Exploring Data and Doing Hypothesis Tests in R

Daniel Vanlunen

Introduction

This report focuses on using a sample of the American National Election Studies survey. This is survey data from a random sample of respondents. It contains information about voters before and after the 2012 presidential election. The primary goal of this report is to address 5 questions with statistical tests. Prior to each test, assumptions are checked and reasons why the test is an appropirate choice are discussed. After the five questions, there are some concluding remarks. Note that because the data contains information about voters in 2012, the conclusions we draw here are applicable to that population.

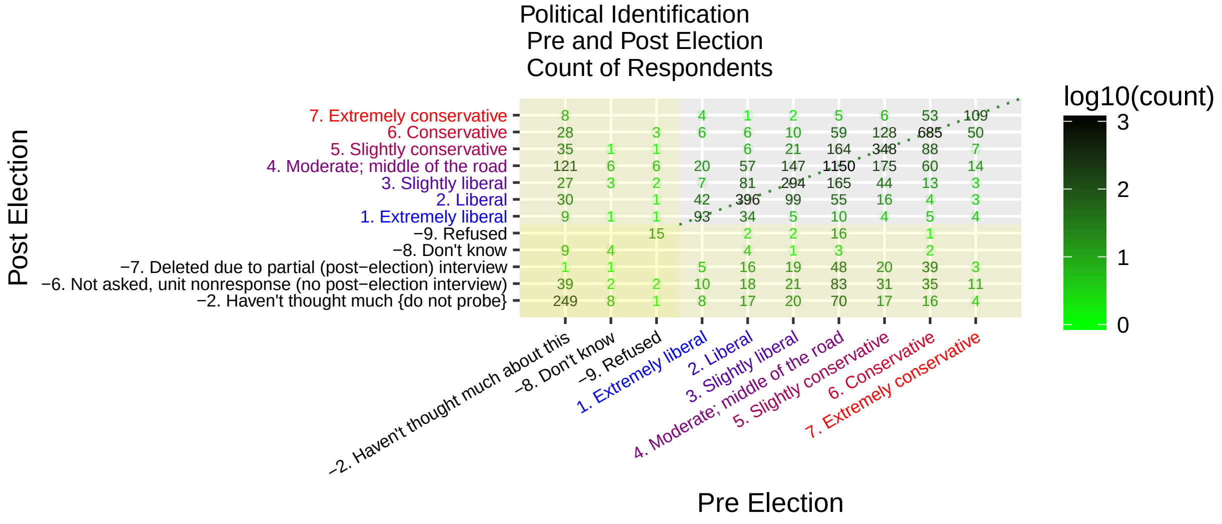
Questions

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

To address this question, we use a non-parametric, rank-based Wilcoxon Sign Test because we only need to see whether the post-election voters tend to rank below or above their pre-election levels. This choice is more appropriate than its parametric counterpart–the paired t-test–because the paired t-test would require us to code the levels of the ordinal political stance as numbers. This would impose a structure on the distance between different values that we do not know exists (e.g. if we coded extemely liberal to extremely conservative to 1:7, that would impose a linear structure).

This test assumes (i) independence between paired observations and (ii) the distribution of the first value in each pair has the same shape and spread as the distribution of the second value (only different in mean). (i) is met because the data is a random sample. We discuss (ii) below.

The table shows the count of respondents by their political identification pre and post election.

