ImpactDataViz 241 Final Project

```r
# load packages
library(data.table)
library(foreign)
library(sandwich)
library(stargazer)
```

```
##
## Please cite as:

##  Hlavac, Marek (2018). stargazer: Well-Formatted Regression and Summary Statistics Tables.

##  R package version 5.2.2. https://CRAN.R-project.org/package=stargazer
```

```r
library(lmtest)
```

```
## Loading required package: zoo

## Warning: package 'zoo' was built under R version 3.6.2

##
## Attaching package: 'zoo'

## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

1. Parse the survey data into a data.table

```r
d <- fread("data/final.csv")
#d <- fread("data/pilot.csv")

# covariates
setnames(d, 'Q3', 'state')
setnames(d, 'Q4', 'gender')
setnames(d, 'Q5', 'age')
setnames(d, 'Q6', 'ethnicity_multi')
setnames(d, 'Q7', 'political_party')
setnames(d, 'Q8', 'education')
setnames(d, 'Q9', 'covid_sick')
setnames(d, 'Q10', 'covid_hospitalized')
setnames(d, 'Q11', 'covid_died')
```

```r
# duration of survey time
setnames(d, 'Duration (in seconds)', 'duration_of_survey')

# which block was active? (did the user see the treatment or
# control data viz)
setnames(d, 'Q15', 'treatment_viz_is_accurate')
setnames(d, 'Q17', 'control_viz_is_accurate')

# outcome questions about COVID attitudes
setnames(d, 'Q18', 'outcome_spread')
setnames(d, 'Q19', 'outcome_death')

# which block was active determines if
# subject received treatment data viz or control
# data viz
d[, treatment := ifelse(is.na(treatment_viz_is_accurate), 0, 1)]
d = d[!is.na(outcome_spread) & !is.na(outcome_death),]

# ethnicity allows for multiple choice
# for covariates, just grab the first one
ethnicity_single = rep(0,nrow(d))
i = 1
for (eth_entry in d[,ethnicity_multi]) {
  eth_tokens = unlist(strsplit(eth_entry, ","))
  ethnicity_single[i] = as.numeric(eth_tokens[1])
  i = i + 1
}
d[, ethnicity := ethnicity_single ]

