Lab 3: Hypothesis Tests about the Mean.

w203: Statistics for Data Science

Michael Alexander

11/15/2016

Brief

The American National Election Studies (ANES) conducts surveys of voters in the United States before and after every presidential election. You are given a small subset of the 2012 ANES survey, contained in the file ANES_2012_sel.csv.

You will use the ANES dataset to address the following questions:

1. Did voters become more liberal or more conservative during the 2012 election?
2. Were Republican voters (examine variable pid_x) older or younger (variable dem_age_r_x), on the average, than Democratic voters in 2012?
3. Were Republican voters older than 51, on the average in 2012?
4. Were Republican voters more likely to shift their political preferences right or left (more conservative or more liberal), compared to Democratic voters during the 2012 election?
5. Select a fifth question that you are interested in investigating.

Introduction

For this analysis we will be using the survey data provided to answer a number of questions about the sampled population. For each question we will need to identify the relevant survey variable and consult the codebook to ensure we have an understanding of what the values really represent. As is common in survey data, we will also need to be careful to filter out missing or non-response values before conducting any analysis. Once we identify and clean the relevant variables we will need to assess things like the type of variable, the distribution, and the variance in order to determine the appropriate statistical test to use to test each hypothesis. Before diving into the hypothesis tests for each question, we should also take a minute to think about where the data is coming from and what it is really saying about respondents. Many of the response variables are asking for self-reported reflections on things like political ideology and party affiliation, and there may be reason to think that people's responses do not necessarily reflect their internal reality. We won't deal directly with this issue but it is important to keep in mind when drawing conclusions about the real world from this kind of data.

1. Did voters become more liberal or more conservative during the 2012 election?

In order to answer the question of whether or not voters became more liberal or conservative during the 2012 election we will look at the variable libcpre_self and libcpo_self, which ask respondents to identify themselves on a 7 point scale from Extremely liberal (1) to Extremely conservative (7).

```r
S = read.csv("ANES_2012_sel.csv")
library(car)
library(lsr)
```

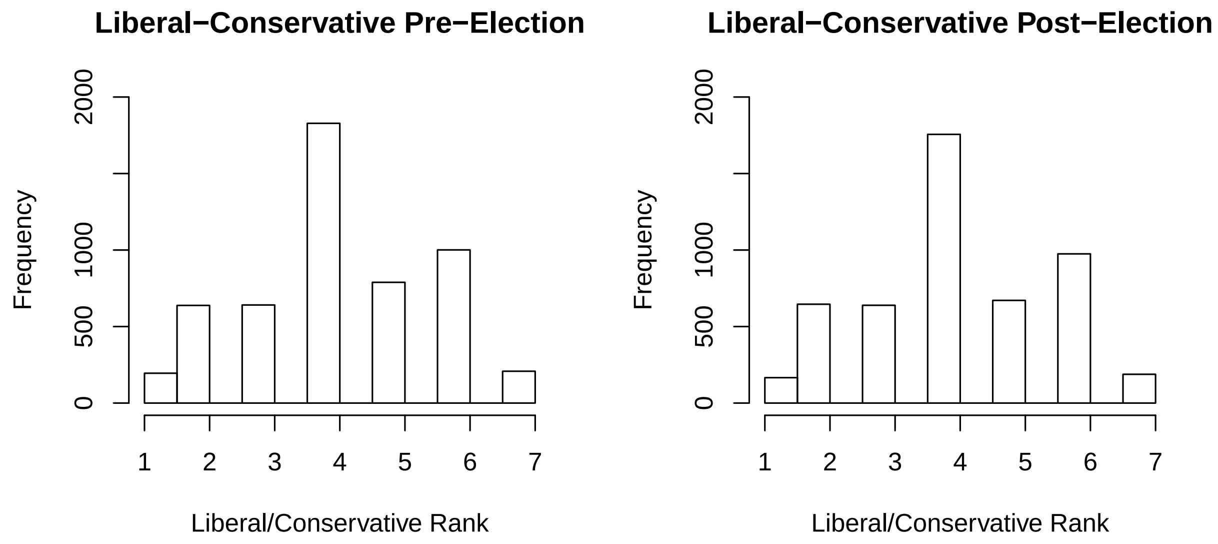
```
## Warning: package 'lsr' was built under R version 3.3.2
```

