Lab 3: Hypothesis Tests about the Mean.

w203: Statistics for Data Science

Introduction

The American National Election Studies (ANES) conducts surveys of voters in the United States before and after every presidential election. Using the data taken from the 2012 elections, we perform various hypothesis tests to answer the following questions of interest via various hypothesis tests:

1. Did voters become more liberal or more conservative during the 2012 election?
2. Were Republican voters older or younger, 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. Right before the 2012 election, were women voters more or less liberal than men voters?

The Data

There are a number of special concerns that arise whenever statisticians work with survey data. In particular, the complete ANES survey data assigns a survey weight to each observation, which corrects for differences in how likely individuals are to be selected, and how likely they are to respond. For the purposes of this assignment, however, we have removed the survey weights and we assume that the observations we have are a random sample from the voting population.

```
S = read.csv("/Users/shanhe/Desktop/w203/Lab/Lab_3/ANES_2012_sel.csv")
```

Analysis

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

We first do an EDA on our variables of interest

```
unique(S$libcpre_self)
```

```
##  [1] 1. Extremely liberal
##  [2] -2. Haven't thought much about this
##  [3] 2. Liberal
##  [4] -8. Don't know
##  [5] 4. Moderate; middle of the road
##  [6] 6. Conservative
##  [7] 5. Slightly conservative
##  [8] 3. Slightly liberal
##  [9] 7. Extremely conservative
## [10] -9. Refused
## 10 Levels: -2. Haven't thought much about this ... 7. Extremely conservative
```

```
unique(S$libcpo_self)
```

```
##  [1] -6. Not asked, unit nonresponse (no post-election interview)
##  [2] 2. Liberal
##  [3] -8. Don't know
##  [4] 4. Moderate; middle of the road
```

```
##  [5] -2. Haven't thought much {do not probe}
##  [6] 6. Conservative
##  [7] 5. Slightly conservative
##  [8] 3. Slightly liberal
##  [9] 7. Extremely conservative
## [10] 1. Extremely liberal
## [11] -7. Deleted due to partial (post-election) interview
## [12] -9. Refused
## 12 Levels: -2. Haven't thought much {do not probe} ...
```

Notice that 1) we have survey responses like "-2. Haven't thought much" and 2) we have different levels of liberalness and conservativeness. Hence, we will 1) add numeric variables that represent the level of conservativeness/liberalness ans 2) exclude answers with no applicability in terms of levels of liberalness and conservativeness.

```
#temporarily assign 0 to the NA answers, but will be excluded in the analysis
S_n <- mutate(
        S,
        libcpre_self_n = as.numeric(ifelse(substr(S$libcpre_self,0,1) == "-", 0, substr(S$libcpre_self,0
        libcpo_self_n = as.numeric(ifelse(substr(S$libcpo_self,0,1) == "-", 0, substr(S$libcpo_self,0,1)
)

# subsetting dataset, exclude voters with non-applicable answers, either pre or post the election
S_n_1 <- subset(S_n, libcpre_self_n != 0 & libcpo_self_n != 0, select= c(libcpre_self_n,libcpo_self_n))
```

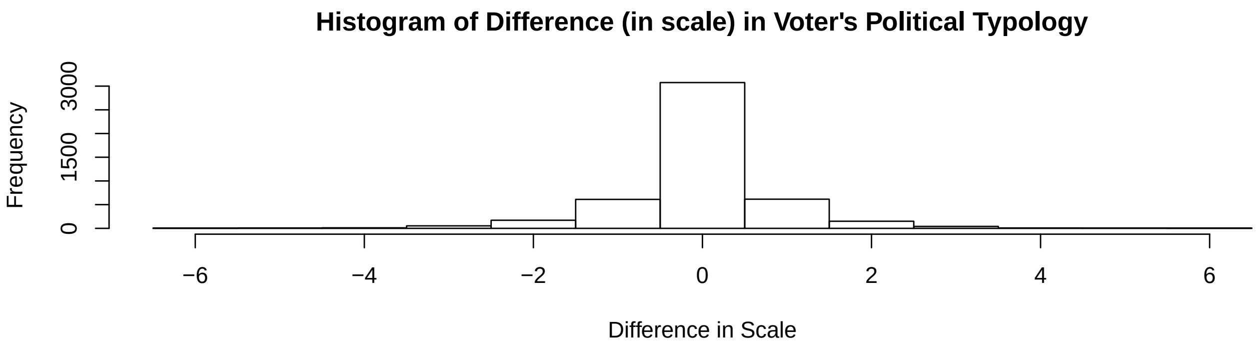
In this case, we have a very natural pairing between our observations. Each voter has a data point before and after the election. Since the variable is ordinal, we will want to use Wilcoxon Signed-Rank test, depending on whether the sample meets its assumption