# counts in control vs treatment
n_control = nrow(d[treatment == 0, ])
n_treatment = nrow(d[treatment == 1, ])
```

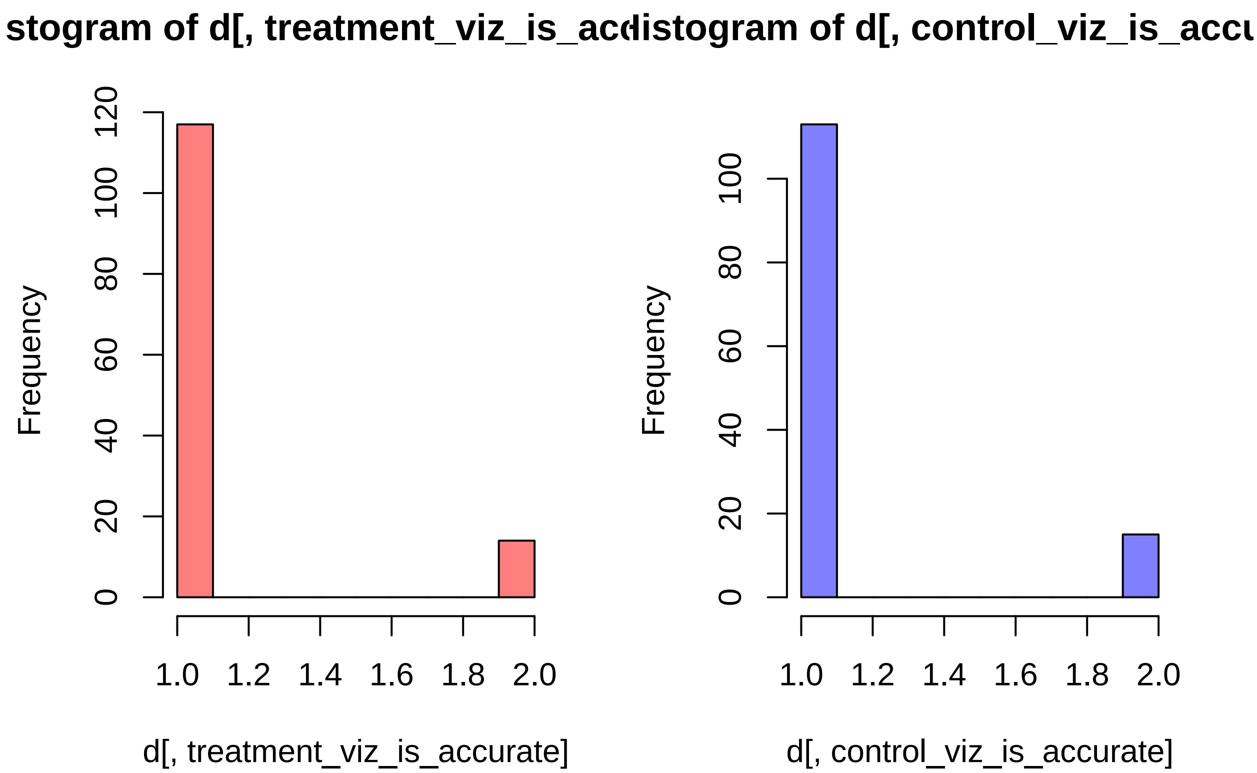
```r
# Multiple plot function
#
# ggplot objects can be passed in ..., or to plotlist (as a list of ggplot objects)
# - cols:   Number of columns in layout
# - layout: A matrix specifying the layout. If present, 'cols' is ignored.
#
# If the layout is something like matrix(c(1,2,3,3), nrow=2, byrow=TRUE),
# then plot 1 will go in the upper left, 2 will go in the upper right, and
# 3 will go all the way across the bottom.
#
multiplot <- function(..., plotlist=NULL, file, cols=1, layout=NULL) {
  library(grid)

  # Make a list from the ... arguments and plotlist
  plots <- c(list(...), plotlist)

  numPlots = length(plots)

  # If layout is NULL, then use 'cols' to determine layout
  if (is.null(layout)) {
```

```
    # Make the panel
    # ncol: Number of columns of plots
    # nrow: Number of rows needed, calculated from # of cols
    layout <- matrix(seq(1, cols * ceiling(numPlots/cols)),
                    ncol = cols, nrow = ceiling(numPlots/cols))
  }

 if (numPlots==1) {
    print(plots[[1]])

  } else {
    # Set up the page
    grid.newpage()
    pushViewport(viewport(layout = grid.layout(nrow(layout), ncol(layout))))

    # Make each plot, in the correct location
    for (i in 1:numPlots) {
      # Get the i,j matrix positions of the regions that contain this subplot
      matchidx <- as.data.frame(which(layout == i, arr.ind = TRUE))

