Lab 3: Hypothesis Tests about the Mean.

w203: Statistics for Data Science

Ted Pham

March 22 2017

Introduction

The American National Election Studies (ANES) conducts surveys of voters in the United States before and after every presidential election. A sample of this survey is given in the file ANES_2012_sel.csv.

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, the survey weights have been removed and it is assumed that the observations in the file ANES_2012_sel.csv comprises a random sample from the voting population.

Let's examine the variables in the csv file.

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

```
##  [1] "X"                     "profile_educ"
##  [3] "profile_gender"        "profile_homeown"
##  [5] "profile_hhincome"      "profile_marital"
##  [7] "dem_age_r_x"           "profile_region9"
##  [9] "pid_x"                 "presapp_track"
## [11] "presapp_job"           "presapp_jobstr"
## [13] "presapp_job_x"         "presapp_econ"
## [15] "presapp_econstr"       "presapp_econ_x"
## [17] "presapp_foreign"       "presapp_foreignstr"
## [19] "presapp_foreign_x"     "presapp_health"
## [21] "presapp_healthstr"     "presapp_health_x"
## [23] "presapp_war"           "presapp_warstr"
## [25] "presapp_war_x"         "libcpre_self"
## [27] "libcpre_choose"        "libcpre_dpc"
## [29] "libcpre_rpc"           "libcpre_ptyd"
## [31] "libcpre_ptyr"          "libcpo_self"
## [33] "libcpo_selfch"         "libcpo_hdc"
## [35] "libcpo_hrc"            "paprofile_libcon_self"
## [37] "paprofile_libcon_pres"
```

There are 37 variables in this sample survey. 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.

From the dataset, we can ask a few interesting questions regarding the voters population during the 2012 election:

1. Did voters become more liberal or more conservative during the 2012 election?
2. Were the Republican voters older or younger, on average, than the Democratic voter?
3. Was the average Republican voter older than 51 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?

Statistical Analysis

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

We use the libcpre_self and libcpo_self variables to answer this question. These record the voters self-identification with the 7 pt scale liberal-conservative placement. There are also non-responders captured in negative values. Some did not specify their political leaning in the pre-election surveys while orthers in the post-election. In the pre-election survey, there were 556 voters in "haven't thought much about this", 26 "Don't Know", and 32 "refused"

```
#first let filter out survey results for voters that are under 18 years old or did not report their age
# to minimize assumption, we subset our these entries which total to only 62, a small size compared to

S = subset(S,S$dem_age_r_x>17)
```

```
# Define a new dataframe SLC that contains only libcpre_self and libcpo_self
SLC <- S

# Look at libcpre_self and libcpo_pro more closely
summary(SLC$libcpre_self)
```

```
## -2. Haven't thought much about this                    -8. Don't know
##                                 541                                25
##                          -9. Refused                1. Extremely liberal
##                                  30                               195
##                           2. Liberal                3. Slightly liberal
##                                 633                               636
##     4. Moderate; middle of the road           5. Slightly conservative
##                                1811                               784
##                      6. Conservative           7. Extremely conservative
##                                 992                               205
```

The same can be observed in the post-election survey but there are "Not Asked–" and "Deleted –".

```
summary(SLC$libcpo_self)
```

```
##                      -2. Haven't thought much {do not probe}
##                                                          396
## -6. Not asked, unit nonresponse (no post-election interview)
##                                                          249
##         -7. Deleted due to partial (post-election) interview
##                                                          151
##                                                -8. Don't know
##                                                           23
##                                                   -9. Refused
##                                                           35
##                                         1. Extremely liberal
##                                                          166
##                                                  2. Liberal
##                                                          641
##                                           3. Slightly liberal
##                                                          634
##                                 4. Moderate; middle of the road
##                                                         1740
##                                        5. Slightly conservative
##                                                          659
##                                              6. Conservative
##                                                          973
##                                       7. Extremely conservative
##                                                          185
```