We then investigate the difference

```
D_1 = S_n_1$libcpo_self_n - S_n_1$libcpre_self_n
summary(D_1)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -6.00000  0.00000  0.00000 -0.01913  0.00000  6.00000
```

```
hist(D_1, breaks = -6:7 - 0.5,
     main = "Histogram of Difference (in scale) in Voter's Political Typology",
     xlab = "Difference in Scale"
     )
```



The sampling distribution of the differences seems to have a symmetric distribution. With an approximately symmetric distribution and a large sample size on ordinal scale, we can use a two-tailed Wilcoxon Signed-Rank test.

Null Hpypothsis = voters didn't become either more liberal or conservative during the 2012 election Alternative

Hypothesis = voters did become either more liberal or conservative during the 2012 election

```r
# Wilcoxon Signed-Rank test
wilcox.test(S_n_1$libcpo_self_n, S_n_1$libcpre_self_n, paired = TRUE)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  S_n_1$libcpo_self_n and S_n_1$libcpre_self_n
## V = 682330, p-value = 0.1662
## alternative hypothesis: true location shift is not equal to 0
```

Notice that we have a p-value of 0.1662, showing weak statistical signifincance. Hence, we can't reject the null hypothesis with more than 84% confidence. We then look at the pratical significance of our hypothesis test:

```r
#Investigate Practical Significance
cohen.d(S_n_1$libcpo_self_n, S_n_1$libcpre_self_n, paired = TRUE)
```

```
##
## Cohen's d
##
## d estimate: -0.02006681 (negligible)
## 95 percent confidence interval:
##         inf         sup
## -0.06025672  0.02012310
```

we have a Cohen's d as -0.02, which indicates small practical significance

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

We first do an EDA on our variables of interest

```r
unique(S_n$pid_x)
```

```
## [1] 1. Strong Democrat            3. Independent-Democrat
## [3] 4. Independent                6. Not very strong Republican
## [5] 5. Independent-Republican     2. Not very strong Democract
## [7] 7. Strong Republican          -2. Missing
## 8 Levels: -2. Missing 1. Strong Democrat ... 7. Strong Republican
```

```r
summary(S_n$dem_age_r_x)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -2.00   35.00   51.00   48.92   62.00   90.00
```

Notice that in pid_x we have "-2. Missing" and different levels of Demoracts and Republicans I will 1) exclude records with "-2. Missing" or "4. Independent" as pid_x since they can't be categrized as either Republican voters or Democratic voters and 2) create a categorical value based on pid_x to categorize Republican voters or Democratic voters

```r
S_n_2 <- mutate(subset(S_n,substr(pid_x,0,1) != '-' & substr(pid_x,0,1) != '4',
                       select = c(pid_x,dem_age_r_x)),
                voter_cat = factor(ifelse(substr(pid_x,0,1) < 4, 'Democrat', 'Republican'))
              )

summary(S_n_2$dem_age_r_x)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   -2.00   35.00   51.00   49.32   62.00   90.00
```

Notice that we see negative age, which doens't make sense and should be excluded

```
S_n_2 <- subset(S_n_2,dem_age_r_x > 0)
```

In this case that we want to compare the average age for Republican voters and Democrat voters, they can be assumed to be independent of each other. Depending our sample distribution, we can use an independent t-test or a Wilcoxon Rank Sum test. We can look at the distribution of dem_age_r_x for both Republican voters and Democrat voters