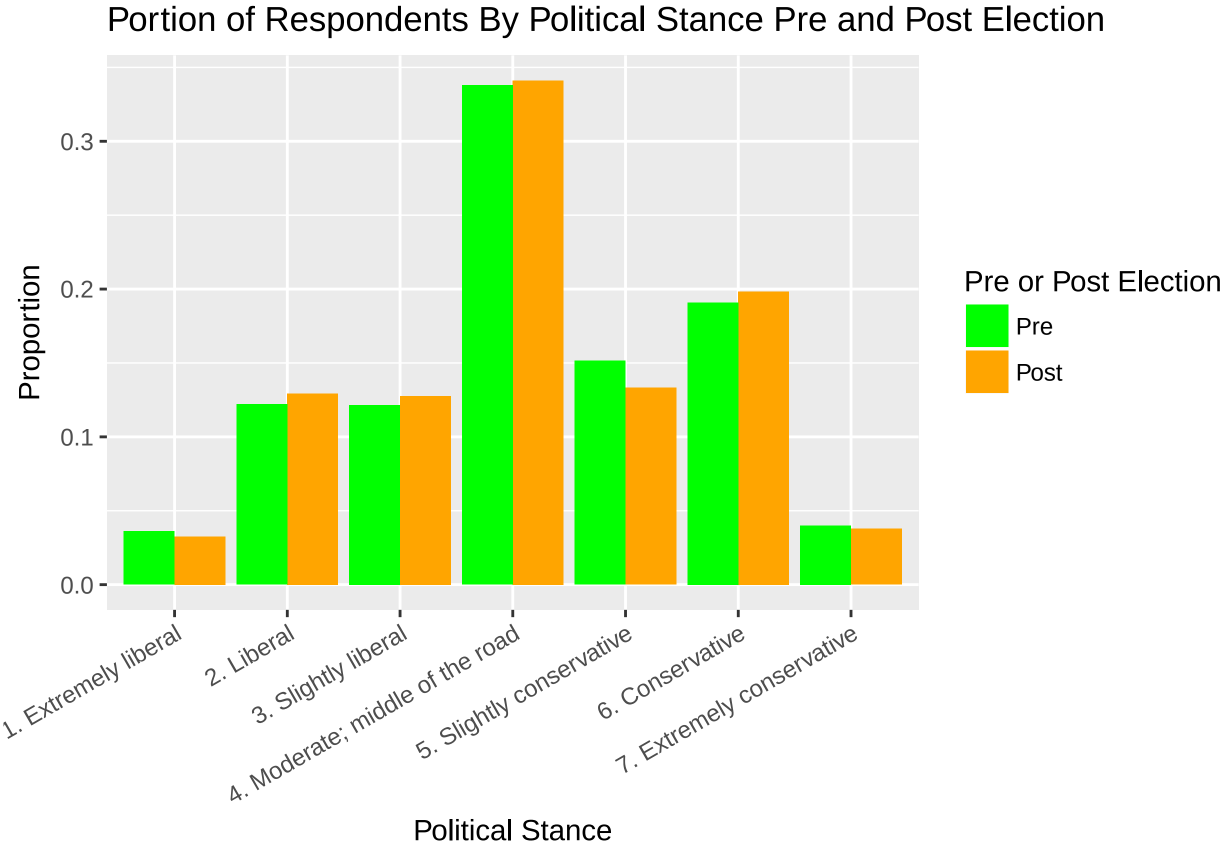


Negative responses (highlighted yellow in the table) either have to be replaced with imputed values or removed to perform our test. Imputing has the benefits of adding extra data. Values might be imputed from other highly correlated attributes like party. However, most negative responses seem to indicate the respondent was wavering on a response, which means their other answers may not be indicative of their true political stance anyway and imputation may reduce our ability to test the actual question. Responses -6 and -7 for post election represent not getting an answer because it was not asked. In this case, imputation would likely be more helpful. However, these only account for 45 respondents so we remove them as well for simplicity. In the test we will use the respondents in the non-highlighted area of the above table

```
# remove negative responses
S1 = filter(S, as.integer(libcpre_self)>3,
            as.integer(libcpo_self)>5)[,c("libcpo_self","libcpre_self")]

#make levels match
S1$libcpo_self=droplevels(S1$libcpo_self)
S1$libcpre_self=droplevels(S1$libcpre_self)
lvls=levels(S1$libcpo_self)
Pre = as.numeric(table(S1$libcpre_self))/length(S1$libcpre_self)
Post = as.numeric(table(S1$libcpo_self))/length(S1$libcpo_self)
```

Now let's check assumption (ii) that the shape before and after election of the distribution looks similar.



Looking at the figure, it appears (ii)) is met: the distribution of political stance before the election has a similar shape and spread to the distribution after.