The variable libcpre_choose and libcpo_selfch made the voters to choose between liberal, moderate, and conservative. It was not clear how these choices would reflect on the 7-pt liberal-conservative scale. For example, a conservative on the libcpre_choose/libcpo_selfch can be either slightly, moderate, or extremely conservative. For this reason, I opted not to include the negative points in both libcpre_self/libcpo_self with an additional assumption that these points were randomly distributed so that no special cluster would be left out. Our analysis here

```
# non-response prompts
pre_non_response = c("-2. Haven't thought much about this","-8. Don't know","-9. Refused")
cpo_non_response = c("-6. Not asked, unit nonresponse (no post-election interview)",
                     "-7. Deleted due to partial (post-election) interview",
                     "-2. Haven't thought much {do not probe}","-9. Refused",
                     "-8. Don't know")

# filter out non response
SLC <- SLC[!(SLC$libcpre_self %in% pre_non_response | SLC$libcpo_self %in% cpo_non_response),]

# original # of voters
dim(S)[1]
```

```
## [1] 5852
```

```
# filtered # of voters
dim(SLC)[1]
```

```
## [1] 4722
```

```
# percent of points ignored
(dim(S)[1]-dim(SLC)[1])/dim(S)[1]*100
```

```
## [1] 19.30964
```

For the purpose of this analysis, only voters who responded with a positive value on both before and after the

election on the liberal-conservative scale are included. Although 19.14% of voters did not fit this criterion, the ones that did accounted for 4697 voters, still a large sample size.

We can run a Wilcoxon signed rank test because they data are ranking-based and the samples are dependent invovling before and after assessment of the same sample population.

```
# redefine the factors
SLC$libcpo_self <- factor(SLC$libcpo_self)
SLC$libcpre_self <- factor(SLC$libcpre_self)

#introduce new variable before and after as the voters self assessment of politcal inclination
SLC["before"] <- as.numeric(SLC$libcpre_self)
SLC["after"] <- as.numeric(SLC$libcpo_self)

# Wilcox signed-rank test
wilcox.test(SLC$before,SLC$after,paired=TRUE)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  SLC$before and SLC$after
## V = 718890, p-value = 0.1725
## alternative hypothesis: true location shift is not equal to 0
```

The practical significance can be calculated as `{r} abs(qnorm(0.172/2))/sqrt(dim(SLC)[1])`. We obtain both stastical and practical insignificance to our hypothesis to whether voters would change their liberal-conservative inclication before or after the 2012 election. There was no evidence to suggest they did and we failed to reject the null hypothesis.

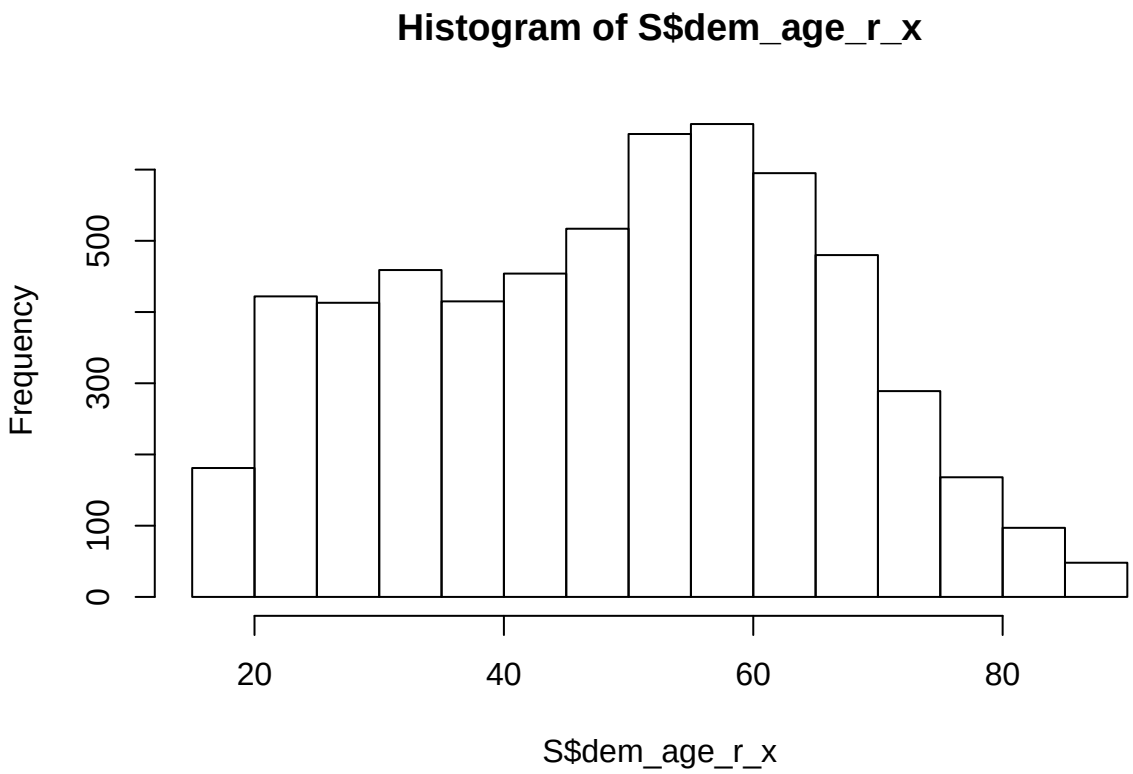
Were Republican voters older or younger , on the average, than Democratic voters in 2012?