      print(plots[[i]], vp = viewport(layout.pos.row = matchidx$row,
                                      layout.pos.col = matchidx$col))
    }
  }
}
```

2. EDA

2.1 Are the data visualizations in control and treatment considered accurate by subjects?

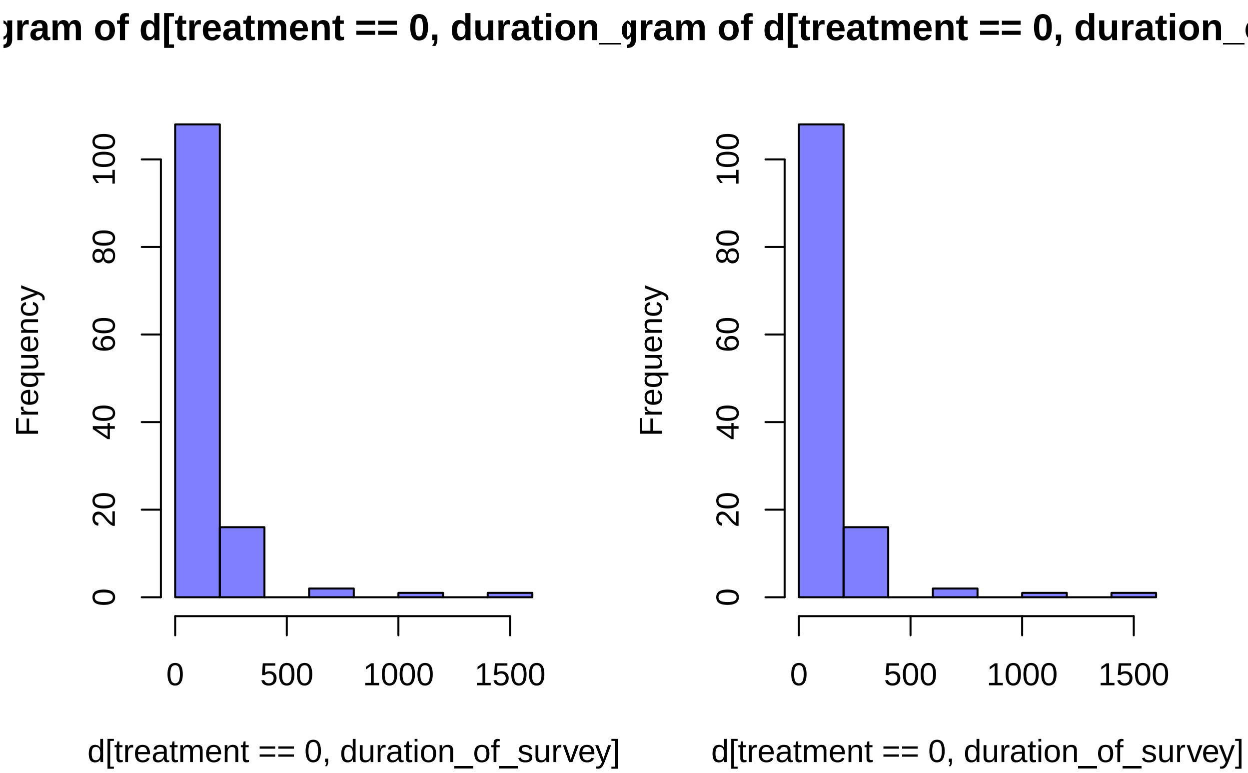
```
par(mfrow=c(1,2))
hist(d[, treatment_viz_is_accurate], col=rgb(1,0,0,0.5), xlim=c(1,2))
hist(d[, control_viz_is_accurate],   col=rgb(0,0,1,0.5), xlim=c(1,2))
mtext("I trust that the information presented in the data visualization is accurate", side=1, outer=TRUI
```



I trust that the information presented in the data visualization is accurate

2.2 Check Time

```
par(mfrow=c(1,2))
hist(d[treatment == 0,duration_of_survey], col=rgb(0,0,1,0.5))
hist(d[treatment == 0,duration_of_survey], col=rgb(0,0,1,0.5))
```




