction and Variable Operationalization

In the era of mass media and increased visibility on police activity, there are high tension emotions toward both journalists and police.

For this test, we operationalized two variables:

Both variables, *ftpolice* and *ftjournal*, indicate user responses between the value of 0 and 100. We choose to operationalize these variables by considering any value between 51 and 100 (the maximum) to indicate a favorable feeling towards journalists/police. We view these favorable feelings as an indication of how much a user respects journalists/police.

We used the delta between a voter's favor for the police and favor for journalists as an indicator of whether the voter has more respect for the police or for journalists.

We deem the variables *ftpolice* and *ftjournal* to not be independent. A journalist's depiction of police activity may change people's feelings toward journalists and police. Vice versa, political figures could be naysayers of journalists and proponents of police. In recent years, with increased reporting on police activity, the association of the two variables leads us to conclude the variables are not independent.

A gap in the way we operationalized these variables is that the term journalist connotes many different types of news avenues because a journalist could be interpreted as a formal news outlet or a freelance blogger. This invites individual perceptions who a journalist is into survey responses which is not controlled for with the way the question was stated in the survey.

Exploratory Data Analysis (EDA) and Data Cleaning

For *ftjournal*, respondents used a widget to input values from 0 to 100. However, some 3 respondents had the value of -7 indicating they skipped the question.

In order to assess only relevant data points, we focused on three variables: *reg*, *ftpolice* and *ftjournal*.

- *reg* = "Are you registered to vote, or not?"
- *ftpolice* = "How would you rate the police?"
- *ftjournal* = "How would you rate journalists?"

We noted no blank values for all of the above columns.

For the purpose of this test, 479 respondents from the original 2,500 were excluded due to: - 2 value of -7 in *ftjournal* and 1 value of "-7" in *reg* because we deemed valid responses as those between 0 and 100 - 476 values of "3" in column *reg* indicating these respondents are not registered to vote

This leaves a sample size of 2,021 that portrays only valid responses to *ftpolice* and *ftjournal* from registered voters.

Next, we assessed the distribution of the data:

- Because we are assessing *ftjournal* versus *ftpolice* (difference between police and journalists), we first leveraged a boxplot for each variable to look at their distribution. The boxplot shows more spread out feelings of warmth toward journalists but indicates a consolidated feeling of warmth in the upper half of the temperature scale for police.

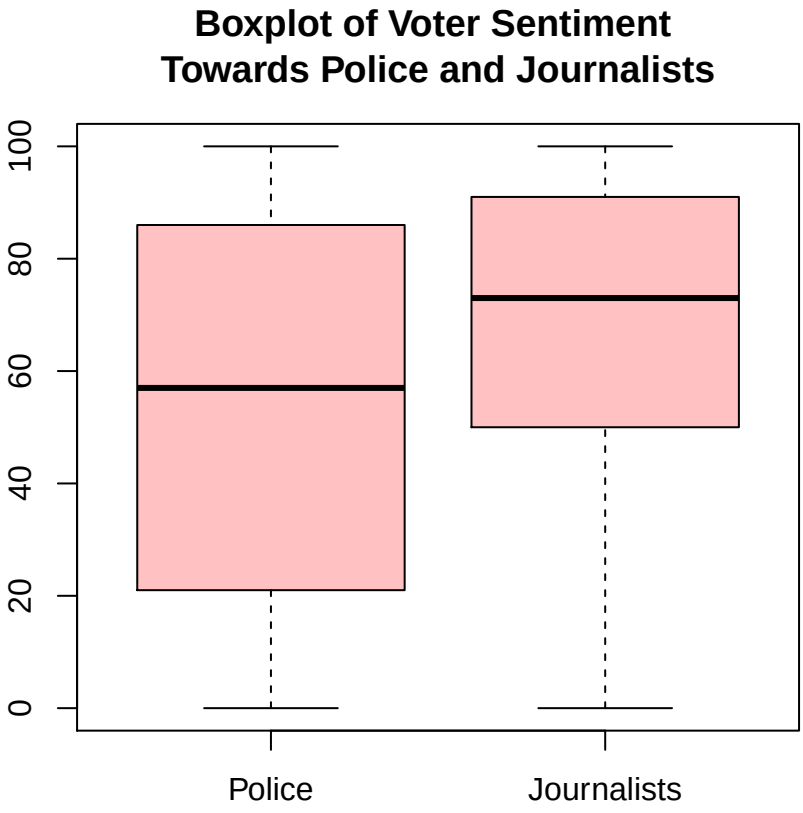

Figure 1: Voter Sentiment towards Police and Journalists

- Next, we assessed the distribution of both variables and deemed the distribution to be not normally distributed as it lacks the bell curve shape.

