Lab 3: Hypothesis Tests about the Mean

w203: Statistics for Data Science

DUE: 11/20/2017, 11:59PM PST

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.

```
library(plyr)
library(dplyr)
```

```
## Warning: package 'dplyr' was built under R version 3.4.2
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:plyr':
##
##     arrange, count, desc, failwith, id, mutate, rename, summarise,
##     summarize
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
S = read.csv("C:\\users\\nishray\\documents\\Berkeley_MIDS\\W203\\Lab3_2017Fall_2\\Lab3_2017Fall\\ANES_
```

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.

WARNING: If you coerce these variables directly into numeric vectors (or if such coercion happens automatically), the levels may not translate into the right numbers. Consider the following code.

```
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
```

You can find explanations for other variables in the ANES codebook.

Assignment

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

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

```
S['libcpre_self_numbers'] = mapvalues(S$libcpre_self, from = levels(S$libcpre_self), to=c("-2","-8","-9

S['libcpo_self_numbers'] = mapvalues(S$libcpo_self, from = levels(S$libcpo_self), to=c("-2","-6","-7","

S['flat_pre'] <- as.numeric(levels(S$libcpre_self_numbers))[as.integer(S$libcpre_self_numbers)]
S['flat_po'] <- as.numeric(levels(S$libcpo_self_numbers))[as.integer(S$libcpo_self_numbers)]
```

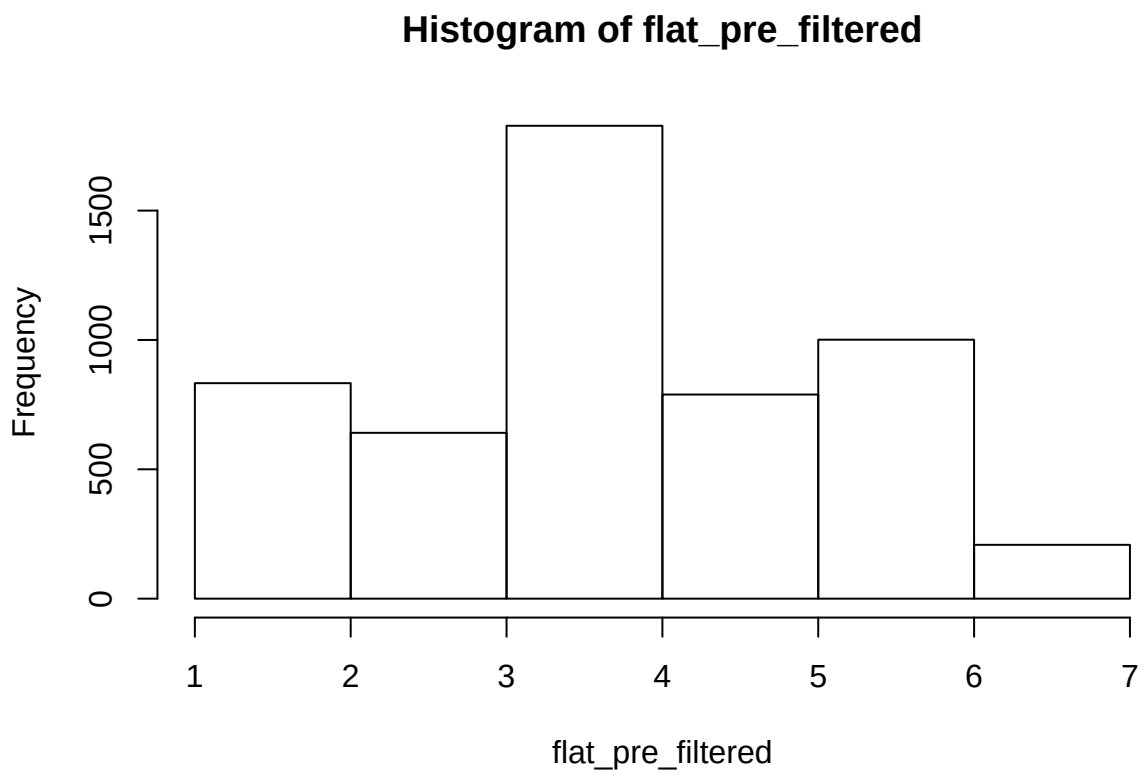
```
flat_pre_filtered <- S[S$flat_pre>0, "flat_pre"]
flat_po_filtered <- S[S$flat_po>0, "flat_po"]

#count of each label
with(S, table(libcpre_self))
```

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

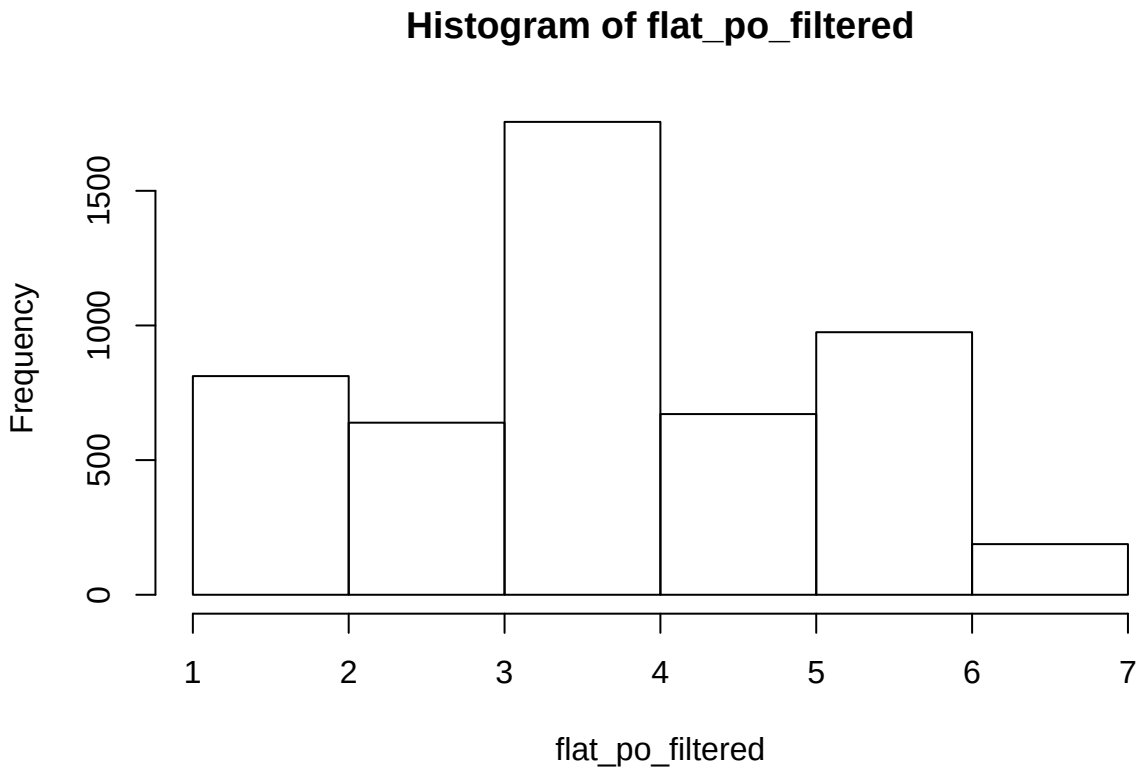
```
#count of each label to confirm that the mapping was correct
with(S, table(flat_pre))
```

```
## flat_pre
##   -9   -8   -2    1    2    3    4    5    6    7
##   32   26  556  195  638  641 1828  789 1001  208
```

```
#visualize data to check t-test assumptions
hist(flat_pre_filtered, breaks=unique(flat_pre_filtered))
```



```r
hist(flat_po_filtered, breaks=unique(flat_po_filtered))
```




```r
t.test(flat_pre_filtered, flat_po_filtered)
```

```
##
##  Welch Two Sample t-test
##
## data:  flat_pre_filtered and flat_po_filtered
## t = 0.77123, df = 10314, p-value = 0.4406
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.03436916  0.07895698
## sample estimates:
## mean of x mean of y
##  4.172264  4.149970
```

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

```r
#read out levels
levels(S$pid_x)
```

