Lab 3: Hypothesis Tests about the Mean.

w203: Statistics for Data Science

Tako Hisada

11/08/2017

Introduction

The Data

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.

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 ask you to assume that the observations you have are a random sample from the voting population.

For a glimpse into some of the intricacies that go into survey design, take a look at the introduction to the ANES User's Guide and Codebook.

```
S = read.csv("ANES_2012_sel.csv")
```

Following is an example of a question asked on the ANES survey:

Where would you place YOURSELF on this scale, or haven't you thought much about this?

Possible answers included:

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

The variable libcpre_self records answers before the election, while libcpo_self records answers after the election.

Analysis

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

Analysis

All survey responses can be forced into numeric values on the scale of 1-7, -2, -8 or -9.

```
table(S$libcpre_self, as.numeric(S$libcpre_self))
```

```
##
##                                      1    2    3    4    5    6    7
##   -2. Haven't thought much about this  556    0    0    0    0    0    0
##   -8. Don't know                         0   26    0    0    0    0    0
##   -9. Refused                            0    0   32    0    0    0    0
##   1. Extremely liberal                   0    0    0  195    0    0    0
##   2. Liberal                             0    0    0    0  638    0    0
##   3. Slightly liberal                    0    0    0    0    0  641    0
##   4. Moderate; middle of the road        0    0    0    0    0    0 1828
##   5. Slightly conservative               0    0    0    0    0    0    0
##   6. Conservative                        0    0    0    0    0    0    0
##   7. Extremely conservative              0    0    0    0    0    0    0
##
##                                      8    9   10
##   -2. Haven't thought much about this    0    0    0
##   -8. Don't know                         0    0    0
##   -9. Refused                            0    0    0
##   1. Extremely liberal                   0    0    0
##   2. Liberal                             0    0    0
##   3. Slightly liberal                    0    0    0
##   4. Moderate; middle of the road        0    0    0
##   5. Slightly conservative             789    0    0
##   6. Conservative                        0 1001    0
##   7. Extremely conservative              0    0  208
```

```
n = length(S$libcpre_self)
```

Our n is 5914.

In order to measure where people stand on the political spectrum, however, the answers -2, -8 or -9 are not informative. Also because of the presence of these values, the numeric values assigned by R are off by 3 from the numbers present in the survey answer choices (i.e. R assigned 4 to "1. Extremely liberal").

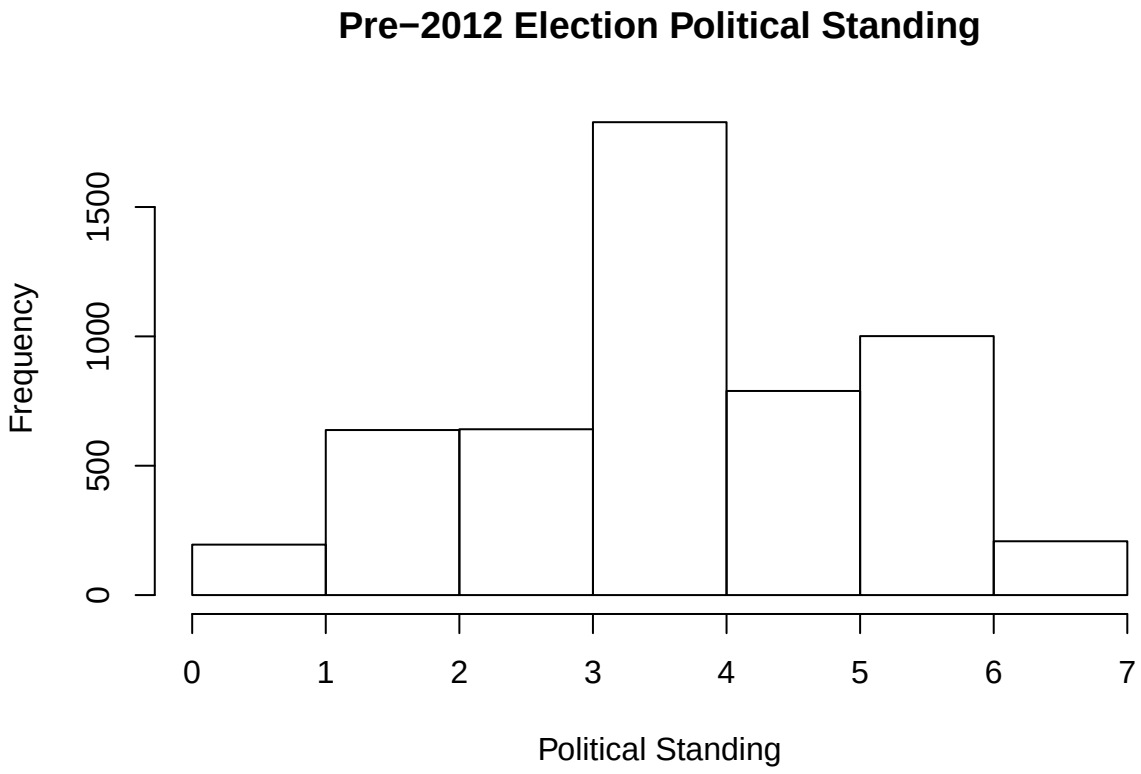
We need to re-assign values to match the survey answer choices and also 0 out the values for the choices -2, -8 and -9.

```
# Define bins for political standing histograms
bins_standing <- seq(from = 0, to = 7, by = 1)

# Function for generating a vector filled with a specified value
generateSeq <- function(v, value) {
  seq = c()
  for(i in 1:length(v)) {
    seq[i] = value
  }
  seq
```

```
}

# Function for generating political weights nulling invalid answers
compilePoliticalWeights <- function(v, null_index, skip_zero = FALSE) {
  num_v = as.numeric(v)
  weights = c()

  for(i in 1:length(unique(num_v))) {
    if(i <= null_index) {
      if(!skip_zero) {
        weights = c(weights, generateSeq(num_v[num_v == i], 0))
      }
    } else {
      weights = c(weights, generateSeq(num_v[num_v == i], i - null_index))
    }
  }
  weights
}

libcpre_weights <- compilePoliticalWeights(S$libcpre_self, 3, TRUE)
summary(libcpre_weights)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   3.000   4.000   4.172   5.000   7.000
```

```
mu_libcpre = mean(libcpre_weights)
hist(libcpre_weights, main = "Pre-2012 Election Political Standing",
     xlab = "Political Standing", breaks = bins_standing)
```