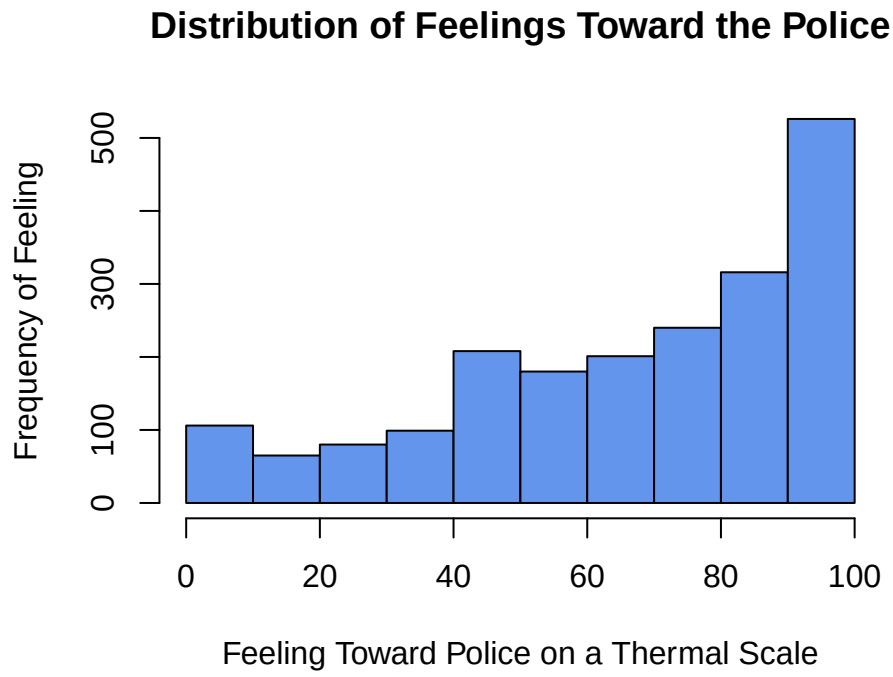

Figure 2: Distribution of Voter Sentiment towards Police

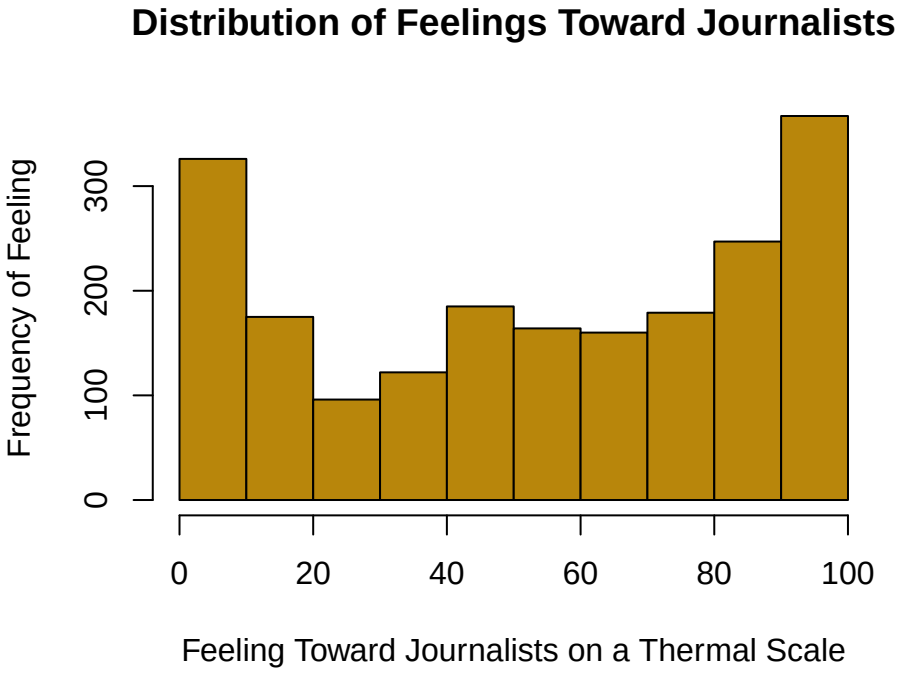

Figure 3: Distribution of Voter Sentiment towards Journalists

- In order to assess the distribution of the delta between journalists and the police, we looked at the distribution of the variable *ftjournal - ftpolice*. The distribution of this delta appears to be not normally distributed due to the fat tail on the left hand side.

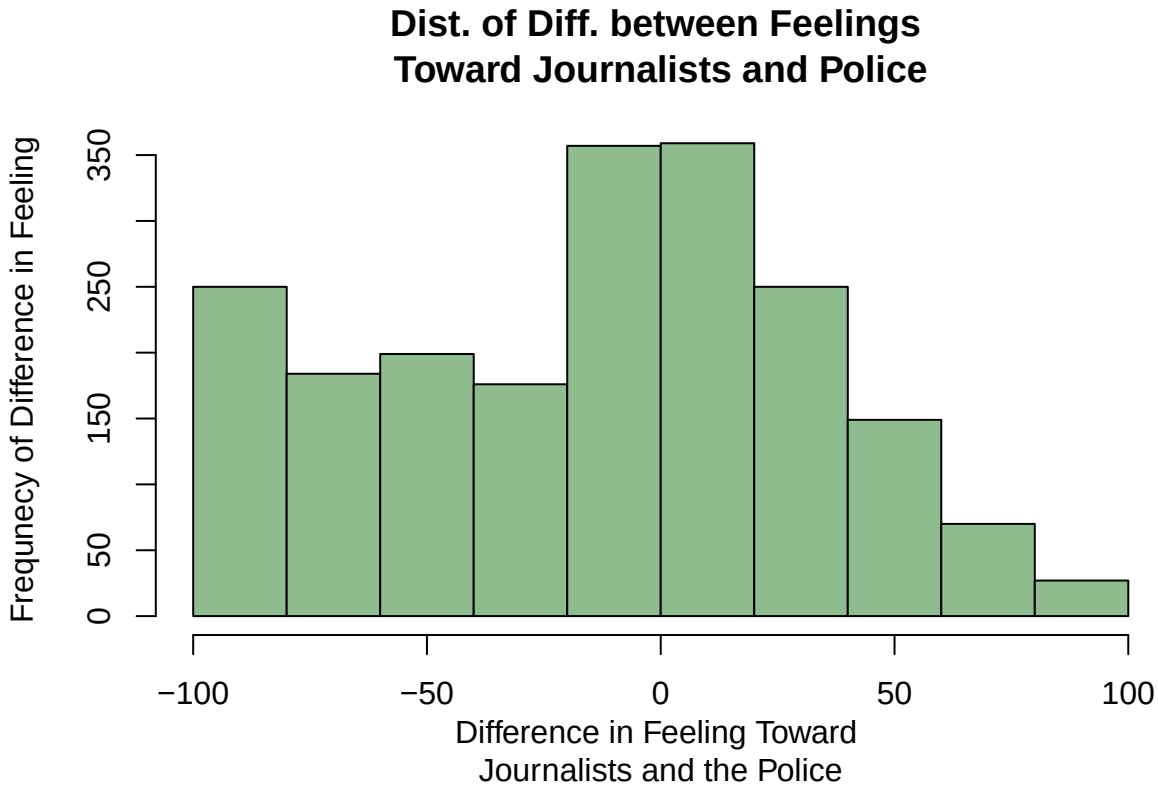

Figure 4: Distribution of Differrence Voter Sentiment of Journalists and Police

Hypothesis Test Selection

We selected to test this question using a sign test because our question and EDA indicate that the recommendations of a sign test are met, which are as follows:

1. The data is ordinal or on a likert scale (0 - 100).
2. The data is paired or dependent (police and journalists) beacause they are the sentiments from the same person. As stated in our introduction, we believe that police sentiments are related to journalist sentiments.
3. Each pair of data is independent from other pairs and from the same distribution. We assumed that everyone who received the survey had no information from how any other person responded and the sample is representative of the underlying population.

Though the sign test has less power than the Wilcoxon Sign Ranked test, we are unable to perform that test because of the limitations of ordinal data.