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

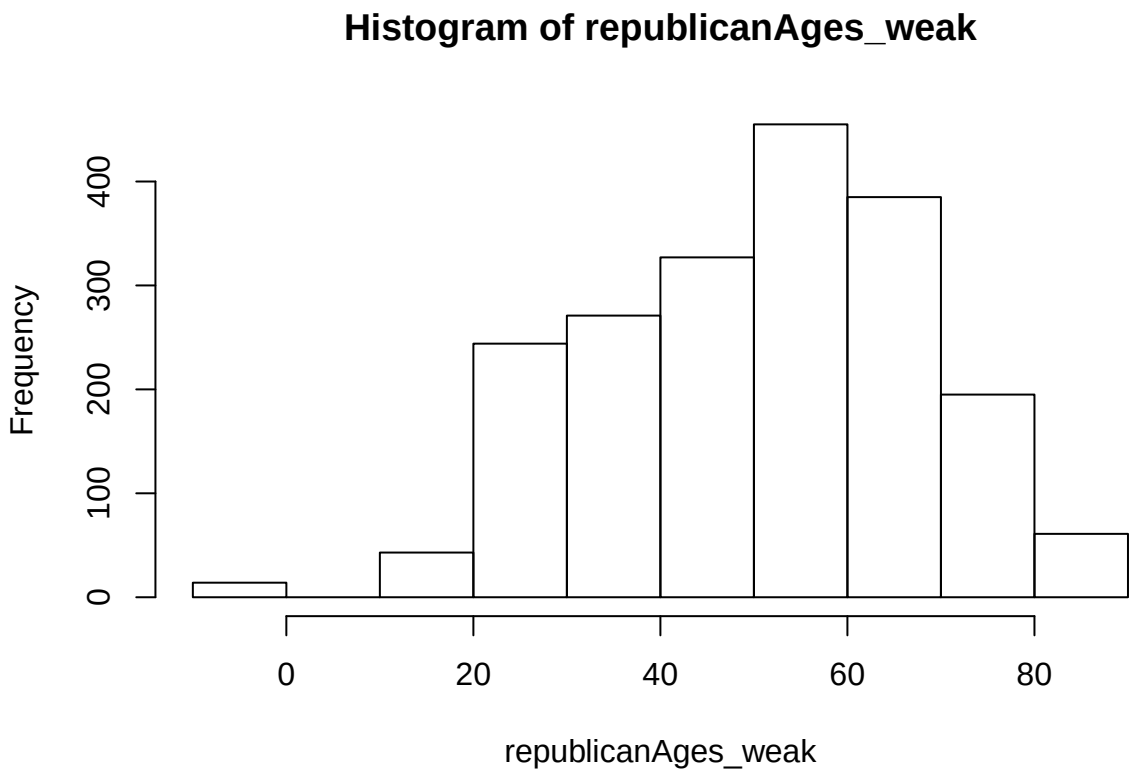
```r
#will have to decide what is "republic" versus "democratic"
S['pid_x_mapped_unparsed'] = mapvalues(S$pid_x, from = levels(S$pid_x), to=c("-2","1","2","3","4","5","
S['pid_x_mapped'] = as.numeric(levels(S$pid_x_mapped_unparsed))[as.integer(S$pid_x_mapped_unparsed)]

#count of each label
with(S, table(pid_x))
```

```
## 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
```

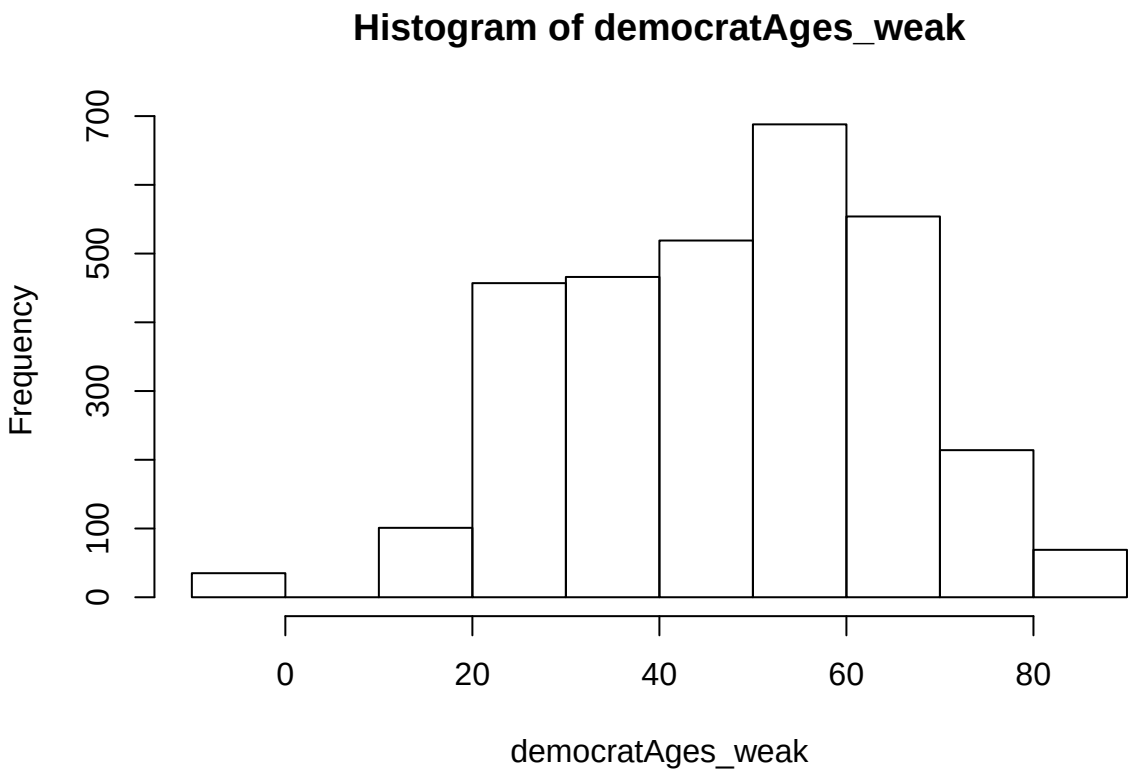
```r
#count of each label to confirm that the mapping was correct
with(S, table(pid_x_mapped))
```

```
## pid_x_mapped
##   -2    1    2    3    4    5    6    7
##   24 1485  871  747  792  610  623  762
```

```r
republicanAges_weak = S[S$pid_x_mapped > 4, "dem_age_r_x"]
democratAges_weak = S[S$pid_x_mapped > 0 & S$pid_x_mapped < 4, "dem_age_r_x"]

#check for normalcy of distributions
hist(republicanAges_weak)
```