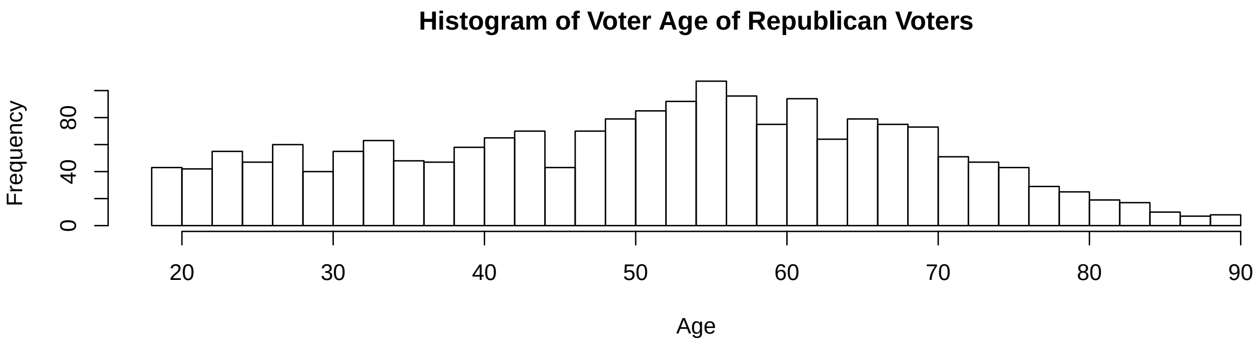
```
# For Republican Voters
length(subset(S_n_2, voter_cat == "Republican")$dem_age_r_x)
```

```
## [1] 1981
```

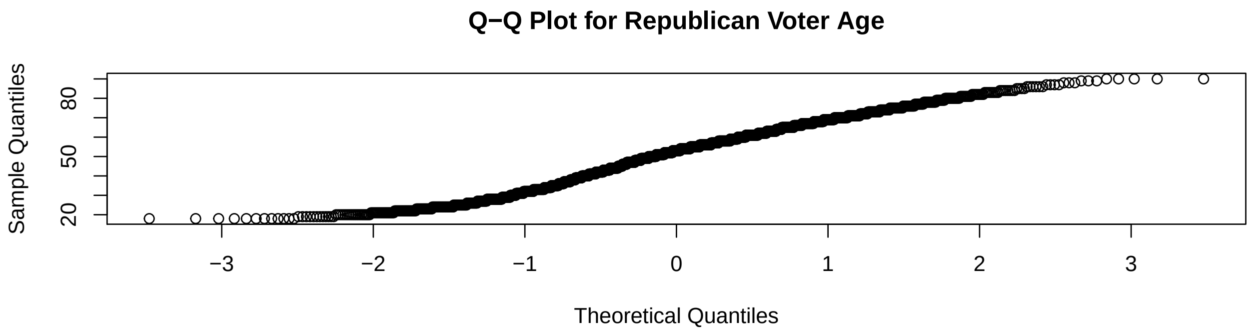
```
summary(subset(S_n_2, voter_cat == "Republican")$dem_age_r_x)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   38.00   53.00   51.33   64.00   90.00
```

```
hist(subset(S_n_2, voter_cat == "Republican")$dem_age_r_x, breaks = 50,
     main = "Histogram of Voter Age of Republican Voters",
     xlab = "Age")
```




```
qqnorm(subset(S_n_2, voter_cat == "Republican")$dem_age_r_x, main = "Q-Q Plot for Republican Voter Age"
```



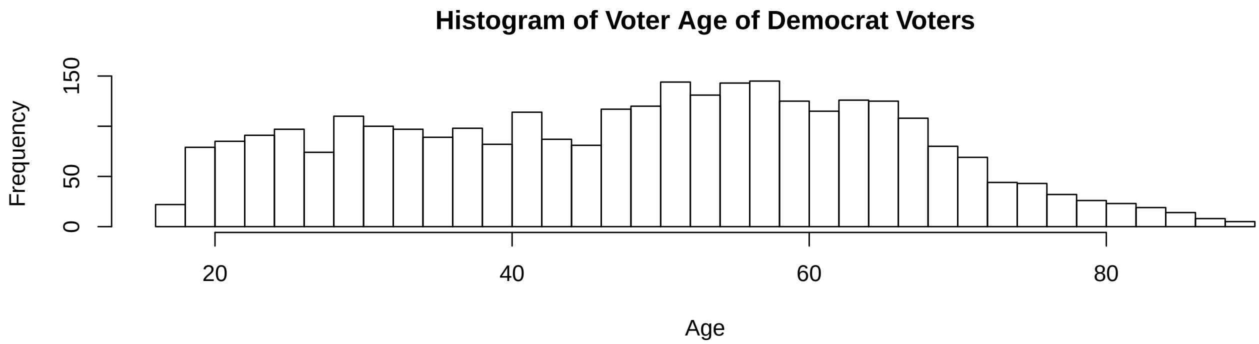

```
# For Democrat Voters
length(subset(S_n_2, voter_cat == "Democrat")$dem_age_r_x)
```

```
## [1] 3068
```

