Problem Set 2

Experiments and Causality

Adam Yang

1. What happens when pilgrims attend the Hajj pilgrimage to Mecca?

On the one hand, participating in a common task with a diverse group of pilgrims might lead to increased mutual regard through processes identified in *Contact Theories*. On the other hand, media narratives have raised the spectre that this might be accompanied by "antipathy toward non-Muslims". Clingingsmith, Khwaja and Kremer (2009) investigates the question.

Using the data here, test the sharp null hypothesis that winning the visa lottery for the pilgrimage to Mecca had no effect on the views of Pakistani Muslims toward people from other countries. Assume that the Pakistani authorities assigned visas using complete random assignment. Use, as your primary outcome the `views` variable, and as your treatment feature `success`. If you're ambitious, write your function generally so that you can also evaluate feelings toward specific nationalities.

```
d <- read.csv("./data/Clingingsmith.2009.csv", stringsAsFactors = FALSE)
```

a. Using either `dplyr` or `data.table`, group the data by `success` and report whether views toward others are generally more positive among lottery winners or lottery non-winners.

Adam Yang: First I would like to state my assumptions. I am assuming that success = 1 means that the person has won the visa lottery for the pilgrimage to Mecca and success = 0 means they have not won, therefore, `success` is the treatment feature. I am also assuming that each of the views columns indicate how the bias of this person towards people from that country. A positive number would equal a positive bias and the a negative number would equal a negative bias against people of that country. The `views` variable indicates the sum of these views.

```
by_success <- d %>% group_by(success)
by_success %>% summarize(avg_views = mean(views), sample_size = length(views))
```

```
## # A tibble: 2 x 3
##   success avg_views sample_size
##     <int>     <dbl>       <int>
## 1       0      1.87         448
## 2       1      2.34         510
```

Adam Yang: From the resulting table shown above, we can see that the average views towards others when success = 0 is **1.87** while the average views towards others when success = 1 is **2.34**. Therefore, views toward others are generally more positive among lottery winners.

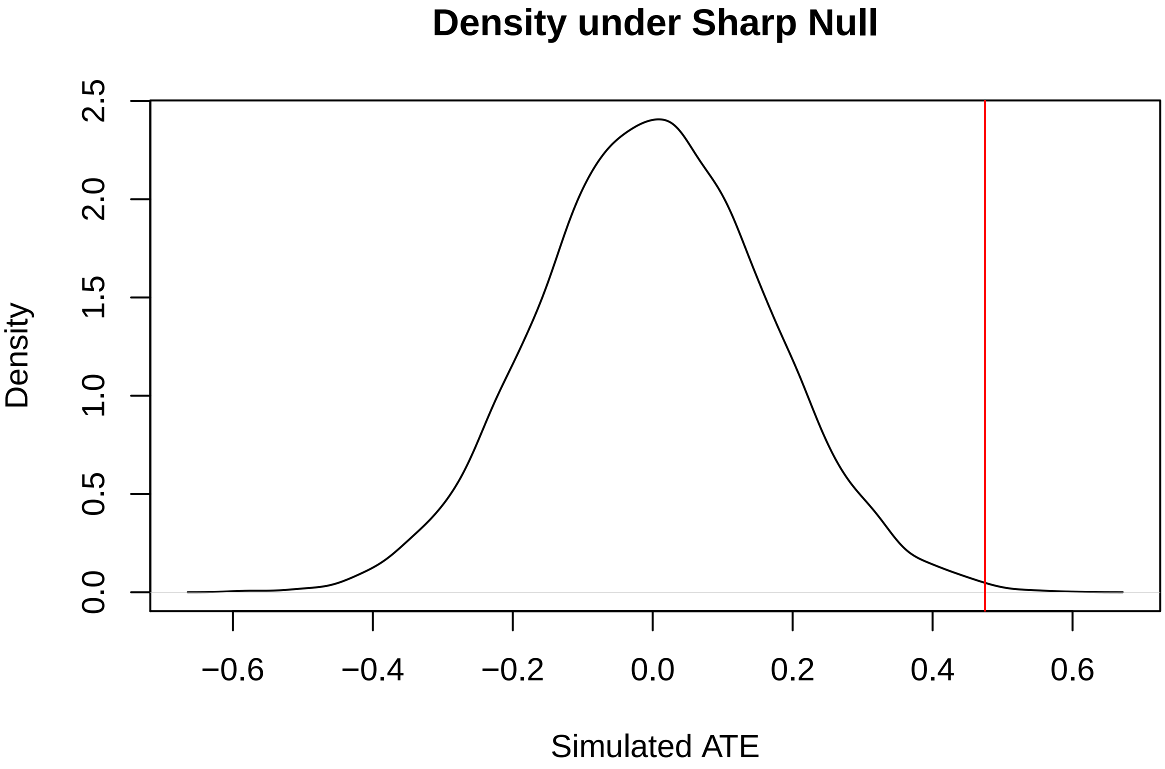
```
# Calculate the average treatment effect of the observed data
exp_ate <- mean(d[d$success == 1,]$views) - mean(d[d$success == 0,]$views)
exp_ate
```

```
## [1] 0.4748337
```

Adam Yang: The experimental estimated average treatment effect that we found is **0.47**.

