Lab 3: Hypothesis Tests about the Mean.

w203: Statistics for Data Science

Chris Fleisch

11/15/2016

Introduction

Using the ANES survey data we will try to answer several questions about the voters from 2012. We'll be looking for changes in voters from before and after the election as well as differences between groups of voters. After a quick exploration of the data we will use the appropriate statistical test for each question depending on what data is used and what we are trying to answer. From there we will draw conclusions using the statistical results and the practical results for each question.

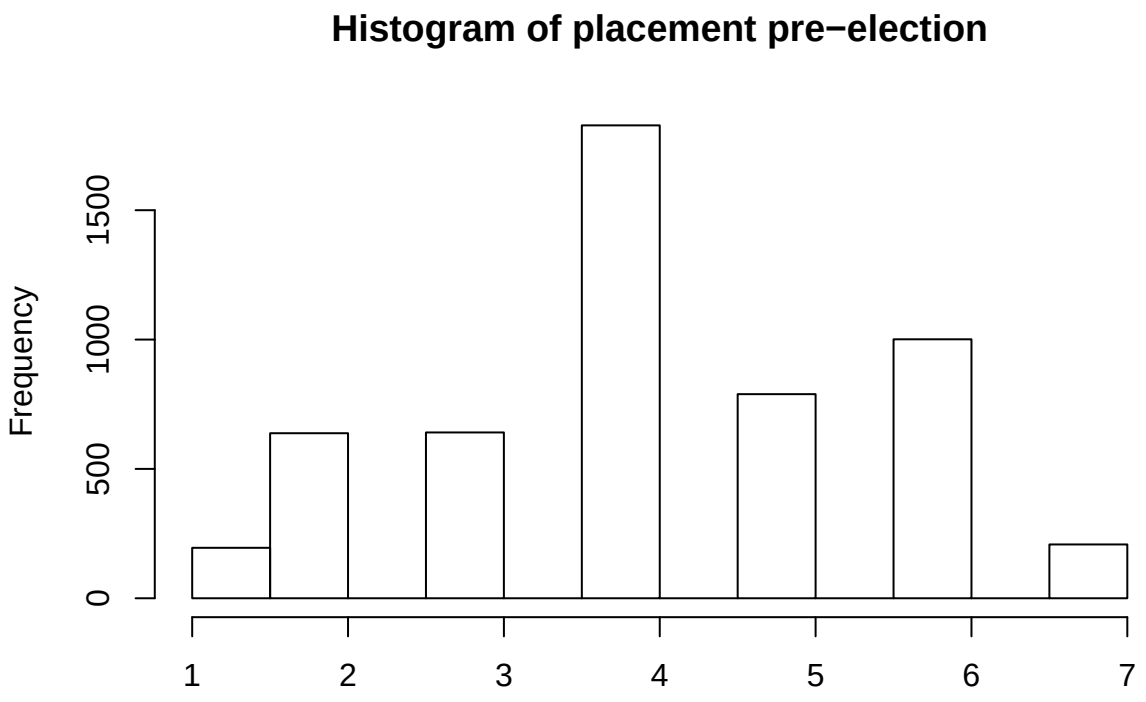
```
S = read.csv("ANES_2012_sel.csv")
```

Analysis

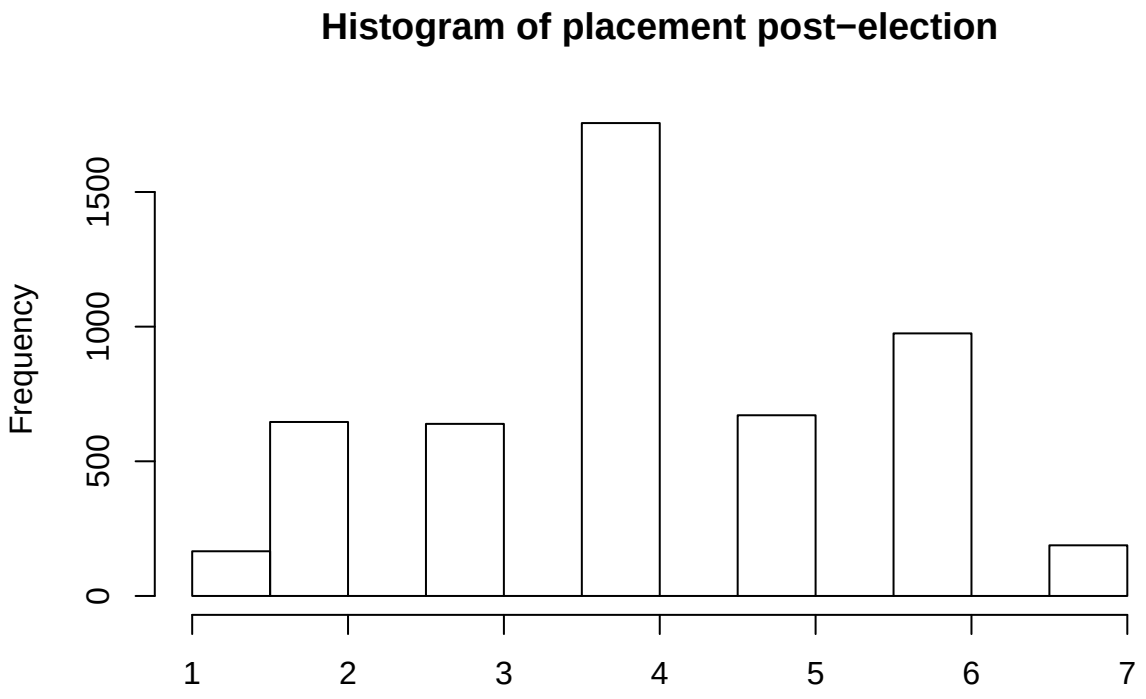
1. Did voters become more liberal or more conservative during the 2012 election?

