# DS-UA 201: Problem Set 2

Jerry Wan

October 13, 2020

This problem set is due at **11:59 pm on Tuesday, October 13th**. The data are on the course website.

Please upload your solutions as a .pdf file saved as `Yourlastname_Yourfirstinitial_pset2.pdf`). In addition, an electronic copy of your .Rmd file (saved as `Yourlastname_Yourfirstinitial_pset2.Rmd`) must be submitted to the course website at the same time. We should be able to run your code without error messages. Please note on your problem set if you collaborated with another student and, if so, whom. In order to receive credit, homework submissions must be substantially started and all work must be shown. Late assignments will not be accepted.

## Problem 1

Do international election monitors reduce the incidence of electoral fraud? Hyde (2007) studies the 2003 presidential election in Armenia, an election that took place during a period where the incumbent ruling party headed by President Robert Kocharian had consolidated power and often behaved in ways that were considered undemocratic.

The full citation for this paper is

Hyde, Susan D. "The observer effect in international politics: Evidence from a natural experiment." *World Politics* 60.1 (2007): 37-63.

At the time of the election, OSCE/ODIHR election monitors reported widespread electoral irregularities that favored the incumbent party such as ballot-box stuffing (pp. 47). However, we do not necessarily know whether these irregularities would have been worse in the absence of monitors. Notably, not all polling stations were monitored – the OSCE/ODIHR mission could only send observers to some of the polling stations in the country. Since in the context of this election only the incumbent party would have the capacity to carry out significant election fraud, Hyde examines whether the presence of election observers from the OSCE/ODIHR mission at polling stations in Armenia reduced the incumbent party's vote share at that polling station.

For the purposes of this problem, you will be using the `armenia2003.dta` dataset

The R code below will read in this data (which is stored in the STATA .dta format)

```r
library(tidyverse)
library(haven)

### Hyde (2007) Armenia dataset
armenia <- read_dta("armenia2003.dta")
```

This dataset consists of 1764 observations polling-station-level election results from the 2003 Armenia election made available by the Armenian Central Election Commission. The election took place over two rounds with an initial round having a large number of candidates and a second, run-off election, between Kocharian and the second-place vote-getter, Karen Demirchyan. We will focus on monitoring and voting in the first round. The specific columns you will need are:

- `kocharian` - Round 1 vote share for the incumbent (Kocharian)
- `mon_voting` - Whether the polling station was monitored in round 1 of the election
- `turnout` - Proportion of registered voters who voted in Round 1
- `totalvoters` - Total number of registered voters recorded for the polling station
- `total` - Total number of votes cast in Round 1
- `urban` - Indicator for whether the polling place was in an urban area (0 = rural, 1 = urban)
- `nearNagorno` - Indicator for whether the polling place is near the Nagorno-Karabakh region (0 = no, 1 = yes)

## Part A

Hyde describes the study as a "natural experiment," stating:

"I learned from conversations with staff and participants in the OSCE observation mission to Armenia that the method used to assign observers to polling stations was functionally equivalent to random assignment. This permits the use of natural experimental design. Although the OSCE/ODIHR mission did not assign observers using a random numbers table or its equivalent, the method would have been highly unlikely to produce a list of assigned polling stations that were systematically different from the polling stations that observers were not assigned to visit. Each team's assigned list was selected arbitrarily from a complete list of polling stations." (p. 48)

What makes this study a "natural experiment" and not a true experiment? What assumption must the study defend in order to identify the causal effect of election monitoring that would be guaranteed to hold in a randomized experiment? What might be some possible reasons why that assumption might not hold in this design?

1.Because the participants are randomly assigned to different polling stations randomly and this makes all the control conditions left are not controlled by the researchers. However it is not a true experiment because as they mention that the mission did not assign obververs using a random numbers table or its equivalent.

2.They need to defend that the result come out from the observation of those assigned polling stations will not have a systematically different from other polling station

3.The available polling station for observation is strictly controlled by the government or the polling stations are only in some certain areas like cities and miss the data from countryside.

## Part B

For the purposes of this part, assume election monitors were assigned as the author describes - in a manner “functionally equivalent to random assignment.” Using the difference-in-means estimator, estimate the average treatment effect of election monitoring on incumbent vote share in round 1. Provide a 95% asymptotic confidence interval and interpret your results. Can we reject the null of no average treatment effect at the $\alpha = 0.05$ level?

```
diff_in_means <- function(treated, control){
  # Point Estimate
  point <- mean(treated) - mean(control)

  # Standard Error
  se <- sqrt(var(treated)/length(treated) + var(control)/length(control))

  # Asymptotic 95% CI
  ci_95 <- c(point - qnorm(.975)*se,
             point + qnorm(.975)*se)

  # P-value
  pval <- 2*pnorm(-abs(point/se))

  # Return as a data frame
  output <- data.frame(meanTreated = mean(treated), meanControl = mean(contro
l), est = point, se = se, ci95Lower = ci_95[1], ci95Upper = ci_95[2], pvalue
= pval)

  return(as_tibble(output))

}
diff_in_means(armenia$kocharian[armenia$mon_voting == 1],
              armenia$kocharian[armenia$mon_voting == 0])

## # A tibble: 1 x 7
##   meanTreated meanControl     est      se ci95Lower ci95Upper        pvalu
e
##         <dbl>       <dbl>   <dbl>   <dbl>     <dbl>     <dbl>         <dbl>
## 1       0.483       0.542 -0.0587 0.00979   -0.0779   -0.0395 0.0000000020
8
```