The first step after looking at the variables is to clean them up by removing the answers that don't fall on the 1-7 scale or are missing/incomplete and then ordering the remaining factor levels.

```
S_1_pre = factor(S$libcpre_self, ordered = TRUE, levels = c("1. Extremely liberal",
                                                            "2. Liberal",
                                                            "3. Slightly liberal",
                                                            "4. Moderate; middle of the road",
                                                            "5. Slightly conservative",
                                                            "6. Conservative",
                                                            "7. Extremely conservative"))

S_1_post = factor(S$libcpo_self, ordered = TRUE, levels = c("1. Extremely liberal",
                                                            "2. Liberal",
                                                            "3. Slightly liberal",
                                                            "4. Moderate; middle of the road",
                                                            "5. Slightly conservative",
                                                            "6. Conservative",
                                                            "7. Extremely conservative"))
```

It is important to note that the intervals between the ranks on the scale here are not necessarily equal. Now that we have cleaned up variables we can look at the histograms for both to get a quick understanding of what the data looks like.

```
hist(as.numeric(S_1_pre), xlab = 'Liberal/Conservative Rank', ylab = 'Frequency' , xlim = c(1,7),
     ylim = c(0,2000), main = 'Liberal-Conservative Pre-Election')

hist(as.numeric(S_1_post), xlab = 'Liberal/Conservative Rank', ylab = 'Frequency' , xlim = c(1,7),
     ylim = c(0,2000), main = 'Liberal-Conservative Post-Election')
```



Looking at the histograms we can see that both variable appear to be somewhat normally distributed. There also doesn't appear to be any major change in the distribution between the variables. Because of the scale used the variables are only ordinal and not continuous so we can not use a parametric test to evaluate the question of whether voters became more/less liberal/conservative. Instead we can use the Wilcoxon signed rank test, since the samples are dependent, to test the null hypothesis that there has been no significant change in Liberal/Conservative rank over the course of the election.

```
wilcox.test(as.numeric(S_1_pre), as.numeric(S_1_post), paired=TRUE, alternative = c("two.sided"))
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  as.numeric(S_1_pre) and as.numeric(S_1_post)
## V = 734760, p-value = 0.1662
## alternative hypothesis: true location shift is not equal to 0
```

```
mean(as.numeric(S_1_pre), na.rm=TRUE)
```

```
## [1] 4.172264
```

```
mean(as.numeric(S_1_post), na.rm=TRUE)
```

```
## [1] 4.14997
```

The p-value returned from the test is only 0.1662 so we are not able to reject the null hypothesis that there was no significant change in voters Liberal/Conservative rank. Looking at the difference of the mean rank we can see that both pre and post election are very close at 4.17 and 4.15 respectively.