We want to know if there was a change in liberal-conservative placement after the election. The null hypothesis is that there is no change and our alternative hypothesis is that there was a change in liberal-conservative placement. We will use a dependent test because this is a paired sample of before and after relationships. We will use a signed rank based test, because we have a Likert (non-parametric) variable and are testing related conditions of before and after the election with the same people.

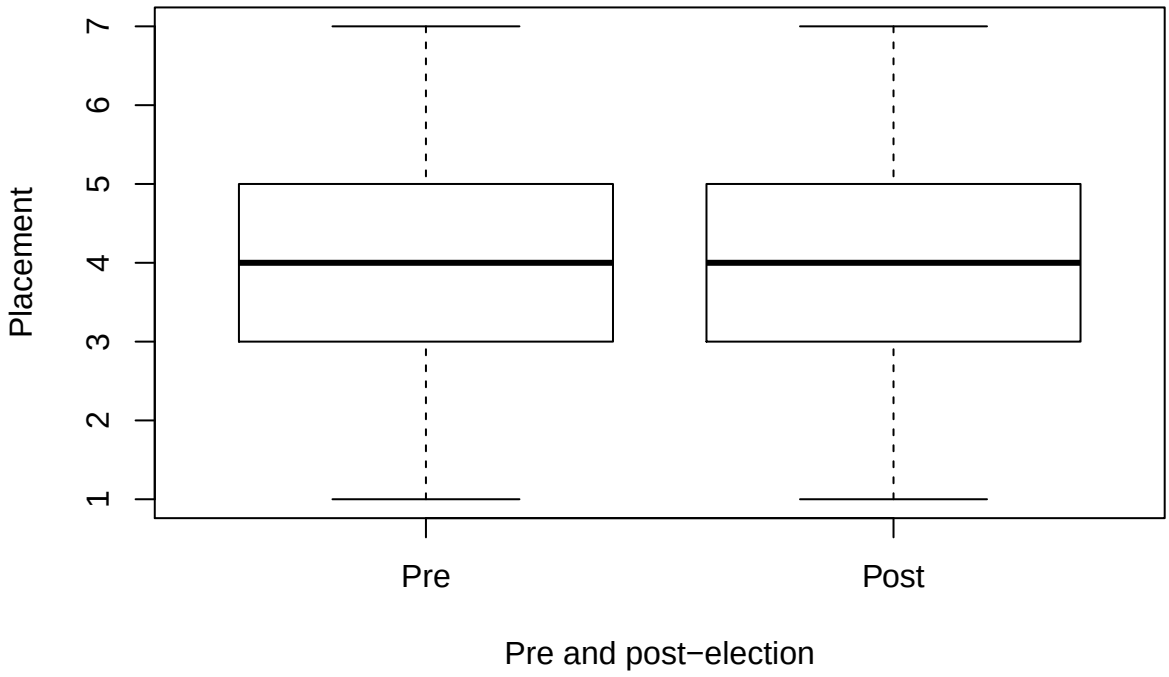
```
# make new variables and recode negative responses as NA
library(data.table)
S$my_libcpre_self <- S$libcpre_self
S$my_libcpo_self <- S$libcpo_self

S$my_libcpre_self[S$my_libcpre_self %like% "-"] <- NA
S$my_libcpo_self[S$my_libcpo_self %like% "-"] <- NA

# refactor
S$my_libcpre_self <- factor(S$my_libcpre_self)
S$my_libcpo_self <- factor(S$my_libcpo_self)

hist(as.numeric(S$my_libcpre_self), main = "Histogram of placement pre-election",
     xlab = NULL, breaks = 20)
```




```
hist(as.numeric(S$my_libcpo_self), main = "Histogram of placement post-election",
     xlab = NULL, breaks = 20)
```




```
boxplot(as.numeric(S$my_libcpre_self ), as.numeric(S$my_libcpo_self),
        names = c('Pre', 'Post'),
        main = "Boxplot of pre-election vs post-election placement",
        xlab="Pre and post-election", ylab = "Placement")
```




```
# Wilcoxon paired test for nonparametric data
(wt <- wilcox.test(as.numeric(S$my_libcpre_self), as.numeric(S$my_libcpo_self),
            paired = TRUE))
```

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

```
# calculate z and r
(z <- qnorm(1 - (wt$p.value/2)))
```

```
## [1] 1.384435
```

```
(n <- length(S$my_libcpre_self[!is.na(S$my_libcpre_self) & !is.na(S$my_libcpo_self)]))
```

```
## [1] 4758
```

```
(r = z / sqrt(n))
```

```
## [1] 0.0200706
```

Looking at the data we had several "don't know or missing" values. Since the question is looking at changes in liberal-conservative placement these negative values are taken out so we can compare changes in any movement

in the liberal-conservative relationship. We want to see if someone moves from liberal to conservative or vice versa.

The Wilcoxon signed rank test gave us a large p-value of .166 that is not statistically significant. We cannot reject our null hypothesis that there is no difference between pre and post-election placement. This adds support to our null hypothesis that there was no change after the election.

The value of our r calculation is very small (.02) which shows that there was not a practical effect size and this was evident in the boxplot. We cannot say that voters became more liberal or more conservative.