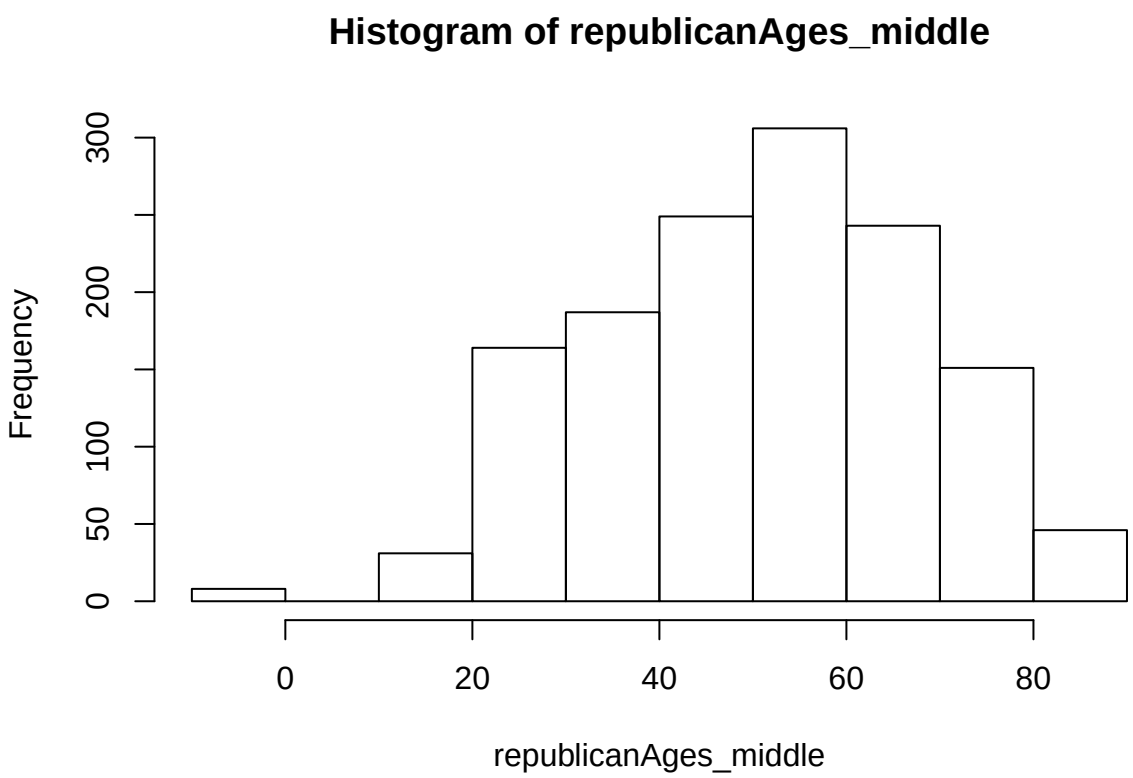

```
hist(democratAges_weak)
```



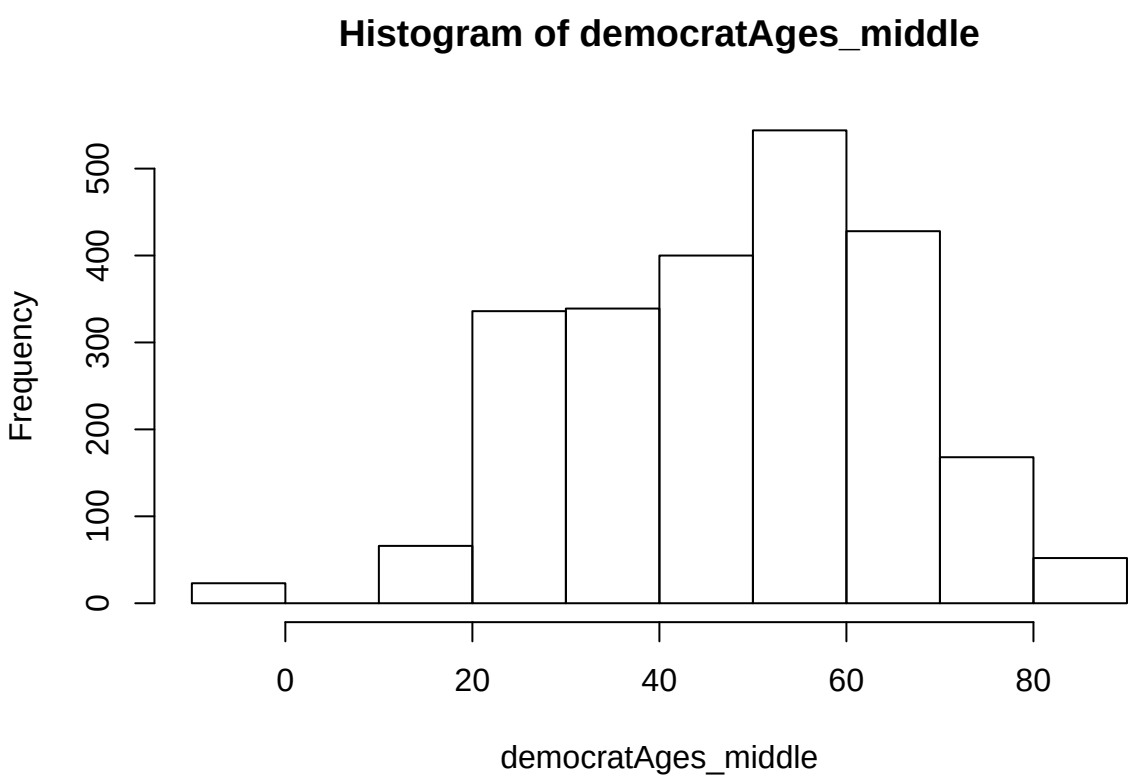

```
t.test(republicanAges_weak, democratAges_weak)
```

```
##
##  Welch Two Sample t-test
##
## data:  republicanAges_weak and democratAges_weak
## t = 5.4043, df = 4274.8, p-value = 6.859e-08
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.715718 3.669187
## sample estimates:
## mean of x mean of y
##  50.95639  48.26394
```

```
#repeat test with stricter definition of republican and democratic

republicanAges_middle = S[S$pid_x_mapped > 5, "dem_age_r_x"]
democratAges_middle = S[S$pid_x_mapped > 0 & S$pid_x_mapped < 3, "dem_age_r_x"]

#check for normalcy of distributions
hist(republicanAges_middle)
```