```r
mean(d[treatment == 0,duration_of_survey])
```

```
## [1] 153.1875
```

```r
mean(d[treatment == 0,duration_of_survey])
```

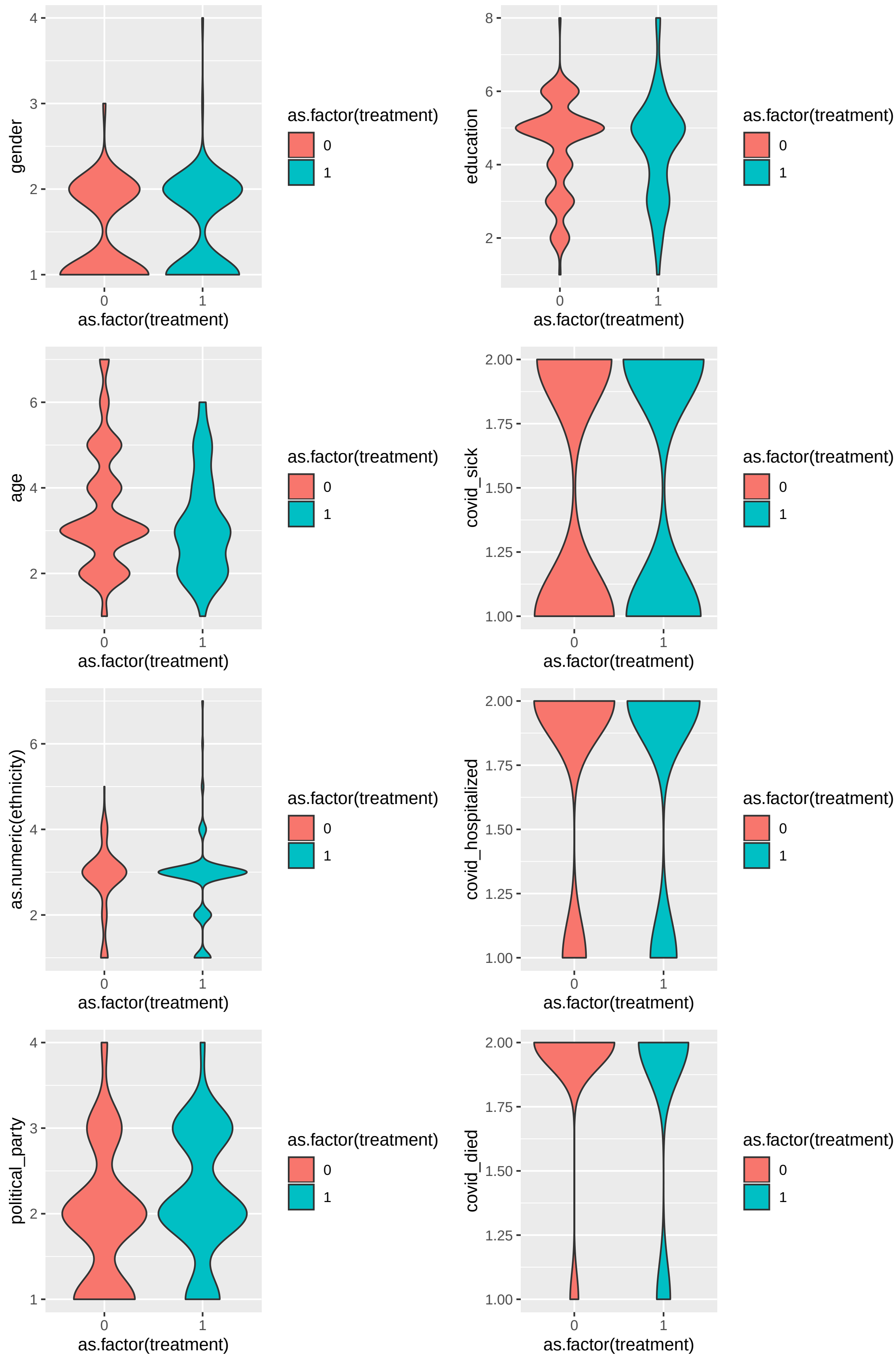
```
## [1] 153.1875
```

3. Covariate balance

3.1 Compare distributions using violin plote

```r
library('ggplot2')

options(repr.plot.width = 14, repr.plot.height = 8)

p1 = ggplot(d, aes(x=as.factor(treatment), y=gender, fill=as.factor(treatment))) +
  geom_violin()

p2 = ggplot(d, aes(x=as.factor(treatment), y=age, fill=as.factor(treatment))) +
  geom_violin()
```

```r
p3 = ggplot(d, aes(x=as.factor(treatment), y=as.numeric(ethnicity),  fill=as.factor(treatment))) +
  geom_violin()

p4 = ggplot(d, aes(x=as.factor(treatment), y=political_party,  fill=as.factor(treatment))) +
  geom_violin()

p5 = ggplot(d, aes(x=as.factor(treatment), y=education,  fill=as.factor(treatment))) +
  geom_violin()

p6 = ggplot(d, aes(x=as.factor(treatment), y=covid_sick,  fill=as.factor(treatment))) +
  geom_violin()

p7 = ggplot(d, aes(x=as.factor(treatment), y=covid_hospitalized,  fill=as.factor(treatment))) +
  geom_violin()

p8 = ggplot(d, aes(x=as.factor(treatment), y=covid_died, fill=as.factor(treatment))) +
  geom_violin()

multiplot(p1, p2, p3, p4, p5, p6, p7, p8, cols=2)
```


gender
as.factor(treatment)
0
1
education
age
covid_sick
as.numeric(ethnicity)
covid_hospitalized
political_party
covid_died
4
3
2
1
8
6
6
4
2
2.00
1.75
1.50
1.25
1.00

4. Estimate ATE

