Lab 1: Question 1

Tal Segal, Xander Snyder, Patrick Old

Importance and Context

Are Democratic voters more enthusiastic about Joe Biden or Kamala Harris?

As Joe Biden's campaign for President of the United States entered August of 2020, the most dreadful year in recent memory, it still lacked one crucial element; a running mate. There was a lot riding on this decision, and his eventual choice on August 11th of Senator Kamala Harris was one that made history. Once elected to office after ousting the Trump campaign by a narrow margin in November, Harris became both the first black woman and South Asian American to be elected second in command. It is a common practice for the leader of a presidential campaign to pick a running mate that appeals to voters they lack connection with, and this held true in this instance. Harris and Biden are similar enough to work together toward shared goals, but different in many readily apparent ways. Kamala, 56, is young compared to her counterpart, who received a large amount of press for his elderly status of 78 years old at the time of taking office. Furthermore, Harris' rich ethnic background and skin color drew much attention during the campaign, compared to Biden, who's background in these regards looks much like that of almost every former president and vice president. These clear differences lead to a question; who are Democratic voters more enthusiastic about?

The answer to this likely holds weight in showing how much of an impact both Biden and Harris had on their campaign winning the election. Furthermore, the results may indicate the future of the Democratic party, such as their readiness of voters to back a candidate with a diverse background compared to that of their more traditional counterpart. Last, it may also provide insight into what we can expect over the next four years, and how Democratic voters may respond to Kamala entering the office of President if an aging Joe Biden were to pass away in office (a difficult but frequently publicized possibility due to his age and health concerns).

Description of Data

We will address this question using pre-election data from the 2020 American National Election Studies (ANES). This is an observational dataset based on a sample of respondents drawn from the YouGov platform. We break down the research question into three key parts: 1) determining what "Democratic" voter means in pre-election data, 2) determining what a voter means in pre-election data, and 3) measuring enthusiasm.

The concept of determining if an individual is "Democratic", or a "voter", is simpler in post-election data than pre-election data. Following the election, it is clear whether an individual voted and for whom. However, pre-election data is more complicated as this data does not yet exist. Instead, we must rely upon what voter's indicate they will do in the following months when surveyed. With this in mind, we define "Democratic" by the respondent's answer to "Generally speaking, do you usually think of yourself as a Democrat, a Republican, or an independent?" being a "Democrat" which represents 34.59% of the total sample. Furthermore, a "voter" is defined in this study as those who answered "Yes" to the question "Do you intend to vote in the November election for President?", 87.82% of the total sample.

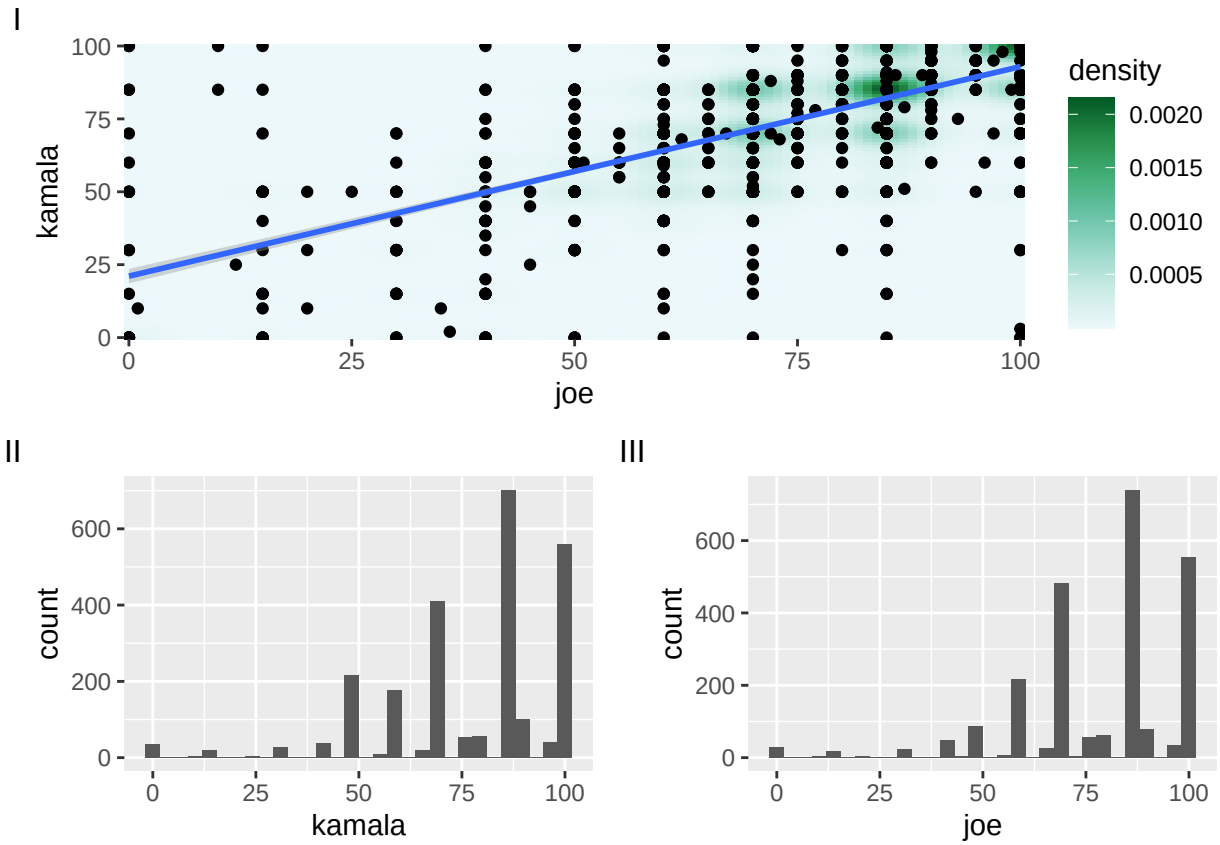
Last, and more complicated, is determining voter enthusiasm for specific candidates. The only metric provided in this study that measured both Joe Biden and Kamala Harris separately on the same topic was a question targetting a respondent's "feeling thermometer"[1] toward a candidate, meant to measure their favorability and warmness toward them on a 0 - 100 scale. Responses of greater than 50 indicate a warmness or favorability