Now we are ready to perform our test. We have no reason to believe respondents should change one way or the other beforehand so we use a two-sided test.

$$H_0 : \mu_{Pre} = \mu_{Post}$$
$$H_a : \mu_{Pre} \neq \mu_{Post}$$

```r
S1$libcpre_self = droplevels(S1$libcpre_self)
wilcox.test(as.integer(S1$libcpre_self)
            ,as.integer(S1$libcpo_self),
            paired=T)
```

```
##
##  Wilcoxon signed rank test with continuity correction
##
## data:  as.integer(S1$libcpre_self) and as.integer(S1$libcpo_self)
## V = 734760, p-value = 0.1662
## alternative hypothesis: true location shift is not equal to 0
```

The test indicates that there is not a significant shift in political stance during the election: we fail to reject the null hypothesis that the mean political stance level before the election equals the mean level after the election.

```r
#Effect Size Correlation Z/N^.5
qnorm(wilcox.test(as.integer(S1$libcpre_self)
            ,as.integer(S1$libcpo_self),
            paired=T)$p.value/2,lower.tail = F)/
  sqrt(2*nrow(S1))
```

```
## [1] 0.01419206
```

```r
# Percent more liberal
nrow(S1[as.integer(S1$libcpo_self)-as.integer(S1$libcpre_self)<0,])/nrow(S1)
```

```
## [1] 0.1803279
```

```r
# Percent more conservative
nrow(S1[as.integer(S1$libcpo_self)-as.integer(S1$libcpre_self)>0,])/nrow(S1)
```

```
## [1] 0.1733922
```

The above calculation of the correlation also indicates that the effect size is very small. To get a sense of the changes, about 18% of respondents became more liberal and 17% became 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?

For this question, we will test the difference in mean age of the two parties with an unpaired t-test. Given our large range of ages, we can make the simplification that age is a continuous variable instead of an ordinal variable. If the assumptions are met, a t-test will be better than a non-parametric alternative. An unpaired t-test is appropriate because the means are not from the same samples and we do not know the population standard deviation.

This test involves a few assumptions. (i) the data come from a random sample, (ii) the data come from normal sampling distributions, and (iii) the variances for the two parties' ages are equal.

Let's examine the data

```r
table(S$dem_age_r_x[S$dem_age_r_x<18 | S$dem_age_r_x>90 |
                      is.na(S$dem_age_r_x)], useNA = "always")
```

```
##
##   -2   17 <NA>
##   60    2    0
```

A frequency table of the age reveals values of -2 (60 respondents) and 17 (2 respondents) which are odd. Given these values are in the pre period 17 is possible because the respondent could turn 18 by the time of the election. Therefore we can keep the two values of 17. The codebook indicates that -2 corresponds to birthdate left blank. To avoid a data dump, we can examine them elsewhere:

```
write.table(S[S$dem_age_r_x==-2,], "missing_age.csv", sep="|",row.names=F)
```

It appears these respondents are missing all the other profile information about the respondent as well like education and martial status. This could be the result of these respondents being the type of people who do not like to give personal information.

The ratio of Democrats to Republicans in this group (35:14) closely matches the population ratio in the sample (3,103:1,995). Therefore, if we can assume the impact of being in this group on age is similar independent of party status, removing the points will not cause signifant bias to our test. Also, the points only represent about 1% of the total sample. Therefore, we will remove them. If the percentage were larger we could look into imputing values.

Next, we have to operationalize the two groups.

```
summary(S$pid_x[S$dem_age_r_x!=-2])
```

```
##                   -2. Missing            1. Strong Democrat
##                            22                          1473
##  2. Not very strong Democract       3. Independent-Democrat
##                           860                           735
##                4. Independent     5. Independent-Republican
##                           783                           604
## 6. Not very strong Republican          7. Strong Republican
##                           621                           756
```

```
write.table(S[S$dem_age_r_x!=-2 & S$pid_x =='-2. Missing',], "missing_dorr.csv", sep="|",row.names=F)
```