```
estimate_ate <- function(dt, outcome, treatment, treat_val) {
  ## This takes a data.table, the name of the outcome variable, and the name
  ## of the treatment indicator.

  g <- dt[ , .(group_mean = mean(get(outcome))), keyby = .(get(treatment))]
  ate <- g[ , diff(group_mean)]

  return(ate)
}
ri <-function(num=10000){

  res <- NA
  for (i in 1:num) {
    res[i] <- d[ , .(group_mean = mean(views)), keyby = .(sample(success))][ , diff(group_mean)]
  }

  return (res)
}
```

4.1 Estimate the ATE for the outcome on how concerned subject is about COVID-19 continued spread

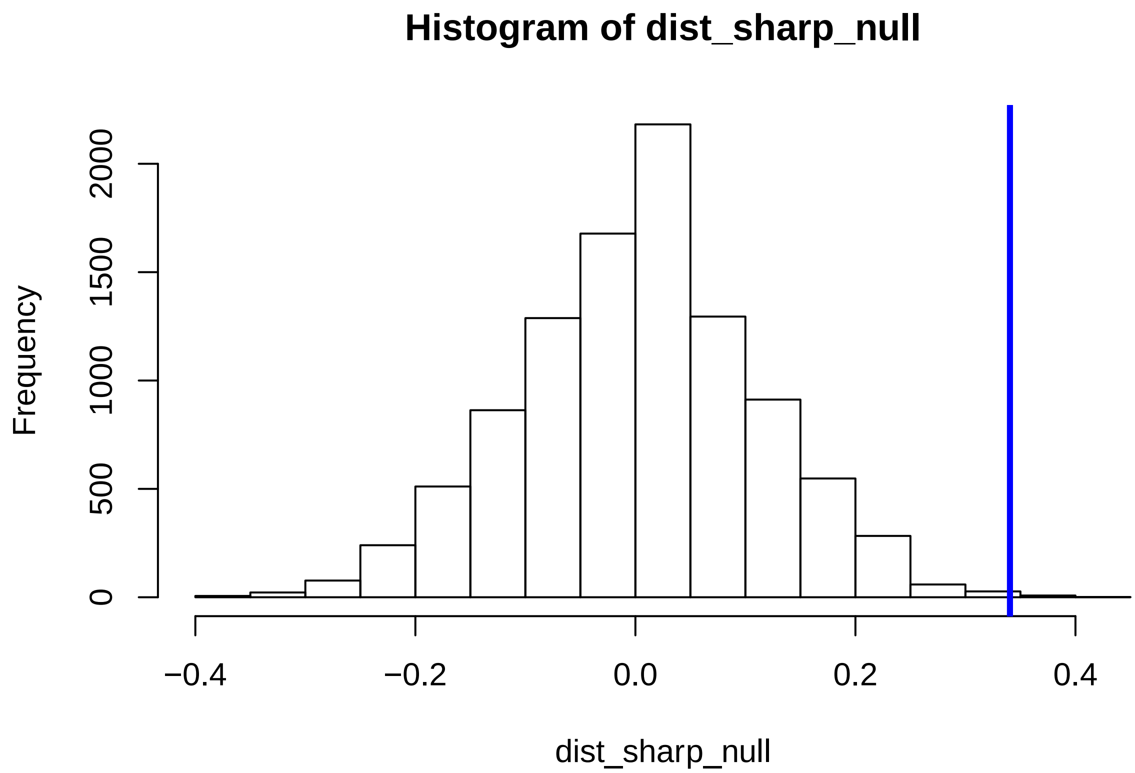
```
g <- d[ , .(group_mean = mean(outcome_spread)), keyby = .(treatment)]
g
```