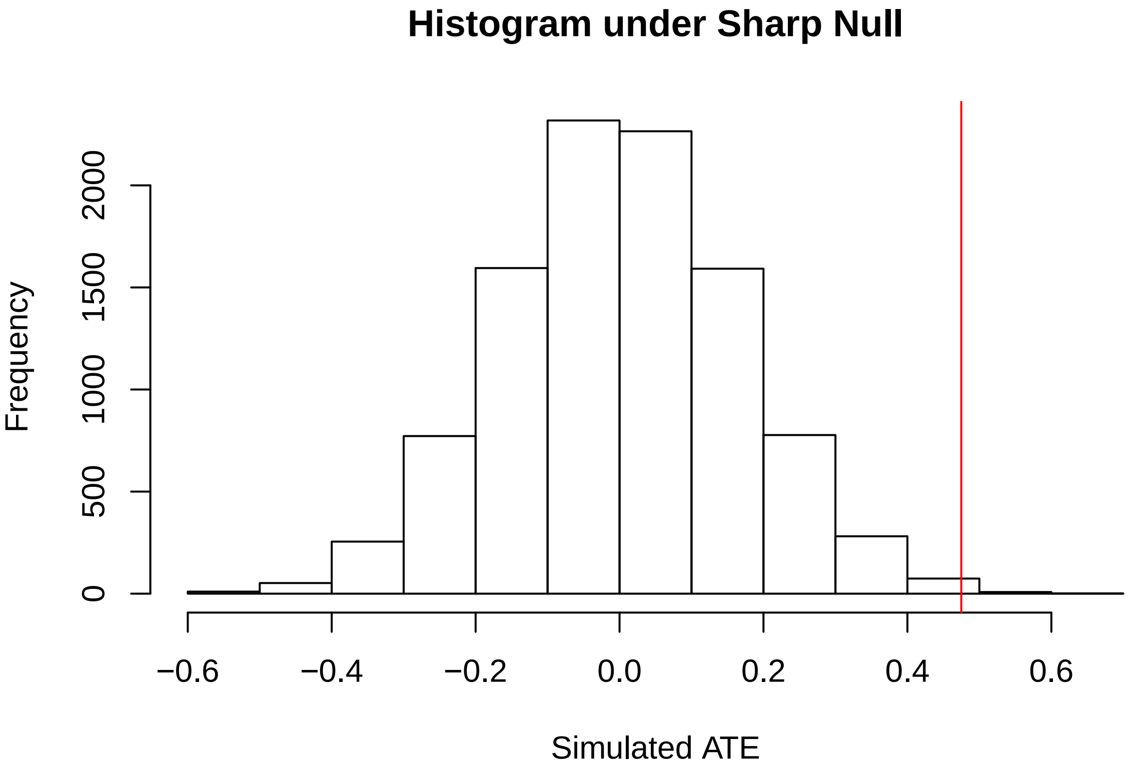
b. But is this a meaningful difference, or could it just be randomization noise? Conduct 10,000 simulated random assignments under the sharp null hypothesis to find out. (Don't just copy the code from the async, think about how to write this yourself.)

```r
# create function to run a simulation of random assignment under the sharp null hypothesis.

simulation <- function(data) {
  # First we randomly create assignments for treatment and control group.
  # The number of 0's and 1's are not always equal but the probability of getting a 0
  # is equal to the probability of getting a 1.
  treatment <- sample(c(0,1), size = length(data$views), replace = TRUE)

  # Next we put the views data into their corresponding groups that we've randomly assigned
  treat.group <- data$views[treatment == 1]
  cont.group <- data$views[treatment == 0]

  # Now we can calculate the estimated ATE for this specific random assignment of
  # control/treatment groups.
  ate <- mean(treat.group) - mean(cont.group)

  return(ate)
}

# Now we run the simulation 10,000 times as directed
distribution.under.sharp.null <- replicate(10000, simulation(d))

# Plot the density distribution and histogram of our simulations
plot(density(distribution.under.sharp.null),
     main = "Density under Sharp Null", xlab = "Simulated ATE")
abline(v = exp_ate, col = "red")
```



```r
hist(distribution.under.sharp.null,
     main = "Histogram under Sharp Null", xlab = "Simulated ATE")
```

```r
abline(v = exp_ate, col = "red")
```



Adam Yang: The resulting density and histogram of our simulated ATE under sharp null is shown above. The red line indicates the ATE we obtained from the actual experimental data.

c. How many of the simulated random assignments generate an estimated ATE that is at least as large as the actual estimate of the ATE?

```r
# The actual estimate of the ATE is stored as exp_ate
prob1c.answ <- sum(distribution.under.sharp.null >= exp_ate)
paste("Answer:", prob1c.answ)
```

```
## [1] "Answer: 17"
```

Adam Yang:

d. What is the implied *one-tailed* p-value?

```r
prob1d.answ <- prob1c.answ/10000
paste("Implied one-tailed p-value:", prob1d.answ)
```

```
## [1] "Implied one-tailed p-value: 0.0017"
```

e. How many of the simulated random assignments generate an estimated ATE that is at least as large *in absolute value* as the actual estimate of the ATE?

```r
prob1e.answ <- sum(abs(distribution.under.sharp.null) >= exp_ate)
paste("Answer:", prob1e.answ)
```

```
## [1] "Answer: 33"
```

f. What is the implied two-tailed p-value?

```
prob1f.answ <- prob1e.answ/ 10000
paste("Implied two-tailed p-value:", prob1f.answ)
```

```
## [1] "Implied two-tailed p-value: 0.0033"
```

2. Term Limits Aren't Good.

Naturally occurring experiments sometimes involve what is, in effect, block random assignment. For example, Rocio Titiunik , in this paper studies the effect of lotteries that determine whether state senators in TX and AR serve two-year or four-year terms in the aftermath of decennial redistricting. These lotteries are conducted within each state, and so there are effectively two distinct experiments on the effects of term length.

The "theory" in the news (such as it is), is that legislators who serve 4 year terms have more time to slack off and not produce legislation. If this were true, then it would stand to reason that making terms shorter would increase legislative production.