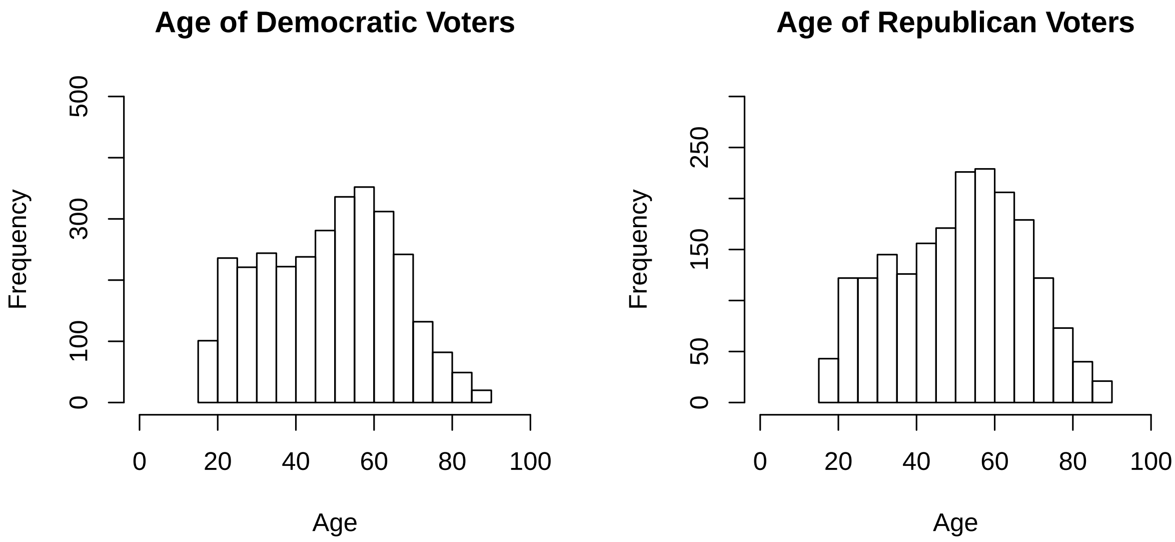
2. Were Republican voters (examine variable pid_x) older or younger (variable dem_age_r_x), on the average, than Democratic voters in 2012?

In order to answer this question we will need to assign a party affiliation based on the variable pid_x so that we can compare age across the Democrat and Republican voters. To do this we will classify all of the survey responses including "Democrat" as Democrats and those containing "Republican" as Republicans. We will also need to account for any records with missing age values.

```
S$party_affiliation = "Other"

S$party_affiliation[S$pid_x %in% c("1. Strong Democrat",
                                    "2. Not very strong Democract",
                                    "3. Independent-Democrat")] = "Democrat"

S$party_affiliation[S$pid_x %in% c("5. Independent-Republican",
                                    "6. Not very strong Republican",
                                    "7. Strong Republican")] = "Republican"

S_2_democrats = subset(S, S$party_affiliation == "Democrat" & S$dem_age_r_x != -2)

S_2_republicans = subset(S, S$party_affiliation == "Republican" & S$dem_age_r_x != -2)

S_2 = subset(S, (S$party_affiliation == "Republican" | S$party_affiliation == "Democrat")
             & S$dem_age_r_x != -2)
```

Now that we have cleaned up variables we can look at the histograms for both to get a quick understanding of what the data looks like.

```
hist(S_2_democrats$dem_age_r_x, xlab = 'Age', ylab = 'Frequency' , xlim = c(0,100),
     ylim = c(0,500), main = 'Age of Democratic Voters')

hist(S_2_republicans$dem_age_r_x, xlab = 'Age', ylab = 'Frequency' , xlim = c(0,100),
     ylim = c(0,300), main = 'Age of Republican Voters')
```



The histograms indicate that there is a skew towards the younger age groups present for both Democrats and Republicans, whose distributions appear to be roughly the same. Because we have a large enough sample, this deviation from the normal distribution shouldn't keep us from using an independent t-Test to test the null hypothesis that Democratic and Republican voters are the same age on average. However, before we conduct the t-Test we need to verify that the variance between the two groups is homogeneous, which can be done using Levene's test.

```
leveneTest(y = S_2$dem_age_r_x, group = as.factor(S_2$party_affiliation))
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value Pr(>F)
## group    1  0.2123  0.645
##       5047
```

The results of the Levene test give a p-value of 0.645, which is actually a little lower than we would like to feel comfortable accepting the assumption of homogeneous variance between the groups. Because of this, we are not able to use a normal t-Test and instead will use a Welch t-Test to correct for this issue.

```
t.test(S_2_democrats$dem_age_r_x, S_2_republicans$dem_age_r_x)
```

```
##
##  Welch Two Sample t-test
##
```

```
## data:  S_2_democrats$dem_age_r_x and S_2_republicans$dem_age_r_x
## t = -5.1653, df = 4206.5, p-value = 2.512e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -3.439637 -1.546939
## sample estimates:
## mean of x mean of y
##  48.83735  51.33064
```