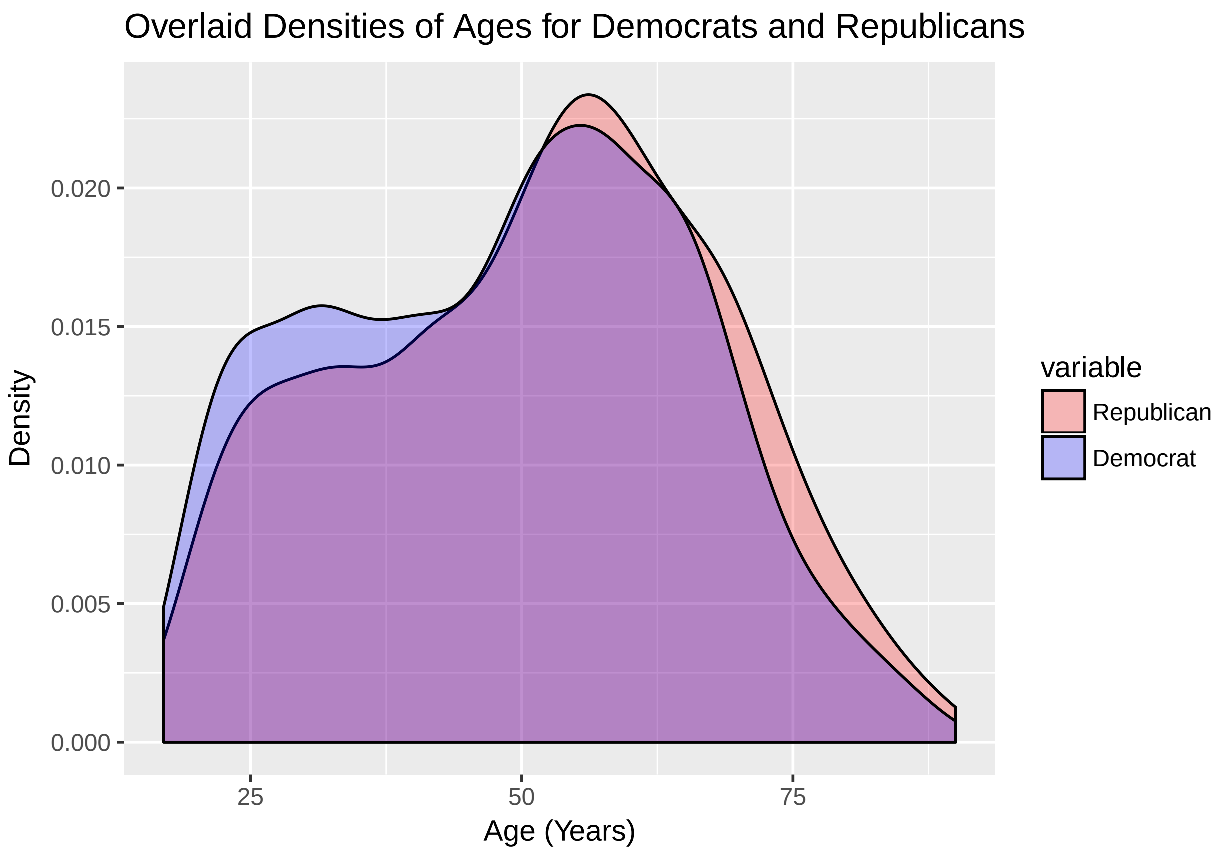
Respondents with pid_x in 1 through 3 can be considered Democrats, while those in 5-7 can be considered republicans. Those who are "Independents" are niether Democrats nor Republicans and are not in our populations of interest, they will therefore not be part of our test. There are 22 missing values. 3 of these 22 have some profile information, but no other questions are answered, while the others have most of their profile information missing (except age). Again 22 is such a small proportion of the sample, so it will not have a large impact on our analysis if it is removed.

```
S$Party =factor(ifelse(S$pid_x=='5. Independent-Republican' |
              S$pid_x=='6. Not very strong Republican' |
              S$pid_x=='7. Strong Republican',"Republican",
              ifelse(S$pid_x=='1. Strong Democrat' |
              S$pid_x=='2. Not very strong Democract' | #sic
              S$pid_x=='3. Independent-Democrat',"Democrat",NA)))

S2=S[S$dem_age_r_x!=-2 & S$pid_x !='-2. Missing',]
R_age = S2[S2$Party=="Republican","dem_age_r_x"]
D_age = S2[S2$Party=="Democrat","dem_age_r_x"]
```

Now we can check assumptions (i) is met because we have a random sample. ii. Normality

```
## Warning: Removed 2653 rows containing non-finite values (stat_density).
```



Both distributions are non-normal. They appear to be bi-modal with one peak around 27 years and another around 64. However, we have a large dataset (1,981 for Republicans and 3,068 for Democrats) and the distributions do not have significant skew meaning our statistic will be a good approximation.

iii. Homogeneity of Variances

```
leveneTest(S2$dem_age_r_x~S2$Party)
```

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

We fail to reject the null hypothesis that the variances of the age of Republicans and Democrats are the same.