To answer this question we focus on 2 variables pid_x and variable dem_age_r_x

```
summary(S$dem_age_r_x)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   18.00   35.00   51.00   49.45   62.00   90.00
```

```
hist(S$dem_age_r_x)
```



Since we already substracted out the voters with age <18 in the previous step, the histogram looks fairly normal for the age variable.

Let's look at pid_x next

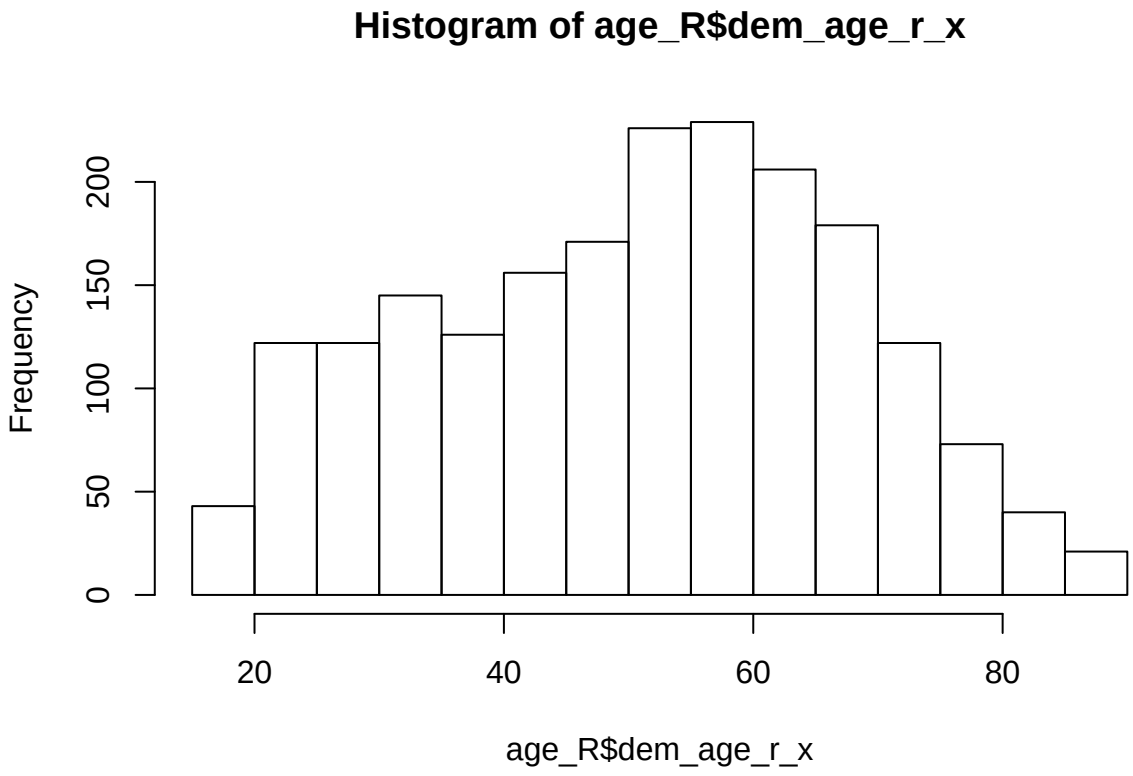
```
summary(S$pid_x)
```

```
##                   -2. Missing            1. Strong Democrat
##                            22                          1472
##  2. Not very strong Democract       3. Independent-Democrat
##                           859                           735
##                4. Independent     5. Independent-Republican
##                           783                           604
## 6. Not very strong Republican          7. Strong Republican
##                           621                           756
```