The result of the Welch t-Test gives us a p-value very close to 0, which suggest that there is a significant difference in age between the two party groups. The mean age of Democratic voters is 48.8 years while the mean age of Republican voters is 51.3. The difference in means is only around 2 years, so while this difference is significant it may not be very practical. We can use Cohen's d to evaluate this effect size in a more quantitative way.

```
cohensD(S_2_democrats$dem_age_r_x, S_2_republicans$dem_age_r_x)
```

```
## [1] 0.149074
```

The value we get for Cohen's d is only 0.149 which confirms the impression that this effect size is small even though it is significant, with Democratic voters about 2 years younger than Republican Voters on average.

3. Were Republican voters older than 51, on the average in 2012?

In order to answer this question we just need to look at our previously created subset of Republican voters and compare the sample to a mu = 51. Since we have already checked the distribution and sample size for the variable we know that we can use a simple one sample t-Test to test the null hypothesis that the average Republican voter is 51 years old.

```
t.test(S_2_republicans$dem_age_r_x, mu=51)
```

```
##
##  One Sample t-test
##
## data:  S_2_republicans$dem_age_r_x
## t = 0.87662, df = 1980, p-value = 0.3808
## alternative hypothesis: true mean is not equal to 51
## 95 percent confidence interval:
##  50.59093 52.07035
## sample estimates:
## mean of x
##  51.33064
```

The results of the one sample t-Test gives a p-value of 0.3808 with a confidence interval from 50.59 to 52.07, which means that we cannot reject the null hypothesis that the average Republican voter is 51 years old. In fact the mean age in our sample is right at 51.33.

4. Were Republican voters more likely to shift their political preferences right or left (more conservative or more liberal), compared to Democratic voters during the 2012 election?

In order to answer this question we need to look at the subset of the data where both the variables libcpre_self and libcpo_self are populated on the 1-7 ranked scale from Extremely liberal to Extremely conservative and then create a new variable to indicate whether this value changes or remains the same over the course of the election cycle.

```
S_4 = subset(S, S$libcpre_self %in% c("1. Extremely liberal",
                                      "2. Liberal",
                                      "3. Slightly liberal",
                                      "4. Moderate; middle of the road",
                                      "5. Slightly conservative",
                                      "6. Conservative",
                                      "7. Extremely conservative")
               & S$libcpo_self %in% c("1. Extremely liberal",
                                      "2. Liberal",
                                      "3. Slightly liberal",
                                      "4. Moderate; middle of the road",
                                      "5. Slightly conservative",
                                      "6. Conservative",
                                      "7. Extremely conservative"))

S_4$libcpre_self = factor(S_4$libcpre_self, ordered = TRUE, levels = c("1. Extremely liberal",
                                                            "2. Liberal",
                                                            "3. Slightly liberal",
                                                            "4. Moderate; middle of the road",
                                                            "5. Slightly conservative",
                                                            "6. Conservative",
                                                            "7. Extremely conservative"))

S_4$libcpo_self = factor(S_4$libcpo_self, ordered = TRUE, levels = c("1. Extremely liberal",
                                                            "2. Liberal",
                                                            "3. Slightly liberal",
                                                            "4. Moderate; middle of the road",
                                                            "5. Slightly conservative",
                                                            "6. Conservative",
                                                            "7. Extremely conservative"))

S_4$shifted = ifelse(S_4$libcpre_self==S_4$libcpo_self,0,1)

mean(S_4$shifted[S_4$party_affiliation == "Republican"])
```

```
## [1] 0.3146453
```

```
mean(S_4$shifted[S_4$party_affiliation == "Democrat"])
```

```
## [1] 0.3850622
```