Our μ pre 2012 election is 4.1722642.

Likewise, we will need to do the same data sanitizing for the post-election data.

```
table(S$libcpo_self, as.numeric(S$libcpo_self))
```

```
##
##                                                                  1    2
##   -2. Haven't thought much {do not probe}                      410    0
##   -6. Not asked, unit nonresponse (no post-election interview)   0  252
##   -7. Deleted due to partial (post-election) interview           0    0
##   -8. Don't know                                                 0    0
##   -9. Refused                                                    0    0
##   1. Extremely liberal                                           0    0
##   2. Liberal                                                     0    0
##   3. Slightly liberal                                            0    0
##   4. Moderate; middle of the road                                0    0
##   5. Slightly conservative                                       0    0
##   6. Conservative                                                0    0
##   7. Extremely conservative                                      0    0
##
##                                                                  3    4
##   -2. Haven't thought much {do not probe}                        0    0
##   -6. Not asked, unit nonresponse (no post-election interview)   0    0
##   -7. Deleted due to partial (post-election) interview         152    0
##   -8. Don't know                                                 0   23
##   -9. Refused                                                    0    0
##   1. Extremely liberal                                           0    0
##   2. Liberal                                                     0    0
##   3. Slightly liberal                                            0    0
##   4. Moderate; middle of the road                                0    0
##   5. Slightly conservative                                       0    0
##   6. Conservative                                                0    0
##   7. Extremely conservative                                      0    0
##
##                                                                  5    6
##   -2. Haven't thought much {do not probe}                        0    0
##   -6. Not asked, unit nonresponse (no post-election interview)   0    0
##   -7. Deleted due to partial (post-election) interview           0    0
##   -8. Don't know                                                 0    0
##   -9. Refused                                                   36    0
##   1. Extremely liberal                                           0  166
##   2. Liberal                                                     0    0
##   3. Slightly liberal                                            0    0
##   4. Moderate; middle of the road                                0    0
##   5. Slightly conservative                                       0    0
##   6. Conservative                                                0    0
##   7. Extremely conservative                                      0    0
##
##                                                                  7    8
##   -2. Haven't thought much {do not probe}                        0    0
##   -6. Not asked, unit nonresponse (no post-election interview)   0    0
##   -7. Deleted due to partial (post-election) interview           0    0
##   -8. Don't know                                                 0    0
##   -9. Refused                                                    0    0
##   1. Extremely liberal                                           0    0
##   2. Liberal                                                   646    0
```