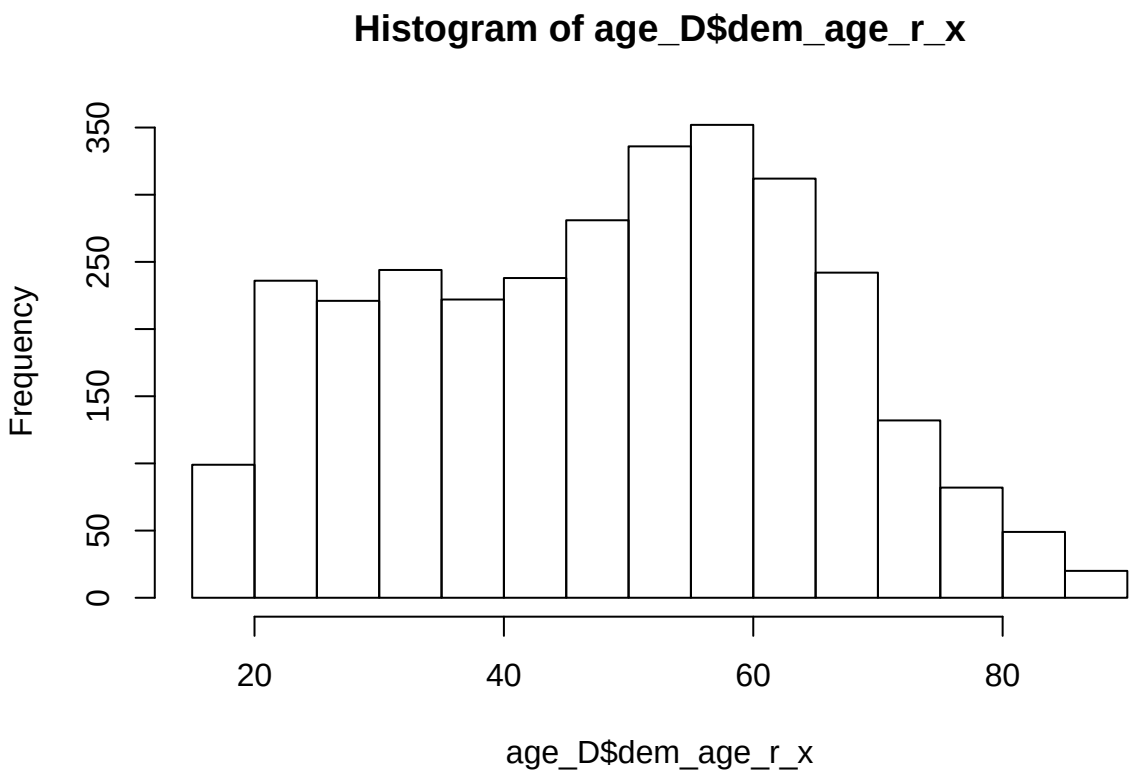
There are 8 categories, 6 of which indicate whether a voter is registered as a Democrat or a Republican.

```
age_R = subset(S, pid_x %in% c("5. Independent-Republican",
                               "6. Not very strong Republican",
                               "7. Strong Republican"))
age_D = subset(S, pid_x %in% c("3. Independent-Democrat",
                               "2. Not very strong Democract",
                               "1. Strong Democrat"))

hist(age_R$dem_age_r_x)
```




```
hist(age_D$dem_age_r_x)
```



The age distribution in Republican and Democratic groups are seemingly normal without any extreme skewness. We need to check if their variances are roughly equal.

```r
print(var(age_R$dem_age_r_x))
```

```
## [1] 281.8245
```

```r
print(var(age_D$dem_age_r_x))
```

```
## [1] 277.8999
```

Since the two groups' variances are similar, we can then perform an independent samples t-test with the null hypothesis as age_R - age_D <= 0 and the alternative hypothesis as age_R - age_D > 0.

```r
t.test(age_R$dem_age_r_x,age_D$dem_age_r_x, alternative = 'greater')
```

```
##
##  Welch Two Sample t-test
##
## data:  age_R$dem_age_r_x and age_D$dem_age_r_x
## t = 5.1233, df = 4204.3, p-value = 1.568e-07
## alternative hypothesis: true difference in means is greater than 0
## 95 percent confidence interval:
##  1.678538      Inf
## sample estimates:
## mean of x mean of y
##  51.33064  48.85812
```