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

Our null hypothesis is that there is no difference in age between republican and democrat voters. We will compare the mean ages of republicans and democrats and do a two-tailed test. The alternative hypothesis is that there is a difference in age between republican and democrat voters.

```
str(S$pid_x)
```

```
##  Factor w/ 8 levels "-2. Missing",..: 2 2 2 2 4 5 2 4 2 7 ...
```

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

```
##                   -2. Missing            1. Strong Democrat
##                            24                          1485
##  2. Not very strong Democract       3. Independent-Democrat
##                           871                           747
##                4. Independent     5. Independent-Republican
##                           792                           610
## 6. Not very strong Republican          7. Strong Republican
##                           623                           762
```

```
# classify republicans as pid_x >= 6
S$my_party[as.numeric(S$pid_x) >= 6] <- "republican"
# classify democrats as 2 <= pid_x <= 4
S$my_party[as.numeric(S$pid_x) >= 2 & as.numeric(S$pid_x) <= 4] <- "democrat"
# independents as pix_x == 5
S$my_party[as.numeric(S$pid_x) == 5] <- "independent"
# NA as pix_x == 1
S$my_party[as.numeric(S$pid_x) == 1] <- NA
# convert to factor
S$my_party <- as.factor(S$my_party)
summary(S$my_party)
```

```
##    democrat independent  republican        NA's
##        3103         792        1995          24
```

```
str(S$dem_age_r_x)
```

```
##  int [1:5914] 86 79 72 54 35 80 50 70 22 49 ...
```

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

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

```r
head(S$dem_age_r_x[order(S$dem_age_r_x)])
```

```
## [1] -2 -2 -2 -2 -2 -2
```

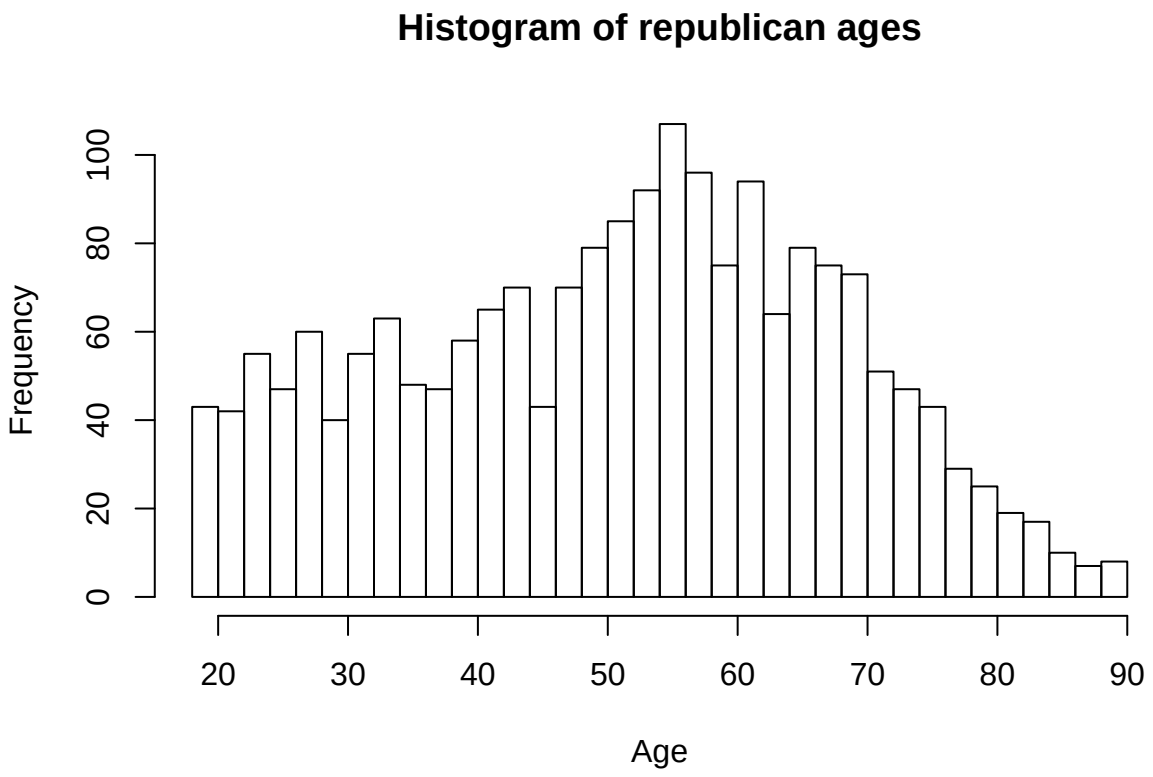
```r
# convert -2 ages to NA
S$my_age <- S$dem_age_r_x
S$my_age[S$my_age == -2] <- NA
summary(S$my_age)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
##   17.00   35.00   51.00   49.44   62.00   90.00      60
```

```r
# count of 17 year olds
length(S$my_age[S$my_age == 17 & !is.na(S$my_age)])
```

```
## [1] 2
```