As such our test is the following:

Null Hypothesis: The number of positive signs is equal to the number of negative signs. P(+) = P(-)

Hypothesis: The number of positive signs is not equal to the number of negative signs. P(+) != P(-)

Interpreting Test Results

From our sign test, we found S = 855 and p-value = 8.465e-09. Additionally, in order to assess significance, we performed an effect size test and note the effect size to be -0.1302.

Statistical Significance of the Test: Due to the p-value of 8.45e-09, which is less than 0.05, we are able to reject the null hypothesis. As such, we conclude that the number of positive signs is not equal to the number of negative signs. We see that there are 855 positive values and 1111 negative values or a 43% probability of success.

Furthermore, the 95% confidence intervals are negative. As such, the difference between the two groups is negative and indicates favor for one group more than the other (police and journalists). In conjunction with our exploratory data analysis that displays a)levels of warmth toward police that are consolidated in the upper half of the thermometer scale of survey respondents while b)levels of warmth toward journalists are normally distributed but with fat tails on both ends of the thermometer scale. In conclusion, we deem that registered voters of the ANES survey have higher levels of warmth for the police and thereby US voters have more respect for police than for journalists. Note: survey weights were applied for the purposes of this test.

Practical Significance of the Test: We believe our test is practically significant due to the ever evolving relationship between the police and journalists. While at a historical point, feelings toward the two groups could have been segregated, in the current era, there is a close-knit and entwined relationship between the two groups that can potentially be polarizing as we saw a glimpse of in this data set.

Question 2: Are republican voters older or younger than Democratic voters?

Introduction

With the increasing polarization of politics, there are stronger stereotypes about who a Democrat and a Republican is. One stereotype is that Democrats are a different age than Republicans. In this section, we investigate whether Democrat voters are older or younger than Republican voters.

Operationalizing Variables

There are 3 variables we operationalized: registered voter, party affiliation, and age. *reg* represents the question "Are you registered to vote, or not?". We used *reg* to remove the responses "No, not registered" in order to ensure our data only contains people who qualify and are prepared to vote. The gap in using *reg* is that someone could not be registered to vote during the survey but could be registered afterwards. This gap mostly applies to young voters who are registering after they turn 18 and skews our sample towards older voters.

For party affiliation, we used *rand_pid*, *pid1d* and *pid1r*. Within the survey, people were randomly assigned to either *pid1d* or *pid1r*. *pid1d* represents the question "Generally, speaking do you usually think of yourself as a Democrat, a Republican, an independent, or what?". *pid1r* was the same question but with the order of Republican and Democrat switched. We used all 3 variables to aggregate party affiliation data. To focus on Republicans and Democrats, we kept "Republican" or "Democrat" responses and removed "Independent" or "something else" responses. A gap here is that those who are in between these two parties were not forced to choose one so we're missing data for borderline voters.

For age, we used *birthyr* which contains the voter's year of birth. We calculated *age* by subtracting birth year from 2018. A gap in defining *age* this way is that we are unable to measure with the months or days.

EDA and Data Cleaning

First, we looked at how people responded to *reg*. One person did not respond (equals -7) and 476 people said "No, not registered to vote." (equals 3). These 477 people were removed and left us with 2023 registered voters.

Next, we looked into party affiliation. 1,006 voters received the Democrat first version and 1,017 voters received the Republican first version. The split between the two versions is even (<1% difference). Both versions received the most responses for Democrat. There were slightly more responses for Independent in the Republican first version which could indicate survey design influence. However, further experimentation is needed to prove this and we assumed both versions had equal and opposing influence. We removed responses

that were not "Republican" or "Democrat" which left us with 1,316 responses. After applying survey weights, we had 1,252 responses (724 Democrats and 528 Republicans). We created a new column, *p_affiliation*, that merged responses from both versions.

We created *age* by subtracting *birthyr* from 2018. We confirmed that all ages met voter requirements (>18 years) and there were no missing ages. The average age in our sample is 53.0 years. The weighted average age is 49.7 years which means that our sample has less younger voters than the population and the weights help us more accurately account for them. Age has a bimodal distribution with 2 peaks (57-62 years, 27-32 years).

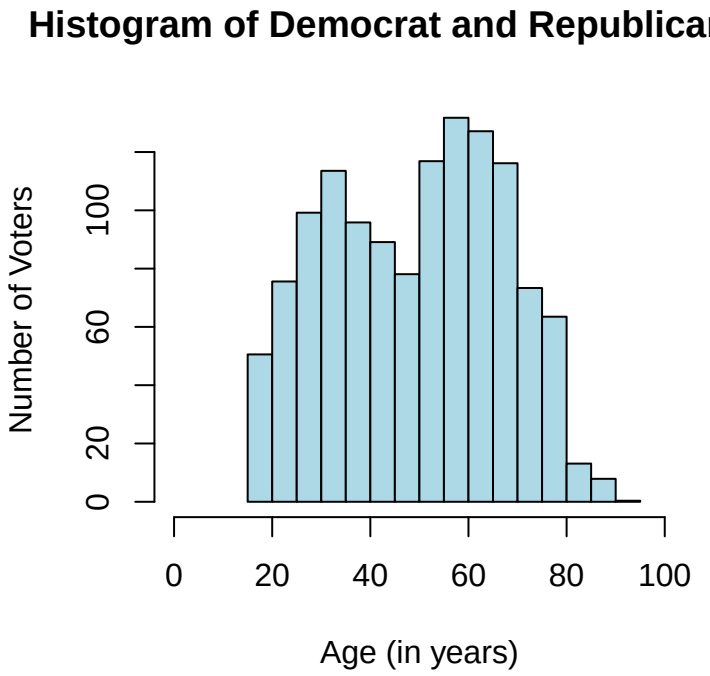

Figure 5: Histogram of Democrat and Republican Registered Voter Age

Since this question focuses on 2 independent samples, we also looked into age for each sample. In our weighted sample, the average Democrat age is 47.8 years with a standard deviation of 17.5. The distribution of Democrat age is bimodal with 2 peaks, 27-32 years and 57-62 years. The average Republican age is 52.4 years with a standard deviation of 17.2. The distribution of Republican age is more normal than Democrats and slightly left skewed with a peak 57-62 years.