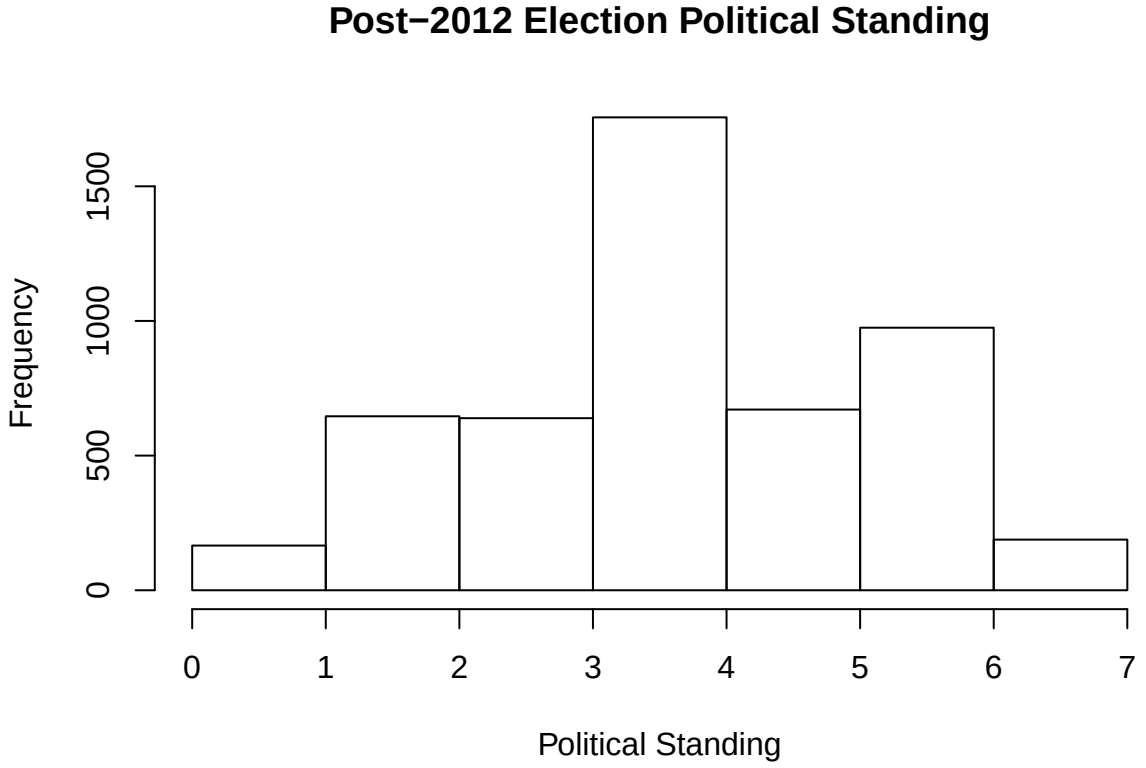
```
##   3. Slightly liberal                                              0  639
##   4. Moderate; middle of the road                                  0    0
##   5. Slightly conservative                                         0    0
##   6. Conservative                                                  0    0
##   7. Extremely conservative                                        0    0
##
##                                                                    9   10
##   -2. Haven't thought much {do not probe}                          0    0
##   -6. Not asked, unit nonresponse (no post-election interview)     0    0
##   -7. Deleted due to partial (post-election) interview             0    0
##   -8. Don't know                                                   0    0
##   -9. Refused                                                      0    0
##   1. Extremely liberal                                             0    0
##   2. Liberal                                                       0    0
##   3. Slightly liberal                                              0    0
##   4. Moderate; middle of the road                               1756    0
##   5. Slightly conservative                                         0  671
##   6. Conservative                                                  0    0
##   7. Extremely conservative                                        0    0
##
##                                                                   11   12
##   -2. Haven't thought much {do not probe}                          0    0
##   -6. Not asked, unit nonresponse (no post-election interview)     0    0
##   -7. Deleted due to partial (post-election) interview             0    0
##   -8. Don't know                                                   0    0
##   -9. Refused                                                      0    0
##   1. Extremely liberal                                             0    0
##   2. Liberal                                                       0    0
##   3. Slightly liberal                                              0    0
##   4. Moderate; middle of the road                                  0    0
##   5. Slightly conservative                                         0    0
##   6. Conservative                                                975    0
##   7. Extremely conservative                                        0  188
```

The data now contains 2 additional invalid values -6 and -7 in addition to -2, -8 and -9 we saw previously.

```
libcpo_weights <- compilePoliticalWeights(S$libcpo_self, 5, TRUE)
summary(libcpo_weights)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1.00    3.00    4.00    4.15    5.00    7.00
```

```
mu_libcpo = mean(libcpo_weights)
hist(libcpo_weights, main = "Post-2012 Election Political Standing",
     xlab = "Political Standing", breaks = bins_standing)
```



Our μ post 2012 election is 4.1499702 which is slightly lower than before the election suggesting there were more people idntified with the liberal-side of the political spectrum.

Hypotheses

From this, we can formulate our hypotheses as below:

μ_1 = Pre-2012 election political standing
μ_2 = Post-2012 election political standing

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_1 : \mu_1 - \mu_2 > 0$$

Statistical Significance

Justificaition

Our distribution is fairly normal and our n is sufficiently large. We can conduct a t-test in determining if the post-election mean is in fact lower than the pre-election.

```
t.test(libcpre_weights, libcpo_weights, alternative = "greater")
```

```
##
##  Welch Two Sample t-test
##
## data:  libcpre_weights and libcpo_weights
## t = 0.77123, df = 10314, p-value = 0.2203
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
```

```
##  -0.02525793         Inf
## sample estimates:
## mean of x mean of y
##  4.172264  4.149970
```

Result

Our p-value is 0.2203. We fail to reject H_0. We are unable to say if people became more liberal nor conservative as a result of this.