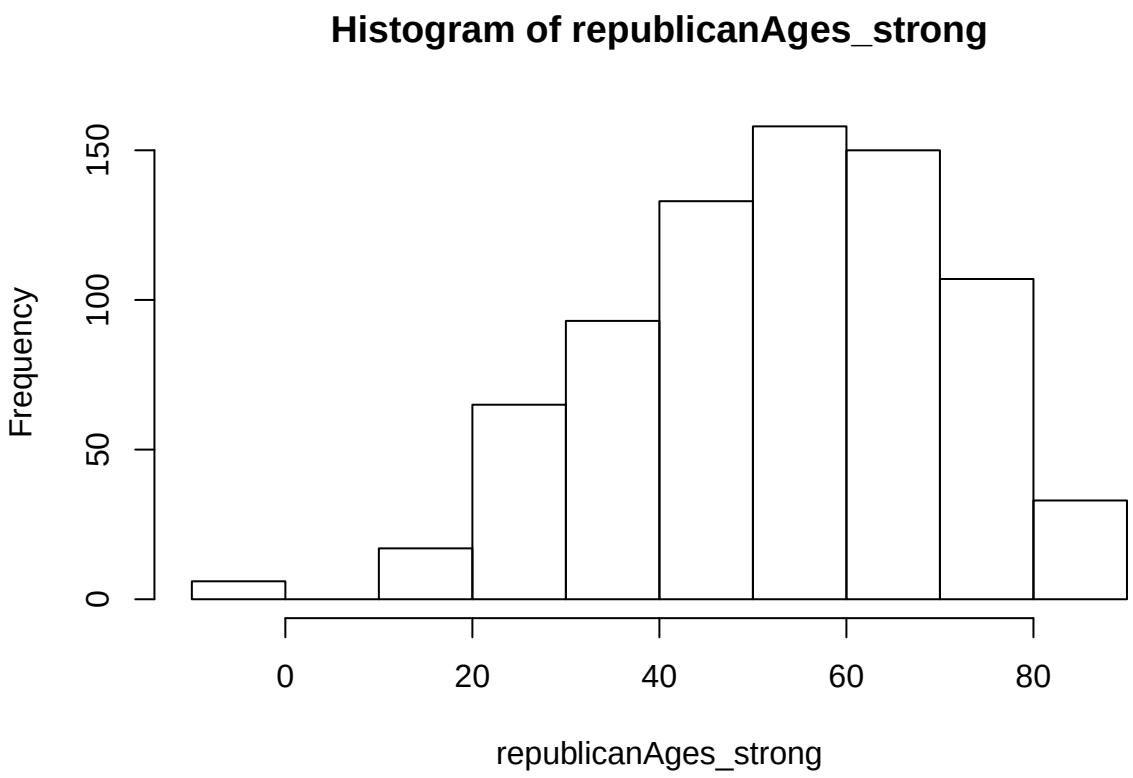

```
hist(democratAges_middle)
```



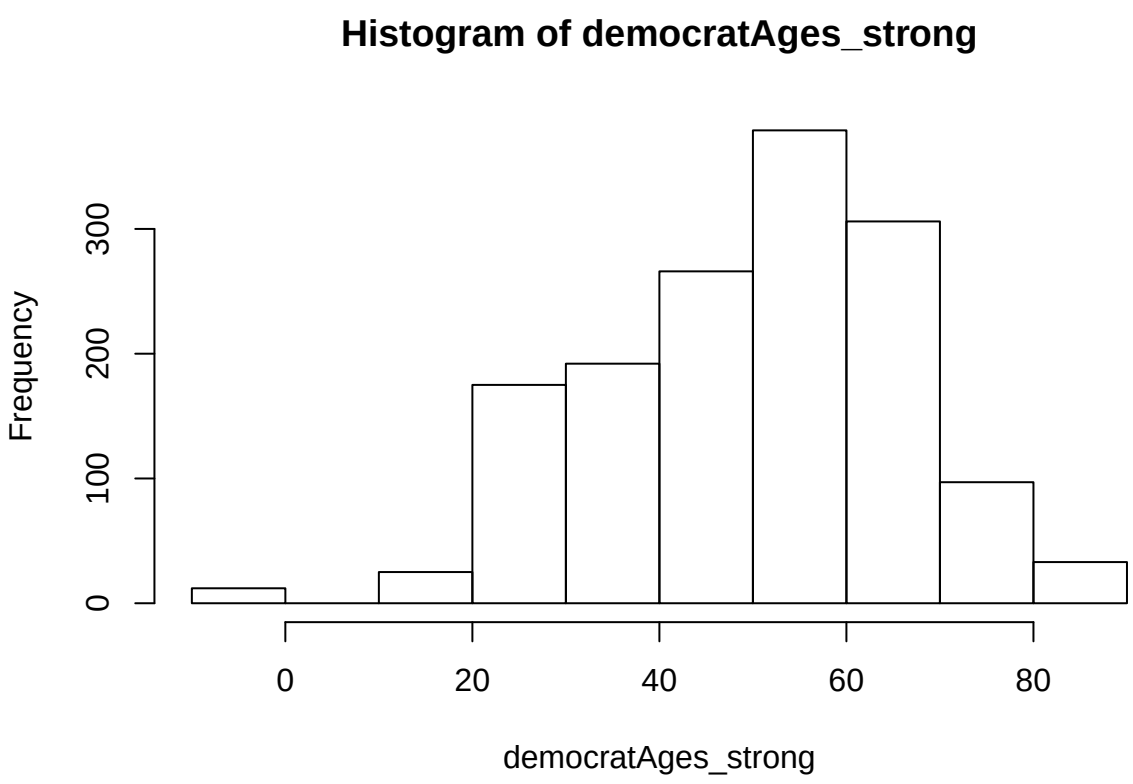

```
t.test(republicanAges_middle, democratAges_middle)
```

```
##
##  Welch Two Sample t-test
##
## data:  republicanAges_middle and democratAges_middle
## t = 3.795, df = 2882.4, p-value = 0.0001507
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.071791 3.363346
## sample estimates:
## mean of x mean of y
##  51.06137  48.84380
```

```
#repeat test with strictest definition of republican and democratic

republicanAges_strong = S[S$pid_x_mapped > 6, "dem_age_r_x"]
democratAges_strong = S[S$pid_x_mapped > 0 & S$pid_x_mapped < 2, "dem_age_r_x"]

#check for normalcy of distributions
hist(republicanAges_strong)
```




```
hist(democratAges_strong)
```

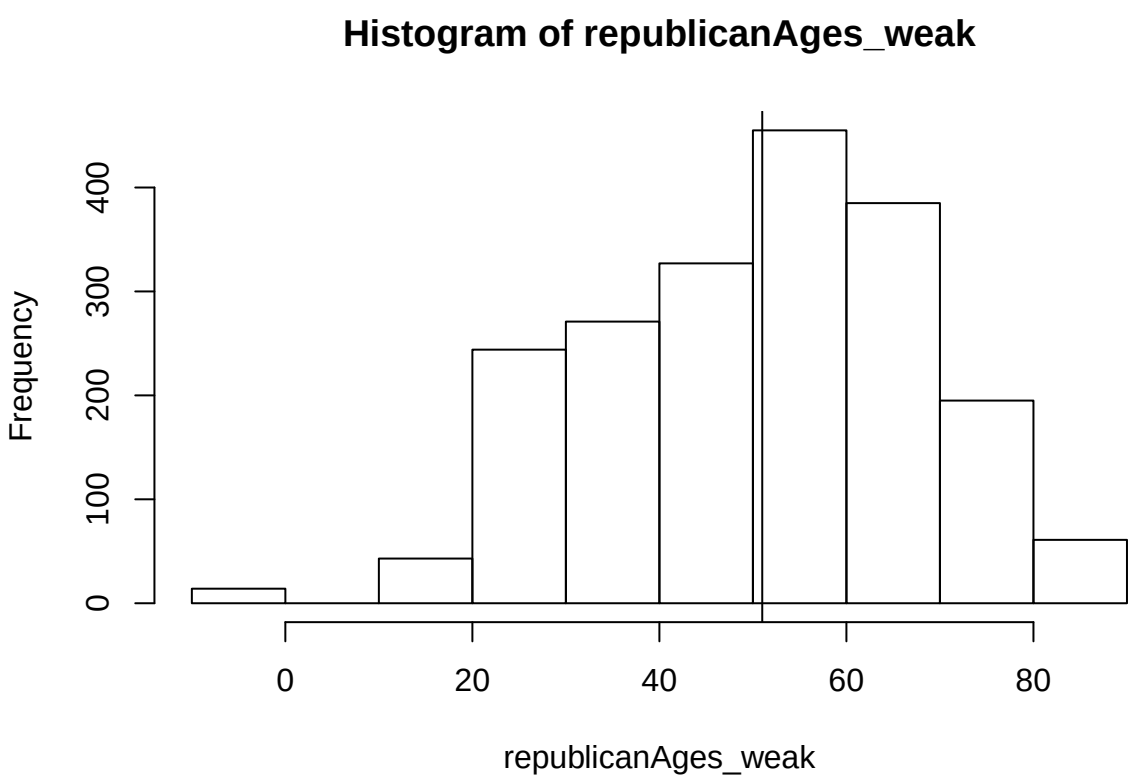



```
t.test(republicanAges_strong, democratAges_strong)
```

```
##
##  Welch Two Sample t-test
##
## data:  republicanAges_strong and democratAges_strong
## t = 3.9296, df = 1437.2, p-value = 8.914e-05
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  1.504200 4.502894
## sample estimates:
## mean of x mean of y
##  53.48031  50.47677
```

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

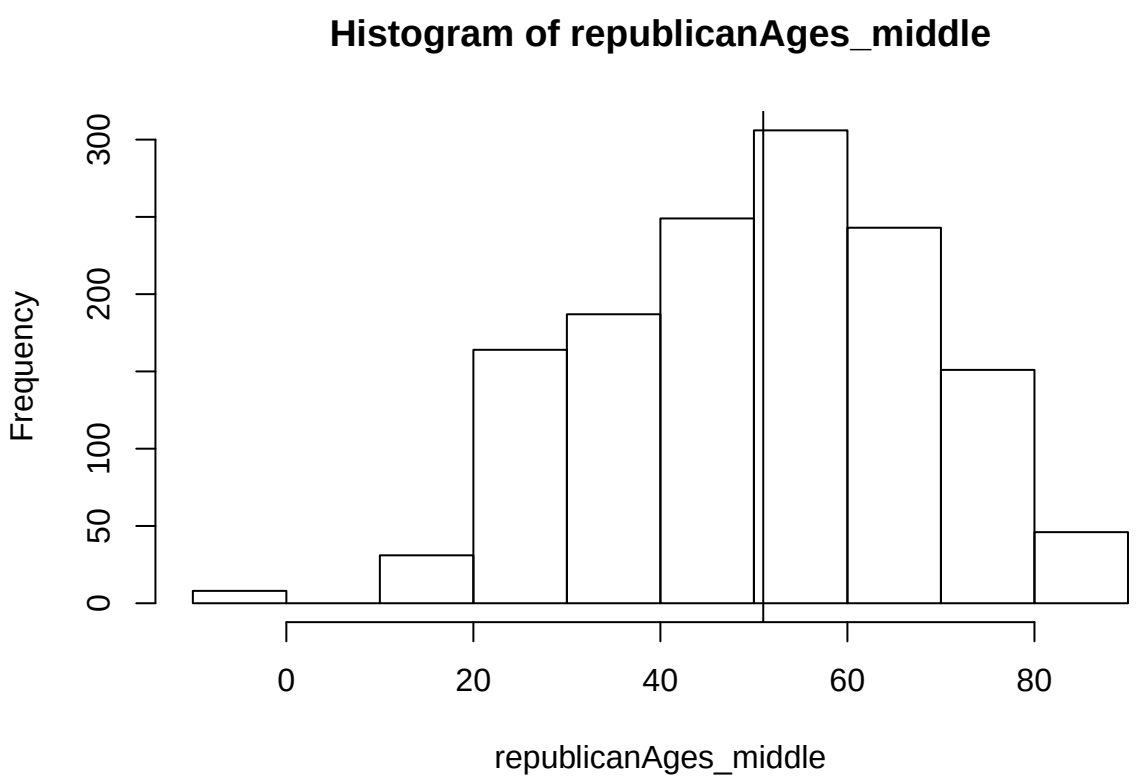
```
#We face the same issues as we did in the previous question.
#How do we define republican versus democrat?
#We'll use the same three tests previously

#loosest definition of republican (independent republican - strong republican)
hist(republicanAges_weak)
abline(v=51)
```




```
t.test(republicanAges_weak, mu=51, alternative="greater")
```

```
##
##  One Sample t-test
##
## data:  republicanAges_weak
## t = -0.11252, df = 1994, p-value = 0.5448
## alternative hypothesis: true mean is greater than 51
## 95 percent confidence interval:
##  50.31859      Inf
## sample estimates:
## mean of x
##  50.95639
```

