Lab 3: Hypothesis Tests about the Mean.

w203: Statistics for Data Science

Tako Hisada

11/19/2017

Introduction

In this assignment, we are analyzing a small subset of the 2012 ANES survey in understanding voter behaviors before and after the US presidential election in 2012.

The Data

The American National Election Studies (ANES) conducts surveys of voters in the United States before and after every presidential election. The survey weights which are usually assigned to observations found in the ANES survey data have been removed and we are assuming that the data in the ANES_2012_sel.csv is a random sample from the voting population.

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

Analysis

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

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

```
summary(factor(substr(S$libcpre_self, 1, 17)))
```

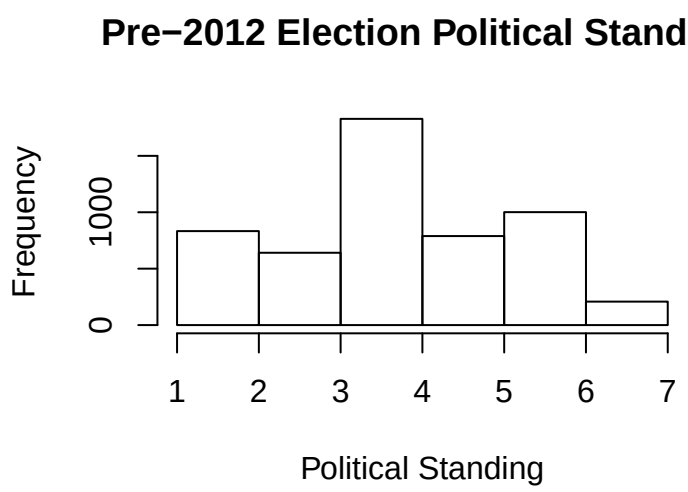
```
## -2. Haven't thoug    -8. Don't know       -9. Refused 1. Extremely libe
##               556                26                32               195
##        2. Liberal 3. Slightly liber 4. Moderate; midd 5. Slightly conse
##               638               641              1828               789
##   6. Conservative 7. Extremely cons
##              1001               208
```

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.

```
compilePoliticalWeights <- function(v, offset, skip_null = TRUE) {
  if(skip_null)
    v = v[grepl("^(\\d{1})", v)]
  as.numeric(v)-offset
}

libcpre_ranks <- compilePoliticalWeights(S$libcpre_self, 3)
```

```r
mu_libcpre <- mean(libcpre_ranks)
bins_standing <- seq(from = min(libcpre_ranks), to = max(libcpre_ranks), by = 1)
hist(libcpre_ranks, main = "Pre-2012 Election Political Standing",
     xlab = "Political Standing", breaks = bins_standing)
```



The pre-2012 election mean is 4.1722642.

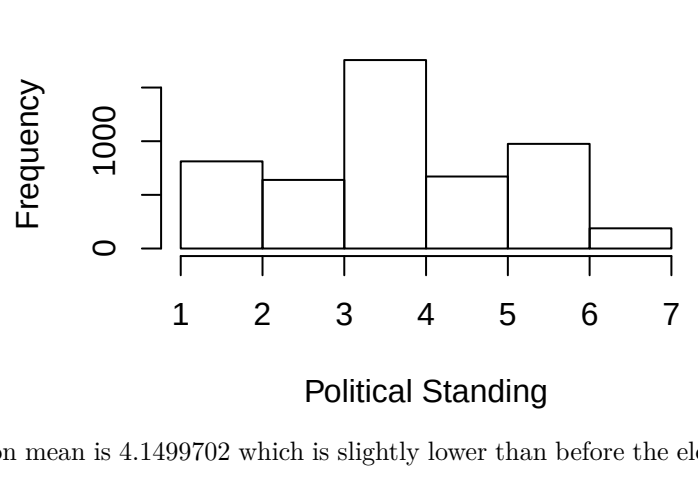
Likewise, we will need to perform the same conversion for the post-election data.

```r
summary(factor(substr(S$libcpo_self, 1, 17)))
```

```
## -2. Haven't thoug -6. Not asked, un -7. Deleted due t    -8. Don't know
##               410               252               152                23
##       -9. Refused 1. Extremely libe        2. Liberal 3. Slightly liber
##                36               166               646               639
## 4. Moderate; midd 5. Slightly conse   6. Conservative 7. Extremely cons
##              1756               671               975               188
```