Practical Significance

We are going to calculate Cohen's d to measure effect size.

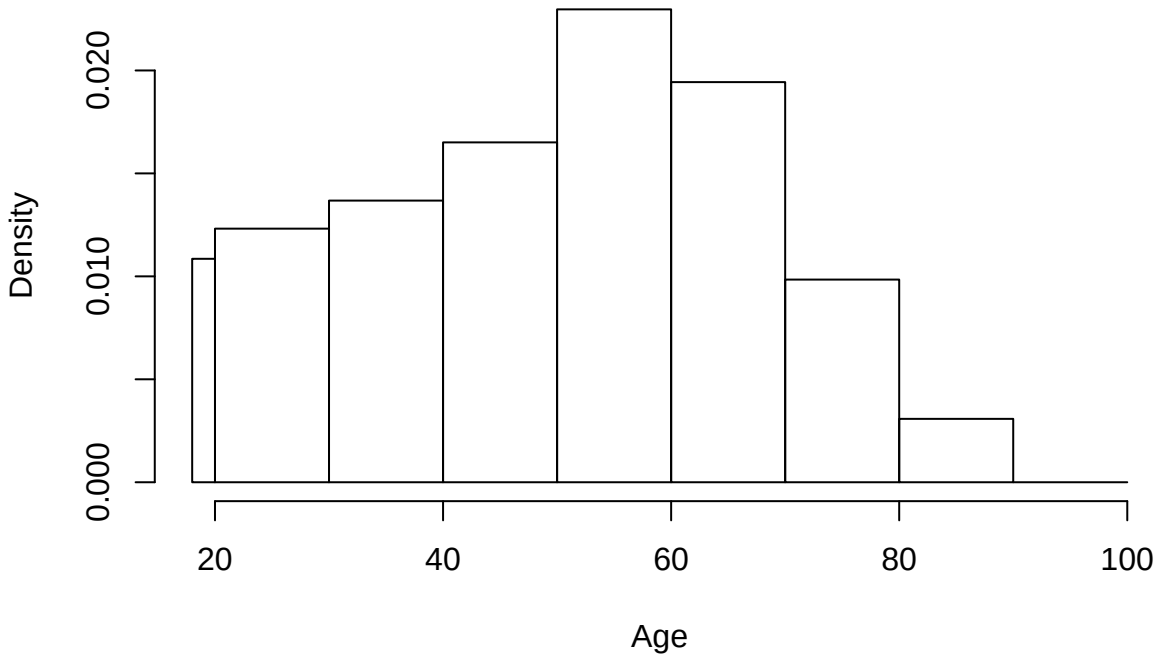
```
calcPooledSd <- function(v1, v2) {
  ((length(v1)-1)*sd(v1)^2 + (length(v2)-1)*sd(v2)^2)/(length(v1) + length(v2) - 2)
}
calcCohensD <- function(v1, v2) {
  (mean(v1) - mean(v2))/calcPooledSd(v1, v2)
}

(d = calcCohensD(libcpre_weights, libcpo_weights))
```

```
## [1] 0.01032583
```

