Political Text Messaging Analysis

Divya Parmar

April 20, 2018

Setup

Clear workspace before starting.

```
rm(list = ls())
```

Install necessary packages.

```
install.packages('dplyr')
library(dplyr)
library(ggplot2)
install.packages('plotrix')
```

Create t-test function to use later on.

```
t.test2 <- function(m1,m2,s1,s2,n1,n2,m0=0,equal.variance=FALSE)
{
  if( equal.variance==FALSE )
  {
    se <- sqrt( (s1^2/n1) + (s2^2/n2) )
    # welch-satterthwaite df
    df <- ( (s1^2/n1 + s2^2/n2)^2 )/( (s1^2/n1)^2/(n1-1) + (s2^2/n2)^2/(n2-1) )
  } else
  {
    # pooled standard deviation, scaled by the sample sizes
    se <- sqrt( (1/n1 + 1/n2) * ((n1-1)*s1^2 + (n2-1)*s2^2)/(n1+n2-2) )
    df <- n1+n2-2
  }
  t <- (m1-m2-m0)/se
  dat <- c(m1-m2, se, t, 2*pt(-abs(t),df))
  names(dat) <- c("Difference of means", "Std Error", "t", "p-value")
  return(dat)
}
```

Exploration

Read in data files.

```
rand_data <- read.csv(file.choose())
sms_data <- read.csv(file.choose())
survey_data <- read.csv(file.choose())
turnout_data <- read.csv(file.choose())
```

Look at data to understand columns.

```
head(rand_data)
```

```
##                              ai_id phone_number first_name     last_name
## 1 f4129a9409fa8c4428f0eefa1ab326d6   4102232802    Randall         Brown
## 2 e19d6a2e31912b7f68c262c7e7c7af4a   4105795833   Mitchell      Yarberry
## 3 eb8ab04d155ce929a6498647e0b4ea42   4106909331   Abhisaar           Rae
## 4 ac56cddc5254d08cbbbff7cfcf58cb3d   4109093490       Jake Pagni-Mugford
## 5 585fdff74a32e9c3180cd9c6c1ac46bb   4108097041    Tiffany       Ornelas
## 6 34373831c557e58f4ca9435fd43518c5   4102330789      James         Birks
##       race age gender marital_status sms_treat
## 1    white  62   male        married Treatment
## 2    white  55   male        married Treatment
## 3    asian  34   male      separated   Control
## 4    white  45   male      separated   Control
## 5 hispanic  41 female        married Treatment
## 6    white  49   male        married Treatment
```

```
head(sms_data)
```

```
##   phone_number message_direction
## 1   4102232802          outbound
## 2   4105795833          outbound
## 3   4108097041          outbound
## 4   4102330789          outbound
## 5   4103853827          outbound
## 6   4109477711          outbound
##
                          message_text
## 1 Hey this is Daniel from VoteReminder, wanted to remind you to vote this Tuesday! How do you
plan to get to your polling location?
## 2 Hey this is Daniel from VoteReminder, wanted to remind you to vote this Tuesday! How do you
plan to get to your polling location?
## 3 Hey this is Daniel from VoteReminder, wanted to remind you to vote this Tuesday! How do you
plan to get to your polling location?
## 4 Hey this is Daniel from VoteReminder, wanted to remind you to vote this Tuesday! How do you
plan to get to your polling location?
## 5 Hey this is Daniel from VoteReminder, wanted to remind you to vote this Tuesday! How do you
plan to get to your polling location?
## 6 Hey this is Daniel from VoteReminder, wanted to remind you to vote this Tuesday! How do you
plan to get to your polling location?
```

```
head(survey_data)
```

```
##   phone_number attempted attempts disposition support_smith
## 1   4102533400      TRUE        1           1             0
## 2   4103375604      TRUE        1           0            NA
## 3   4106196906      TRUE        2           1             0
## 4   4105982410      TRUE        1           1             1
## 5   4108920128      TRUE        2           1             0
## 6   4109957322      TRUE        1           1             1
##   environment_thermometer
## 1                       5
## 2                    <NA>
## 3                       7
## 4                       7
## 5                       4
## 6                       8
```

```
head(turnout_data)
```

```
##                              ai_id turnout2017
## 1 f4129a9409fa8c4428f0eefa1ab326d6           1
## 2 e19d6a2e31912b7f68c262c7e7c7af4a           1
## 3 eb8ab04d155ce929a6498647e0b4ea42           1
## 4 ac56cddc5254d08cbbbff7cfcf58cb3d           0
## 5 585fdff74a32e9c3180cd9c6c1ac46bb           0
## 6 34373831c557e58f4ca9435fd43518c5           1
```

Look at control/treatment breakdown for entire universe of voters.