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

```r
libcpo_ranks <- compilePoliticalWeights(S$libcpo_self, 5)
mu_libcpo = mean(libcpo_ranks)
hist(libcpo_ranks, main = "Post-2012 Election Political Standing",
     xlab = "Political Standing", breaks = bins_standing)
```



Our post 2012 election mean is 4.1499702 which is slightly lower than before the election 4.1722642 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

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

Statistical Significance

Justificaition

Our data is ordinal. Therefore we should conduct a Wilcoxon test in determining if the post-election mean is meaningfully different from the pre-election one.

```
(rs <- wilcox.test(libcpre_ranks, libcpo_ranks, alternative = "two.sided"))
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  libcpre_ranks and libcpo_ranks
## W = 13505000, p-value = 0.3219
## alternative hypothesis: true location shift is not equal to 0
```

Result

Our p-value is 0.321896. We fail to reject H_0. We are unable to say if people became more liberal nor conservative after the election.

Practical Significance

We are now going to compute the correlation r to measure the effect size.

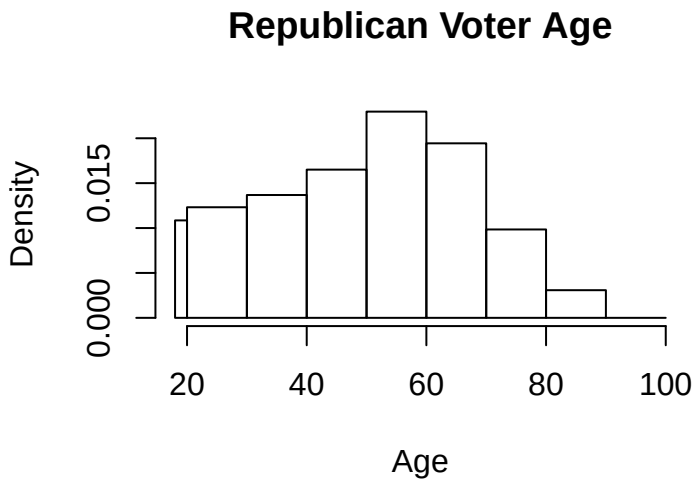
```
calcEffectSize <- function(stat, n) {
  r = qnorm(stat)/sqrt(n)
}

n = length(libcpre_ranks)+length(libcpo_ranks)
(r = calcEffectSize(rs$p.value, n))
```

```
## [1] -0.004547155
```

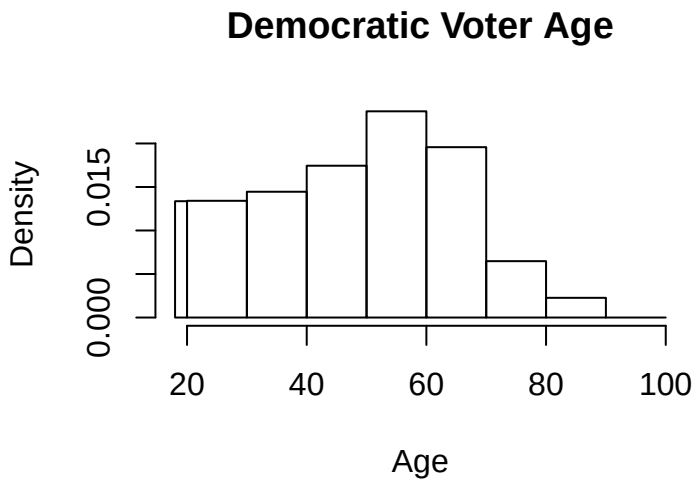
The r correlation for this case is -0.0045472 which signifies a small negative correlation.

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
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)
bins_ages <- c(18, seq(from = 20, to = 100, by = 10))
hist(rep_ages, main = "Republican Voter Age", xlab = "Age", breaks = bins_ages)
```




```r
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 slightly higher than 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's are sufficiently high in this case: 1981 for Republicans and 2207 for Democrats. The distributions are fairly normal as seen above. 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. We can say with confidence that the average Republican voter age was higher than that of the Democratic voters.