By looking at the mean of the new shifted variable for both groups we can get an initial idea of how likely voters were to shift there political preferences. The mean for Republican voters at 0.315 is slightly lower than the mean for Democratic voters at 0.385, which suggests that in our sample Democrats were around 7% more likely to have some kind of shift in their preferences than Republicans. To test the significance we can't use a parametric test, because our shifted variable is binary and does not have a normal distribution. Instead, we will use a Wilcoxon Rank-Sum to compare the two independent groups.

```
wilcox.test(S_4$shifted[S_4$party_affiliation == "Republican"],
            S_4$shifted[S_4$party_affiliation == "Democrat"],
            paired=FALSE, alternative = c("two.sided"))
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  S_4$shifted[S_4$party_affiliation == "Republican"] and S_4$shifted[S_4$party_affiliation == "
## W = 1958000, p-value = 2.84e-06
## alternative hypothesis: true location shift is not equal to 0
```

The results of the Wilcoxon signed rank test give us a small p-value very close to 0, which suggests that the difference in preference shifting between Democrats and Republicans is significant. In terms of practical significance, we saw above that the difference between the means of the two groups is only around 7%, which doesn't seem very large. We can again use Cohen's d to evaluate this effect size in a more quantitative way.

```
cohensD(S_4$shifted[S_4$party_affiliation == "Republican"],
        S_4$shifted[S_4$party_affiliation == "Democrat"])
```

```
## [1] 0.147469
```

The value we get for Cohen's d is only 0.147 which confirms the impression that this effect size is small even though it is significant.

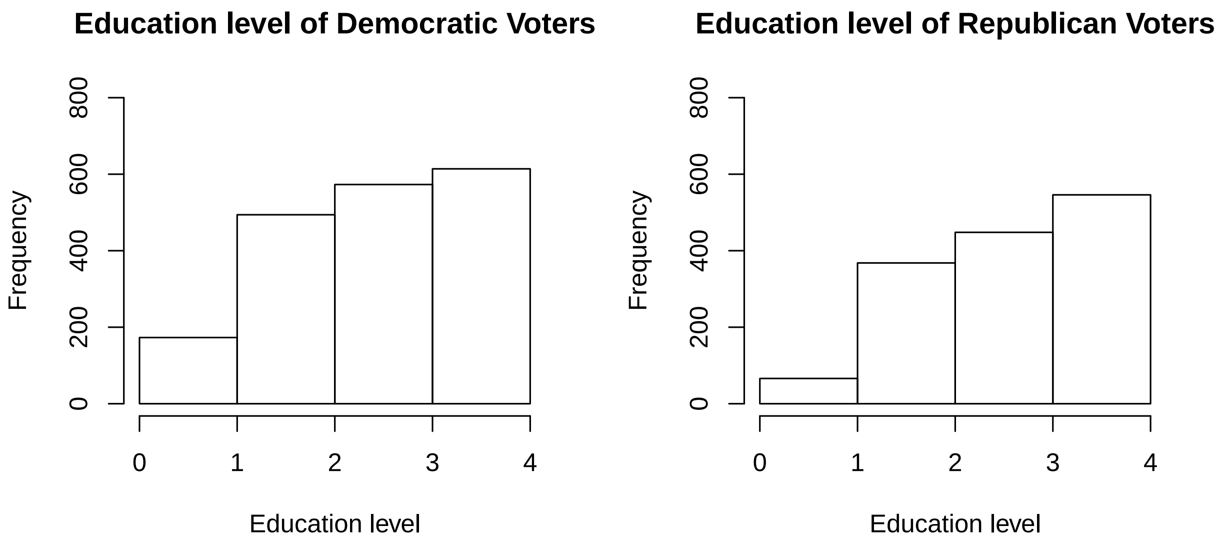
5. Do Democratic voters have the same level of education as Republican voters, on average?