- Treatment has 12,461 individuals

- Control has 12,539 individuals

- It looks like an evenly balanced A/B test.

```
table(rand_data$sms_treat)
```

```
## 
##   Control Treatment 
##     12539     12461
```

Look at texting data to make sure experiment was executed correctly.

- There are 12,461 voters in the treatment group.
- All of the phone numbers got a voting reminder.
- All of the phone numbers got a environment text message.
- This means the treatment is working as intended.

```
#How many unique phone numbers got sent text messages -> 12,461
outbound_messages = sms_data[sms_data$message_direction == 'outbound',]
length(unique(outbound_messages$phone_number))
```

```
## [1] 12461
```

```
#How many unique phone numbers got the voting reminder text message -> 12,461
#We can conclude that all individuals in the treatment group got reminder text messages as intended
reminder_messages = sms_data[sms_data$message_text == 'Hey this is Daniel from VoteReminder, wanted to remind you to vote this Tuesday! How do you plan to get to your polling location?',]
length(unique(outbound_messages$phone_number))
```

```
## [1] 12461
```

```
#We can use the grep function to see how many people got a message about the environment
#That number is 12,461 so it seems like everyone got it
env_messages = sms_data[grep("environment",sms_data$message_text),]
length(unique(env_messages$phone_number))
```

```
## [1] 12461
```

Combine treatment data with survey data to answer additional questions.

```
#Combine turnout data with treatment group data
df <- merge(rand_data, turnout_data, by = "ai_id")

#Add survey data to dataframe above
df <- merge(df, survey_data, by = "phone_number")
```

Did the messaging program boost turnout?

In looking at the data, it seems that the treatment group had higher turnout. And the difference is statistically significant.

```
#Create a crosstab of treatment group and turnout
turnout_breakout <- table(df$sms_treat,df$turnout2017)

#Look at crosstab for turnout
#Reorder columns so first column is voters and second column is non-voters
#From first look, it seems like treatment group had higher turnout
turnout_breakout <- turnout_breakout[,c(2,1)]
turnout_breakout
```

```
##
##                 1    0
##   Control   1358 2119
##   Treatment 1452 2043
```

```
#We must statistically test this by doing a proportions test on our crosstab table
#We find that the Treatment group has stastically higher turnout than the Control group with a p-value of 0.03
prop.test(turnout_breakout, correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  turnout_breakout
## X-squared = 4.4858, df = 1, p-value = 0.03418
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.047903604 -0.001864523
## sample estimates:
##    prop 1    prop 2
## 0.3905666 0.4154506
```

Did the messaging program convince voters to vote for Jane Smith?

In looking at the data, we find no statistical difference between the treatment and control groups. It seems as the messaging program did not convince voters to vote for Jane Smith.

```
#Create crosstab of treatment group and support for Smith
smith_support_breakout <- table(df$sms_treat,df$support_smith, exclude=NULL)
smith_support_breakout <- smith_support_breakout[,c(2,1)]
smith_support_breakout
```

```
##
##                1    0
##   Control   1442  945
##   Treatment 1445  879
```

```
#Do a proportion test to see if treatment group has statistically higher rate of support for Smith
#We find that there is no statistical difference in support for Smith between the two groups
prop.test(smith_support_breakout, correct=FALSE)
```

```
##
##  2-sample test for equality of proportions without continuity
##  correction
##
## data:  smith_support_breakout
## X-squared = 1.5491, df = 1, p-value = 0.2133
## alternative hypothesis: two.sided
## 95 percent confidence interval:
##  -0.04548119  0.01014672
## sample estimates:
##    prop 1    prop 2
## 0.6041056 0.6217728
```

Did the message program increase how much voters care about protecting the environment?