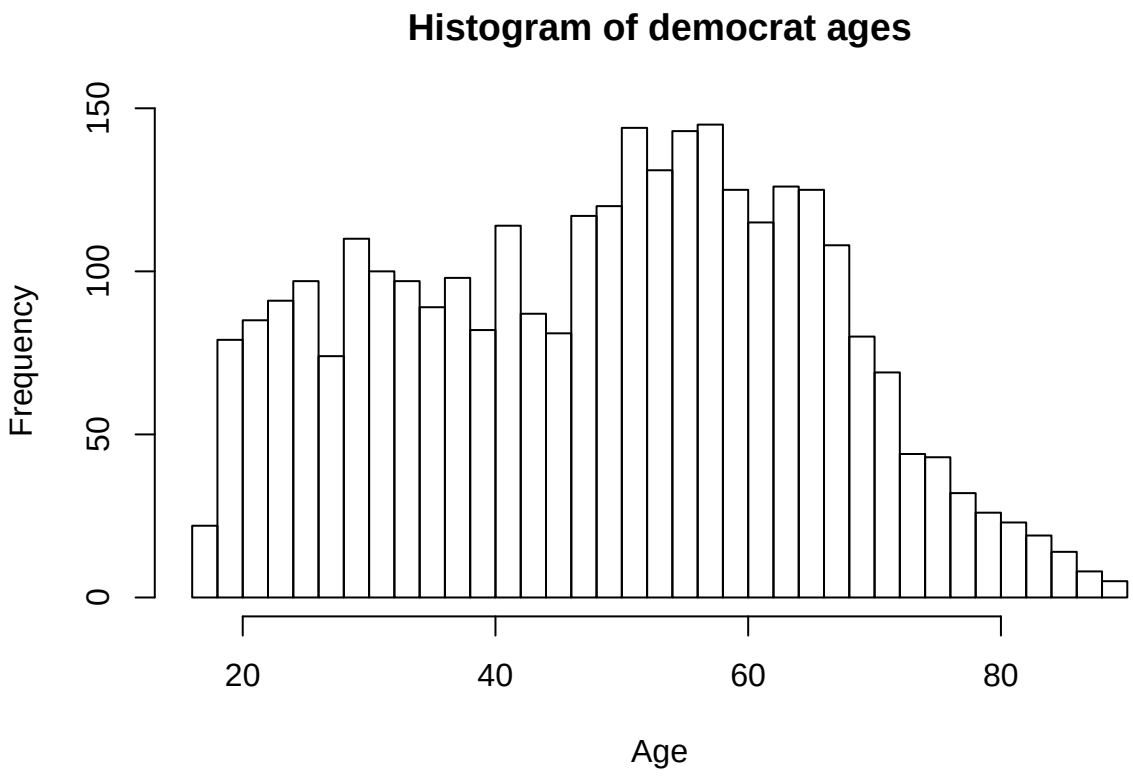
```r
# two are 17, we'll let them slide. Maybe they turn 18 by the time they vote.

# look at histogram of republican ages
repub_ages <- S$my_age[S$my_party == "republican"]
hist(repub_ages, breaks = 50, main = "Histogram of republican ages",
     xlab = "Age")
```



```r
length(repub_ages)
```

```
## [1] 2019
```

```r
# look at histogram of democrat ages
dem_ages <- S$my_age[S$my_party == "democrat"]
hist(dem_ages, breaks = 50, main = "Histogram of democrat ages",
     xlab = "Age")
```




```r
length(dem_ages)
```

```
## [1] 3127
```

The two histograms are not approaching normal, but they don't look too bad and we have a large n for each one so we can rely on the central limit theorem. We are going to compare two groups so we'll use the t.test.

```r
# test for equal variance
library(car)
# put the ages into a stacked list for leveneTest
ages = stack(list(repub_ages=repub_ages, dem_ages=dem_ages))
leveneTest(values ~ ind, ages, center = median)
```

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

```r
# perform t.test without welch's correction since variances are similar
t.test(repub_ages, dem_ages, var.equal = TRUE)
```

```
##
##  Two Sample t-test
##
## data:  repub_ages and dem_ages
## t = 5.1721, df = 5047, p-value = 2.404e-07
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.548236 3.438340
## sample estimates:
## mean of x mean of y
##  51.33064  48.83735
```

```r
library(effsize)
cohen.d(repub_ages, dem_ages, na.rm = TRUE)
```

```
##
## Cohen's d
##
## d estimate: 0.149074 (negligible)
## 95 percent confidence interval:
##        inf        sup
## 0.09248318 0.20566476
```

```r
age_diff <- mean(repub_ages, na.rm = TRUE) - mean(dem_ages, na.rm = TRUE)
age_diff
```

```
## [1] 2.493288
```