Since p = 1.34 e-7 is small, we can reject the null hypothesis and the t-test is statistically significant. On average, Republican voters were older than Democrats. But this test is not very practically significant with r small effect size `{r}t/sqrt(t^2+1969+3046-2)` and small Cohen's d effect size calculated in the following code:

```r
n1 = 1969
n2 = 3046
s1s = var(age_R$dem_age_r_x)
s2s = var(age_D$dem_age_r_x)
s2p = sqrt(((n1-1)*s1s + (n2-1)*s2s)/(n1+n2-2))
(mean(age_R$dem_age_r_x) - mean(age_D$dem_age_r_x))/s2p
```

```
## [1] 0.1479091
```

3. Were Republican voters older than 51, on the average in 2012? The sample size is significantly large, and the distribution is fairly normal as shown in the histogram for age_R. This is a one-sample t-test with null hypothesis: mu <= 51 alternative hypothesis: mu > 51

```r
t.test(age_R$dem_age_r_x,mu=51, paired = FALSE, alternative = 'greater')
```

```
##
##  One Sample t-test
##
## data:  age_R$dem_age_r_x
## t = 0.87662, df = 1980, p-value = 0.1904
## alternative hypothesis: true mean is greater than 51
## 95 percent confidence interval:
##  50.70995      Inf
## sample estimates:
## mean of x
##  51.33064
```

with p-value of 0.07733, the t-test is statistically insignificant so we cannot reject the null hypothesis. The republican voters, on average, were not older than 51 years old.

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 reuse our SLC dataset here. SLC is a modified version of S with political leaning before and after the election processed.

We create a new binary variable pshift to indicate whether a voter shifted their political leaning during the election. If the difference between the after and before scores on the 7-pt scale is not zero, then the voter shifted their political preference.

```
# sub set republican and democratic voters
SLC_R = subset(SLC, pid_x %in% c("5. Independent-Republican",
                                 "6. Not very strong Republican",
                                 "7. Strong Republican"))
SLC_D = subset(SLC, pid_x %in% c("3. Independent-Democrat",
                                 "2. Not very strong Democract",
                                 "1. Strong Democrat"))

#defind variable political shift
SLC_R['pshift'] = (SLC_R$after - SLC_R$before) != 0
SLC_D['pshift'] = (SLC_D$after - SLC_D$before) != 0
```

The pshift variable is binary with TRUE meaning that there was a shift in preference. We use a test of equal proportions (Chi-squared test) with:

null hypothesis: Republican voters are more than or equally like to shift political preference compared to Democratic voters alternative hypothesis: republican voters are less likely to shift politcal preference compared to Democratic voters

We use a prop.test to calculate the value of chi-square because we cannot make any assumption on whether the probability of a voter would shift their political preferences during the election is 0.5. We also assume that because the Democrats are generally more liberal hence they are more likely to shift their preferences.

```
R_True_shift = dim(subset(SLC_R,pshift == TRUE))[1]
R_False_shift = dim(SLC_R)[1] - R_True_shift
D_True_shift = dim(subset(SLC_D,pshift = TRUE))[1]
D_False_shift = dim(SLC_D)[1] - D_True_shift

test_matrix = matrix(c(R_True_shift,D_True_shift,R_False_shift,D_False_shift),2,2)
prop.test(test_matrix, alternative = 'less')
```

```
##
##  2-sample test for equality of proportions with continuity
##  correction
##
## data:  test_matrix
## X-squared = 2309.1, df = 1, p-value < 2.2e-16
## alternative hypothesis: less
## 95 percent confidence interval:
##  -1.0000000 -0.6689969
## sample estimates:
##   prop 1   prop 2
## 0.312212 1.000000
```

With p-value small, our test is statistically significant and on average, the Democratic voters are more likely to shift their political preferences during the election.