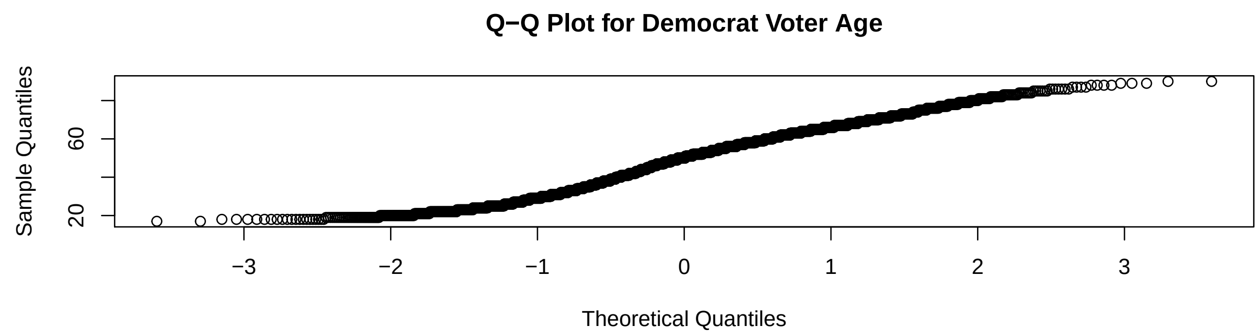
```
summary(subset(S_n_2, voter_cat == "Democrat")$dem_age_r_x)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   17.00   35.00   50.00   48.84   62.00   90.00
```

```r
hist(subset(S_n_2, voter_cat == "Democrat")$dem_age_r_x, breaks = 50,
     main = "Histogram of Voter Age of Democrat Voters",
     xlab = "Age")
```




```r
qqnorm(subset(S_n_2, voter_cat == "Democrat")$dem_age_r_x, main = "Q-Q Plot for Democrat Voter Age")
```



Notice that both distributions are slightly skewed. Here we want to compare the age of the Democrat voters to that of the Republican voters. Although we don't have normal distributions, we do have large sample sizes allowing us to perform a two-sample t-test for the continuous vairble, age, with:

Null Hypothesis = the average age of Democrat voters and Republican voters were the same in 2012 election
Alternative Hypothsis = the average age of Democrat voters and Republican voters were different in 2012 election

```r
t.test(dem_age_r_x ~ voter_cat, data = S_n_2)
```

```
##
##  Welch Two Sample t-test
##
## data:  dem_age_r_x by voter_cat
## 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 in group Democrat mean in group Republican
##                 48.83735                 51.33064
```

Notice that we have a p-value as 2.512e-07, showing strong statistical significance. We hence are confident to reject the null hypothesis. Moreover, the confidence interval shows us a negative direction with an negative upper bound, indicating strong evidence that the average age of the Democrat voters was smaller than that of the Repulican voters in 2012 election. We then look at the practical significance for our hypothesis test:

```r
cohen.d(dem_age_r_x ~ voter_cat, data = S_n_2)
```

```
##
## Cohen's d
##
## d estimate: -0.149074 (negligible)
## 95 percent confidence interval:
##         inf         sup
## -0.20565352 -0.09249442
```

Although our sample statistics show strong statistical significance, the effect size is quite small. With a Cohen's d value of -0.149, it shows small practicial significance.