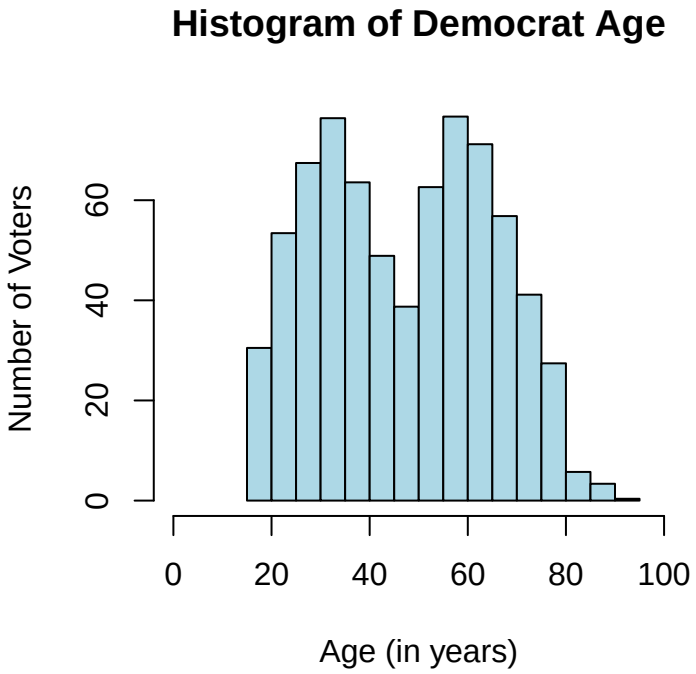

Figure 6: Histogram of Democrat Registered Voter Age

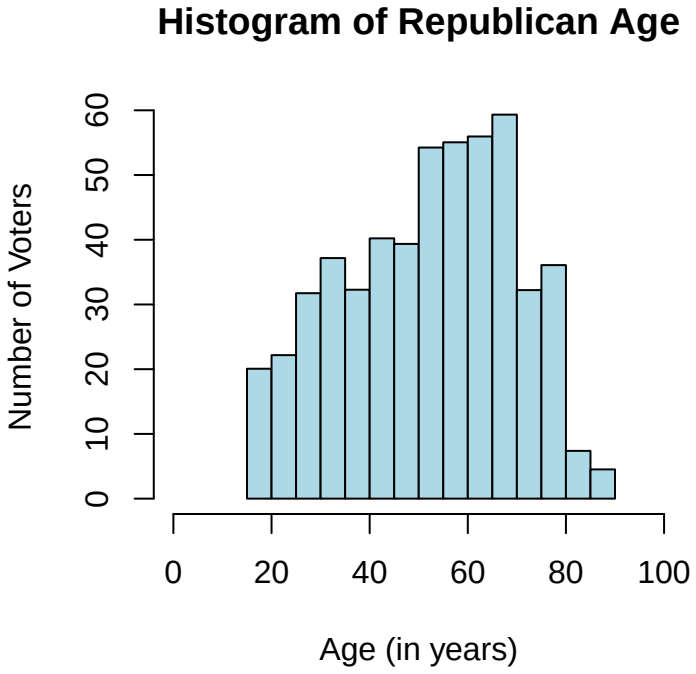

Figure 7: Histogram of Republican Registered Voter Age

Hypothesis Test Selection

We selected to use a 2 sample pooled t-test. The following assumptions were met:

- 2 Independent Samples: Democrats and Republicans are 2 independent groups because each datapoint represents a unique voter who could only select either Republican or Democrat, not both.
- Metric Scale: Age data is continuous being that is is a range of integers.
- Parameter: The population parameter is average age because averages are good for summarizing a group's age and is easy to communicate.
- Large Sample Size: Both samples are large (has >30 voters). Since at large sample sizes the distributions of the z-statistic and t-statistic are similar, we can use a t-test.
- Sample Normal Distribution: Our EDA showed there were some deviations from normality for age. However, overall we found that age's distribution was normal enough for a t-test.
- Equal Sample Variances: The weighted variance of the 2 samples are equal ($var_{Democrat} = 304.8$ and $var_{Republican} = 296.2$) because neither one is double the other. We used pooled sample variance to have a more precise variance estimator.
- IID is met: Everyone who received the survey had no information from how the person before responded and because we applied the survey weights, this gives us a sample that is representative of the underlying population.

We performed a two-tailed test because it requires higher significance to reject the null hypothesis than a one-tailed test. We can understand if Democrats are older or younger than Republicans by further interpretting the sample test statistic. For this test, our null hypothesis is that there is no different between the average age of Democrats and Republicans ($mean_{Democrat} - mean_{Republican} = 0$) and our alternate hypothesis is that there is a difference between the average age of Democrats and Republicans ($mean_{Democrat} - mean_{Republican} \neq 0$).

Interpreting Test Results

The t-statistic is -4.6554. In a 95% significance level, the t-statistic must be between -1.96 and 1.96 to accept the null hypothesis. Since the sample t-statistic is outside of this range, we rejected the null hypothesis. Also, the t-statistic is negative indicating a negative difference in average age. Cohen's D is -0.266 which is a small negative effect size. From these two observations of the t-test, we can conclude that on average Democrats are younger than Republicans.

Question 3: Do a majority of independent voters believe that the federal investigations of Russian election interference are baseless?

Introduction

Voter trust in the integrity of elections is the cornerstone of successful democracies. We are interested in studying American voter perceptions around the integrity of the 2016 election outcome regardless of their political affiliation. The question focuses specifically on Americans who self-identified as independent in 2018 and whether they believed there was Russian interference in the 2016 Presidential election.

Key concepts we have identified are :

- The Definition of a Voter
- Voter Independence
- Belief in Russian interference

We are excluding the Mueller survey question because his name may raise bias and inaccurately influence our findings.

Operationalizing Variables

The independent variable in the study is the independent voter group in 2018. Voters are those who self-identified as being registered to vote in 2018 based on the *reg* question in the survey. Independent voters are those who self-identified to be "3. Independent" in their response to the *pid1r* or *pid1d* question in the survey. We decided to exclude all free text and blank responses (6% of responses) from the study due to the subjectivity involved in interpretation. The dependent variable being studied is *russia16* where voters state whether they think it was probable that Russia interfered in the 2016 Presidential election. Weight was a supporting variable that was used to reflect the true population frequency of observations.

Gaps

A gap in the study exists since the question *pid1r* and *pid1d* is ambiguous about what being independent means. Survey respondents confuse social groups with political affiliation. This is the reason for the subjective data found in the free text *pid2r* and *pid2d* follow-on responses.