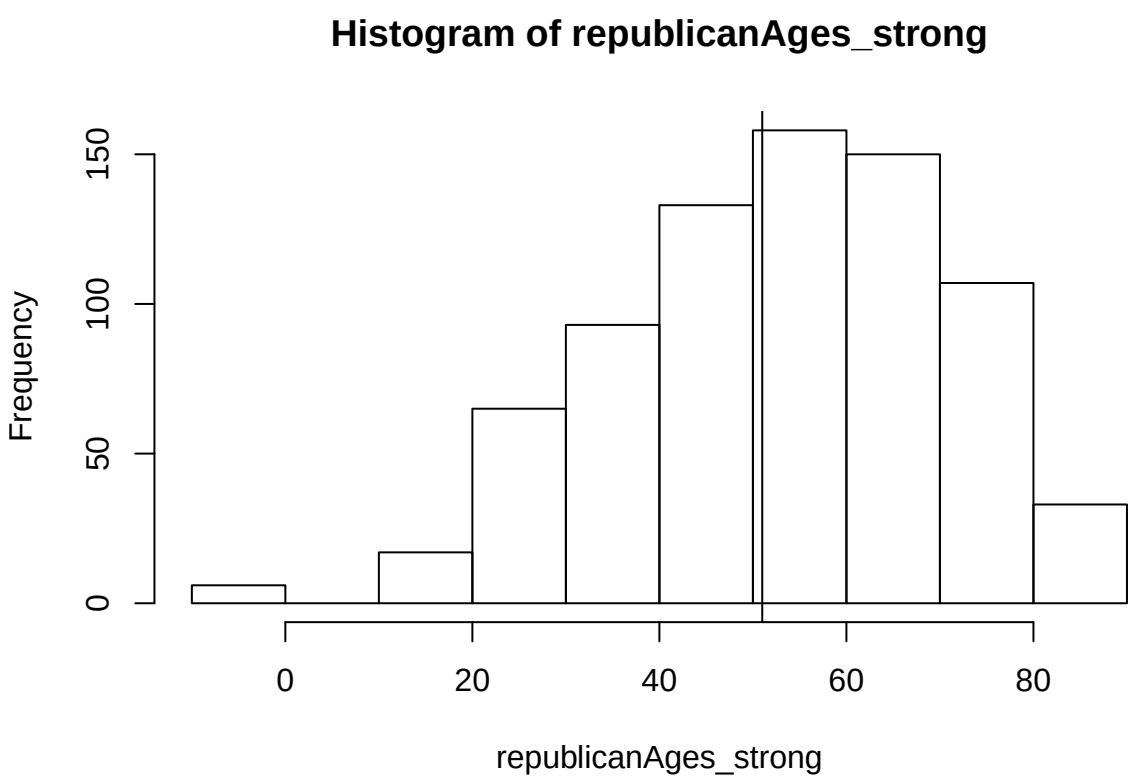
```
#middle definition of republican (slightly republican - strong republican)
hist(republicanAges_middle)
abline(v=51)
```




```
t.test(republicanAges_middle, mu=51, alternative="greater")
```

```
##
##  One Sample t-test
##
## data:  republicanAges_middle
## t = 0.13197, df = 1384, p-value = 0.4475
## alternative hypothesis: true mean is greater than 51
## 95 percent confidence interval:
##  50.29592      Inf
## sample estimates:
## mean of x
##  51.06137
```

```
#strict definition of republican (strong republican)
hist(republicanAges_strong)
abline(v=51)
```

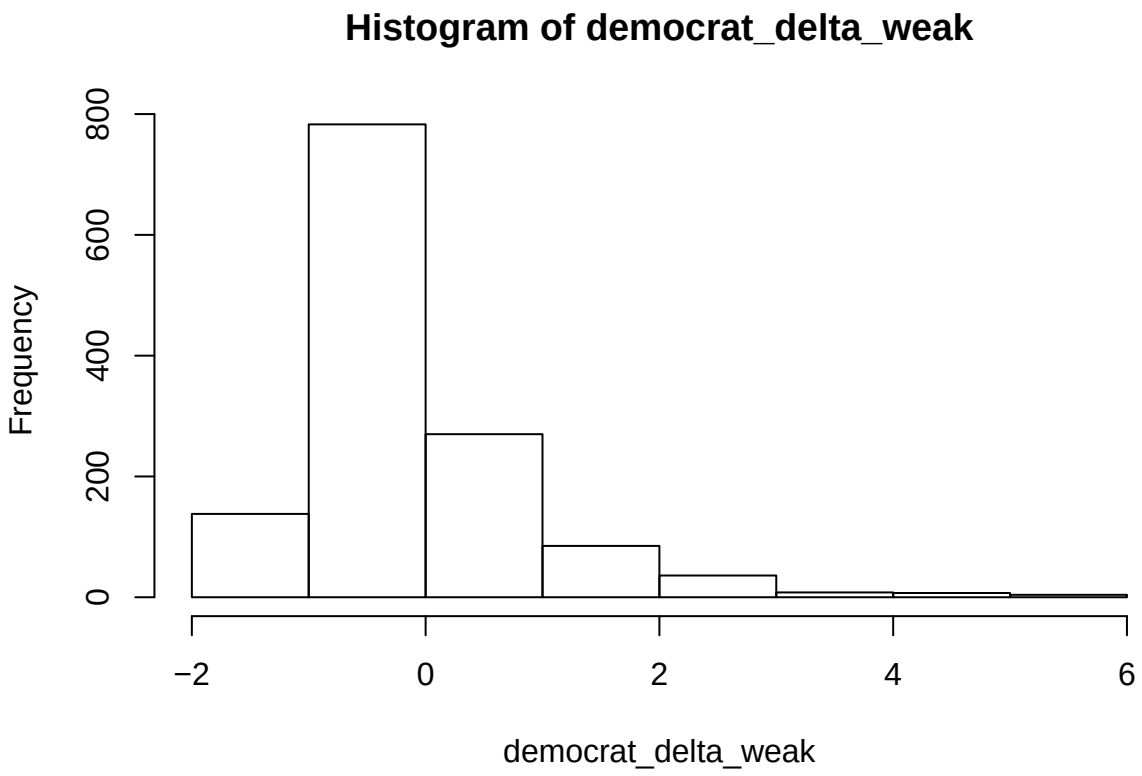



```
t.test(republicanAges_strong, mu=51, alternative="greater")
```