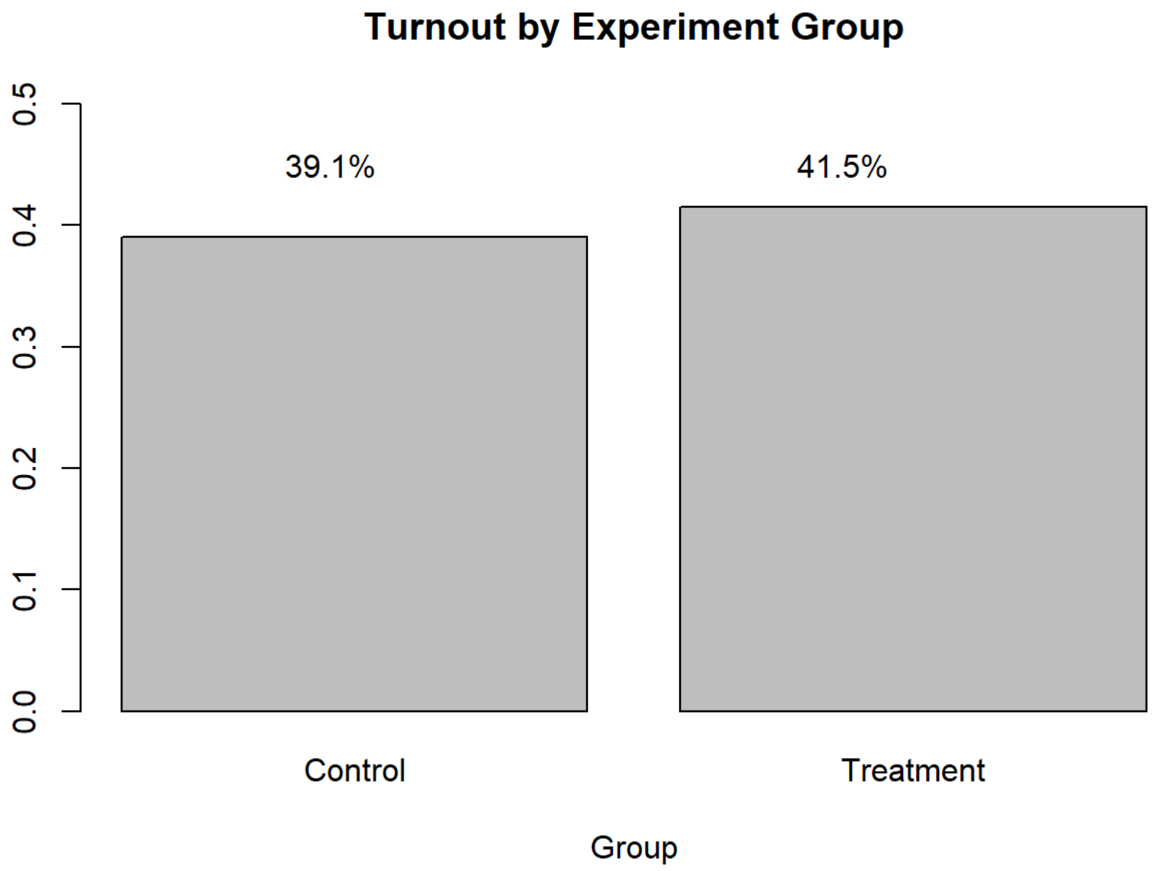
The treatment group has an higher average support for the environment, and it is statistically significant.

```
#Use dplyr package to calculate mean, standard deviation, and sample size for environmental prot
ection
#group_by(df[!is.na(df$environment_thermometer),], sms_treat) %>%
#  summarise(mean_value = mean(as.numeric(environment_thermometer)),standard_dev = sd(as.numeric
(environment_thermometer)), n_values = n(),
#            std_error = sd(as.numeric(environment_thermometer))/sqrt(n()))

# Mean, SD, N, SE
#1 Control         7.16         2.37     2471     .0477
#2 Treatment       7.33         2.51     2492     .0502

#Use our t-test function to see if the two means are different
#Our function turns a difference of means of .17 with a p-value of 0.015
#We can say that the Treatment group does care more about protecting the environment
t.test2(7.16, 7.33, 2.37, 2.51, 2471, 2492)
```

```
## Difference of means           Std Error                   t
##         -0.17000000          0.06929111         -2.45341709
##             p-value
##          0.01418486
```