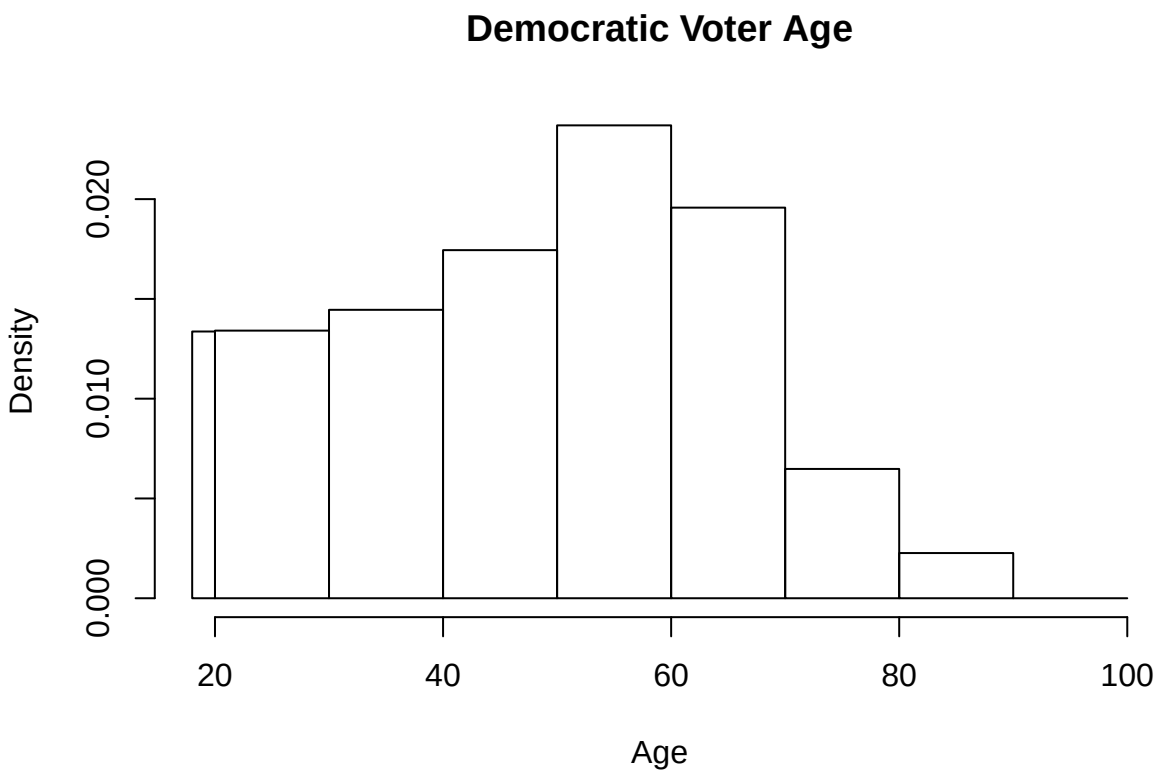
The Cohen's d value for this case is 0.0103258 which is < 0.8. There is no practical significance observed between μ_1 and μ_2.

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

```
# Define bins for voter age histograms
bins_ages <- seq(from = 20, to = 100, by = 10)
bins_ages <- c(18, bins_ages)

# Extract Republican voters
rep_ages <- S$dem_age_r_x[grep("republican", S$pid_x, ignore.case = T)]
rep_ages <- rep_ages[is.numeric(rep_ages) & rep_ages >= 18]
rep_ages_n = length(rep_ages)
rep_age_mean = mean(rep_ages)
hist(rep_ages, main = "Republican Voter Age", xlab = "Age", breaks = bins_ages)
```




```
# Extract Democratic voters
dem_ages <- S$dem_age_r_x[grep("democrat", S$pid_x, ignore.case = T)]
dem_ages <- dem_ages[is.numeric(dem_ages) & dem_ages >= 18]
dem_ages_n = length(dem_ages)
dem_age_mean = mean(dem_ages)
hist(dem_ages, main = "Democratic Voter Age", xlab = "Age", breaks = bins_ages)
```



The average Republican voter age is 51.3306411 which is in a little bit higher that that of the Democratic voeters' 49.6946081.

Hypotheses

From this, we can formulate our hypotheses as below:

μ_1 = Average Republican voter age
μ_2 = Average Democratic voter age

$$H_0 : \mu_1 - \mu_2 = 0$$
$$H_1 : \mu_1 - \mu_2 > 0$$

Statistical Significance

Justificaition

Our n for Republicans is 1981 and for Democrats is 2207. The distributions are fairly normal as seen above and our n's are sufficiently large. We can conduct a t-test in determining if μ_1 is in fact greater than μ_2.

```
rs <- t.test(rep_ages, dem_ages, alternative = "greater")
p <- rs$p.value
```

Result

Our p-value is 0.0007. We can reject H_0 at a significance level of 0.05, 0.01 and 0.001.

Practical Significance

We are going to calculate Cohen's D to measure effect size.

```
(d = calcCohensD(rep_ages, dem_ages))
```

```
## [1] 0.00600131
```

The Cohen's d value for this case is 0.0060013 which is < 0.8. There is no practical significance observed between μ_1 and μ_2.

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

The average Republican voter is 51.3306411.

Hypotheses

From this, we can formulate our hypotheses as below:

$$H_0 : \mu = 51$$
$$H_1 : \mu > 51$$

Statistical Significance

Justificaition

Our n is 1981. The distribution is fairly normal as seen in Q2 and our n is sufficiently large. We can conduct a t-test in determining if the average is in fact greater than 51 or not.

```
rs <- t.test(rep_ages, alternative = "greater", mu = 51)
p <- rs$p.value
```

Result

Our p-value is 0.1904006. We fail to reject H_0 at a significance level of 0.05 or anything smaller.

Practical Significance

We are going to calculate correlation r to measure effect size.

```
computeEffectSizeCorrelationR <- function(t, df) {
  t/sqrt(t^2+df)
}
r = computeEffectSizeCorrelationR(rs$statistic, rep_ages_n-1)
```

The effect size correlation r for this case is 0.0196967 which is < 0.8. There is no practical significance observed.