[1] See the ANES 2020 User Guide and Codebook page 87 for a full description

toward the candidate, where as zero to 50 indicate the opposite, and a score of 50 is neutral. Responses in this study are filtered upon valid responses (those that fall in the zero to 100 range), leaving 2483 responses (29.99% of the total sample) to analyze after applying all filtering mentioned to this point. While the feeling thermometer metric does not directly mention the term "enthusiasm" in the prompt, the definition of enthusiasm is "Having or showing intense and eager enjoyment, interest, or approval." Thus, it is not a far stretch to measure what is largely a synonymous description, but one should nonetheless have care in reasoning with the results of this study because of this.

Both distributions of Harris and Biden thermometer have a left skew (figure X), which results in a strong positive correlation when both are plotted on each axis of a scatter plot (figure X). The plots are of no surprise considering each member is of the democratic party and those are the only voters in the filtered sample for this study. However, it is notable that there is a dip toward the upper end of each candidate's histogram, where the mode and median come in at 85 out of 100. Furthermore, respondents tended to round to the nearest 5th, with the most common responses coming in at 85, 100, and 75 in that order for both candidates. This created clearly marked density zones in the scatter plot (figure X). Here, all points are plotted, but due to the overlap seen in a few regions that would not be visible with points alone, a density gradient is added to assist in visualizing the data.

- ADD FOOTER



Most appropriate test

I then test whether this subgroup reports having greater enthusiasm for Kamala Harris or Joe Biden. There are three different statistical tests that are possible to run in this place; the paired t-test, the Wilcoxon Signed-Rank Test, and the Sign Test. A two-sample paired t-test would assume that 1) the data is metric, 2) the data is paired, and 3) the sample size is large and/or follows a normal distribution. While the data appears quantitative, it is in fact ordinal, not metric. The reason for this is that if we were to state it is

metric, this would mean that there is an equal difference between two pairs of datapoints on the scale (i.e. the difference between 0 and 10 is the same as that between 90 and 100). However, because this data is being judged on a "feeling thermometer" scale, I do not believe we have the basis to assume that this holds true. Respondents are likely not familiar with this scale prior to this test, and due to their unfamiliarity with it, likely do not treat all of the points throughout the scale equally or equidistant from their surrounding points.

The second possibility is the Wilcoxon Signed-Rank Test. The assumptions for this test are that 1) the data is of an interval scale, 2) the data is paired and drawn IID, and 3) the data is symmetric (in this case as it is a paired test, the symmetry is based upon the paired differences). While this data meetings the second assumption clearly and largely meets the 3rd assumption (box plot showing symmetry ommited due to space concerns), the first assumption of a continuous distribution fails.

This leaves us with the Sign Test. The assumptions for this test are that 1) the data is ordinal, 2) the data is paired and drawn IID. This simple test relies upon finding the sign of the difference between pairs, and does not worry about the magnitude of the difference between each pair. While this test is easy to apply to a variety of situations, one must often have a great deal of data to find significant results due it relying upon such basic information. Much of the data is discarded for this test, which is why it is not a commonly used test as many others, but nonetheless is applicable for this study and should work well enough considering there are over 2400 data points to work with.

Test, results and interpretation

```r
library(BSDA)

SIGN.test(x = jk$joe,
          y = jk$kamala,
          md=0,
          alternative = "two.sided",
          conf.level = 0.95)

joe_greater <- as.integer(as.logical(jk$joe > jk$kamala))

joe_perc <- sum(joe_greater) / length(joe_greater) * 100
kamala_perc <- 100 - joe_perc
#print (paste("Joe Biden Percentage:", round(joe_perc, 2)))
#print (paste("Kamala Harris Percentage:", round(kamala_perc, 2)))

binom.test(sum(joe_greater), length(joe_greater), p = 0.5,
           alternative = c("two.sided"),
           conf.level = 0.95)
```

The null hypothesis is that there is no difference in means between the enthusiasm backing Joe Biden or Kamala Harris by democratic voters. Rejecting the null hypothesis would therefore mean that voters had a greater enthusiasm - OR perhaps this should be that