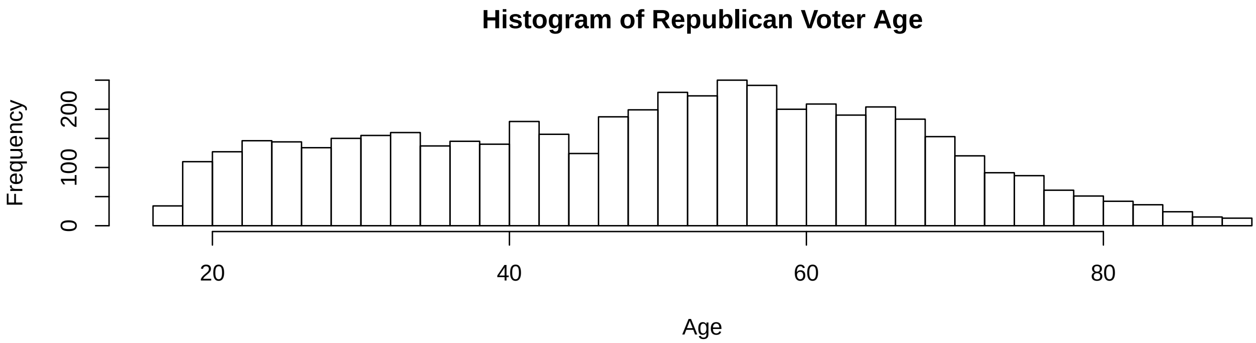
3. Were Republican voters older than 51, on the average in 2012?

We first create a subset of the data used in question 2 to only contain the age for Republican voters and perform a EDA on the variables of interest

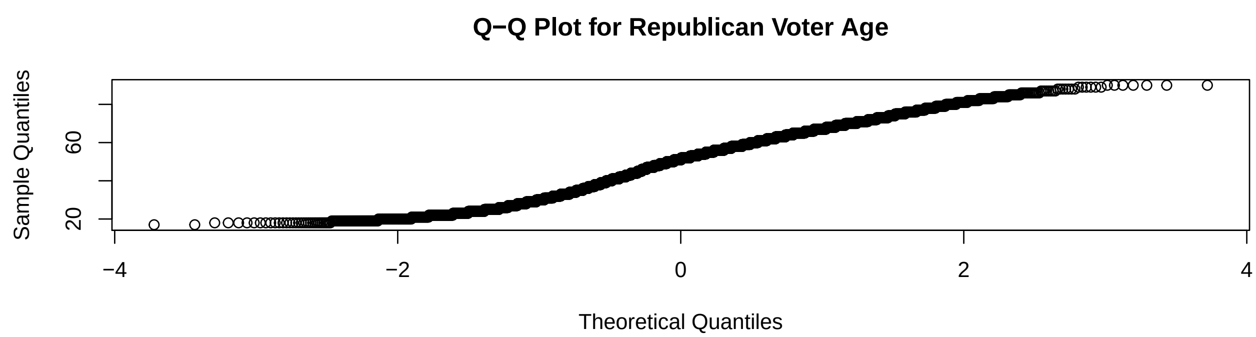
```r
S_n_3 <- subset(S_n_2, voter_cat == "Republican", select = c(voter_cat, dem_age_r_x))

summary(S_n_3$dem_age_r_x)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   38.00   53.00   51.33   64.00   90.00
```

```r
hist(S_n_2$dem_age_r_x, breaks = 50,
     main = "Histogram of Republican Voter Age",
     xlab = "Age")
```




```r
qqnorm(S_n_2$dem_age_r_x, main = "Q-Q Plot for Republican Voter Age")
```



Notice that the distribution is skewed with a peak at around 55. Here we want to compare the age of the Republican voters to a constant. Although we don't have a normal sample distribution, we do have a large sample size that invokes Central Limite Theorem allowing us to perform a parametric t-test:

Null Hypothesis = In 2012, the average age of the Republican voters was 51 Alternative Hypothesis = In 2012, the average age of the Republican voters was greater than 51

```
t.test(S_n_3$dem_age_r_x - 50)
```

```
##
##  One Sample t-test
##
## data:  S_n_3$dem_age_r_x - 50
## t = 3.5279, df = 1980, p-value = 0.0004284
## alternative hypothesis: true mean is not equal to 0
## 95 percent confidence interval:
##  0.5909323 2.0703498
## sample estimates:
## mean of x
##  1.330641
```

Notice our t-statistic shows strong statistical significance rejecting the null hypothesis with up to 99.97% confidence. Moreover, the two-tailed test 95% Confidence Interval has a lower bound greater than 0, showing strong evidence that the mean age of the Republican voters was greater than 51 in the 2012 election. We then look at the practical significance for our hypothesis test:

```
(mean(S_n_3$dem_age_r_x) - 50) / sd(S_n_3$dem_age_r_x)
```

```
## [1] 0.0792632
```

We have an effect size of 0.08, which means a small practical significance.