The rationale behind selecting independent voters for the study is that they will have an unbiased view since they have no party loyalty. Gaps in this assumption exist since independent candidates are just as likely as those who have party affiliations to have strong biases against certain candidates that may cloud their judgment.

We identified voters as those registered to vote in 2018. However, it is possible that people registered to vote may not actually vote.

The time lag in data collection and timing of this test in 2018 limits the relevance of the results for 2020 in terms of registration status or political affiliation.

EDA and Data Cleaning

To ensure that the data is clean we checked and confirmed that there were no duplicates in case ids.

Data transformation was performed to merge *pid1* and *pid1r* into a column *p_affiliation* to collate responses and to drop free text and blanks. In *pid2r* and *pid2d*, the variability of responses made interpretation and categorization challenging. 3 non-responses were removed in the *russia16* question . We dropped observations where people were not registered to vote under *reg*. We also checked the variable data for uniformity and unique values.

Survey weights were applied to our data to ensure that the sample is representative of the true population characteristics. The data was then plotted in a bar chart to get an overall sense of data proportion, direction, reasonableness, and sample size. We observed that more than 50% of independent voters think Russian interference was probable and used the used the directional weight in the plot to design the hypothesis test.

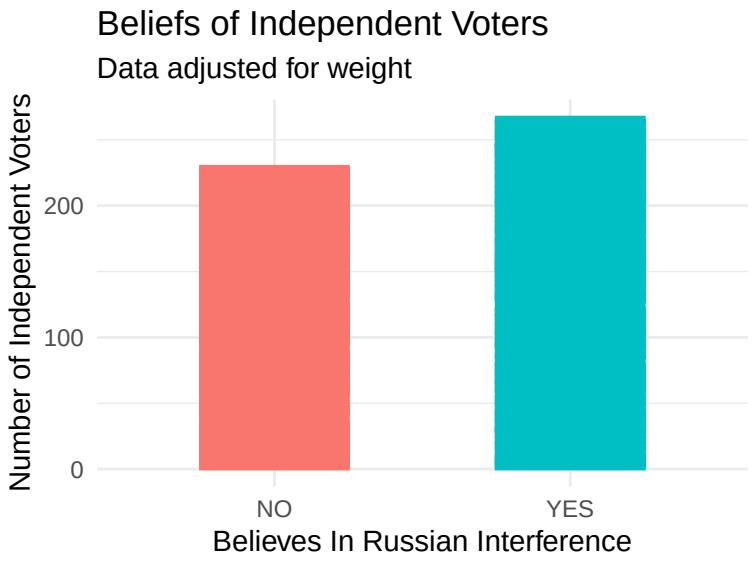

Figure 8: Bar Chart of Belief in Russian Interference

Hypothesis Test

The results of the EDA show that a higher proportion of independent voters believe there was Russian interference. As such, we believe that it will be more appropriate to state the tests in the following manner to ensure the power of the test is maximized.

We will evaluate the following hypotheses to show whether majority of independent voters believe that there is basis for investigations of Russian interference. If the answer is yes, then it provides a logical counterpart to answer the question on the proportion of independent voters who believe there is no basis.

A majority is defined as having strong evidence from the sample that more than 50% of independent voters believe there was Russian interference. The significance level set for the fair binomial test is 5% based on conventional norms.

The "Probable" or "Not Probable" response to the *russia16* question gives us dichotomous nominal data in a binomial distribution after removing non-responses. An answer of 'Probable' is deemed a success and 'Not Probable' is not.

Null Hypothesis Independent voters are 50% likely to believe that federal investigations of Russian interference had basis.

Alternative Hypothesis Independent voters are more than 50% likely to believe that federal investigations of Russian interference had basis.

A one-tailed binomial test with 50% probability and 5% significance is conducted to reject or not reject the null hypothesis.

The assumptions are that each observation is independent and identically distributed. We believe that observations are independent of each other given the diversity of views and beliefs independent voters may have on any issue. Knowing how one independent voter thinks about an issue does not necessarily give us more information about how another independent voter would think under the same circumstances After applying the survey weights we are confident that the sample is a fair representation of the true population and identically distributed.

Interpreting Test Results

The p-value is 0.0531 and is slightly larger than the significance level of 5%. Using Cohen's G (Cohen, 1988) to interprete the effect size shows that the difference in sample proportion and expected proportion

is negligible at 0.037. This means we do not reject the null hypothesis. We conclude that the result is not statistically and practically significant and it is not true that majority of independent voters believe federal investigations of Russian interference are baseless.

Question 4: Was anger or fear more effective at driving increases in voter turnout?

Introduction

"Fear is the path to the dark side. Fear leads to anger. Anger leads to hate. Hate leads to suffering." - Yoda

Understanding the emotions and psychology that drives voting behavior in a population is very important. It can be influenced by current events, news, and propaganda. The degree to which emotions influence voting behavior can be unique to any point in time given the circumstances. Nonetheless, we can gain insights from understanding voter sentiments at the time of voting. Thus to answer our question with the data avaialble, we will be looking at the prevalance of anger and fear in voting and non-voting populations.

Operationalizing Variables & Data Cleansing

We chose to consider a number of key variables:

- Voter Registration Status (*reg*)
- 2018 Turnout (*turnout18*)
- 2016 Turnout (*turnout16*)
- General Anger Metric (*geangry*)
- General Fear Metric (*geangry*)
- Weight (*weight*)

We chose to focus on the 2018 turnout data and disregard the 2016 turnout data for 2 reasons:

- The survey was conducted in December 2018, and thus most indicative of sentiments of voters in 2018.
- No data was provided around voter sentiments in 2016.

Thus, we are considering whether anger was more prevalent than fear in voters vs. non-voters in the 2018 election.

In our data cleansing, we ensured that there no duplicate case IDs, removed unregistered voters from our voting population, and ensured that there were no cases of said voters being underage. From the 2018 voter turnout data, we disregarded values of "5" (unsure if they voted) and "-7" (no response) to give strict voter and non-voter populations.