5. Was there a difference in opinion between Democratic and Republican voters on how Barrack Obama

had handled foreign relations? We use pid_x and presapp_foreign_x variables here.

```
summary(S$presapp_foreign_x)
```

```
##             -8. Don't know                 -9. Refused
##                        208                          40
##        1. Approve strongly    2. Approve not strongly
##                       2200                        1027
## 4. Disapprove not strongly     5. Disapprove strongly
##                        593                        1784
```

```
na = c("-8. Don't know","-9. Refused")
SPF = S[!S$presapp_foreign_x %in% na,]
SPF$presapp_foreign_x <- factor(SPF$presapp_foreign_x )

SPF["foreign"] <- as.numeric(SPF$presapp_foreign_x)

# republican
SPF_R = subset(SPF, pid_x %in% c("5. Independent-Republican",
                                 "6. Not very strong Republican",
                                 "7. Strong Republican"))
SPF_D = subset(SPF, pid_x %in% c("3. Independent-Democrat",
                                 "2. Not very strong Democract",
                                 "1. Strong Democrat"))
```

We use Wilcoxon two independent sample test here because the data are ranking (ordinal) and no assumption about normality. The null hypothesis is Republican voters have the same regard to Barrack Obama's handling of foreign affairs The alternative hypothesis is Republican have a different regard to Barrack Obama's handling of foreign affairs.

```
wilcox.test(SPF_R$foreign,SPF_D$foreign, alternative = 'greater')
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  SPF_R$foreign and SPF_D$foreign
## W = 5059600, p-value < 2.2e-16
## alternative hypothesis: true location shift is greater than 0
```

So on average, Republican voters differ in opinion about Barrack Obama's handling of foreign affairs. However this was not practically significant.

```
qnorm(2.2e-16)/sqrt(dim(SPF)[1])
```

```
## [1] -0.108563
```

Did they disapprove of Barrack Obama on this? We can run a simple t.test

```
print(t.test(SPF_R$foreign,mu=3,alternative='greater'))
```

```
##
##  One Sample t-test
##
## data:  SPF_R$foreign
## t = 21.369, df = 1921, p-value < 2.2e-16
## alternative hypothesis: true mean is greater than 3
## 95 percent confidence interval:
##  3.412512      Inf
## sample estimates:
```

```
## mean of x
##   3.44693
```

```
# practical significance
21.369/sqrt(21.369^2 + 1921)
```

```
## [1] 0.4382397
```

So on average, Republican voters disapprove of Barrack Obama's handling of foreing affairs and this has moderate practical significant.

We can try to do another analysis for the approval rating on Barrack Obama's handling of the economy.

```
na = c("-8. Don't know","-9. Refused")
SPE = S[!S$presapp_econ_x %in% na,]
SPE$presapp_econ_x <- factor(SPE$presapp_econ_x )

SPE["econ"] <- as.numeric(SPE$presapp_econ_x)

# republican
SPE_R = subset(SPE, pid_x %in% c("5. Independent-Republican",
                                 "6. Not very strong Republican",
                                 "7. Strong Republican"))
SPE_D = subset(SPE, pid_x %in% c("3. Independent-Democrat",
                                 "2. Not very strong Democract",
                                 "1. Strong Democrat"))
wilcox.test(SPE_R$econ,SPE_D$econ, alternative = 'greater')
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  SPE_R$econ and SPE_D$econ
## W = 5338800, p-value < 2.2e-16
## alternative hypothesis: true location shift is greater than 0
```

```
qnorm(2.2e-16)/sqrt(dim(SPF)[1])
```

```
## [1] -0.108563
```

The results are similar to the assessment of the president's handling of foreign affairs.

A few high-level takeaways.

The analysis of a excerpt of ANES data reveals several insights:

1. Voters did not change political preference during the 2012 election.
2. On average, Republican voters were older than Democratic counterparts. But the average age for Republican voters were not more than 51 years old.
3. The Democratic voters were more likely to shift political preference than the Republican voters.
4. The Republican voters tend to have a negative views on incumbent president's handling of foreign affairs and economics. An effort to check if there's something that Barrack Obama did that Republicans would agree to seems futile given the dataset.

Although these insights are somewhat predictable and none are practically significant, they expose certain areas that party politicians should be worry about (in 2012 and perhaps still relevant today):

1. The Republican need to think about how to recruit younger demographics.
2. The Democratic need to account for their base's unpredictability and devise an agenda that would appeal to conservative folks.