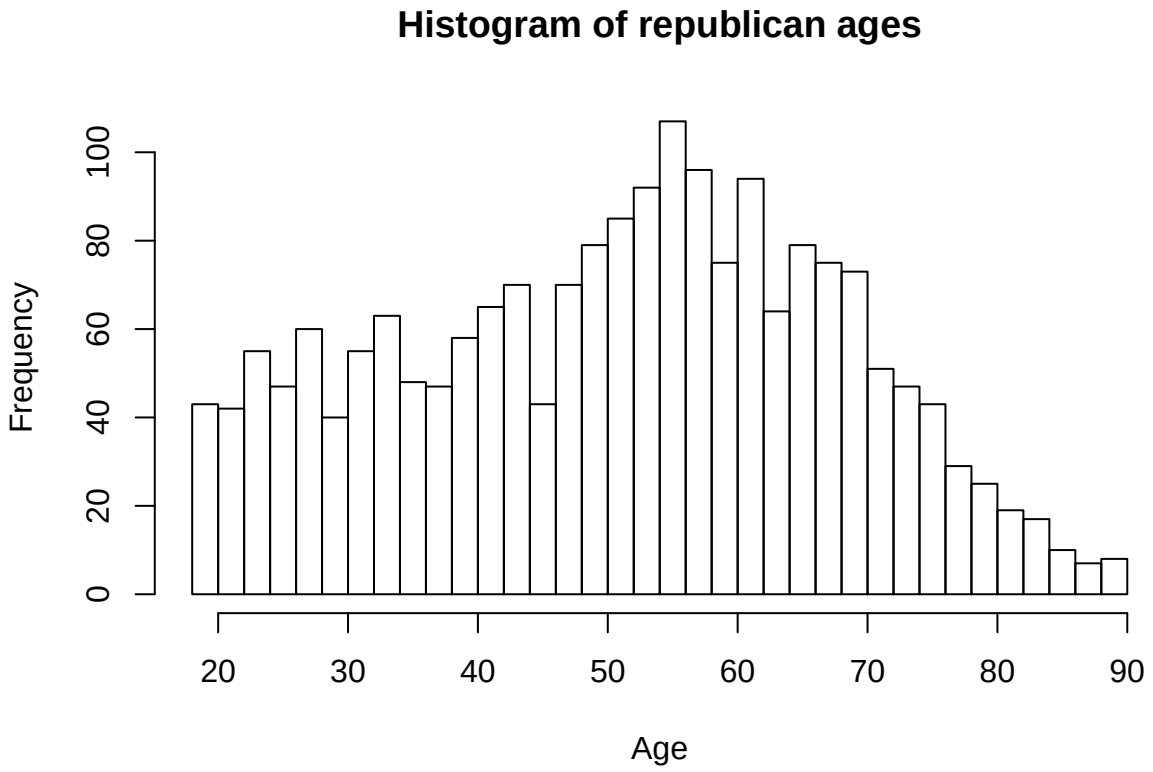
We have a very small p-value which means we can reject the null hypothesis that there is no difference between the mean ages of the two groups. This adds support to our hypothesis that there is a difference between the two groups. The t.test shows the mean of republican ages to be greater than that of democrat ages.

However, the cohen.d of .15 suggests that this difference is not a significant practical effect. And calculating the difference between the two means shows a difference of about 2.5 years with the republicans being older. A couple years' difference in age doesn't seem to matter that much in life unless you're trying to buy beer. For our two groups of voters the age difference is unlikely to make much of a practical difference.

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

Our null hypothesis is that there is no difference between the republican voters age and 51. We will do a one sample two-tailed t-test. Our alternative hypothesis is that there is a difference in the mean age of republicans and 51.

```r
# shows a distribution not approaching normal
hist(repub_ages, breaks = 50, main = "Histogram of republican ages",
     xlab = "Age")
```




```
length(repub_ages)
```

```
## [1] 2019
```

```
# we have a large enough n that we can rely on the CLT
t.test(repub_ages, mu = 51)
```

```
##
##  One Sample t-test
##
## data:  repub_ages
## 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
```

We have a p-value of .38 which shows that this is not statistically significant and we cannot reject the null hypothesis that there is no difference between the mean ages and 51. This adds support to our null hypothesis that the mean age is 51.

We can see that the mean age of 51.33 is pretty close to 51 and is practically the same for our purposes. We cannot say that republican voters were older than 51 on average.

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?

Our null hypothesis is that there is no difference in shift between republican voters and democratic voters for the pre and post-election placement. Our alternative hypothesis is that there was a difference in shift between the two groups in the pre and post-election placement.

Before we make our comparison we will test to see if there's a statistically significant shift in each group. Our null hypothesis is that there was no shift for each group and our alternative hypothesis is that there was a shift for each group. We will use a signed rank sum test (two-tailed) because we are looking at before and after pairings within each group.

We'll then need to calculate the shift for each group and then compare the means of those shifts. We will use the ranked sum test for comparing the differences of Likert data with a two-tailed test.

```
# test republican shift
(wt <- wilcox.test(as.numeric(S$my_libcpre_self[S$my_party == "republican"]),
            as.numeric(S$my_libcpo_self[S$my_party == "republican"]),
            paired = TRUE))
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  as.numeric(S$my_libcpre_self[S$my_party == "republican"]) and as.numeric(S$my_libcpo_self[S$m
## V = 72582, p-value = 0.3623
## alternative hypothesis: true location shift is not equal to 0
```

```
# calculate r effect size
(z = qnorm(1 - (wt$p.value/2)))
```

```
## [1] 0.9109407
```

```
(n <- length(S$my_libcpre_self[!is.na(S$my_libcpre_self) & !is.na(S$my_libcpo_self)
                               & S$my_party == "republican"]))
```

```
## [1] 1757
```

```
(r = z / sqrt(n))
```

```
## [1] 0.02173223
```

This non-parametric paired test shows that the republicans did not have a statistically significant shift in pre/post-election placement. And there is no practical effect with a very small r calculation.

```
# calculate the differences (shift) for republicans for later use
repub_placement_pre <- as.numeric(S$my_libcpre_self[S$my_party == "republican"])
repub_placement_post <- as.numeric(S$my_libcpo_self[S$my_party == "republican"])
repub_shift <- repub_placement_pre - repub_placement_post
```

```
# test for democrat shift
(wt <- wilcox.test(as.numeric(S$my_libcpre_self[S$my_part == "democrat"]),
            as.numeric(S$my_libcpo_self[S$my_party == "democrat"]),
            paired = TRUE))
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  as.numeric(S$my_libcpre_self[S$my_part == "democrat"]) and as.numeric(S$my_libcpo_self[S$my_pa
## V = 232530, p-value = 0.02921
## alternative hypothesis: true location shift is not equal to 0
```

```
# calculate r effect size
(z = qnorm(1 - (wt$p.value/2)))
```

```
## [1] 2.180668
```

```
(n <- length(S$my_libcpre_self[!is.na(S$my_libcpre_self) & !is.na(S$my_libcpo_self)
                               & S$my_party == "democrat"]))
```