One way to measure legislative production is to count the number of bills (legislative proposals) that each senator introduces during a legislative session. The table below lists the number of bills introduced by senators in both states during 2003.

```
library(foreign)

d <- read.dta("./data/Titiunik.2010.dta")
head(d)
```

```
##   term2year bills_introduced texas0_arkansas1
## 1         0               18                0
## 2         0               29                0
## 3         0               41                0
## 4         0               53                0
## 5         0               60                0
## 6         0               67                0
```

a. Using either `dplyr` or `data.table`, group the data by state and report the mean number of bills introduced in each state. Does Texas or Arkansas seem to be more productive? Then, group by two- or four-year terms (ignoring states). Do two- or four-year terms seem to be more productive? **Which of these effects is causal, and which is not?** Finally, using `dplyr` or `data.table` to group by state and term-length. How, if at all, does this change what you learn?

```
# First I will group the data by state to show the mean number of bills introduced.
by_state <- d %>% group_by(texas0_arkansas1)
by_state %>% summarise(avg_num_bills = mean(bills_introduced))
```

```
## # A tibble: 2 x 2
##   texas0_arkansas1 avg_num_bills
##              <int>         <dbl>
## 1                0          68.8
## 2                1          25.5
```

Adam Yang: From the table shown above, we can see that Texas seems to be more productive.

```
# Now I will group by two- or four-year terms to show the mean number of bills introduced.
by_term <- d %>% group_by(term2year)
by_term %>% summarise(avg_num_bills = mean(bills_introduced))
```

```
## # A tibble: 2 x 2
##   term2year avg_num_bills
##       <int>         <dbl>
## 1         0          53.1
## 2         1          38.6
```

Adam Yang: From the table shown above, 4 year terms seems to be more productive. I am assuming that the question is asking which of these effects we are experimenting causality for. In that case, since the random assignment is directed towards the length of the terms, that effect is tested for causality. The state category is used for blocking.

```
# Now I will group by both state and term length
by_state_term <- d %>% group_by(texas0_arkansas1, term2year)
blocked_dt <- by_state_term %>% summarise(avg_num_bills = mean(bills_introduced))
blocked_dt
```

```
## # A tibble: 4 x 3
## # Groups: texas0_arkansas1 [?]
##   texas0_arkansas1 term2year avg_num_bills
##              <int>     <int>         <dbl>
## 1                0         0          76.9
## 2                0         1          60.1
## 3                1         0          30.7
## 4                1         1          20.6
```

Adam Yang: This new information did not change what I learned by much. It seems like Texas has a larger average number of bills produced than Arkansas regardlses of term length. Furthermore, for both states, the 4-year term senators produced a larger average number of bills than the 2-year term senators The magnitude of the ATEs are different.

b. For each state, estimate the standard error of the estimated ATE.

Adam Yang: The instructions for this question does not seem to be very clear. I am assuming that they want us to first calculate the ATE for each of the states. Then using equation (3.6), to estimate the standard error of each of the estimated ATE. Also, I am assuming that the 2 year term is the treatment group.

```
# First calculate the ATE for each state.
tx.ate <- blocked_dt[2,3] - blocked_dt[1,3]
ar.ate <- blocked_dt[4,3] - blocked_dt[3,3]
paste("The ATE for the state of Texas is:", tx.ate)
```

```
## [1] "The ATE for the state of Texas is: -16.7416666666667"
```

```
paste("The ATE for the state of Arkansas is:", ar.ate)
```

```
## [1] "The ATE for the state of Arkansas is: -10.0947712418301"
```

```
# Method using equation 3.6 in Textbook

# First create 4 separete data tables for the control/treatment group of TX and AR.
tx.4year <- by_state_term %>% filter(texas0_arkansas1 == 0, term2year == 0)
tx.2year <- by_state_term %>% filter(texas0_arkansas1 == 0, term2year == 1)
ar.4year <- by_state_term %>% filter(texas0_arkansas1 == 1, term2year == 0)
ar.2year <- by_state_term %>% filter(texas0_arkansas1 == 1, term2year == 1)

# Calculate Var(Y(1)) using equation 3.7 for Texas
tx.4year.m <- length(tx.4year$bills_introduced)
tx.4year.mean <- sum(tx.4year$bills_introduced)/tx.4year.m
tx.var.Y0 <- sum((tx.4year$bills_introduced - tx.4year.mean)^2)/(tx.4year.m - 1)
```

```r
# Calculate Var(Y(0)) using equation 3.8 for Texas
tx.2year.m <- length(tx.2year$bills_introduced)
tx.2year.mean <- sum(tx.2year$bills_introduced)/tx.2year.m
tx.var.Y1 <- sum((tx.2year$bills_introduced - tx.2year.mean)^2)/(tx.2year.m - 1)

# Calculate SE of the ATE using equation 3.6 for Texas
tx.SE.ate <- sqrt(tx.var.Y0/tx.4year.m + tx.var.Y1/tx.2year.m)
paste("The estimated SE of the ATE for Texas is", tx.SE.ate)
```

```
## [1] "The estimated SE of the ATE for Texas is 9.3458711493195"
```

```r
# Calculate Var(Y(1)) using equation 3.7 for Arkansas
ar.4year.m <- length(ar.4year$bills_introduced)
ar.4year.mean <- sum(ar.4year$bills_introduced)/ar.4year.m
ar.var.Y0 <- sum((ar.4year$bills_introduced - ar.4year.mean)^2)/(ar.4year.m - 1)

# Calculate Var(Y(0)) using equation 3.8 for Arkansas
ar.2year.m <- length(ar.2year$bills_introduced)
ar.2year.mean <- sum(ar.2year$bills_introduced)/ar.2year.m
ar.var.Y1 <- sum((ar.2year$bills_introduced - ar.2year.mean)^2)/(ar.2year.m - 1)

# Calculate SE of the ATE using equation 3.6 for Arkansas
ar.SE.ate <- sqrt(ar.var.Y0/ar.4year.m + ar.var.Y1/ar.2year.m)
paste("The estimated SE of the ATE for Arkansas is", ar.SE.ate)
```

```
## [1] "The estimated SE of the ATE for Arkansas is 3.3959791516933"
```