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(rep_ages, dem_ages))
```

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

The Cohen's d value for this case is 0.0060013. There is very small practical significance observed.

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. We are unable to say Republican voters were older than 51 on average.

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)
```

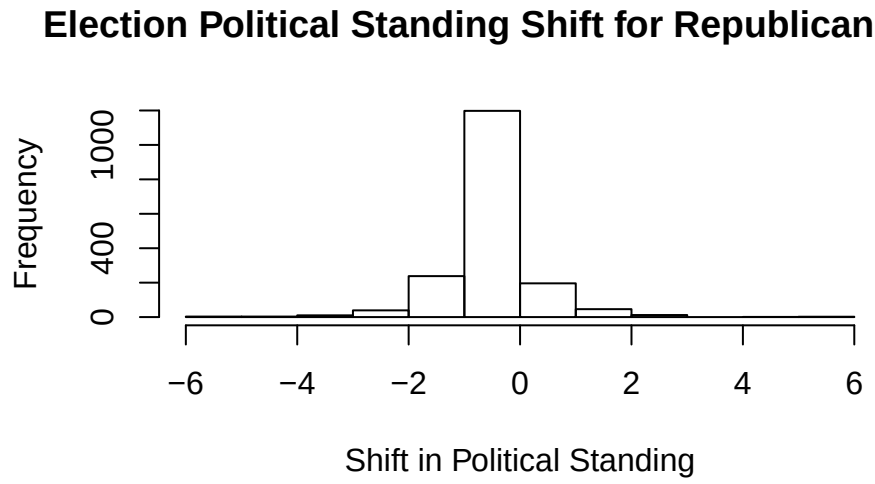
The effect size correlation r for this case is 0.0196967 which is < 0.1. The practical significance is very small.

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 change for Republican voters.

```
calcDiff <- function(v1, v2, skip_null = TRUE) {
  if(length(v1) != length(v2)) {
    FALSE
  }
  v = c()
  for(i in 1:length(v1)) {
    if((v1[i] < 1 || v2[i] < 1) && skip_null) {
      next
    }
    v = c(v, v1[i]-v2[i])
  }
  v
}

libcpre_rep_ranks <- compilePoliticalWeights(
  S$libcpre_self[grep("republican", S$pid_x, ignore.case = T)], 3, FALSE)
mu_libcpre_rep = mean(libcpre_rep_ranks)
libcpo_rep_ranks <-compilePoliticalWeights(
  S$libcpo_self[grep("republican", S$pid_x, ignore.case = T)], 5, FALSE)
mu_libcpo_rep = mean(libcpo_rep_ranks)
lib_diff_rep = calcDiff(libcpre_rep_ranks, libcpo_rep_ranks)
mu_lib_diff_rep = mean(lib_diff_rep)
```

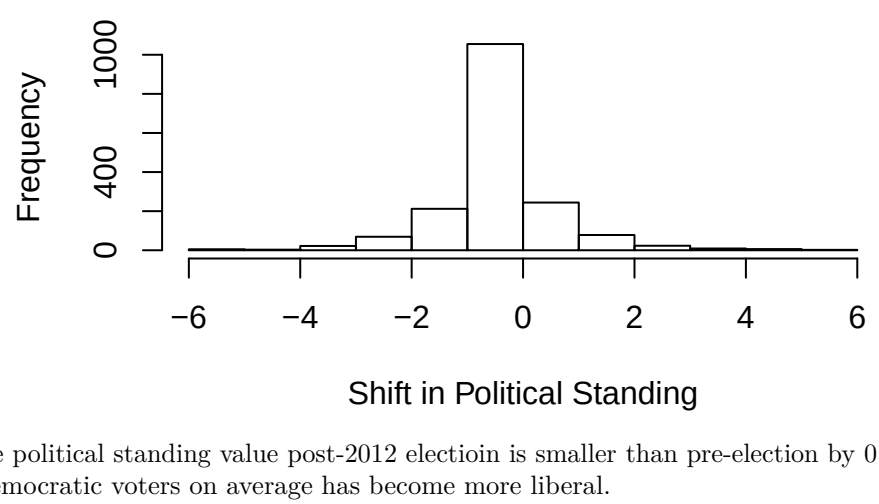
```r
bins_standing <- seq(from = min(lib_diff_rep), to = max(lib_diff_rep), by = 1)
hist(lib_diff_rep,
     main = "Election Political Standing Shift for Republican Voters",
     xlab = "Shift in Political Standing", breaks = bins_standing)
```



The average political standing value post-2012 electioin is greater than pre-election by -0.0188787. This means that the Republican voters on average has become more conservative.