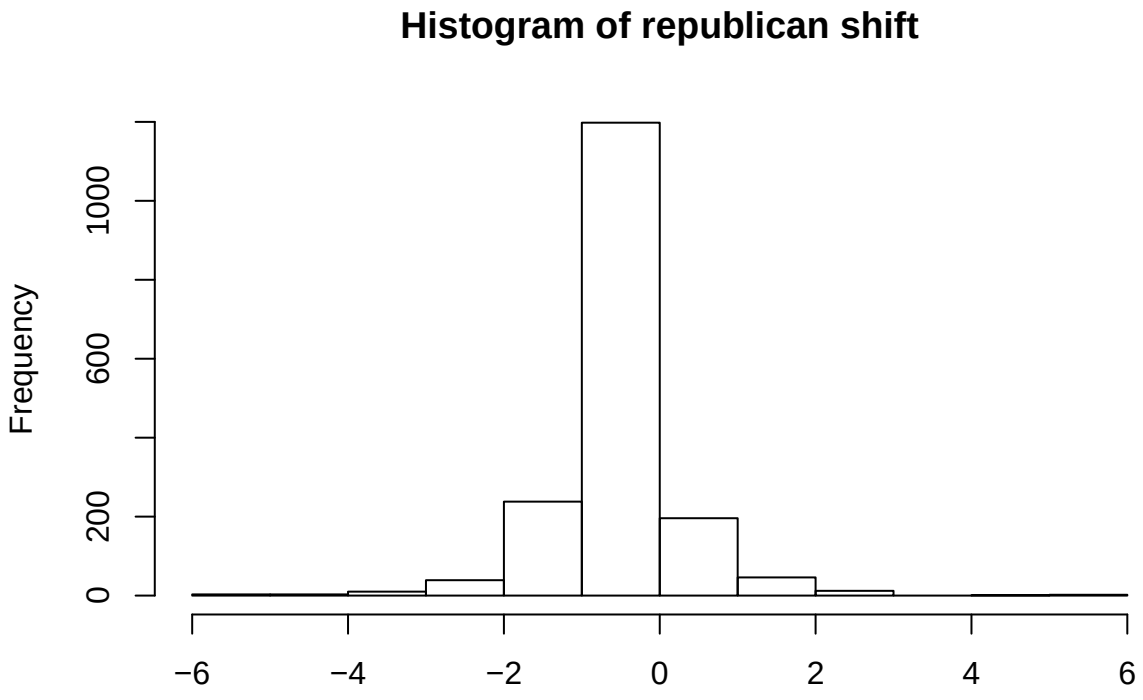
```
## [1] 2419
```

```
(r = z / sqrt(n))
```