With all three assumptions met we are ready to conduct our test. Given that we have no strong reason to believe one group is older than the other, the alternative will be a two-sided test that the ages are different.

$$
\begin{aligned}
H_0 :& \mu_R = \mu_D \\
H_a :& \mu_R \neq \mu_D
\end{aligned}
$$

Now we can run our test.

```
t.test(R_age, D_age, var.equal = T)
```

```
##
##  Two Sample t-test
##
## data:  R_age and D_age
## 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
```

Here the results are very significant with a p value well below .01. We reject the null hypothesis that the age of the two groups are the same. Even though the results are very significant, they have little practical significance. This is because the difference between the two mean sample ages is only about 2.5 years. This is not a very practically significant difference when someone is already about 50 years old.

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

For this question we can use the same two variables we used for the previous question. Because we are specifically asking if they were older than 51, we can use a one-sided, single-sample t-test.

$$
\begin{aligned}
H_0 &: \mu_R = 51 \\
H_a &: \mu_R > 51
\end{aligned}
$$

For this test to be valid the data need to be normally distributed and from a random sample. We are assuming they came from a random sample. Even though the ages of Republicans is not normally distributed, we have a large sample (1,981) and the distribution is not very skewed. Therefore, the test will be a good approximation.

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

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

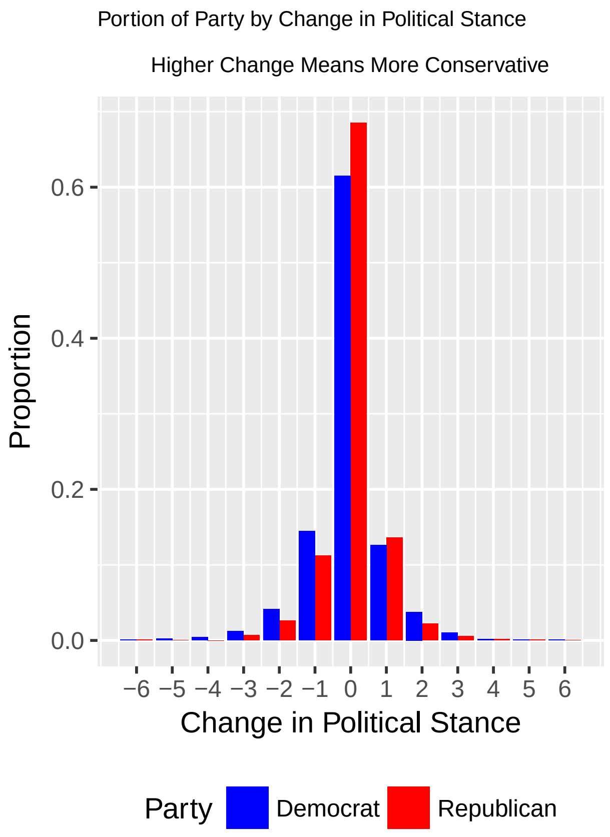
We fail to reject the null hypothesis that the population mean is equal to 51. It also appears there is very little practical significance to this result because the average in our sample is very close to the null hypothesis value.

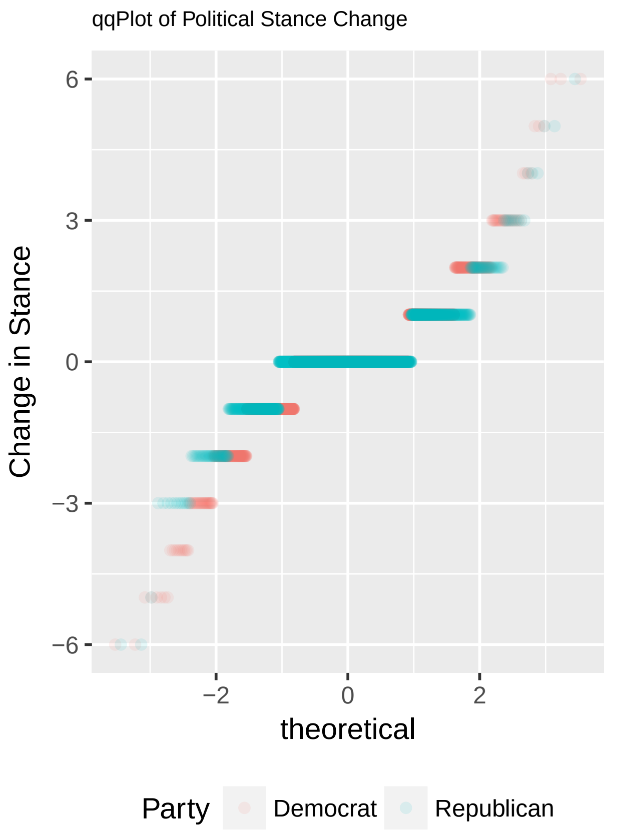
4 Were Republican voters more likely to shift their political preferences right or left (more conservative or more liberal), compared to Democratic voters during the 2012 election?