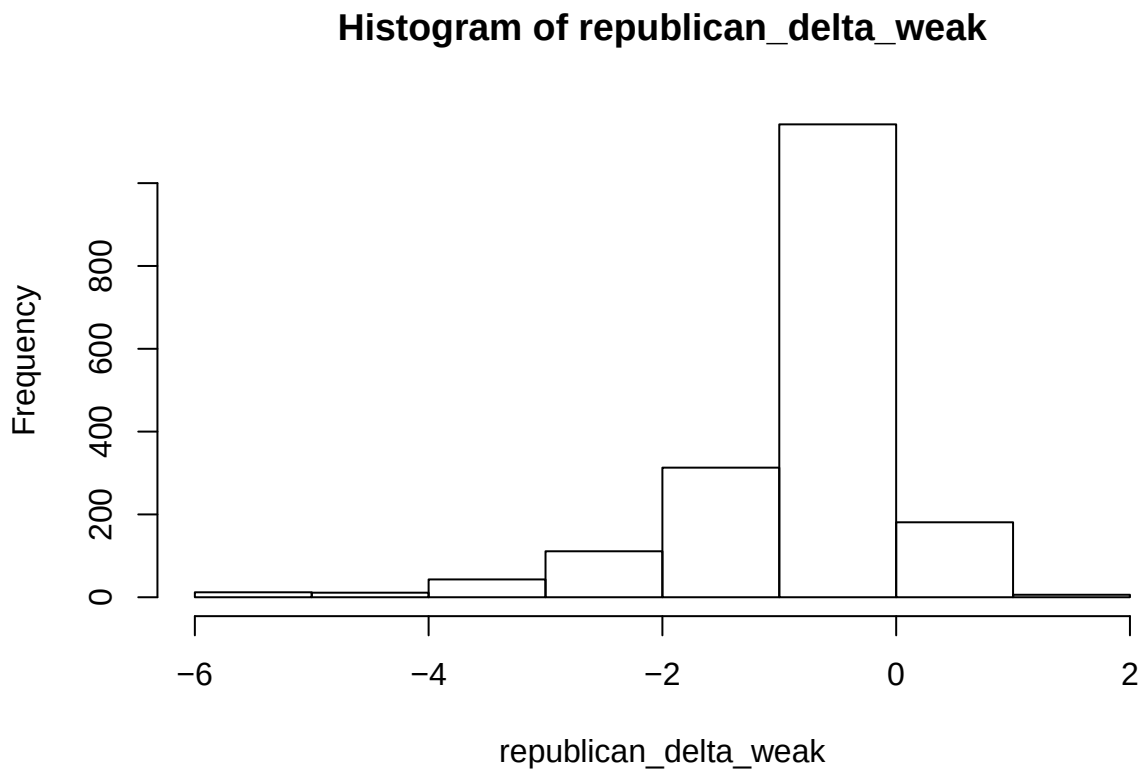
```
##
##  One Sample t-test
##
## data:  republicanAges_strong
## t = 3.8957, df = 761, p-value = 5.327e-05
## alternative hypothesis: true mean is greater than 51
## 95 percent confidence interval:
##  52.43179      Inf
## sample estimates:
## mean of x
##  53.48031
```

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?

```
both_responses <- S[S$flat_pre > 0 & S$flat_po > 0, c("flat_pre", "flat_po")]
both_responses["differences"] <- both_responses$flat_po - both_responses$flat_pre

#weak definition of republican and democratic
democrat_delta_weak = both_responses[both_responses$flat_pre < 4, "differences"]
republican_delta_weak = both_responses[both_responses$flat_pre > 4, "differences"]

hist(democrat_delta_weak, breaks=unique(democrat_delta_weak))
```



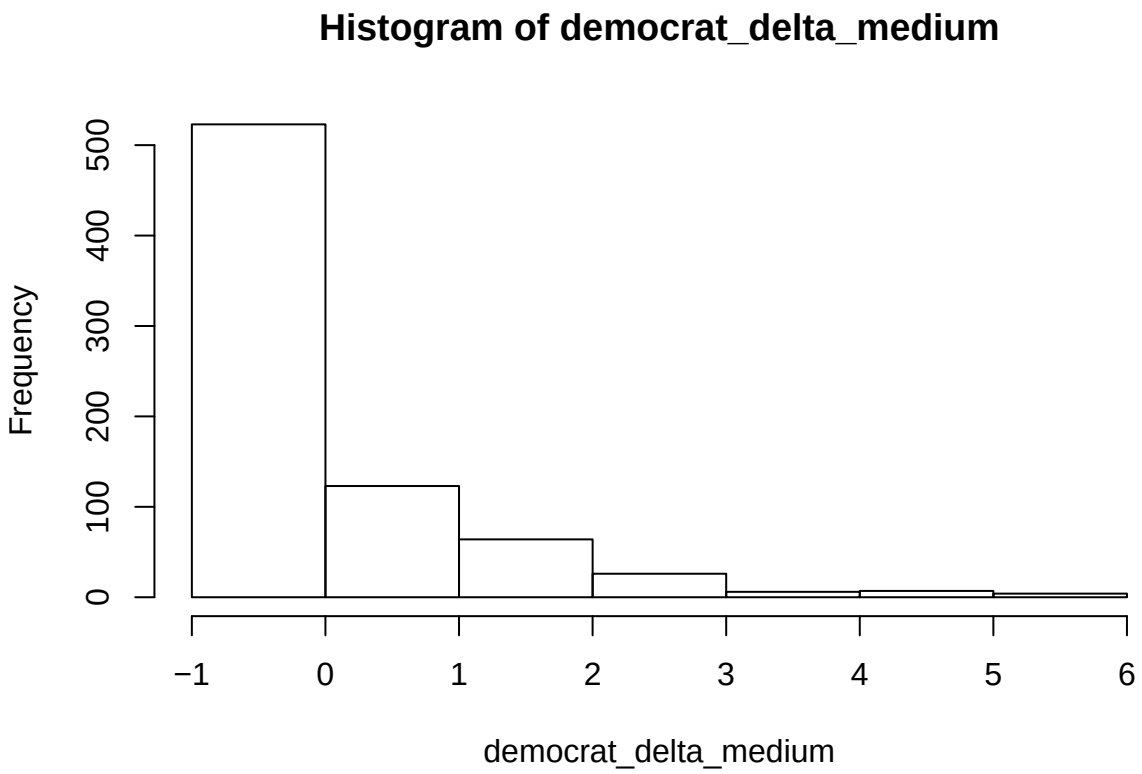

```
hist(republican_delta_weak, breaks=unique(republican_delta_weak))
```