```
##    treatment group_mean
## 1:         0   3.132812
## 2:         1   3.473282
```

```
ate_spread <- g[ , diff(group_mean)]
ate_spread
```

```
## [1] 0.3404699
```

```
res <- NA
for (i in 1:10000) {
    res[i] <- d[ , .(group_mean = mean(outcome_spread)), keyby = .(sample(treatment))][ , diff(group_mea
}
dist_sharp_null <- res
hist(dist_sharp_null)
abline(v=ate_spread, lwd=3, col='blue')
abline(v=abs(ate_spread), lwd=3, col='blue')
```




```r
p_value_one_tailed <- mean(dist_sharp_null >= ate_spread)
p_value_one_tailed
```

```
## [1] 0.0015
```

```r
p_value_two_tailed <- mean(abs(dist_sharp_null) >= abs(ate_spread))
p_value_two_tailed
```

```
## [1] 0.0021
```

4.2 Estimate the ATE for the outcome on emotional reaction to COVID-19 deaths in US

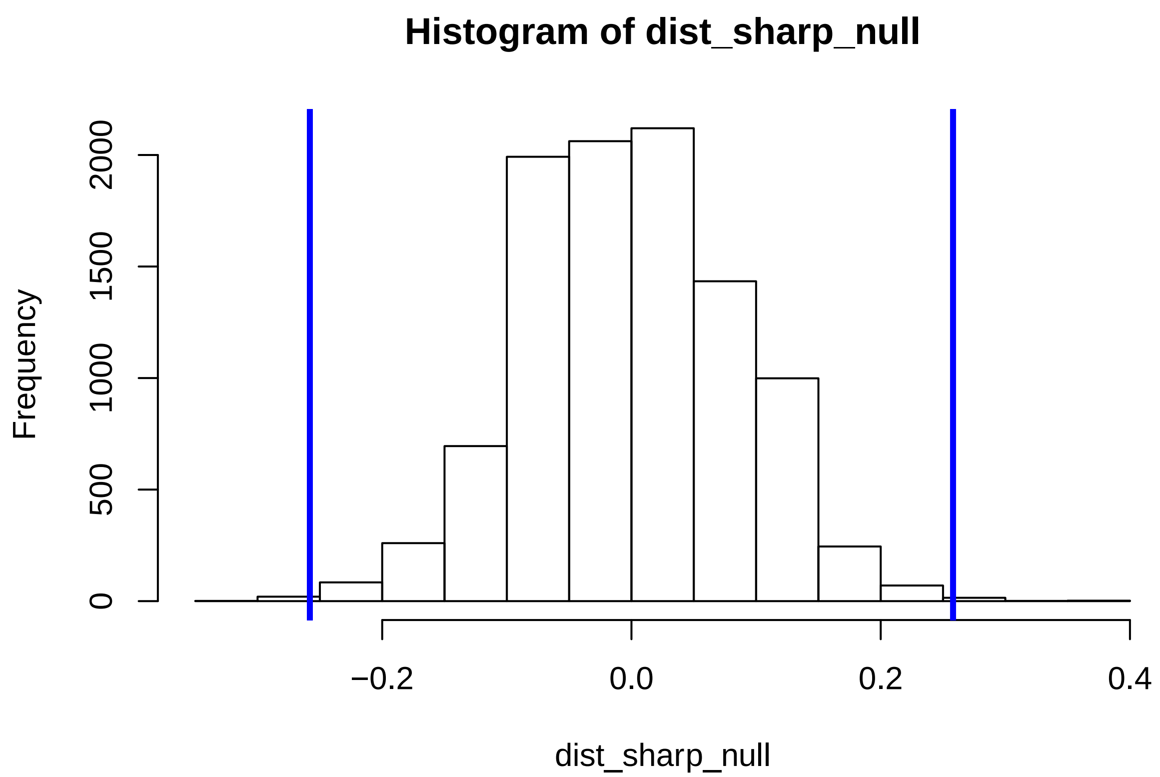
```r
g <- d[ , .(group_mean = mean(outcome_death)), keyby = .(treatment)]
g
```