Adam Yang: Another way to solve this question is to use a regression as explained in the ASYNC. The results are similar to the ones found using equation (3.6) but slightly different.

```r
# Method using regression

# First create a data table for only texas samples
tx.data <- by_state_term %>% filter(texas0_arkansas1 == 0)
# Do a linear regression
tx.regression <- summary(lm(tx.data$bills_introduced ~ tx.data$term2year))
# SE given in the regression summary
tx.se <- tx.regression$coefficients[2,2]
paste("The estimated SE of the ATE for Texas is", tx.se)
```

```
## [1] "The estimated SE of the ATE for Texas is 9.47015869198299"
```

```r
# First create a data table for only Arkansas samples
ar.data <- by_state_term %>% filter(texas0_arkansas1 == 1)
# Do a linear regression
ar.regression <- summary(lm(ar.data$bills_introduced ~ ar.data$term2year))
# SE given in the regression summary
ar.se <- ar.regression$coefficients[2,2]
paste("The estimated SE of the ATE for Arkansas is", ar.se)
```

```
## [1] "The estimated SE of the ATE for Arkansas is 3.34687166973233"
```

c. Use equation (3.10) to estimate the overall ATE for both states combined.

```r
tx.N <- length(tx.data$bills_introduced)
ar.N <- length(ar.data$bills_introduced)
N <- length(by_state_term$bills_introduced)
```

```r
tx.ate <- as.numeric(tx.ate)
ar.ate <- as.numeric(ar.ate)

Overall.ATE <- (tx.N/N)*tx.ate + (ar.N/N)*ar.ate
Overall.ATE
```

```
## [1] -13.2168
```

Adam Yang: Using equation (3.10), I found that the overall ATE is **-13.22** assuming the 2 year term is the treatment group.

d. Explain why, in this study, simply pooling the data for the two states and comparing the average number of bills introduced by two-year senators to the average number of bills introduced by four-year senators leads to biased estimate of the overall ATE.

Adam Yang: If we simply pool the data for the two states and compare the average number of bills introduced by two-year senators to the average number of bills introduced by four-year senators, we get an estimated ATE of -14.51. This is larger in magnitude than the ATE we obtained by applying blocking to the two states (-13.22). This is because there is an omitted variable bias with the states. Senators in Texas introduce many more bills on average than senators in Arkansas which means the effect to our outcome by senators in Texas will outweight the effect to our outcome by senators in Arkansas. In order to reduce the bias in our estimate of the overall ATE, we need to block by state so that difference between the two states will not bias our outcome.

e. Insert the estimated standard errors into equation (3.12) to estimate the standard error for the overall ATE.

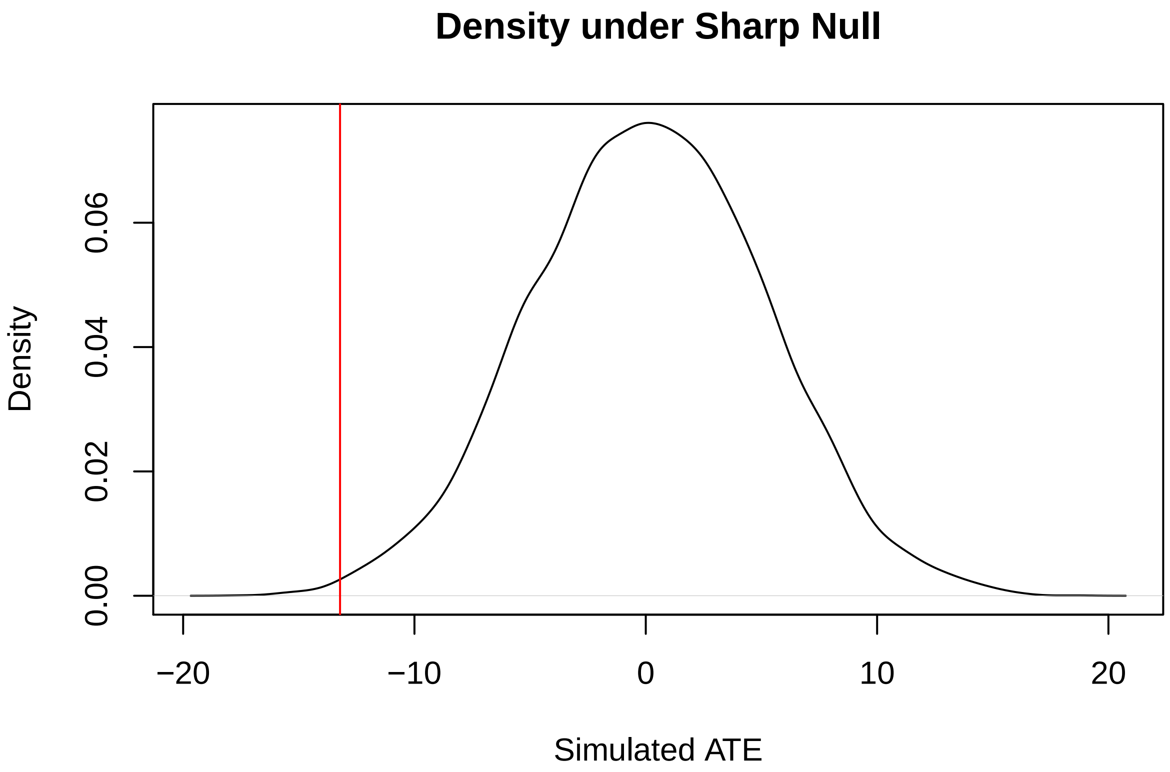
```r
# I will be using the standard errors obtained from the regression method
sqrt((tx.se^2)*(tx.N/N)^2 + (ar.se^2)*(ar.N/N)^2)
```

```
## [1] 4.789128
```