This question involves two independent groups (Republicans and Democrats) at two points in time (pre and post election). We need a measure of shift pre and post election and then we can perform a test to compare the mean shift between the two independent groups. One way to do this is to code the 7 values from extremly liberal to extremely conservative as numbers 1:7. This imposes a linear structure on the scale that might not be exactly appropriate, however it is a reasonable structure given that we need some measure of difference.

The figure below shows the distribution of change in political stance for Republicans and Democrats (after having removed negative values in the same manner as previous questions)

```
S4= filter(S, as.integer(libcpre_self)>3,
           as.integer(libcpo_self)>5,
           Party %in% c("Democrat","Republican"))
S4$libcpo_self=droplevels(S4$libcpo_self)
S4$libcpre_self=droplevels(S4$libcpre_self)
S4$stancechange=as.numeric(S4$libcpo_self)-as.numeric(S4$libcpre_self)
```





Now we can choose to compare the means of the two groups using a Wilcoxon Rank Sum Test or an independent, two-sample t-test. The nonparametric Wilcoxon Rank Sum Test works well with ordinal variables, but here we have already imposed a linear structure to calculate the difference. Thus, we will use a t-test.

The assumptions are (i) random sample, (ii) normal distribution of each group, and (iii) homogeneity of variance between groups. (i) is met. (ii) From the qqplot we see the data are not perfectly normal. However, since we have such a large sample and the distributions are not extremely skewed, the t-test will still be robust to deviations from normality.

We use Levene's test to test (iii)

```
leveneTest(S4$stancechange~S4$Party)
```

```
## Levene's Test for Homogeneity of Variance (center = median)
##         Df F value    Pr(>F)
## group    1  32.408 1.336e-08 ***
##       4156
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

We reject the null hypothesis that the variances of the groups are equal. Thus, we will instead perform Welch's t-test to account for this.

Now we will perform our test. We use a two-sided test because we have no reason to believe either party will change their stance more in one direction.

$$H_0 : \mu_R = \mu_D$$
$$H_a : \mu_R \neq \mu_D$$

```r
t.test(S4$stancechange~S4$Party)
```

```
##
##  Welch Two Sample t-test
##
## data:  S4$stancechange by S4$Party
## t = -2.3071, df = 4110.1, p-value = 0.0211
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.12318873 -0.01000439
## sample estimates:
##   mean in group Democrat mean in group Republican
##              -0.04771784               0.01887872
```

Our test is significant at the .05 level, though not at the .01 level. We reject the null hypothesis that Republicans shifted their stance more or less than Democrats. However, both means are extremely close to 0 and their difference is only about a twentieth of a group shift. This has very little practical significance.

5 Do a larger or smaller portion of Republican Voters rent their homes for cash than Democrat Voters in 2012?

This question involves comparing proportions of two independent groups (Republicans and Democrats).

```r
S5 = S[S$Party=="Democrat"|S$Party=="Republican",]
table(S5$profile_homeown, useNA = "always")
```

```
##
##                                               -1. Inapplicable
##                                                           1816
## 1. Owned or being bought by you or someone in your household
##                                                           2432
##                                             2. Rented for cash
##                                                            761
##                           3. Occupied without payment of cash rent
##                                                             89
##                                                           <NA>
##                                                            816
```

```r
write.table(S5[S5$profile_homeown=="-1. Inapplicable"|
                 is.na(S5$profile_homeown),],
            "missing_homeown.csv", sep="|",row.names=F)
```

We remove the inapplicable cases because they are missing all profile information. Also, we are only interested in Rented for cash vs other groups, so we combine 1. Owned with 3.Occupied without payment into a non-Rented for Cash Group.