```
##    treatment group_mean
## 1:         0   2.398438
## 2:         1   2.656489
```

```r
ate_death <- g[ , diff(group_mean)]
ate_death
```

```
## [1] 0.258051
```

```
res <- NA
for (i in 1:10000) {
    res[i] <- d[ , .(group_mean = mean(outcome_death)), keyby = .(sample(treatment))][ , diff(group_mea
}
dist_sharp_null <- res
hist(dist_sharp_null)
abline(v=ate_death, lwd=3, col='blue')
abline(v=ate_death*-1, lwd=3, col='blue')
```




```
p_value_one_tailed <- mean(dist_sharp_null >= ate_death)
p_value_one_tailed
```

```
## [1] 0.0018
```

```
p_value_two_tailed <- mean(abs(dist_sharp_null) >= abs(ate_death))
p_value_two_tailed
```

```
## [1] 0.003
```

5. Linear Regression

5.1 Perform linear regression on concern over COVID-19 spread outcome

```
model_spread = lm(outcome_spread ~ treatment, d)
model_spread1 = lm(outcome_spread ~ treatment
                   + as.factor(gender)
                   + as.factor(age)
                   + as.factor(ethnicity)
                   + as.factor(political_party)
                   + as.factor(education)
                   + as.factor(covid_sick)
                   + as.factor(covid_hospitalized)
                   + as.factor(covid_died), d)
stargazer(model_spread, model_spread1, type="text")
```

```
##
## ==============================================================================
##                                             Dependent variable:
##                                 ----------------------------------------------
##                                                outcome_spread
##                                          (1)                     (2)
## ------------------------------------------------------------------------------
## treatment                              0.340***                0.339***
##                                        (0.107)                 (0.104)
##
## as.factor(gender)2                                             -0.063
##                                                                (0.102)
##
## as.factor(gender)3                                             0.886*
##                                                                (0.508)
##
## as.factor(gender)4                                              0.646
##                                                                (0.795)
##
## as.factor(age)2                                                 0.323
##                                                                (0.347)
##
## as.factor(age)3                                                 0.295
##                                                                (0.344)
##
## as.factor(age)4                                                 0.181
##                                                                (0.362)
##
## as.factor(age)5                                                 0.370
##                                                                (0.364)
##
## as.factor(age)6                                                 0.393
##                                                                (0.425)
##
## as.factor(age)7                                                1.156**
##                                                                (0.494)
##
```

```
## as.factor(ethnicity)2                                     0.195
##                                                          (0.219)
##
## as.factor(ethnicity)3                                    -0.087
##                                                          (0.170)
##
## as.factor(ethnicity)4                                     0.230
##                                                          (0.233)
##
## as.factor(ethnicity)5                                    -0.471
##                                                          (0.508)
##
## as.factor(ethnicity)6                                   -2.519***
##                                                          (0.799)
##
## as.factor(ethnicity)7                                     0.715
##                                                          (0.824)
##
## as.factor(political_party)2                             0.675***
##                                                          (0.126)
##
## as.factor(political_party)3                              0.357**
##                                                          (0.149)
##
## as.factor(political_party)4                              -0.222
##                                                          (0.324)
##
## as.factor(education)2                                    -0.050
##                                                          (0.494)
##
## as.factor(education)3                                     0.194
##                                                          (0.483)
##
## as.factor(education)4                                     0.017
##                                                          (0.492)
##
## as.factor(education)5                                     0.188
##                                                          (0.477)
##
## as.factor(education)6                                     0.337
##                                                          (0.492)
##
## as.factor(education)7                                    -0.166
##                                                          (0.906)
##
## as.factor(education)8                                     0.502
##                                                          (0.569)
##
## as.factor(covid_sick)2                                   -0.200
##                                                          (0.121)
##
## as.factor(covid_hospitalized)2                           -0.166
##                                                          (0.165)
##
```