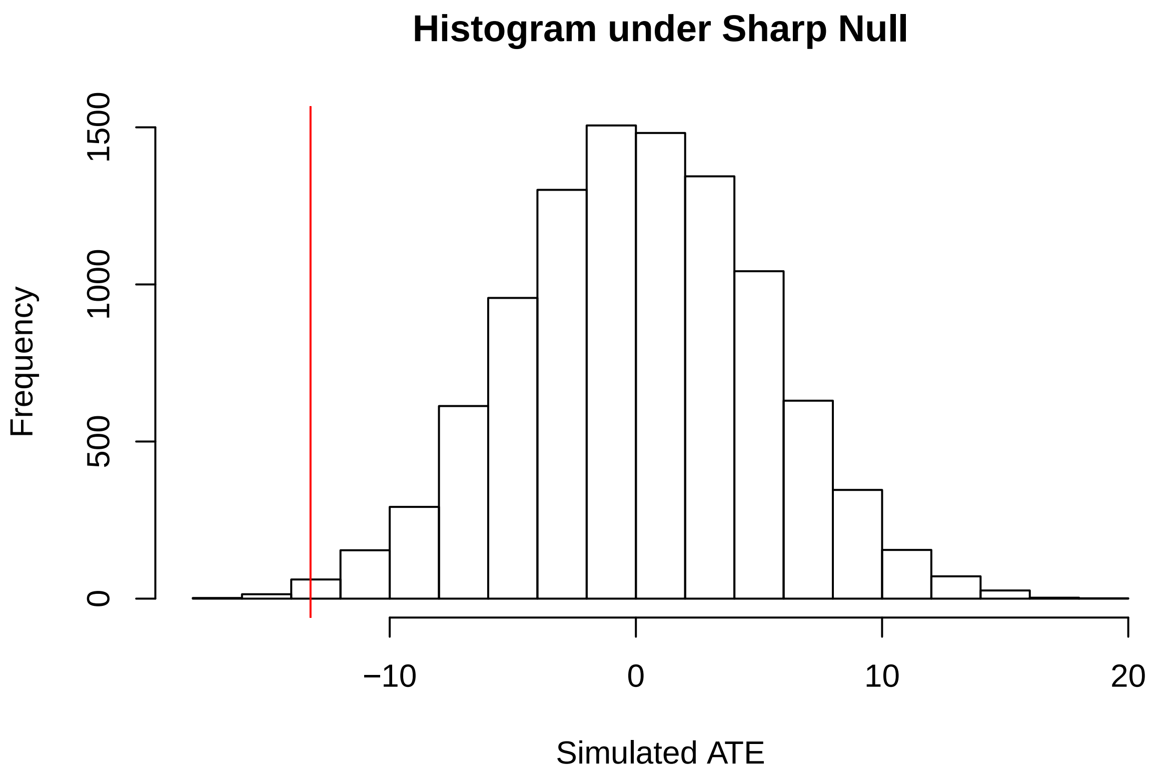
Adam Yang: The estimate of the standard error for the averall ATE is **4.79**.

f. Use randomization inference to test the sharp null hypothesis that the treatment effect is zero for senators in both states.

```r
# create function to run a simulation of random assignment under the sharp null hypothesis.

simulation <- function(data) {
  # First we randomly create assignments for treatment and control group.
  # The number of 0's and 1's are not always equal but the probability of getting a 0
  # is equal to the probability of getting a 1.
  treatment <- sample(c(0,1), size = length(data$bills_introduced), replace = TRUE)

  # Next we put the views data into their corresponding groups that we've randomly assigned
  treat.group <- data$bills_introduced[treatment == 1]
  cont.group <- data$bills_introduced[treatment == 0]

  # Now we can calculate the estimated ATE for this specific random assignment of
  # control/treatment groups.
  ate <- mean(treat.group) - mean(cont.group)

  return(ate)
}

# Now we create a function to find the overall ATEs for our simulations
Overall_ATE_Sim <- function(txdata, ardata) {
  tx.sim.ate <- simulation(txdata)
  ar.sim.ate <- simulation(ardata)
  tx.N <- length(tx.data$bills_introduced)
  ar.N <- length(ar.data$bills_introduced)
  N <- length(by_state_term$bills_introduced)
  OverallATE <- (tx.N/N)*tx.sim.ate + (ar.N/N)*ar.sim.ate
  return(OverallATE)
}

# Now we run the simulation 10,000 times as directed
distribution.under.sharp.null <- replicate(10000, Overall_ATE_Sim(tx.data, ar.data))

# Plot the density distribution and histogram of our simulations
plot(density(distribution.under.sharp.null),
     main = "Density under Sharp Null", xlab = "Simulated ATE")
abline(v = Overall.ATE, col = "red")
```



```r
hist(distribution.under.sharp.null,
     main = "Histogram under Sharp Null", xlab = "Simulated ATE")
abline(v = Overall.ATE, col = "red")
```



```r
# Find the implied one-tailed p-value
onetail <- sum(distribution.under.sharp.null <= Overall.ATE)/10000
paste("Implied one-tailed p-value:", onetail)
```

```
## [1] "Implied one-tailed p-value: 0.0029"
```