Now let's do the same for Democratic voters.

```r
libcpre_dem_ranks <-compilePoliticalWeights(
  S$libcpre_self[grep("democrat", S$pid_x, ignore.case = T)], 3, FALSE)
mu_libcpre_dem <- mean(libcpre_dem_ranks)
libcpo_dem_ranks <-compilePoliticalWeights(
  S$libcpo_self[grep("democrat", S$pid_x, ignore.case = T)], 5, FALSE)
mu_libcpo_dem <- mean(libcpo_dem_ranks)
lib_diff_dem = calcDiff(libcpre_dem_ranks, libcpo_dem_ranks)
mu_lib_diff_dem = mean(lib_diff_dem)
hist(lib_diff_dem,
     main = "Election Political Standing Shift for Democratic Voters",
     xlab = "Shift in Political Standing", breaks = bins_standing)
```



The average political standing value post-2012 electioin is smaller than pre-election by 0.052662. This means that the Democratic voters on average has become more liberal.

Hypotheses

From this, we can formulate our hypotheses as below:

μ_1 = Average change in political standing for Republican voters
μ_2 = Average change in political standing for Democratic voters

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

Statistical Significance

Justificaition

As we are dealing with ordinal data, we will use the nonpametric Wilcoxon rank-sum test as we did for Q1.

```
rs <- t.test(lib_diff_rep, lib_diff_dem, alternative = "two.sided")
p <- rs$p.value
```

Result

Our p-value is 0.0288 We can reject H_0 at a significance level of 0.05, however not at 0.01. From this, we can say that the Republican voters were more likely to shift their political standing after the election than the Democratic voters.

Practical Significance

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

```
(d = calcCohensD(lib_diff_dem, lib_diff_rep))
```

```
## [1] 0.07721797
```

The Cohen's d value for this case is 0.077218. There is very small practical significance observed.

5. Were flyover state voters less likely to approve Barack Obama's job as President compared to coastal state voters?

Approval rating of Barack Obama as President amongst Republican voters

```
summary(S$presapp_job_x)
```

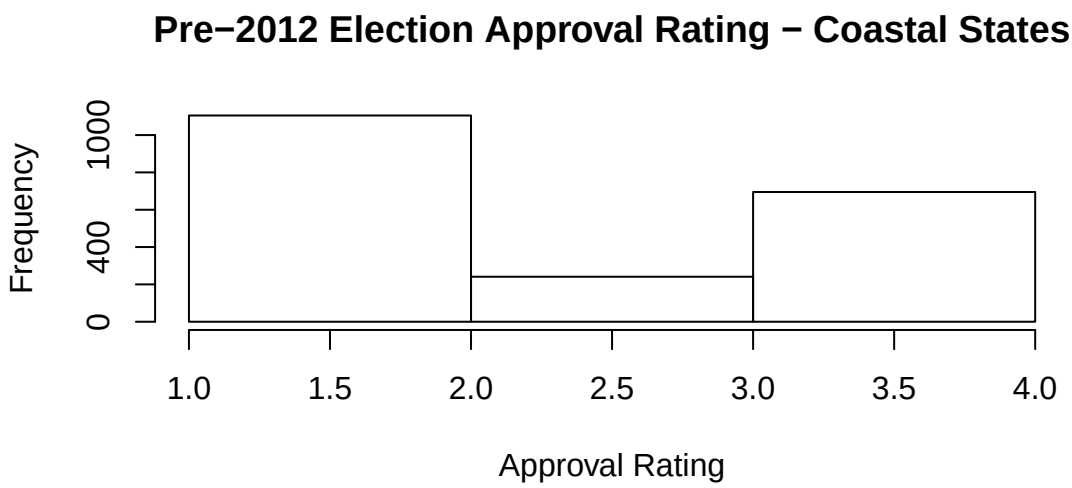
```
##             -8. Don't know                -9. Refused
##                         93                         44
##        1. Approve strongly    2. Approve not strongly
##                       2179                       1103
## 4. Disapprove not strongly     5. Disapprove strongly
##                        606                       1889
```

```
approval_rate = 100*length(S$presapp_job_x[grep('Approve', S$presapp_job_x)])/length(S$presapp_job_x)
```

Overall, it appears there were more people approved Barack Obama's job as President before the election in 2012 at the approval rate of 55%. Now we will break this down by which region the voters reside.