```
## as.factor(covid_died)2                                       0.092
##                                                             (0.181)
##
## Constant                          3.133***                  2.450***
##                                    (0.076)                  (0.634)
##
## ---------------------------------------------------------------------------------
## Observations                         259                       259
## R2                                  0.038                     0.314
## Adjusted R2                         0.034                     0.227
## Residual Std. Error           0.858 (df = 257)        0.768 (df = 229)
## F Statistic               10.183*** (df = 1; 257) 3.614*** (df = 29; 229)
## =================================================================================
## Note:                                                 *p<0.1; **p<0.05; ***p<0.01
```

5.2 Perform linear regression on reaction to COVID-19 deaths outcome

```r
model_death = lm(outcome_death ~ treatment, d)
model_death1 = lm(outcome_death ~ treatment
                  + as.factor(gender)
                  + as.factor(age)
                  + as.factor(ethnicity)
                  + as.factor(political_party)
                  + as.factor(education)
                  + as.factor(covid_sick)
                  + as.factor(covid_hospitalized)
                  + as.factor(covid_died),
                  d)

stargazer(model_death, model_death1, type="text")
```

```
##
## ================================================================================
##                                              Dependent variable:
##                                   ----------------------------------------------
##                                                  outcome_death
##                                          (1)                     (2)
## --------------------------------------------------------------------------------
## treatment                             0.258***                0.250***
##                                       (0.084)                 (0.085)
##
## as.factor(gender)2                                             0.050
##                                                               (0.084)
##
## as.factor(gender)3                                             0.573
##                                                               (0.415)
##
## as.factor(gender)4                                             0.437
##                                                               (0.649)
##
## as.factor(age)2                                               -0.371
##                                                               (0.283)
```

```
##
## as.factor(age)3                                      -0.323
##                                                     (0.281)
##
## as.factor(age)4                                      -0.345
##                                                     (0.296)
##
## as.factor(age)5                                      -0.258
##                                                     (0.297)
##
## as.factor(age)6                                      -0.279
##                                                     (0.348)
##
## as.factor(age)7                                       0.170
##                                                     (0.404)
##
## as.factor(ethnicity)2                                 0.041
##                                                     (0.179)
##
## as.factor(ethnicity)3                                -0.103
##                                                     (0.139)
##
## as.factor(ethnicity)4                                -0.019
##                                                     (0.190)
##
## as.factor(ethnicity)5                                -0.261
##                                                     (0.415)
##
## as.factor(ethnicity)6                               -1.667**
##                                                     (0.653)
##
## as.factor(ethnicity)7                                 0.374
##                                                     (0.674)
##
## as.factor(political_party)2                         0.543***
##                                                     (0.103)
##
## as.factor(political_party)3                         0.340***
##                                                     (0.122)
##
## as.factor(political_party)4                          -0.107
##                                                     (0.265)
##
## as.factor(education)2                                 0.233
##                                                     (0.404)
##
## as.factor(education)3                                 0.216
##                                                     (0.395)
##
## as.factor(education)4                                 0.135
##                                                     (0.402)
##
## as.factor(education)5                                 0.088
##                                                     (0.390)
```

```
##
## as.factor(education)6                                           0.263
##                                                                (0.402)
##
## as.factor(education)7                                          -0.475
##                                                                (0.741)
##
## as.factor(education)8                                           0.290
##                                                                (0.465)
##
## as.factor(covid_sick)2                                         -0.102
##                                                                (0.099)
##
## as.factor(covid_hospitalized)2                                 -0.084
##                                                                (0.135)
##
## as.factor(covid_died)2                                          0.067
##                                                                (0.148)
##
## Constant                              2.398***                 2.327***
##                                       (0.059)                  (0.518)
##
## ------------------------------------------------------------------------------
## Observations                            259                      259
## R2                                     0.036                    0.252
## Adjusted R2                            0.032                    0.157
## Residual Std. Error               0.672 (df = 257)        0.628 (df = 229)
## F Statistic                    9.533*** (df = 1; 257) 2.657*** (df = 29; 229)
## ==============================================================================
## Note:                                              *p<0.1; **p<0.05; ***p<0.01
```