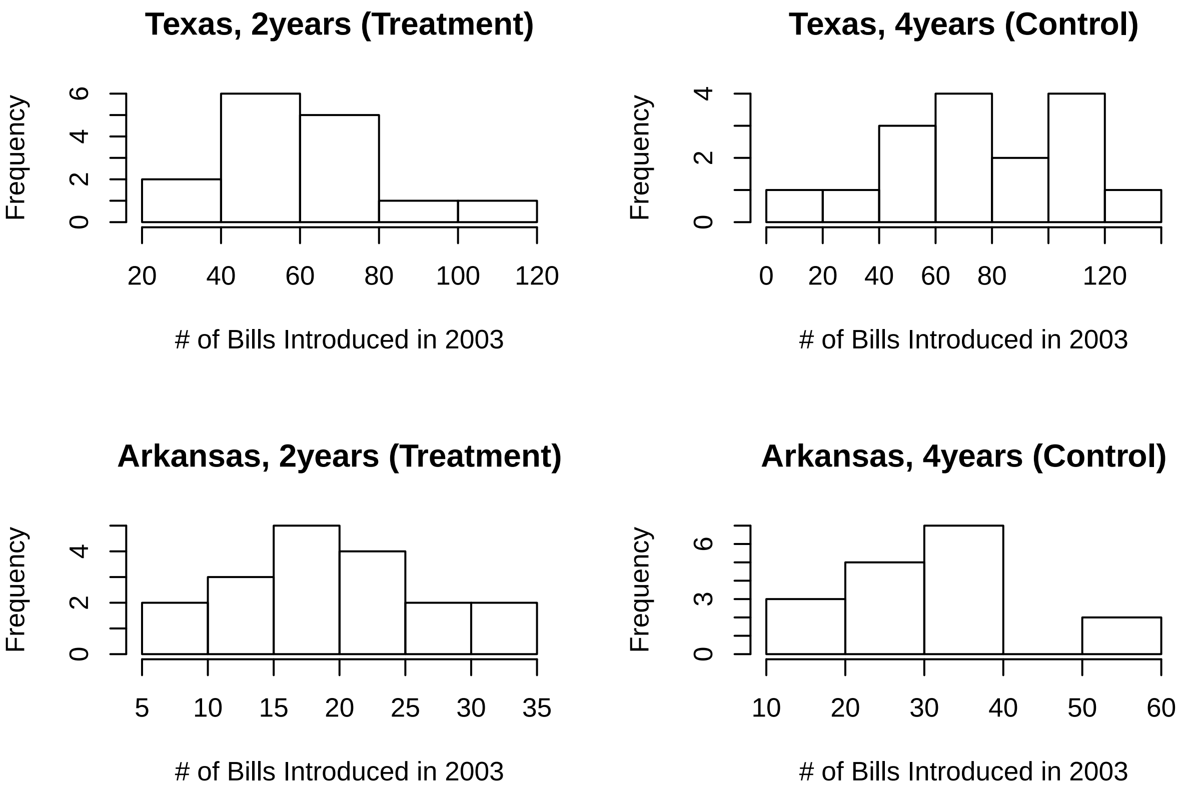
```r
#Find the implied two-tailed p-value
twotail <- sum(abs(distribution.under.sharp.null) >= abs(Overall.ATE))/10000
paste("Implied two-tailed p-value:", twotail)
```

```
## [1] "Implied two-tailed p-value: 0.008"
```

Adam Yang: Both the one tailed and two tailed p-values are much smaller than 0.05 which means we can reject the sharp null hypothesis that the treatment effect is zero for senators in both states.

g. **In Addition:** Plot histograms for both the treatment and control groups in each state (for 4 histograms in total).

```r
layout(matrix(c(1,2,3,4), 2, 2, byrow = T))
hist(tx.2year$bills_introduced, main = "Texas, 2years (Treatment)", xlab = "# of Bills Introduced in 20
hist(tx.4year$bills_introduced, main = "Texas, 4years (Control)", xlab = "# of Bills Introduced in 2003
hist(ar.2year$bills_introduced, main = "Arkansas, 2years (Treatment)", xlab = "# of Bills Introduced in
hist(ar.4year$bills_introduced, main = "Arkansas, 4years (Control)", xlab = "# of Bills Introduced in 2
```



Texas, 2years (Treatment)

Texas, 4years (Control)

Arkansas, 2years (Treatment)

Arkansas, 4years (Control)

3. Cluster Randomization

Use the data in *Field Experiments* Table 3.3 to simulate cluster randomized assignment. (*Notes: (a) Assume 3 clusters in treatment and 4 in control; and (b) When Gerber and Green say* `simulate'', they do not mean`*run simulations with R code'', but rather, in a casual sense "take a look at what happens if you do this this way." There is no randomization inference necessary to complete this problem.*)

```
## load data
d <- read.csv("./data/ggChapter3.csv", stringsAsFactors = FALSE)
```

a. Suppose the clusters are formed by grouping observations {1,2}, {3,4}, {5,6}, . . . , {13,14}. Use equation (3.22) to calculate the standard error assuming half of the clusters are randomly assigned to treatment.

```
# First we label each of the villages by their clusters
d$clusters <- c(1,1,2,2,3,3,4,4,5,5,6,6,7,7)
# Then we group by the clusters and greate a cluster level dataset
by_clusters <- d %>% group_by(clusters)
cluster.data <- by_clusters %>% summarise(cluster_Y = mean(Y), cluster_D = mean(D))

# Now define all the variables
k <- length(cluster.data$clusters) # number of clusters
N <- length(d$Village) # Total number of subjects
m <- 6 # size of treatment group (assume 3 clusters in treatment)
Ybar0 <- cluster.data$cluster_Y
Ybar1 <- cluster.data$cluster_D

# I created a function to calculate the variance the same way as equation 3.2
```

```r
# This is because the var() function in R has N-1 in the denominator rather than N
variance <- function(data) {
  var(data)*(length(data)-1)/length(data)
}

# A function is also created for covariance because the R version isn't what we want
covariance <- function(a,b) {
  cov(a,b)*(length(a)-1)/length(a)
}

# Now to calcuate the SE with equation 3.22
SE.ATE <- sqrt((1/(k-1))*(m*variance(Ybar0)/(N-m)
                       + (N-m)*variance(Ybar1)/m
                       + 2*covariance(Ybar0, Ybar1)))
SE.ATE
```

```
## [1] 4.554065
```

Adam Yang: Using equation 3.22 we calcuated a standard error of **4.55**.

b. Suppose that clusters are instead formed by grouping observations {1,14}, {2,13}, {3,12}, . . . , {7,8}. Use equation (3.22) to calculate the standard error assuming half of the clusters are randomly assigned to treatment.

```r
# First we label each of the villages by their clusters
d$clusters <- c(1,2,3,4,5,6,7,7,6,5,4,3,2,1)
# Then we group by the clusters and greate a cluster level dataset
by_clusters <- d %>% group_by(clusters)
cluster.data <- by_clusters %>% summarise(cluster_Y = mean(Y), cluster_D = mean(D))

# Now define all the variables
k <- length(cluster.data$clusters) # number of clusters
N <- length(d$Village) # Total number of subjects
m <- 6 # size of treatment group (assume 3 clusters in treatment)
Ybar0 <- cluster.data$cluster_Y
Ybar1 <- cluster.data$cluster_D

# Now to calcuate the SE with equation 3.22
SE.ATE <- sqrt((1/(k-1))*(m*variance(Ybar0)/(N-m)
                       + (N-m)*variance(Ybar1)/m
                       + 2*covariance(Ybar0, Ybar1)))
SE.ATE
```

```
## [1] 1.171092
```

Adam Yang: For this case, using equation 3.22 we calcuated a standard error of **1.171**.

c. Why do the two methods of forming clusters lead to different standard errors? What are the implications for the design of cluster randomized experiments?