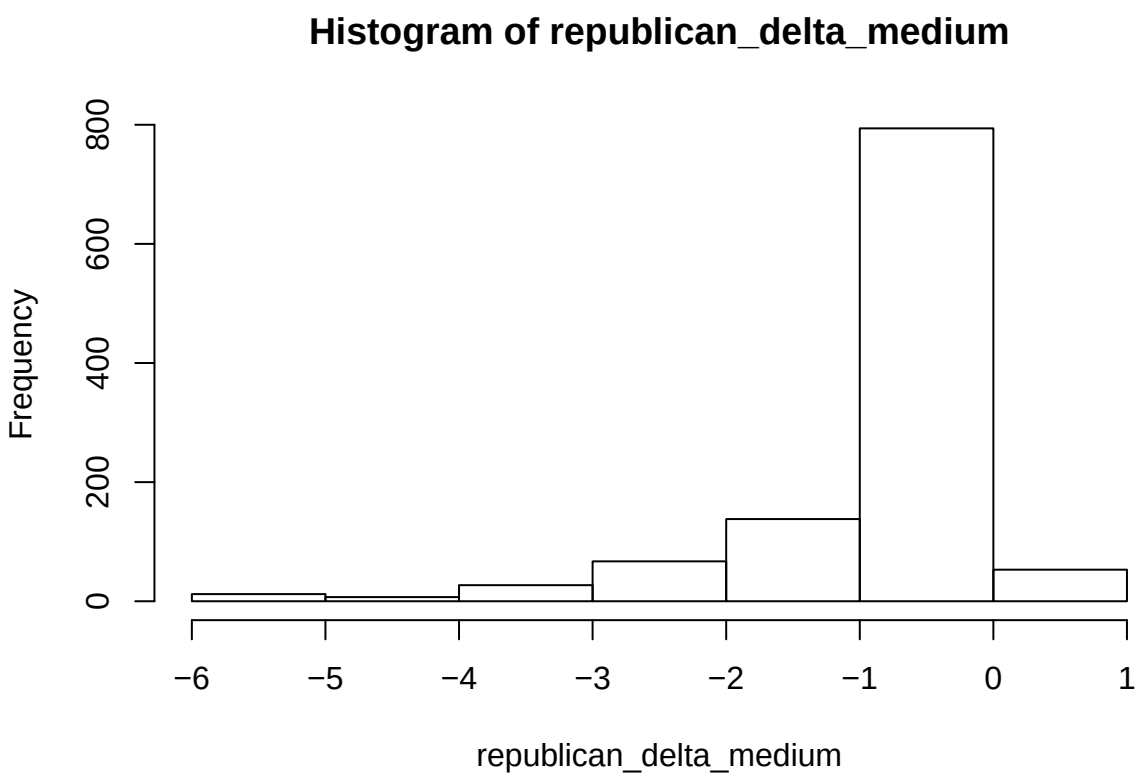

```
#our data is roughly normal and large sample (1330 in democrat, 1820 in republican)
#we can assume iid

t.test(democrat_delta_weak, republican_delta_weak)
```

```
##
##  Welch Two Sample t-test
##
## data:  democrat_delta_weak and republican_delta_weak
## t = 19.377, df = 2790.2, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.6210377 0.7608801
## sample estimates:
##  mean of x  mean of y
##  0.3726521 -0.3183068
```

```
#medium definition of republican and democratic
democrat_delta_medium = both_responses[both_responses$flat_pre < 3, "differences"]
republican_delta_medium = both_responses[both_responses$flat_pre > 5, "differences"]

hist(democrat_delta_medium, breaks=unique(democrat_delta_medium))
```



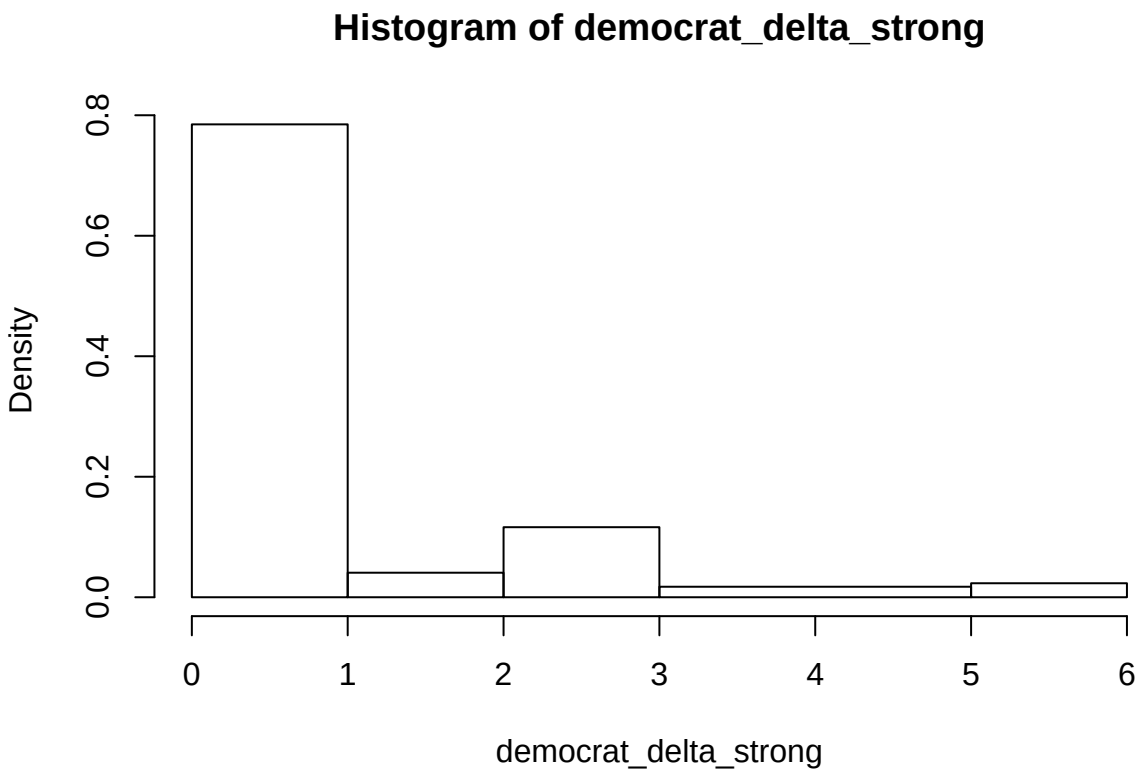

```
hist(republican_delta_medium, breaks=unique(republican_delta_medium))
```



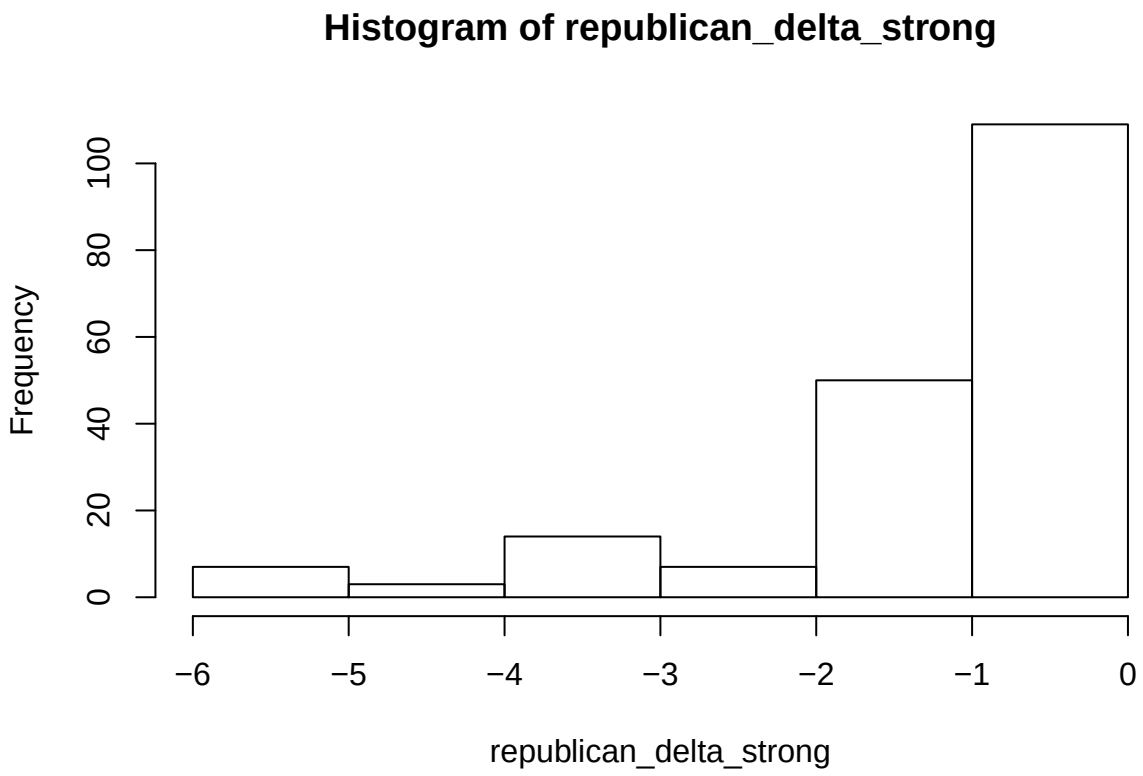

```
#our data is roughly normal and large sample (753 in democrat, 1098 in republican)
#we can assume iid - though the data is paired, we are comparing republican and democrat pairs
#no longer paired

t.test(democrat_delta_medium, republican_delta_medium)
```

```
##
##  Welch Two Sample t-test
##
## data:  democrat_delta_medium and republican_delta_medium
## t = 17.587, df = 1495.1, p-value < 2.2e-16
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  0.7631979 0.9548116
## sample estimates:
##  mean of x  mean of y
##  0.5019920 -0.3570128
```

```
#strict definition of republican and democratic
democrat_delta_strong = both_responses[both_responses$flat_pre < 2, "differences"]
republican_delta_strong = both_responses[both_responses$flat_pre > 6, "differences"]

hist(democrat_delta_strong, breaks=unique(democrat_delta_strong))
```




```
hist(republican_delta_strong, breaks=unique(republican_delta_strong))
```