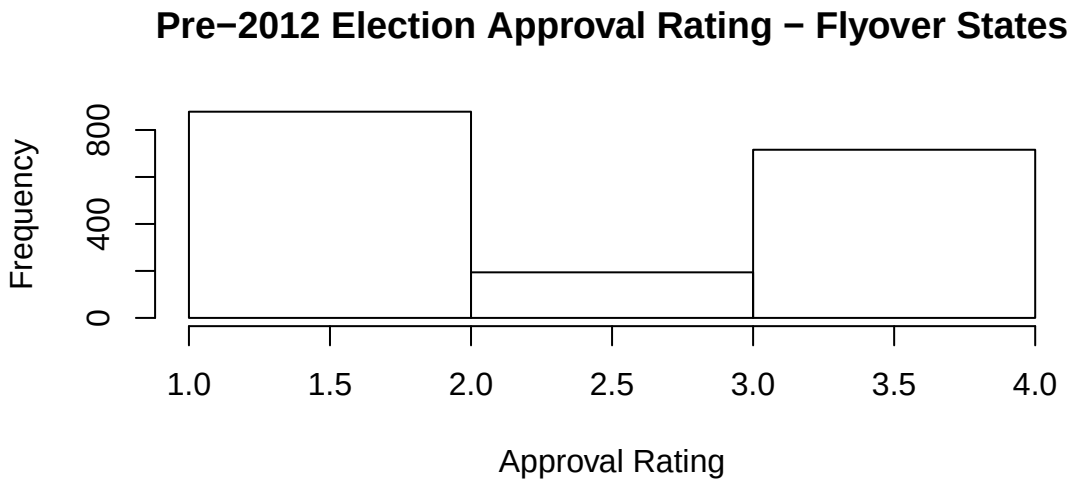
We'll first look at the coastal states voters.

```r
presapp_job_x_coast <- compilePoliticalWeights(
  S$presapp_job_x[grep(
    "New England|atlantic|Pacific", S$profile_region9, ignore.case = T)], 2)
mu_presapp_job_x_coast <- mean(presapp_job_x_coast)
bins_standing <- seq(from = min(presapp_job_x_coast), to = max(presapp_job_x_coast), by = 1)
hist(presapp_job_x_coast, main = "Pre-2012 Election Approval Rating - Coastal States",
     xlab = "Approval Rating", breaks = bins_standing)
```



Now, we'll compile the approval rate for the flyover state voters.

```r
presapp_job_x_flyover <- compilePoliticalWeights(
  S$presapp_job_x[grep("central|Mountain", S$profile_region9, ignore.case = T)], 2)
mu_presapp_job_x_flyover <- mean(presapp_job_x_flyover)
hist(presapp_job_x_flyover, main = "Pre-2012 Election Approval Rating - Flyover States",
     xlab = "Approval Rating", breaks = bins_standing)
```



Hypotheses

From this, we can formulate our hypotheses as below:

μ_1 = Pre-2012 election approval rate of Barack Obama as President amongst flyover-state voters
μ_2 = Pre-2012 election approval rate of Barack Obama as President amongst coastal-state voters

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

Statistical Significance

Justificaition

Our data is ordinal. Therefore we should conduct a Wilcoxon test in determining if the means are different.

```
(rs <- wilcox.test(presapp_job_x_flyover, presapp_job_x_coast, alternative = "greater"))
```

```
##
##  Wilcoxon rank sum test with continuity correction
##
## data:  presapp_job_x_flyover and presapp_job_x_coast
## W = 1925700, p-value = 0.0009538
## alternative hypothesis: true location shift is greater than 0
```

Result

Our p-value is 0.000954. We can reject H_0 at a significance level of 0.05 and 0.01. We can say with confidence that the flyover state voters were less likley to approve Barack Obama's job as President.

Practical Significance

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

```
n = length(presapp_job_x_flyover) + length(presapp_job_x_coast)
(r = calcEffectSize(rs$p.value, n))
```

```
## [1] -0.05016673
```

The correlation r for this case is -0.0501667 which indicates a small negative effect size. There is no practical significance observed.

Conclusion

When we look at overall numbers, Barack Obama had a positive approval rating of 55% before the election. However, looking further, we see a noticeable discrepancy between regions in terms of approval for his job as President as we saw in Q5. It is also notable from Q1 and Q4 that even after Obama's victory, people for the most part remained divided and stuck with their political views. The analysis reveals a very divided view of the voters across the United States in terms of age (Q2 and Q3), political standings and where they live (Q5) which we saw further-accelerated and exploded in the subsequent presidential election in 2016.