Adam Yang: The two methods of forming clustsers lead to different standard errors because the resulting cluster level means are very different. In part a, the potential outcomes within the clusters are pretty similar while in part b, the potential outcomes within the clusters are very different. For example the first cluster in part a consists of observations {1,2} which have {Y(0),Y(1)} values of {0,0} and {1,0} respectively. Therefore, the cluster level means are {0.5,0}. The first cluster in part b consists of observations {1,14} which have {Y(0),Y(1)} values of {0,0} and {18,17} respectively. Therefore, the cluster level means are {9,8.5}. As it turns out, the cluster level means for part b are very similar in value as opposed to part a where the cluster level means are very different in value. That is why part a has a larger standard error than part b.The implications for the design of cluster randomized experiments is to be very careful with working with clustered data. The penalty associated with clustering depends on the variability of the cluster level means. Ideally, if the cluster level means are very similar, then our standard error won't be too big. If the difference between the cluster level means are very big, then we should make some adjustments.

4. Sell Phones?

You are an employee of a newspaper and are planning an experiment to demonstrate to Apple that online advertising on your website causes people to buy iPhones. Each site visitor shown the ad campaign is exposed to $0.10 worth of advertising for iPhones. (Assume all users could see ads.) There are 1,000,000 users available to be shown ads on your newspaper's website during the one week campaign.

Apple indicates that they make a profit of $100 every time an iPhone sells and that 0.5% of visitors to your newspaper's website buy an iPhone in a given week in general, in the absence of any advertising.

a. By how much does the ad campaign need to increase the probability of purchase in order to be "worth it" and a positive ROI (supposing there are no long-run effects and all the effects are measured within that week)?

Adam Yang:

$$RequiredWeeklyProfit = \frac{\$100}{purchase} * \frac{x * 1,000,000 purchases}{week}$$

$$WeeklyCostForAds = \frac{\$0.10}{person} * \frac{1,000,000 people}{week}$$

$$RequiredWeeklyProfit >= WeeklyCost$$

$$\frac{\$100}{purchase} * \frac{x * 1,000,000 purchases}{week} >= \frac{\$0.10}{person} * \frac{1,000,000 people}{week}$$

$$x >= \frac{\$0.10}{\$100} = 0.001 \; or \; 0.1\%$$

Adam Yang: The ad campaign needs to increase the probability of purchase by at least **0.1%** in order to be "worth it".

b. Assume the measured effect is 0.2 percentage points. If users are split 50:50 between the treatment group (exposed to iPhone ads) and control group (exposed to unrelated advertising or nothing; something you can assume has no effect), what will be the confidence interval of your estimate on whether people purchase the phone?

```r
# Sample sizes in control and treatment groups
n1 <- 1000000/2
n2 <- 1000000/2
# Number of purchases
x1 <- 0.005*n1 # control
x2 <- (0.005 + 0.002)*n2 #treatment
# Calculate p
p <- (x1 + x2)/(n1 + n2)
# Calculate standard error for a two-sample proportion test
```

```r
se <- sqrt(p*(1-p)*(1/n1 +1/n2))
# Calculate the left and right bound of the confidence interval
left.bound <- 0.002 - 1.96*se
right.bound <- 0.002 + 1.96*se

paste("The 95% confidence interval is (", left.bound,",",right.bound,")")
```

```
## [1] "The 95% confidence interval is ( 0.00169727040184349 , 0.00230272959815651 )"
```

Adam Yang: The 95% confidence interval is **(0.0017, 0.0023)** or **(0.17%, 0.23%)**.

- **Note:** The standard error for a two-sample proportion test is $\sqrt{p(1-p)*(\frac{1}{n_1}+\frac{1}{n_2})}$ where $p = \frac{x_1+x_2}{n_1+n_2}$, where x and n refer to the number of "successes" (here, purchases) over the number of "trials" (here, site visits). The length of each tail of a 95% confidence interval is calculated by multiplying the standard error by 1.96.

c. Is this confidence interval precise enough that you would recommend running this experiment? Why or why not?

As shown in part b, we measured an effect size of 0.2 percentage point increase in the probability of purchase of iphones once our ads are introduced. The 95% confidence interval lies between 0.17 and 0.23 percentage increase. That means if we run the same experiment 100 times, 97.5 of those experiments will have an effect size larger than 0.17 percentage points. The lower bound of that confidence interval is relatively much higher than the 0.1 percentage point increase that we need to break even. Therefore, I would recommend running this experiment.

d. Your boss at the newspaper, worried about potential loss of revenue, says he is not willing to hold back a control group any larger than 1% of users. What would be the width of the confidence interval for this experiment if only 1% of users were placed in the control group?

```r
# Sample sizes in control and treatment groups
n1 <- 1000000 * 0.01 # control
n2 <- 1000000 * 0.99 # treatment
# Number of purchases
x1 <- 0.005*n1 # control
x2 <- (0.005 + 0.002)*n2 #treatment
# Calculate p
p <- (x1 + x2)/(n1 + n2)
# Calculate standard error for a two-sample proportion test
se <- sqrt(p*(1-p)*(1/n1 +1/n2))
# Calculate the left and right bound of the confidence interval
left.bound <- 0.002 - 1.96*se
right.bound <- 0.002 + 1.96*se

paste("The 95% confidence interval is (", left.bound,",",right.bound,")")
```

```
## [1] "The 95% confidence interval is ( 0.000359994958391588 , 0.00364000504160841 )"
```

Adam Yang: The 95% confidence interval in this case would be **(0.00036, 0.00364)** or **(0.036%, 0.364%)**. This resulting confidence interval is much wider than the previous confidence interval we calculated and therefore does not seem precise enough for me to recommend running this experiment. Especially because the 0.1 percentage increase is well within this 95% confidence interval.