In order to answer this question we will use our previously established party affiliations and compare the groups using the profile_educ variable. The first step is to clean up the education variable by removing the inapplicable responses and then ordering the remaining factor levels.

```
S_5 = S

S_5$profile_educ = factor(S_5$profile_educ, ordered = TRUE, levels =
                                                          c("1. Less than high school",
                                                            "2. High school",
                                                            "3. Some college",
                                                            "4. Bachelor's degree or higher"))
```

The education variable ranges on a scale from 1-4 with increasing education levels. It is important to note that the intervals between the ranks here are not equal. Now that we have cleaned up variables we can look at the histograms for both to get a quick understanding of what the data looks like.

```
hist(as.numeric(S_5$profile_educ[S_5$party_affiliation == "Democrat"]),
     xlab = 'Education level', ylab = 'Frequency',
     breaks = seq(0,4,by=1), ylim = c(0,800), main = 'Education level of Democratic Voters')

hist(as.numeric(S_5$profile_educ[S_5$party_affiliation == "Republican"]),
     xlab = 'Education level', ylab = 'Frequency',
     breaks = seq(0,4,by=1), ylim = c(0,800), main = 'Education level of Republican Voters')
```



The histograms indicate a skew towards higher levels of education for both groups, with the distributions looking fairly similar. Because our education variable is ordinal and the distribution is not normal we will need to use a non-parametric test to test the null hypothesis that Republican and Democratic voters have the same education level. Since the two group samples are independent we will use the Wilcoxon signed rank test again.

```
wilcox.test(as.numeric(S_5$profile_educ[S_5$party_affiliation == "Democrat"]),
            as.numeric(S_5$profile_educ[S_5$party_affiliation == "Republican"]),
            paired=FALSE, alternative = c("two.sided"))
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  as.numeric(S_5$profile_educ[S_5$party_affiliation == "Democrat"]) and as.numeric(S_5$profile_
## W = 1215400, p-value = 2.374e-05
## alternative hypothesis: true location shift is not equal to 0
```

```
mean(as.numeric(S_5$profile_educ[S_5$party_affiliation == "Democrat"]), na.rm=TRUE)
```

```
## [1] 2.878101
```

```
mean(as.numeric(S_5$profile_educ[S_5$party_affiliation == "Republican"]), na.rm=TRUE)
```

```
## [1] 3.032213
```

The results of the Wilcoxon signed rank test give us a small p-value very close to 0, which suggests that the difference in education level between Democrats and Republicans is significant. When we look at the difference in means we can see that Republican voters actually have a higher education level on average at 3.03 compared to the Democrat's 2.87. In terms of practical significance, this doesn't seem very large. We can again use Cohen's d to evaluate this effect size in a more quantitative way.

```
cohensD(as.numeric(S_5$profile_educ[S_5$party_affiliation == "Democrat"]),
        as.numeric(S_5$profile_educ[S_5$party_affiliation == "Republican"]))
```

```
## [1] 0.1625312
```

The value we get for Cohen's d is only 0.16 which confirms the impression that this effect size is small even though it is significant.

Conclusion

After conducting analysis on these five research questions, the overall conclusion is that while there are some statistically significant differences between Democrat and Republican voters, for the most part the effect sizes of these differences seem to be relatively small. The second and third hypothesis tests we conducted show that Democrats are about 2 years younger on average than Republicans, but this effect size is small and doesn't suggest a huge divide between the groups. The last question asks whether there is a difference in education level between Democrats and Republicans, and again we found a statistically significant difference showing Republicans are slightly more educated than Democrats but with a very small effect size. An interesting question for further research along these lines would be to create a binary education variable for Undergraduate degree and above. It is possible that the difference in education between Republicans and Democrats would be reversed if we defined the variable in this way.

The first and fourth questions ask about how Liberal/Conservative political preferences change over time. Interestingly, our first test suggested that there is not a significant shift when looking at the entire population. However, in the fourth test when we narrowed our focus to specifically Democrats and Republicans we saw between 30-40 % of the respondents make some kind of shift in preference, with Republican voters significantly more likely to change (again with a small effect size). In order to explain this discrepancy, an avenue for further research would be to look more closely at the shifting preferences of Non party identified voters, which our findings suggest should remain fairly stable.