EDA

Once selecting the appropriate data set, we looked at the distribution of ratings for anger and fear for both voters and non-voters. In the below figure, we can intuitively see, that when we compare anger between voters and non-voters, we see a more normal-like distribution for non-voters, with the highest frequency around 3, dropping off for 1 and 2, but significantly dropping off at 4 and 5. Where as for voters, we see a more even distribution. From the EDA, it seems that voters are more likely to rank anger at a 4 or 5 than non-voters. When comparing anger and fear in the voter population, we see that voters are less likely to rank fear as high as anger. Lastly, when comparing fear in the voting and non-voting population, we see that the voting population is more normal-like, while the non-voting population is heavily skewed left or toward the lower ratings for fear.

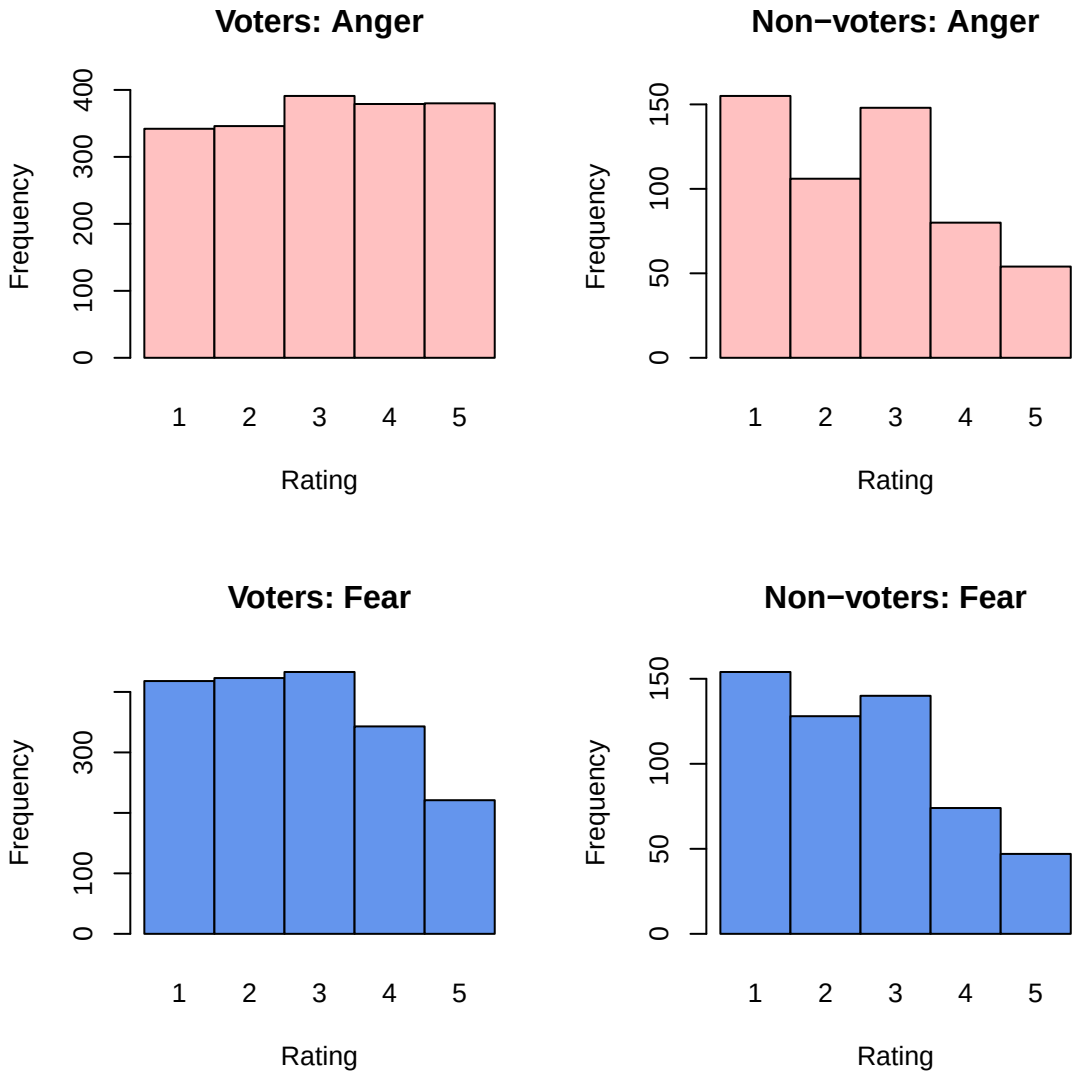

Figure 9: Anger and Fear Sentiment of Voters and non-Voters

To further our intuition on whether fear or anger was more prevalent in the voting population, we created a spineplot which shows more information of the data pairwise. It shows for each anger rating the proportion which selected each fear rating, while also showing the proportion relative to the whole population (width of the bars). Here we see that in the voting population, those who ranked 1 (not at all) or 2 (a little) for anger had the highest proportion also rank 1 for fear, and we see this same pattern for a ranking of 5 (extremely). This shows that in the extremes, the emotions tend to be related. We see the same matching patterns for 3 (somewhat), and 4 (very) but to a lesser extent.