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?

We first create a subset of the data to exclude non applicable survey answers and create a variable to identify Demcrat or Republican voters, similar to what's been done previously

```
S_n_4 <- mutate(subset(S_n, libcpre_self_n != 0 & libcpo_self_n != 0 & substr(pid_x,0,1) != '-' & subst
                       select = c(pid_x,libcpre_self_n,libcpo_self_n)),
                voter_cat = factor(ifelse(substr(pid_x,0,1) < 4, 'Democrat', 'Republican'))
              )
```

We then examine the differences between the libcpre_self_n and libcpo_self_n, representing voter's swing in political typology before and after election, for both Republican and Democrat voters

```
D_4_R <- subset(S_n_4, voter_cat == 'Republican')$libcpo_self_n - subset(S_n_4, voter_cat == 'Republica
D_4_D <- subset(S_n_4, voter_cat == 'Democrat')$libcpo_self_n - subset(S_n_4, voter_cat == 'Democrat')$

# For Republican Voters
length(D_4_R)
```

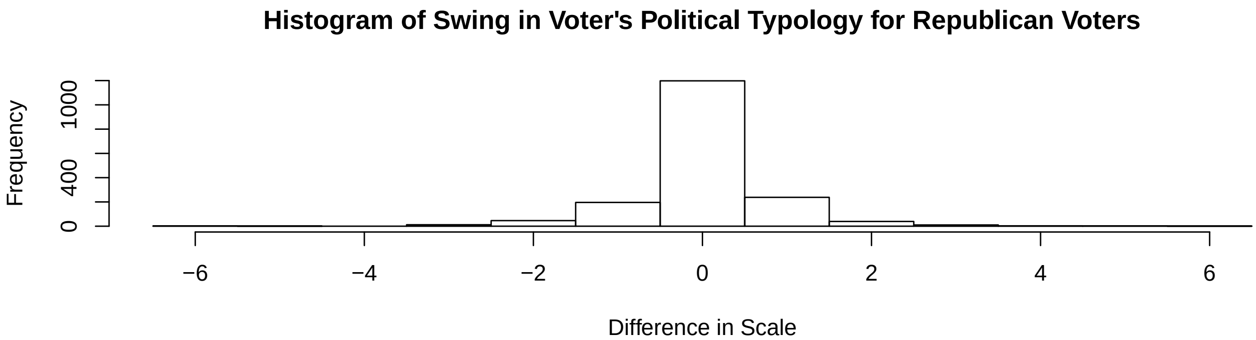
```
## [1] 1748
```

```
summary(D_4_R)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -6.00000  0.00000  0.00000  0.01888  0.00000  6.00000
```

```
hist(D_4_R, breaks = -6:7 - .5,
     main = "Histogram of Swing in Voter's Political Typology for Republican Voters",
```

```r
     xlab = "Difference in Scale"
     )
```



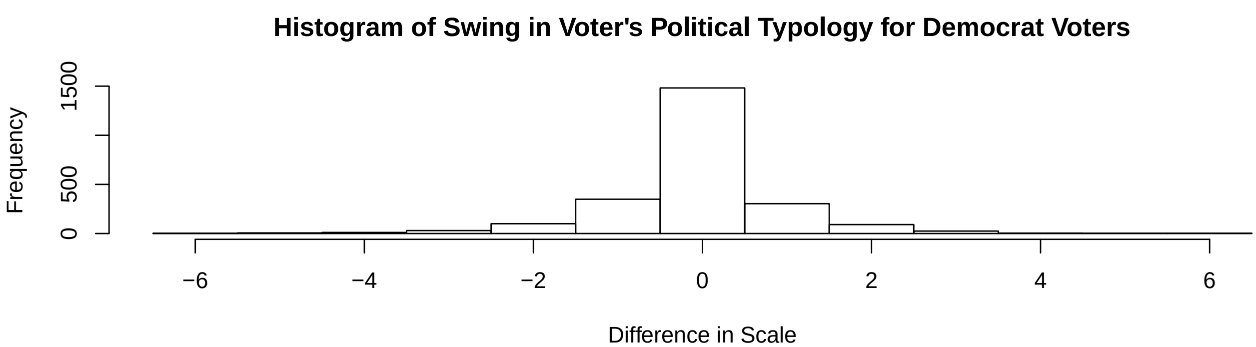

```r
# For Democrat Voters
length(D_4_D)
```

```
## [1] 2410
```

```r
summary(D_4_D)
```

```
##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
## -6.00000  0.00000  0.00000 -0.04772  0.00000  6.00000
```

```r
hist(D_4_D, breaks = -6:7 - .5,
     main = "Histogram of Swing in Voter's Political Typology for Democrat Voters",
     xlab = "Difference in Scale"
     )
```