5. Sports Cards

Here you will find a set of data from an auction experiment by John List and David Lucking-Reiley (2000).

```
d2 <- read.csv("./data/listData.csv", stringsAsFactors = FALSE)
head(d2)
```

```
##   bid uniform_price_auction
## 1   5                     1
## 2   5                     1
## 3  20                     0
## 4   0                     1
## 5  20                     1
## 6   0                     1
```

In this experiment, the experimenters invited consumers at a sports card trading show to bid against one other bidder for a pair trading cards. We abstract from the multi-unit-auction details here, and simply state that the treatment auction format was theoretically predicted to produce lower bids than the control auction format. We provide you a relevant subset of data from the experiment.

a. Compute a 95% confidence interval for the difference between the treatment mean and the control mean, using analytic formulas for a two-sample t-test from your earlier statistics course.

```
control <- d2$bid[d2$uniform_price_auction == 0]
treatment <- d2$bid[d2$uniform_price_auction == 1]
t.test(treatment,control)
```

```
##
##  Welch Two Sample t-test
##
## data:  treatment and control
## t = -2.8211, df = 61.983, p-value = 0.006421
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -20.854624  -3.557141
## sample estimates:
## mean of x mean of y
##  16.61765  28.82353
```

Adam Yang: The 95% confidence interval as show above is between **(-20.85, -3.56)**.

b. In plain language, what does this confidence interval mean?

The 95% confidence interval we calculated means that if we run the experiment 100 times, we should expect 95 of those times to have a difference in means (treatment minus control) to be between -20.85 and -3.56. Since 0 does not lie within our 95% confidence interval, we can reject the null hypothesis that the true difference in means is equal to 0.

c. Regression on a binary treatment variable turns out to give one the same answer as the standard analytic formula you just used. Demonstrate this by regressing the bid on a binary variable equal to 0 for the control auction and 1 for the treatment auction.

```
# First create the regression
regression <- summary(lm(d2$bid~d2$uniform_price_auction))
regression
```

```
##
## Call:
## lm(formula = d2$bid ~ d2$uniform_price_auction)
```

```
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -28.824 -11.618  -3.221   8.382  58.382
##
## Coefficients:
##                           Estimate Std. Error t value Pr(>|t|)
## (Intercept)                 28.824      3.059   9.421 7.81e-14 ***
## d2$uniform_price_auction   -12.206      4.327  -2.821  0.00631 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 17.84 on 66 degrees of freedom
## Multiple R-squared:  0.1076, Adjusted R-squared:  0.09409
## F-statistic: 7.959 on 1 and 66 DF,  p-value: 0.006315
```

Adam Yang: The ATE calculated from part a is 16.61765 - 28.82353 = -12.206 which is equal to the slope of the regression we found above. Furthermore, the p-value is equivalent (0.00631) as well as the t-value (-2.821).

d. Calculate the 95% confidence interval you get from the regression.

```r
# Calculate the 95% confidence interval from the Regression
# Use slope +- (t-value)(std. error)
slope <- regression$coefficients[2,1]
Tval <- qt(0.975, 66) # using 66 dof
se <- regression$coefficients[2,2]

# Calculate the bounds
lower.bound <- slope - Tval*se
upper.bound <- slope + Tval*se

paste("The 95% confidence interval is (", lower.bound,",",upper.bound,")")
```

```
## [1] "The 95% confidence interval is ( -20.8441620300328 , -3.5676026758496 )"
```

Adam Yang: The confidence interval we got is **(-20.84, -3.57)** which is very similar to what we got in part a.

e. On to p-values. What p-value does the regression report? Note: please use two-tailed tests for the entire problem.

```r
regression$coefficients[2,4]
```

```
## [1] 0.006314796
```

Adam Yang: The p-value reported from the regression is **0.006315**.

f. Now compute the same p-value using randomization inference.

```r
# create function to run the simulation

simulation <- function(data) {
  # First we randomly create assignments for treatment and control group.
  # The number of 0's and 1's are not always equal but the probability of getting a 0
  # is equal to the probability of getting a 1.
  treatment <- sample(c(0,1), size = length(data$bid), replace = TRUE)

  # Next we put the views data into their corresponding groups that we've randomly assigned
  treat.group <- data$bid[treatment == 1]
  cont.group <- data$bid[treatment == 0]

  # Now we can calculate the estimated ATE for this specific random assignment of
  # control/treatment groups.
  ate <- mean(treat.group) - mean(cont.group)

  return(ate)
}

# Now we run the simulation 10,000 times
distribution.under.sharp.null <- replicate(10000, simulation(d2))

# Now calculate the two-tail pvalue
pval <- sum(abs(distribution.under.sharp.null) >= 12.206)/10000
pval
```

```
## [1] 0.0059
```

```r
paste("P-value from randomization inference is:", pval)
```

```
## [1] "P-value from randomization inference is: 0.0059"
```

g. Compute the same p-value again using analytic formulas for a two-sample t-test from your earlier statistics course. (Also see part (a).)

```r
control <- d2$bid[d2$uniform_price_auction == 0]
treatment <- d2$bid[d2$uniform_price_auction == 1]
t.test(treatment,control)
```

```
##
##  Welch Two Sample t-test
##
## data:  treatment and control
## t = -2.8211, df = 61.983, p-value = 0.006421
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -20.854624  -3.557141
## sample estimates:
## mean of x mean of y
##  16.61765  28.82353
```

Adam Yang: The p-value given above from the t-test is **0.006421**.

h. Compare the two p-values in parts (e) and (f). Are they much different? Why or why not? How might your answer to this question change if the sample size were different?

Adam Yang: The p-values in parts e and f can be quite different. Many simulations I get a p-value in part f around 0.0066 but there are times when I get a p-value above 0.007 and below 0.005. This is because the p-value calculated from random inference is more sensitive to asymmetrical distributions. If we increased the sample size, the p-values will become closer in value.