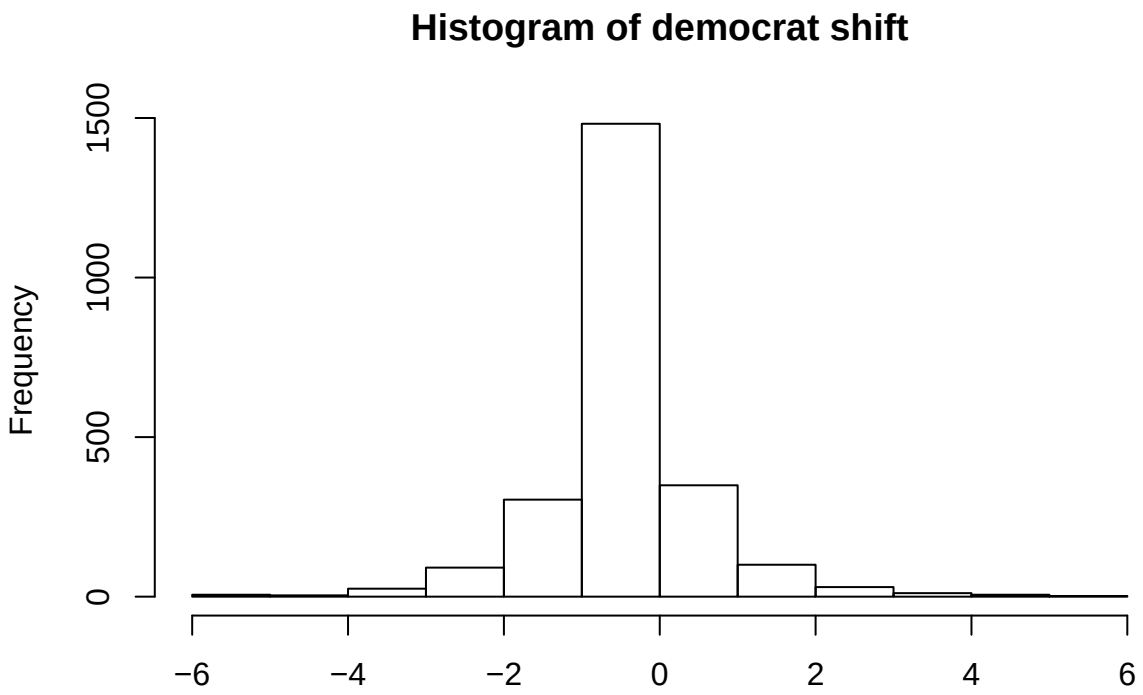
```
## [1] 0.04433753
```

This non-parametric paired test shows us that the democrats had a statistically significant shift in pre/post-election placement with a p-value of .029. There was no practical effect as noted by the small r calculation.

```
# calculate the differences (shift) for democrats
dem_placement_pre <- as.numeric(S$my_libcpre_self[S$my_part == "democrat"])
dem_placement_post <- as.numeric(S$my_libcpo_self[S$my_party == "democrat"])
dem_shift <- dem_placement_pre - dem_placement_post

# histograms of shift
hist(repub_shift, main = "Histogram of republican shift", xlab = NULL)
```



```r
hist(dem_shift, main = "Histogram of democrat shift", xlab = NULL)
```




```r
# very similar distributions

# test the two groups shifts
(wt <- wilcox.test(repub_shift, dem_shift))
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  repub_shift and dem_shift
## W = 2023400, p-value = 0.01096
## alternative hypothesis: true location shift is not equal to 0
```

```r
# calculate r effect size
(z = qnorm(1 - (wt$p.value/2)))
```

```
## [1] 2.544017
```

```r
(r = z / sqrt(length(repub_shift) + length(dem_shift)))
```

```
## [1] 0.03546378
```

```r
# repub mean shift toward the right
(repub_mean_shift <- mean(repub_shift, na.rm = TRUE))
```

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

```r
# dem mean shift toward the right
(dem_mean_shift <- mean(dem_shift, na.rm = TRUE))
```

```
## [1] 0.04771784
```

```r
# difference between republican and democrat shift
(diff <- repub_mean_shift - dem_mean_shift)
```

```
## [1] -0.06659656
```

We have a p-value of .01 that shows a statistically significant value in the amount of shift between the two parties. We can reject the null hypothesis which adds support that there was a difference in shift for the two parties.

The republicans shifted towards the right (republican: -0.019) although that wasn't statistically significant. The democrats shifted towards the left (democrat: .048) and that was statistically significant. The difference between their shifts was statistically significant, but it looks like this was mostly due to the democratic shift to the left and not because of a republican shift to the right.

The practical effect size between the two shifts is very small with an r of .03. The difference between the two means is -.07 which also doesn't show much practical difference on our Likert scale. So while the difference is statistically significant there's no practical difference between the shift in republican placement and democrat placement during the election. We cannot say that the republicans are more likely to shift in one direction. And the shifts for republicans and democrats has no practical effect either.

5. Select a fifth question that you are interested in investigating.

Is there a difference between republicans and democrats and their approval of the president's handling of the war?

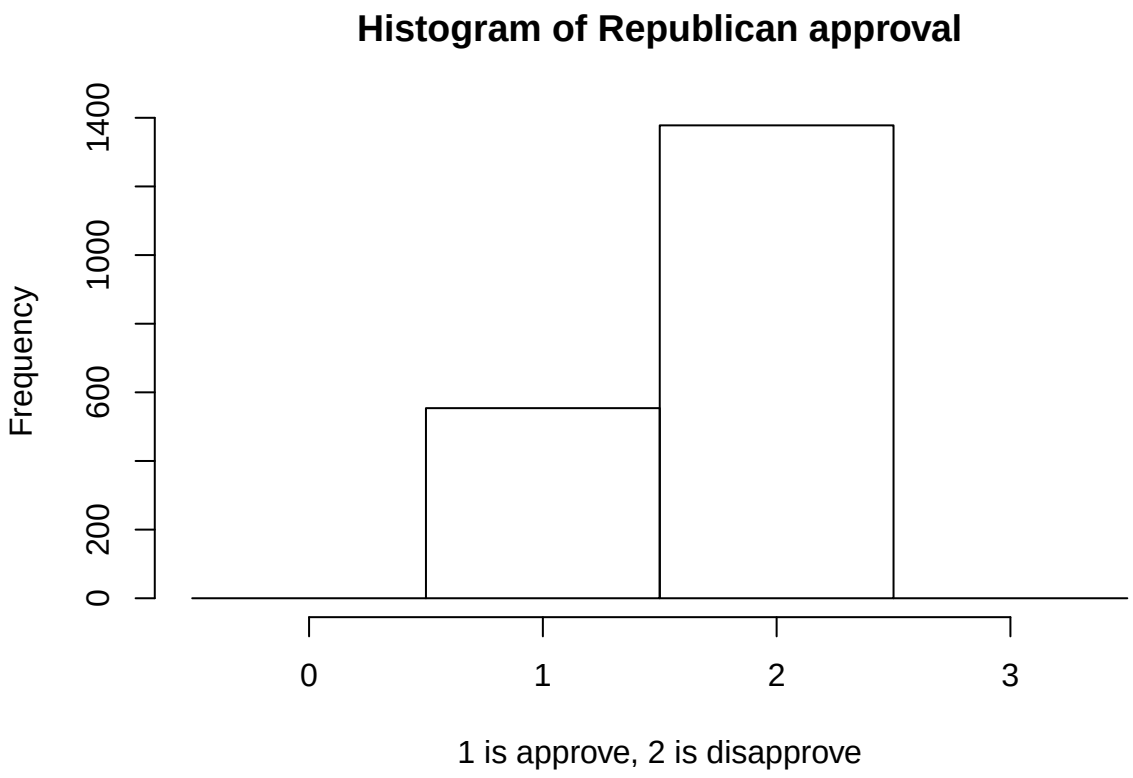
We will use the presapp_war variable to determine each groups approval.

Our null hypothesis is that there is no difference between republicans and democrats who approve of the President's handling of the war. Our alternative hypothesis is that there is a difference between the two groups and their war handling approval.