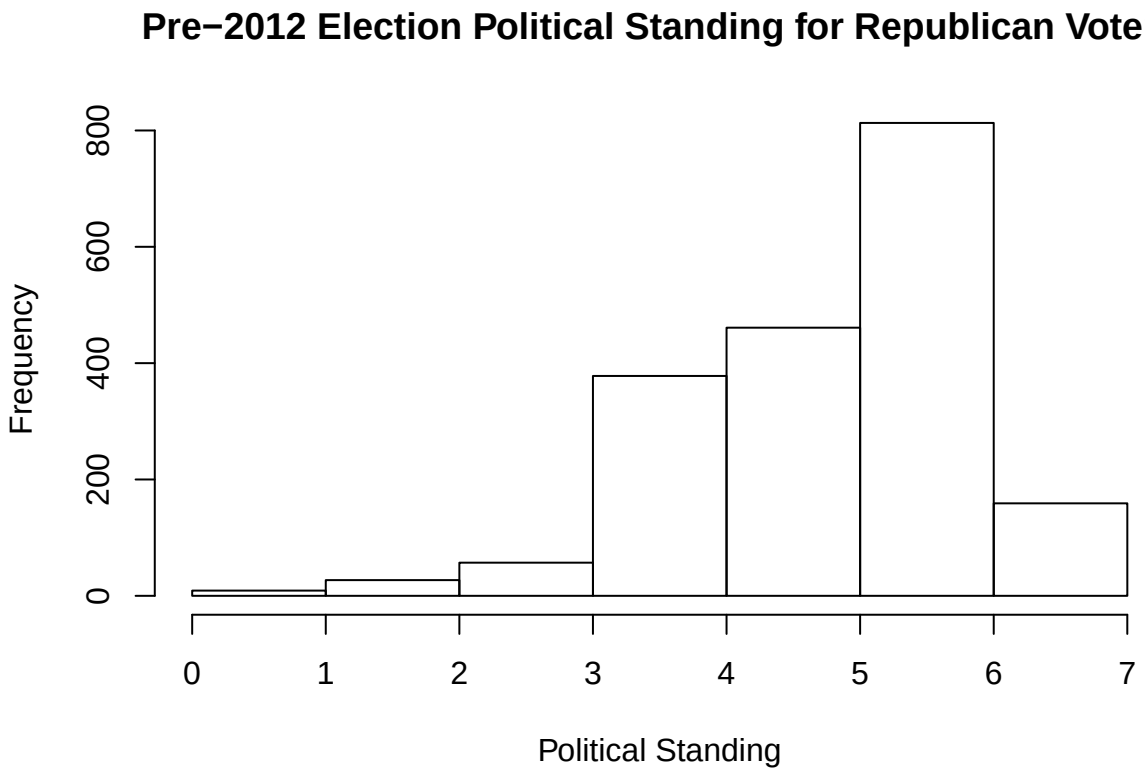
4. Were Republican voters more likely to shift their political preferences right or left (more conservative or more liberal), compared to Democratic voters during the 2012 election?

Let's compute how the average political standing changed for Republican voters.

```
# Pre-2012 election Republican voters
libcpre_rep_weights <-compilePoliticalWeights(
  S$libcpre_self[grep("republican", S$pid_x, ignore.case = T)], 3, TRUE)
summary(libcpre_rep_weights)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   5.000   6.000   5.274   6.000   7.000
```