```
S5 = S5[S5$profile_homeown!="-1. Inapplicable" &
          !is.na(S5$profile_homeown),]
S5$profile_homeown=droplevels(S5$profile_homeown)
S5$rent = ifelse(S5$profile_homeown=="2. Rented for cash",1,0)
table(S5$Party)
```

```
##
##   Democrat Republican
##       1854       1428
```

```
(pDhat=mean(S5$rent[S5$Party=="Democrat"]))
```

```
## [1] 0.2993528
```

```
(pRhat=mean(S5$rent[S5$Party=="Republican"]))
```

```
## [1] 0.1442577
```

A z test will be appropriate to test the difference in proportions if its assumptions are met. It assumes (i) the data come from a random sample, and (ii) the sample is large enough that the normal distribution is a good approximation of the binomal outcome. (i) is met because we have a random sample. (ii) is also met because we have sample sizes of 1,854 Democrats and 1,428 Republicans while the estimated proportions indicate the distributions are not that skewed.

Now we perform our test. We use a two sided test because we do not know beforehand which group we expect to have a larger portion of Renters.

$$H_0 : \mu_R = \mu_D$$
$$H_a : \mu_R \neq \mu_D$$

```
# Z score
(Z = (pDhat-pRhat)/
   sqrt(mean(S5$rent)*(1-mean(S5$rent))*
          (1/length(S5$rent[S5$Party=="Democrat"])+1/
             length(S5$rent[S5$Party=="Republican"]))))
```

```
## [1] 10.43776
```

```
(p_value = 2*pnorm(Z,lower.tail=F))
```

```
## [1] 1.666922e-25
```

```
(Estimated_diff = pDhat-pRhat)
```

```
## [1] 0.155095
```

Given our extremely low p value, this test has a significant result. We reject the null hypothesis that the portion of Republicans who rent for cash is equal to the portion of Democrats.

In terms of practical significance, the portion of Democrats that rent for cash (29.9%) is more than double the portion of Republicans (14.4%). The portion of Democrats that rent for cash is 15.5% (in absolute terms) greater than the portion of Republicans.

Conclusion and Takeaways

In this report we addressed 5 different questions using statistical tests. The table below summarizes the questions, the tests used to address them, the p-value signficance of the tests, and the practical significance of the tests.

Our results indicate that

1. There is insignificant evidence to reject the claim that voters didn't not change their political stance.
2. There is support for that claim that Republican voters were not the same age on average as Democrat voters in 2012. However, our sample of Republican voters was only 2.5 years older on average and the averages were centered near 50 years.
3. There is not significant evidence to reject the claim that Repulibcan voters in 2012 were on average 51 years old.
4. There is support for the claim that Republican voters in 2012 had a different level of political stance shift than Democrats. However, the shifts in political stance were very small for both groups.
5. There is support for the claim that a larger portion of Democrat Voters in 2012 rented for cash than of Republicans.

Table 1: Report Summary

Question	Test	PValue	Practical_Significance
1.Did voters become more liberal or more conservative during the 2012 election?	Wilcoxon Signed Rank Test comparing voters' dependent pre and post election political stance	.17	r=.01, 18% more liberal, 17% more conservative (no effect)
2.Were Republican voters older or younger than Democrat voters in 2012?	t-test comparing independent mean age of voters by party	<.001**	Mean difference in age of about 2.5 years
3.Were Republican voters older than 51, on the average in 2012?	Single sample t-test using Republican voters' age	.19	Very little (sample mean extremely close to 51)
4.Were Republican voters more likely to shift their political stance right or left than Democrats	t-test comparing independent mean stance shift by party (after coding ordinal political stance 1:7)	.02*	Republicans only move 1/20th of a level more conservative than democrats
5.Do a larger or smaller portion of Republicans Voters rent their homes for cash than Democrats?	z-test comparing proportions of voters that rent for cash by party	<.001**	The portion of Democrats that rent is more than double (15.5 absolute percentage points greater) than the portion of Republicans who rent

* Statistically Significant

** Very Statistically Signficant

Though our analyses may be impacted in some regards by the fact that we needed to remove negative (e.g. incomplete, invalid) values and operationalize concepts (e.g., how to define "Republican" and a "shift in political stance"), overall we have learned a lot applying practical solutions with the data we have.