We don't see any natural pairing between the samples and we assume them to be independent samples. It seems like the differences between the libcpre_self_n and libcpo_self_n for both Republican and Democrat voters are symmetrically distributed with similar shapes, on top of having large sample sizes, allowing us to perform the Wilcoxon Rank Sum test.

Since our data is in ordinal scale, we will do the Wilcoxon Rank Sum test:

Null Hypothese = Republican voters weren't more likely to shift their political preferences right or left compared to Democratic voters during the 2012 election

Alternative Hypothese = Republican voters were more likely to shift their political preferences right or left compared to Democratic voters during the 2012 election

```r
wilcox.test(D_4_R, D_4_D)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
```

```
## data:  D_4_R and D_4_D
## W = 2189300, p-value = 0.01096
## alternative hypothesis: true location shift is not equal to 0
```

```
c(mean(D_4_R), mean(D_4_D))
```

```
## [1]  0.01887872 -0.04771784
```

Notice that we have a p-value of 0.01096 which shows strong statistical significance, rejecting the null hypothese at ~99% confidence level.This means that our sample data is poorly explained by the null hypothese compared to our alternative hypothesis. Moreover, from comparing the average swing for both Republican and Democrat voters, combined with our hypothesis test, we have strong evidence that the Repulican voters were more likely to swing left during the 2012 election. We then investigate the practical significance of our hypothesis test:

```
# Investigate Practical Significance
cohen.d(D_4_R, D_4_D)
```

```
##
## Cohen's d
##
## d estimate: 0.0700727 (negligible)
## 95 percent confidence interval:
##         inf         sup
## 0.008460392 0.131685002
```

We see an effect size of 0.07, which shows small practical significance.

5. Right before the 2012 election, were women voters equally liberal as men voters?

We first do an EDA on variables of interest:

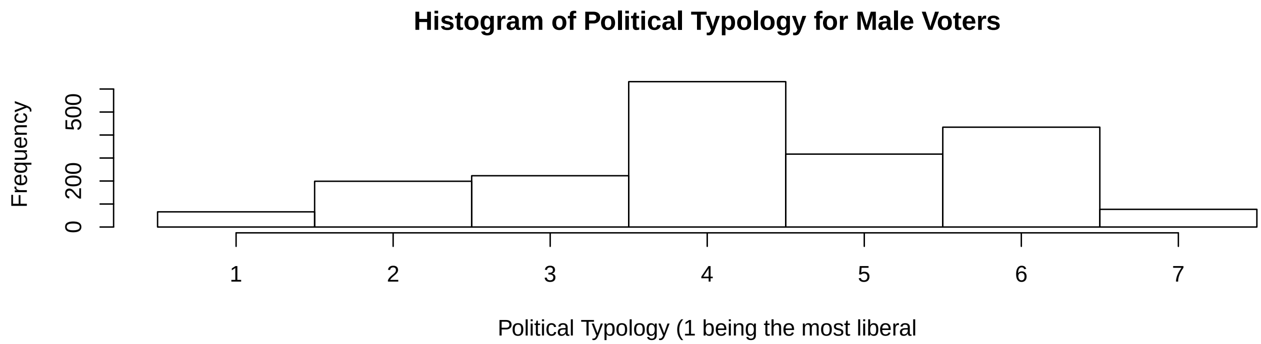
```
# Create subset that includes records with applicable gender and libcpre_self_n response
S_n_5 <- mutate(subset(S_n, libcpre_self_n != 0 & substr(profile_gender,0,1) != "-",
                       select = c(libcpre_self_n, profile_gender)),
                gender = factor(ifelse(profile_gender == '1. Male', "Male", "Female"))
              )

# For Male Voters
length(subset(S_n_5, gender == "Male")$libcpre_self_n)
```

```
## [1] 1948
```

```
summary(subset(S_n_5, gender == "Male")$libcpre_self_n)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   3.000   4.000   4.306   6.000   7.000
```

```
hist(subset(S_n_5, gender == "Male")$libcpre_self_n, breaks = 1:8 - 0.5,
     main = "Histogram of Political Typology for Male Voters",
     xlab = "Political Typology (1 being the most liberal")
```