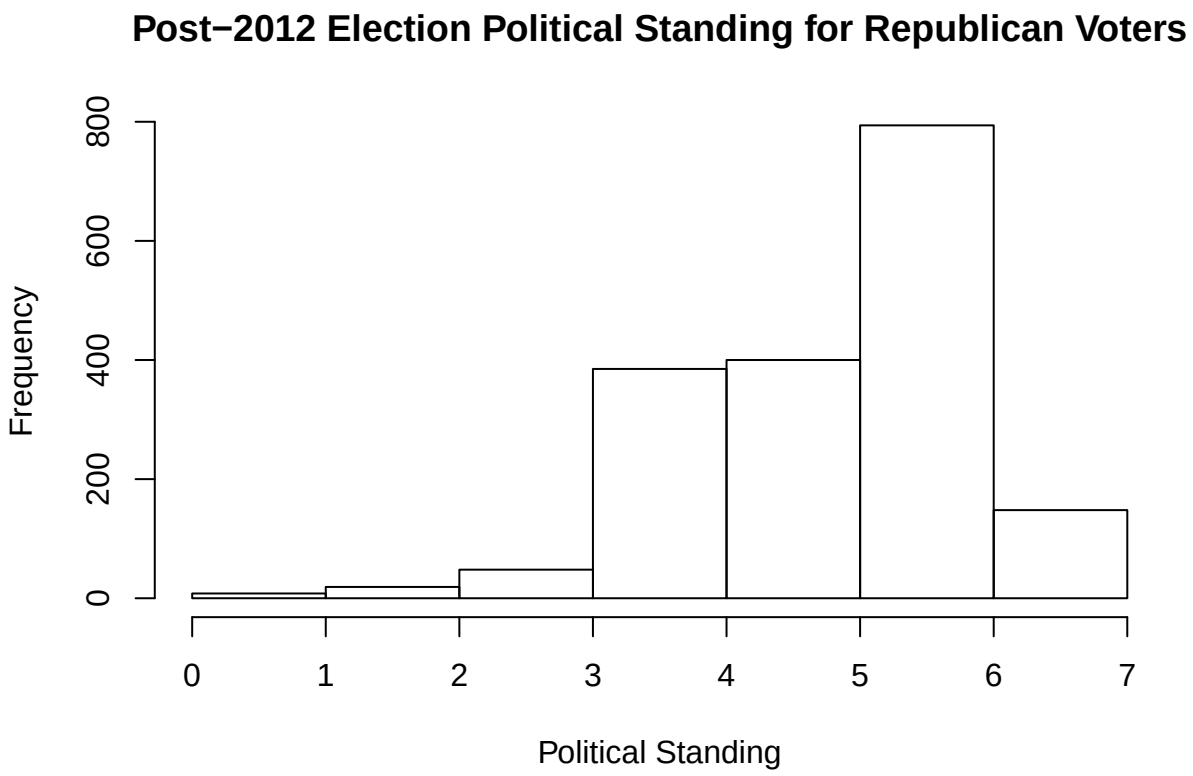
```
mu_libcpre_rep = mean(libcpre_rep_weights)
hist(libcpre_rep_weights, main = "Pre-2012 Election Political Standing for Republican Voters",
     xlab = "Political Standing", breaks = bins_standing)
```




```
# Post-2012 election Republican voters
libcpo_rep_weights <-compilePoliticalWeights(
  S$libcpo_self[grep("republican", S$pid_x, ignore.case = T)], 5, TRUE)
summary(libcpo_rep_weights)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   4.000   6.000   5.289   6.000   7.000
```

```
mu_libcpo_rep = mean(libcpo_rep_weights)
hist(libcpo_rep_weights, main = "Post-2012 Election Political Standing for Republican Voters",
     xlab = "Political Standing", breaks = bins_standing)
```




```
(diff_mu_lib_rep = mu_libcpo_rep - mu_libcpre_rep)
```

```
## [1] 0.01440859
```

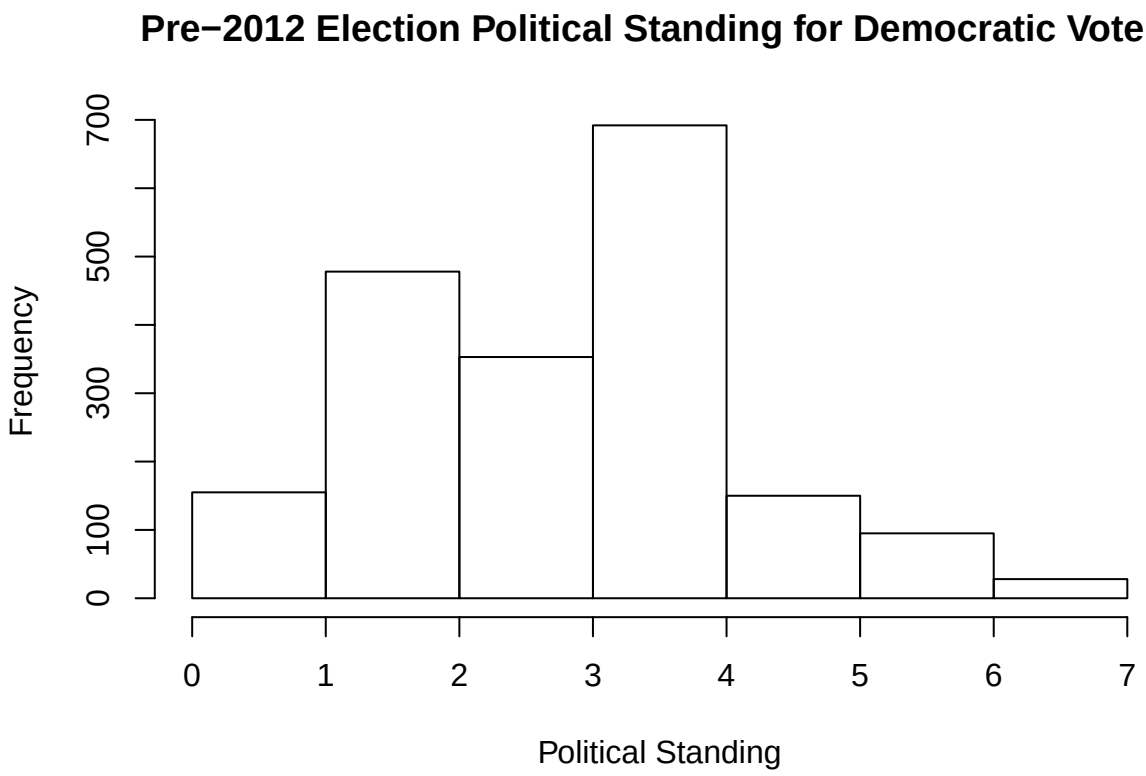
The average political standing value post-2012 electioin is greater than pre-election by 0.0144086. This means that the Republican voters on average has become more conservative.