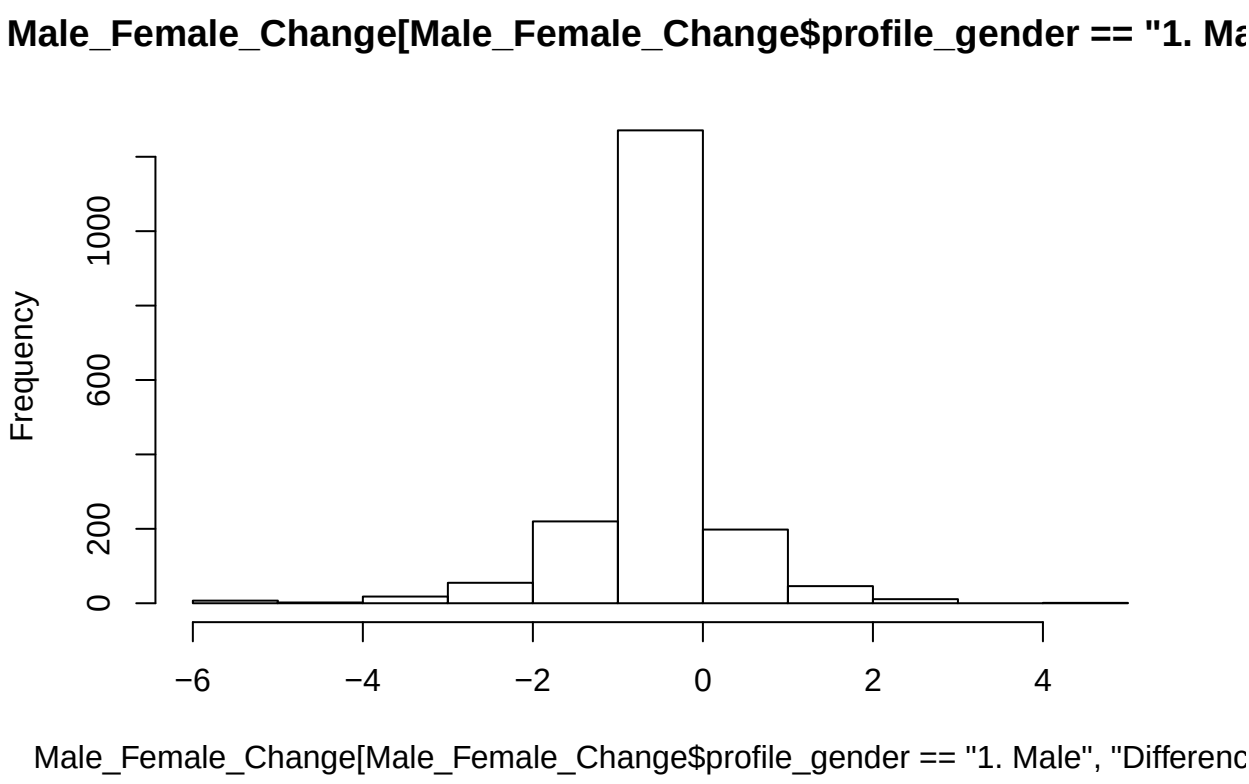



```
#our data is non normal and fairly small sample (n=172, 190 for democratic and republican respectively)
#will use a non-parametric test

wilcox.test(democrat_delta_strong, republican_delta_strong)
```

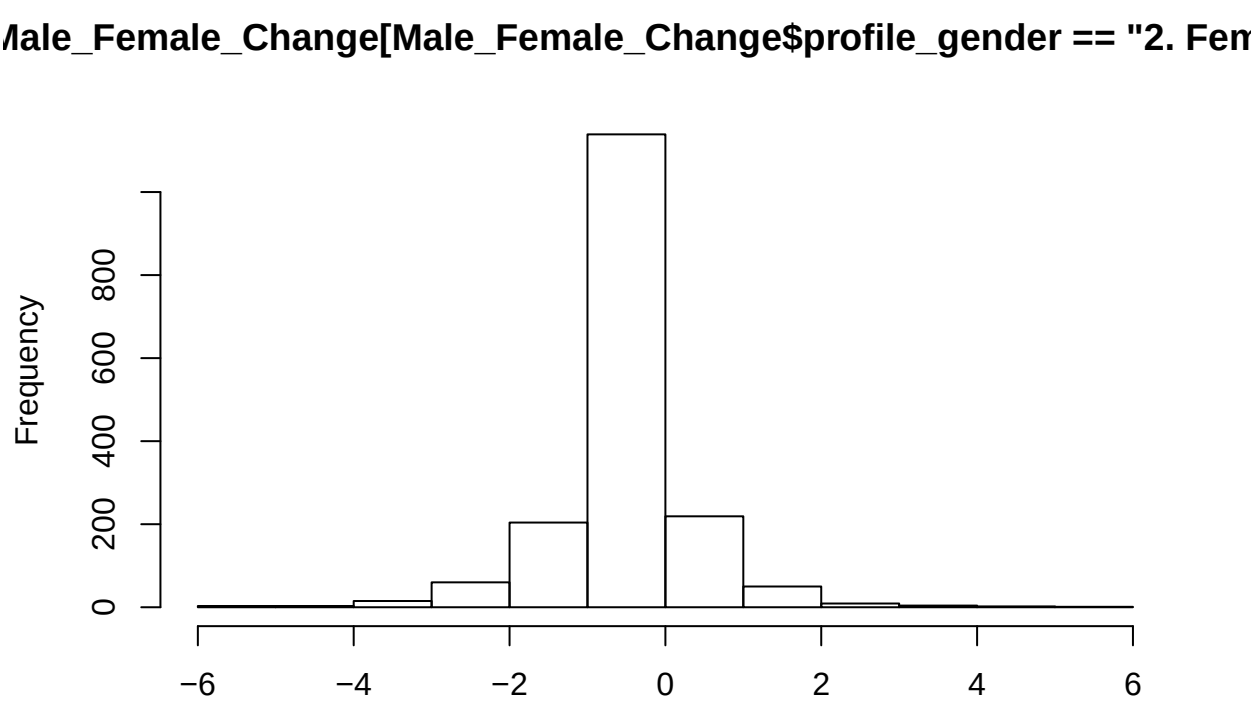
```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  democrat_delta_strong and republican_delta_strong
## W = 27612, p-value < 2.2e-16
## alternative hypothesis: true location shift is not equal to 0
```

5. Select a fifth question that you are interested in investigating. Did female or male voters change their political leanings more during the 2012 election?

```
Male_Female_Change <- S[S$flat_pre > 0 & S$flat_po > 0, c("profile_gender","flat_pre", "flat_po")]

Male_Female_Change["Difference"] <- Male_Female_Change$flat_po - Male_Female_Change$flat_pre

hist(Male_Female_Change[Male_Female_Change$profile_gender == "1. Male", "Difference"])
```




```
hist(Male_Female_Change[Male_Female_Change$profile_gender == "2. Female", "Difference"])
```




```
#approximately normally distributed

length(Male_Female_Change[Male_Female_Change$profile_gender == "1. Male", "Difference"])
```

```
## [1] 1829
```

```
length(Male_Female_Change[Male_Female_Change$profile_gender == "2. Female", "Difference"])
```

```
## [1] 1709
```

```
# large enough sample sizes

t.test(Male_Female_Change[Male_Female_Change$profile_gender == "1. Male", "Difference"], Male_Female_Cha
```

```
##
##  Welch Two Sample t-test
##
## data:  Male_Female_Change[Male_Female_Change$profile_gender == "1. Male",  and Male_Female_Change[Ma
## t = -1.4376, df = 3481.3, p-value = 0.1506
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.10019624  0.01542264
## sample estimates:
##   mean of x   mean of y
## -0.05467469 -0.01228789
```

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.