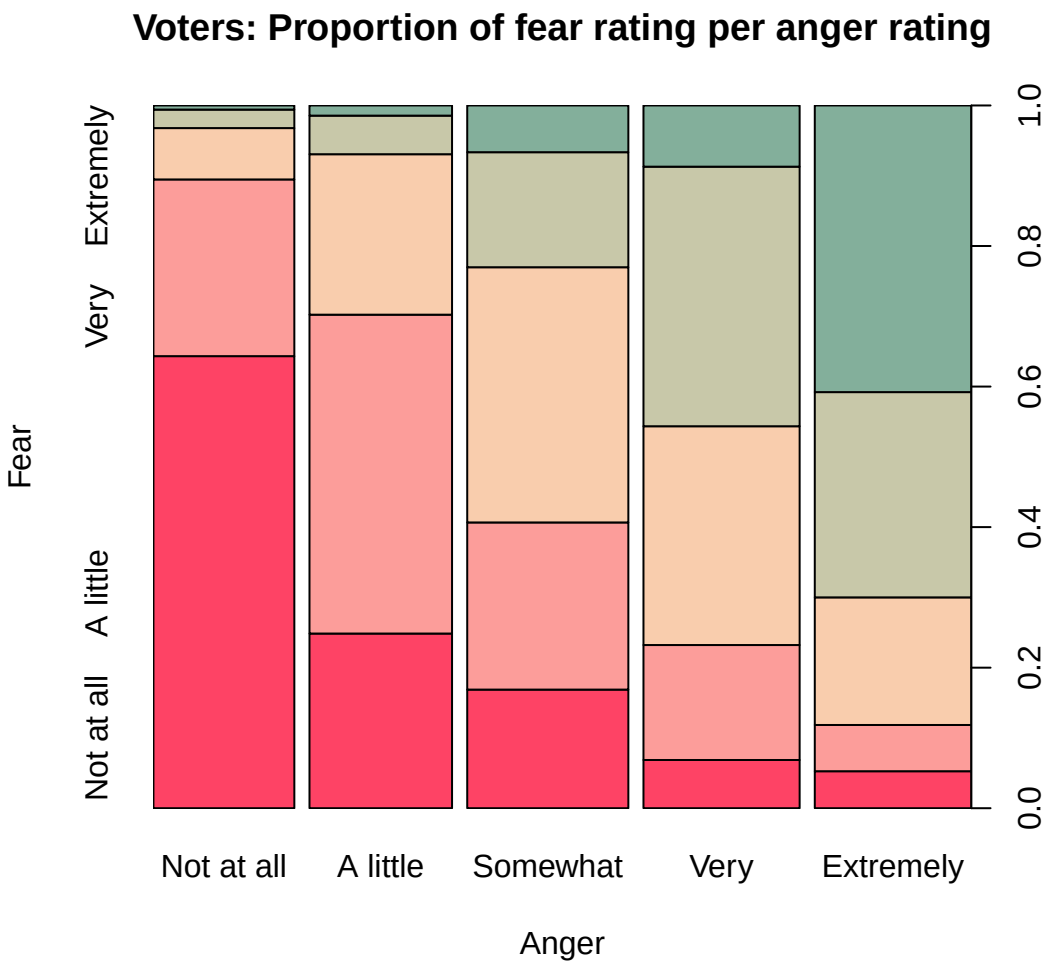

Figure 10: Voters: Proportion of fear rating per anger rating

In the non-voting population, we see the similar trend of anger and fear ratings matching, but also see a higher proportion of pairs around 3 (somewhat), where as for voters, they were more evenly spread.

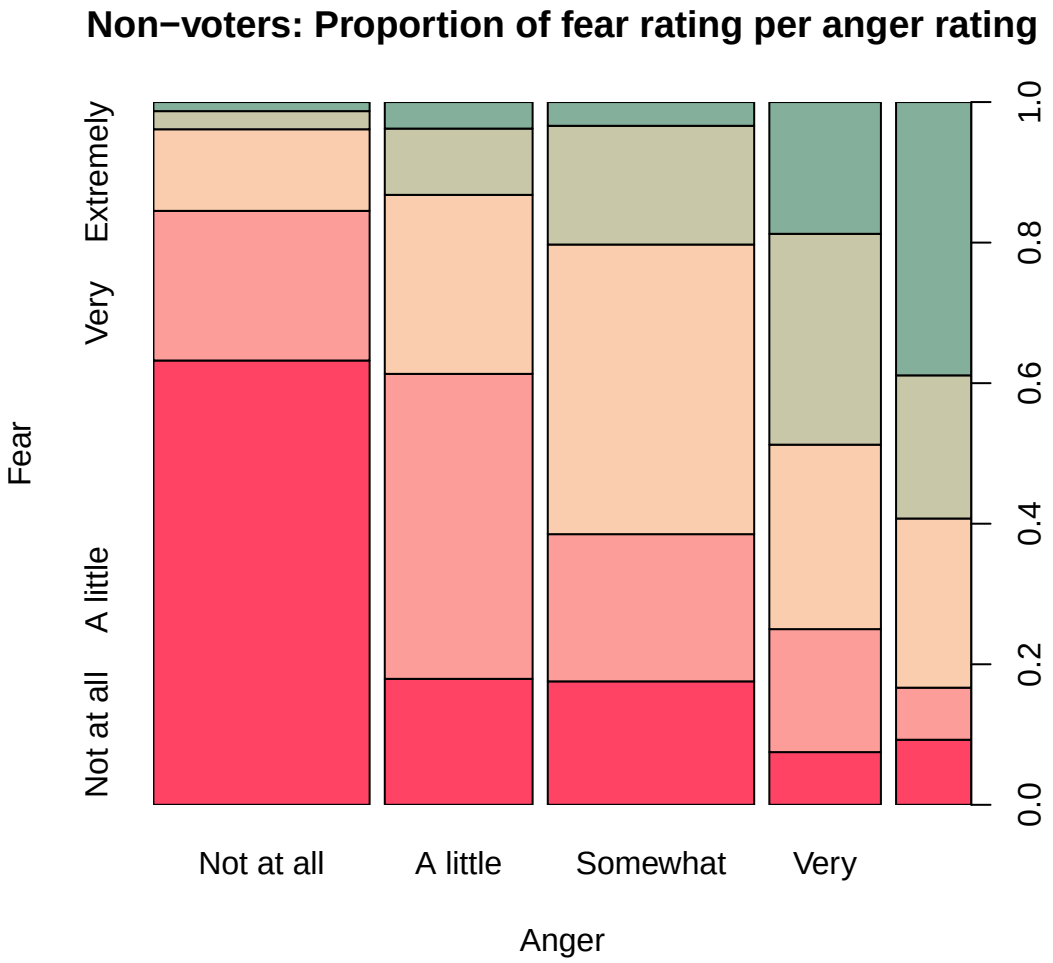

Figure 11: Non-Voters: Proportion of fear rating per anger rating

Hypothesis Testing

The non-parametric sign test was chosen for the 2018 voters and non-voters to answer our question because:

1. The data is ordinal or on a likert scale ("not at all" to "extremely") thus the intervals of the data are not clear. If the scale was interval, we could have used the more powerful Wilcoxon Sign Ranked test.
2. The data is paired or dependent beacause they are both negative sentiments from the same person, and thus can be considered related. Likewise in our EDA, we see that levels of anger and fear tend to be similar.
3. Each pair of data is independent from other pairs and from the same distribution due to the nature of ANES study design which chose sampled folks identically to represent the overall population.
4. The sample size is large enough even when looking at strictly positive and negative pairs for both voters (1024) and non-voters (293).

As such our test is the following:

Null Hypothesis: The probability of a decreasing pair (Fear>Anger) is the same as the probability of an increasing pair (Fear<Anger).

Alternative Hypothesis: The probability of an increasing pair is greater than the probability of a decreasing pair.

Interpreting Results

In order to conduct the sign test, we used the "BSDA" R package, but also confirmed our results and calculated the effect size by finding the signs of the pairs and using the binom.test function.