Now let's do the same for Democratic voters.

```
# Pre-2012 election Democratic voters
libcpre_dem_weights <-compilePoliticalWeights(
  S$libcpre_self[grep("democrat", S$pid_x, ignore.case = T)], 3, TRUE)
summary(libcpre_dem_weights)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   3.308   4.000   7.000
```

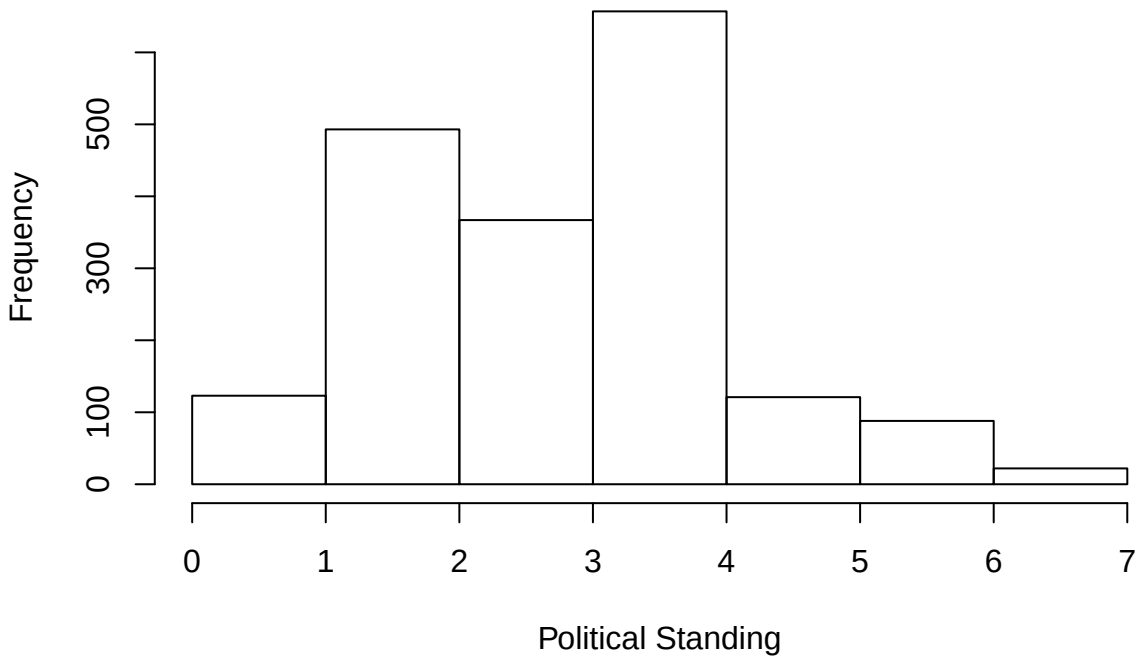
```
mu_libcpre_dem <- mean(libcpre_dem_weights)
hist(libcpre_dem_weights, main = "Pre-2012 Election Political Standing for Democratic Voters",
     xlab = "Political Standing", breaks = bins_standing)
```




```
# Post-2012 election Democratic voters
libcpo_dem_weights <-compilePoliticalWeights(
  S$libcpo_self[grep("democrat", S$pid_x, ignore.case = T)], 5, TRUE)
summary(libcpo_dem_weights)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   2.000   3.000   3.274   4.000   7.000
```

```
mu_libcpo_dem <- mean(libcpo_dem_weights)
hist(libcpo_dem_weights, main = "Post-2012 Election Political Standing for Democratic Voters",
     xlab = "Political Standing", breaks = bins_standing)
```




```
(diff_mu_lib_dem = mu_libcpo_dem - mu_libcpre_dem)
```

```
## [1] -0.0343967
```

The average political standing value post-2012 electioin is smaller than pre-election by 0.0343967. This means that the Democratic voters on average has become more liberal.

S$libcpo_self ## 5. Select a fifth question that you are interested in investigating.

Prepare a report addressing these questions. A successful submission should include:

1. A brief introduction.
2. A suitable hypothesis test for each question above.
3. For each test, include:
    - A brief exploratory analysis targetted to check the assumptions needed for your test.
    - A justification for why the test is the most appropriate choice
    - An explanation of test results, including *BOTH* statistical significance and practical significance.
4. A brief conclusion with a few high-level takeaways.

Please limit your submission to 10 pages. Be sure to submit both your pdf report as well as your source file.