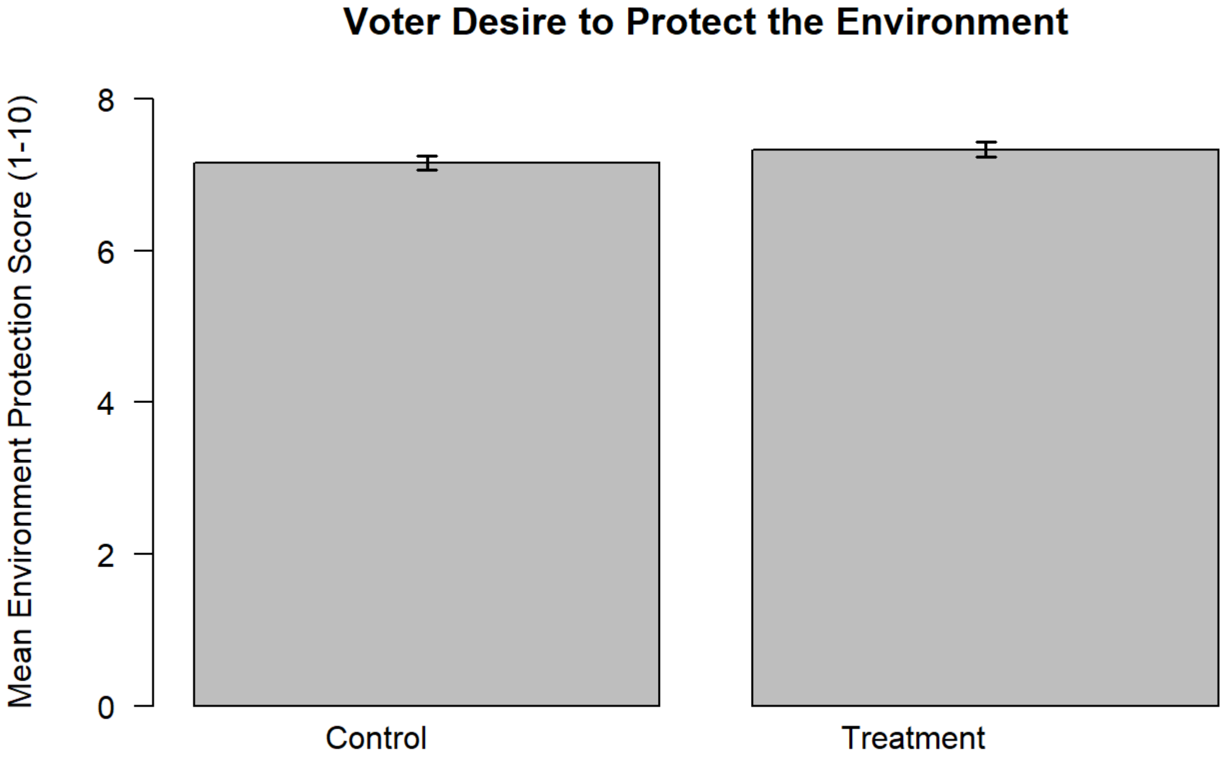
Graph of turnout by two groups.

```
#Graph for turnout
barplot(c(1358/(1358+2119),1452/(1452+2043)),main="Turnout by Experiment Group", xlab="Group",na
mes.arg=c("Control","Treatment"),
        ylim=c(0,0.5))
text(.65, 0.45, "39.1%")   #1358/(1358+2119)
text(1.75, 0.45, "41.5%")  #1452/(1452+2043)
```



Graph of environmental support by two groups.

```
#Graph for environment thermometer
#create dataframe specifically for this plot
names_vector <- c("Control","Treatment")
means_vector <- c(7.16,7.33)
sd_vector    <- c(2.37,2.51)
n_vector     <- c(2471, 2492)
env_df <- data.frame(names_vector,means_vector,sd_vector, n_vector)
env_df <- do.call(data.frame,env_df)

#Create base plot
env_plot <- barplot(height = env_df$means_vector,
                      beside = true, las = 2,
                      ylim = c(0, 8),
                      cex.names = 0.75, xaxt = "n",
                      main = "Voter Desire to Protect the Environment",
                      ylab = "Mean Environment Protection Score (1-10)",
                      border = "black", axes = TRUE)
#Manually add text labels
text(x = env_plot, y = par("usr")[3]-0.4, srt = 0,
     adj = 1, labels = env_df$names_vector, xpd = TRUE)

#Add confidence bands by manually calculating standard error using standard deviation and sample
 size
segments(env_plot, env_df$means_vector - env_df$sd_vector/sqrt(env_df$n_vector) * 1.96, env_plo
t,
         env_df$means_vector + env_df$sd_vector/sqrt(env_df$n_vector * 1.96), lwd = 1.5)
arrows(env_plot, env_df$means_vector - env_df$sd_vector/sqrt(env_df$n_vector) * 1.96, env_plot,
       env_df$means_vector + env_df$sd_vector/sqrt(env_df$n_vector) * 1.96, lwd = 1.5, angle = 9
0,
       code = 3, length = 0.05)
```


Voter Desire to Protect the Environment
Mean Environment Protection Score (1-10)
8
6
4
2
0
Control
Treatment