For 2018 voters we found a p-value $<$ 2.2e-16 and proportion effect size of ~0.32 thus giving us a highly statistically significant result with medium effect size. Rejecting the null hypothesis, it shows that in 2018 voters anger was more prevalent than fear. Furthermore, looking at our probability of success of our binomial test, it shows that about ~66% of those with unequal anger and fear were more angry than fearful. Lastly, the 95% confidence interval was [0.635, 1.0].

For 2018 non-voters, statistical significance was not found (p-value = 0.726) thus we fail to reject the null hypothesis that anger and fear are equally prevalent. Of the unequal pairs of anger and fear in the non-voting population, about ~49% were pairs where anger was greater than fear. A low effect size was found in this test of about ~0.024. This could be due to the smaller sample size in the non-voter population. Lastly, the 95% confidence interval was [0.429, 0.547].

When comparing the outcomes for the 2018 voting and non-voting population, higher proportions of voters had higher ratings for anger than the non-voting population, and furthermore higher proportions rated anger greater than fear than the non-voting population when comparing unequal ratings. Thus, this indicates that anger was a greater driving force in voter turnout.

Question 5: Are a majority of mail-in voters from the 2018 election more likely to support Biden than other candidates for the 2020 election?

Introduction

Due to public safety guidance to prevent COVID-19 spread, the 2020 general election has a larger focus on mail-in voting than in previous elections. There are concerns raised about wide-scale usage of mail-in voting which has led to bi-partisan debate on the topic. We would like to investigate whether mail-in voters are more likely to be Biden voters. This is an important question to explore because the suppression of mail-in voting could alter the election outcome.

Operationalizing Variables

The independent variable we operationalized is whether a voter submitted their ballot by mail in 2018 using *turnout18*. Those who responded as "Definitely voted by mail" were included in our test. Other responses were removed.

The dependent variable is whether a voter supports Joe Biden or another candidate. We used data from *vote20jb*. Voters who said they support Trump, Joe Biden or neither in the 2020 election were included. Responses stating "probably not vote" were removed.

A majority is defined as a sample proportion above 50%.

The gap is that behavior from the 2018 election may not be completely indicative of 2020 behavior, given the changes that have occurred (e.g. pandemic, economic change).

Another potential gap is that there exists a possibility that more people may turn up to vote in person on Nov 3, 2020 but we would not know until the actual event.

EDA and Data Cleaning

Variables used were *reg*, *turnout18*, *weight* and *vote20jb*. To ensure that the data is clean we checked and confirmed that there were no duplicates in case ids. From *reg*, we removed "-7" i.e. no response and removed

"3. No, not registered to vote".

Next, we looked at *turnout2018* and removed everyone who did not vote by mail.

Using *vote20jb*, we removed the 3 who said they probably would not vote. To better align this variable with our hypothesis, we transformed *vote20jb* to signal whether someone would vote for Biden (equal to 1) or would not vote for Biden (equal to 0 if Trump or another candidate).In order to ensure the sample reflects the underlying population characteristics, we applied survey weights and found that 188 (44.4%) weighted number of people did not support Biden and 293 (55.6%) did. We plotted a bar chart to visualize the directional weight of our sample and to plan our hypothesis test.

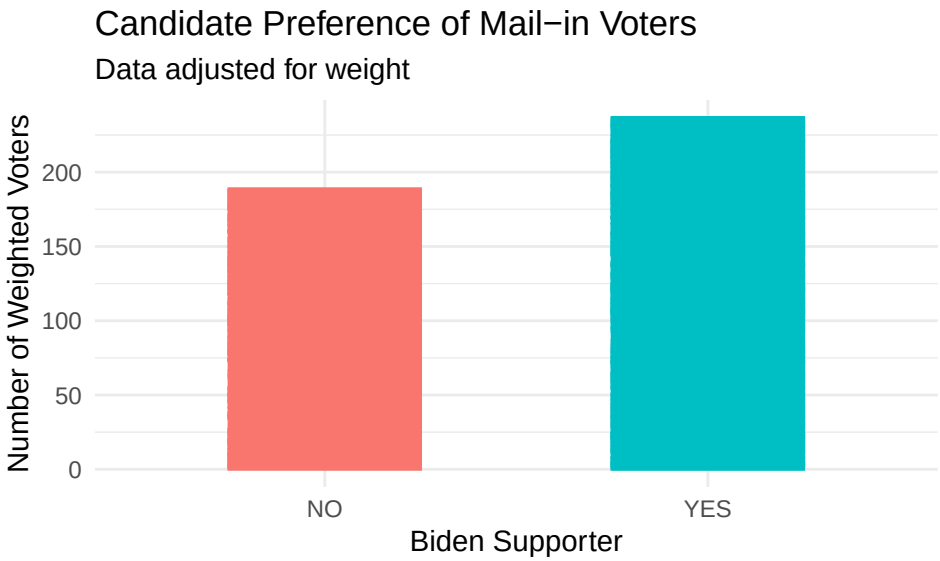

Figure 12: Bar Chart of Candidate Preference for Mail-in Voters

Hypothesis Test

In order to determine whether 2018 voters are more likely to vote for Biden in 2020, we chose a Fair Bionomial Test.

Fair Binomial Test Assumptions :

- IID: We assumed IID is met. Everyone who received the survey had no information from how any other person responded and because we applied the survey weights, this gives us a sample that is representative of the underlying population.

Other Factors:

- Sample Size: Weighted sample size is 425 and is sufficient
- Variable: We transformed vote20jb to be a binary variable therefore we can represent supporting Biden as a success, and not supporting Biden as a failure and use the binomial distribution and outcome.

Null Hypothesis: People who voted by mail in 2018 are equally likely to support Biden versus another candidate in 2020.

Alternative Hypothesis: People who voted by mail in 2018 are more likely to support Biden in 2020. (Probabilility $> 50\%$)

Interpreting Results

The p-value is 0.0128 and is less than the significance level of 5%. This means we reject the null hypothesis and accept the alternative hypothesis that 2018 mail-in voters are more likely to support support Biden

in 2020. In order to analyze the practical significance of the result, we apply Cohen's G (Cohen, 1988) to interprete the effect size. This tells us that the difference in sample proportion and expected proportion is negligible at 0.055. Hence, we conclude that while the test shows statististical significance, the result has no practical significance.