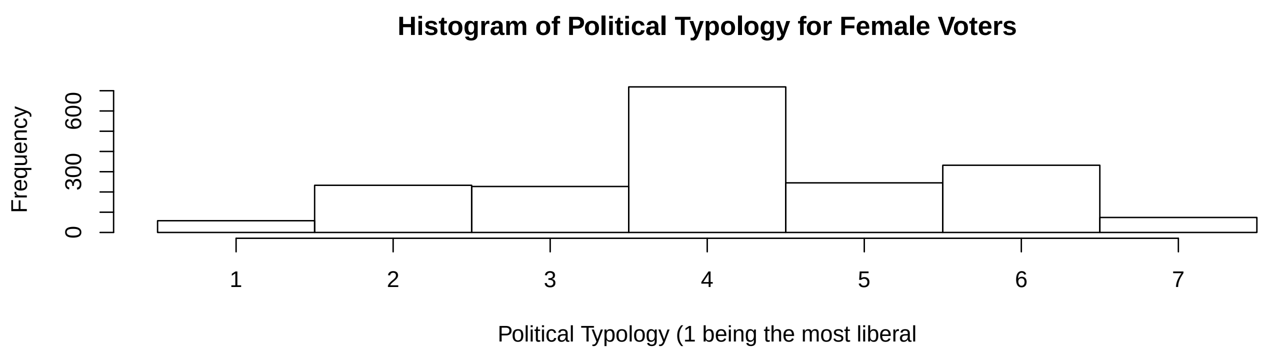

```r
# For Female Voters
length(subset(S_n_5, gender == "Female")$libcpre_self_n)
```

```
## [1] 1888
```

```r
summary(subset(S_n_5, gender == "Female")$libcpre_self_n)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    3.00    4.00    4.14    5.00    7.00
```

```r
hist(subset(S_n_5, gender == "Female")$libcpre_self_n, breaks = 1:8 - 0.5,
     main = "Histogram of Political Typology for Female Voters",
     xlab = "Political Typology (1 being the most liberal")
```



Although they don't seem to be normally distributed, the large sample sizes involes Central Limit Theorem and approximate the sampling distributions for the sample means to be normally distributed. With that, we can run a Wilcoxon Rank Sum test for the ordinal value libcpre_self_n:

Null Hypothesis: Right before the 2012 election, women voters were equally liberal as men voters Alternative Hypothesis: Right before the 2012 election, women voters were more or less liberal than men voters

```r
wilcox.test(libcpre_self_n ~ gender, data = S_n_5)
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  libcpre_self_n by gender
## W = 1707700, p-value = 8.057e-05
## alternative hypothesis: true location shift is not equal to 0
```

Notice that we have a p-value of 8.057e-05, which shows strong statistical significance of our test statistic, rejecting the null hypothese at > 99% confidence level. This means that our sample data is poorly explained by the null hypothese compared to our alternative hypothesis. And we have strong evidence that right before the 2012 election, women voters were not equally liberal as men voters. We then look at the practical significance of our hypothesis test

```r
# Investigate Practical Significance
cohen.d(libcpre_self_n ~ gender, data = S_n_5)
```

```
##
## Cohen's d
##
## d estimate: 0.1149731 (negligible)
## 95 percent confidence interval:
##        inf        sup
## 0.05160251 0.17834361
```

We see an effect size of 0.11, which shows small practical significance.

Conclusions

After performing hypothesis testings, whether parametric or non parametric, for the questions of interests. We 1) rejected null hypotheses that provides poor explanations of the data against alternative hypotheses and 2) didn't reject the null hypotheses that provided good explanation of the data.

Although we saw strong statistical significances for a few of the hypothesis tests, all of them had small practical significances. This is due to a large sample size that we have for our dataset. But please note that we are using a simplied version of the ANES survey data with the survey weight removed. The results of the analyses above are done based purely on the simplied sample and may not reflect actual situations for 2012 election.