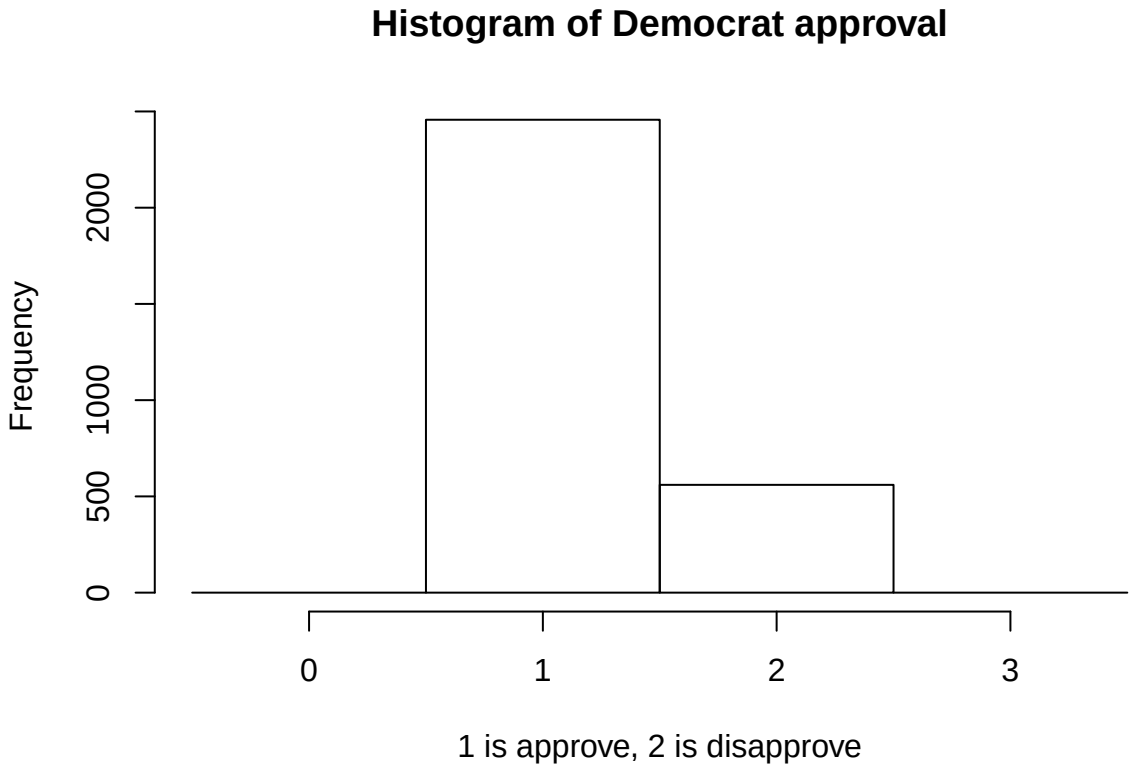
```r
# create a new variable to hold the approve or disapprove only
S$my_app_war[as.numeric(S$presapp_war) == 3] <- "approve"
S$my_app_war[as.numeric(S$presapp_war) == 4] <- "disapprove"

S$my_app_war <- as.factor(S$my_app_war)
str(S$my_app_war)
```

```
##  Factor w/ 2 levels "approve","disapprove": 1 1 1 2 1 1 1 1 1 1 ...
```

```r
repub_app <- as.numeric(S$my_app_war[S$my_party == "republican"])
dem_app <- as.numeric(S$my_app_war[S$my_party == "democrat"])

hist(repub_app, main = "Histogram of Republican approval",
     breaks = -1:3 + .5, xlab = "1 is approve, 2 is disapprove")
```




```
hist(dem_app, main = "Histogram of Democrat approval",
     breaks = -1:3 + .5, xlab = "1 is approve, 2 is disapprove")
```



These are not approaching normal, but n is large enough we can rely on the CLT for use in our two-tailed t-test.

```r
# put the party approvals into a stacked list for leveneTest
party_app = stack(list(repub_app=repub_app,
                       dem_app=dem_app))
leveneTest(values ~ ind, party_app, center = median)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value    Pr(>F)
## group    1  70.013 < 2.2e-16 ***
##       4947
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```r
# we need to correct for the differences in variances
t.test(repub_app, dem_app)
```

```
##
##  Welch Two Sample t-test
##
## data:  repub_app and dem_app
## t = 42.24, df = 3665.4, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.5031448 0.5521265
## sample estimates:
## mean of x mean of y
##  1.713251  1.185615
```

```r
cohen.d(repub_app, dem_app, na.rm = TRUE)
```

```
##
## Cohen's d
##
## d estimate: 1.272009 (large)
## 95 percent confidence interval:
##      inf      sup
## 1.209613 1.334405
```

```r
(diff = mean(repub_app, na.rm = TRUE) - mean(dem_app, na.rm = TRUE))
```

```
## [1] 0.5276357
```

This is a very small p-value that shows a highly significant statistical value. We can reject the null hypothesis that there is no difference. This adds support to our alternative hypothesis that there is a difference between republicans and democrats' approval of the President's handling of the war. Our t-test shows that republicans disapprove more and democrats approve more. This was also displayed in our histograms.

There's also a practical difference between the two party's approvals. We have large cohen's d value of 1.27 that supports this practical effect. The difference in means is about .5 that shows about a half point difference between the two parties approval rating for handling the war.

Conclusion

We see that there was not much of a change in voters' placement from before and after the election. Splitting the voters into republican and democrat groups show that there were some statistically significant shifts in placement between those groups, but there was not much evidence of any practical difference between the groups.

Republicans are slightly older from a statistical point of view than democrats, but their mean age difference was only a couple years which is not very much. And the republicans' mean age was close to 51.

Lastly, the voters are very split in their approval of the presidents handling of the war. Democrats mostly approve and republicans mostly disapprove.