## Part C

Evaluate the author’s identification assumptions by examining whether the treatment is balanced on three pre-treatment covariates: the total number of registered voters, whether a polling place was in an urban area, and whether the polling place was located near the Nagorno-Karabakh region (Kocharian’s home region and a disputed territory between

Armenia and Azerbaijan). Discuss your results. Are they consistent with the author's description of "as-if random" assignment?

```
treatedVoters = mean(armenia$totalvoters[armenia$mon_voting == 1])
controlVoters = mean(armenia$totalvoters[armenia$mon_voting == 0])

treatedUrban = mean(armenia$urban[armenia$mon_voting == 1])
controlUrban = mean(armenia$urban[armenia$mon_voting == 0])

treatdNear = mean(armenia$nearNagorno[armenia$mon_voting == 1])
controlNear = mean(armenia$nearNagorno[armenia$mon_voting == 0])

treatedVoters

## [1] 1467.539

controlVoters

## [1] 1069.408

treatedUrban

## [1] 0.618984

controlUrban

## [1] 0.507874

treatdNear

## [1] 0.04679144

controlNear

## [1] 0.05807087
```

With the data shown above, we can believe that the total number of voters can make the result treated as a random result. However, we found that the proportion of polling station in urban areas are higher than the suburb area. But I think it is understandable and can have little effect on the final result. However, we can see that the potion of station near Nagorno-Karabakh region is smaller than the total number of polling station, this may cause the bias in the final result.

## Part D

Divide the sample into five strata based on the total number of registered voters at each polling station (`totalvoters`):

| Stratum | Total Registered Voters |
|---|---|
| Tiny | `totalvoters` < 430 |
| Small | 430 ≤ `totalvoters` < 1192 |

| | |
|---|---|
| Medium | $1192 \leq \texttt{totalvoters} < 1628$ |
| Large | $1628 \leq \texttt{totalvoters} < 1879$ |
| Huge | $1879 \leq \texttt{totalvoters}$ |

Estimate the average treatment effect of election monitoring in round 1 on incumbent vote share using a stratified difference-in-means estimator, stratifying on the total number of registered voters. Provide a 95% asymptotic confidence interval and interpret your results. Can we reject the null of no average treatment effect at the $\alpha = 0.05$ level? Compare your answer to your estimate from Question 2 and discuss any differences you see.

```
armenia$stratum = 0
armenia$stratum[armenia$totalvoters < 430] = 1
armenia$stratum[armenia$totalvoters >= 430 & armenia$totalvoters < 1192] = 2
armenia$stratum[armenia$totalvoters >= 1192 & armenia$totalvoters < 1628] = 3
armenia$stratum[armenia$totalvoters >= 1879] = 4

armenia %>% group_by(stratum) %>% summarize(meanMoniteredVoting = mean(mon_vo
ting), Ng = n())

## `summarise()` ungrouping output (override with `.groups` argument)

## # A tibble: 5 x 3
##   stratum meanMoniteredVoting    Ng
##     <dbl>               <dbl> <int>
## 1       0               0.516   351
## 2       1               0.185   352
## 3       2               0.326   353
## 4       3               0.523   354
## 5       4               0.571   354

diff_in_means_stratum <- function(treated, control, stratumname="all"){
  # Point Estimate
  point <- mean(treated) - mean(control)

  # Standard Error
  se <- sqrt(var(treated)/length(treated) + var(control)/length(control))

  # Asymptotic 95% CI
  ci_95 <- c(point - qnorm(.975)*se,
             point + qnorm(.975)*se)

  # P-value
  pval <- 2*pnorm(-abs(point/se))

  # Return as a data frame
  output <- data.frame(stratum = stratumname, meanTreated = mean(treated), me
anControl = mean(control), est = point, se = se, ci95Lower = ci_95[1], ci95Up
per = ci_95[2], pvalue = pval, N= length(treated) + length(control))
```

```
  return(as_tibble(output))

}

stratum_ests <- bind_rows(map(unique(armenia$stratum), function(x) diff_in_me
ans_stratum(armenia$kocharian[armenia$mon_voting == 1&armenia$stratum == x],
                                                                 armenia$kocharian
[armenia$mon_voting == 0&armenia$stratum == x], x))) %>%
  select(stratum, est, se, N_g = N)

stratum_ests

## # A tibble: 5 x 4
##   stratum      est     se   N_g
##     <dbl>    <dbl>  <dbl> <int>
## 1       3 -0.0119  0.0206   354
## 2       1  0.00749 0.0287   352
## 3       4 -0.0224  0.0178   354
## 4       2 -0.0188  0.0245   353
## 5       0 -0.0394  0.0187   351

# Aggregate the estimates in the four strata to an estimate of the ATE

# Point estimate
N_all = sum(stratum_ests$N_g)
point_stratum = sum(stratum_ests$est*(stratum_ests$N_g/N_all))
point_stratum

## [1] -0.01699158

# Standard error
var_stratum = sum((stratum_ests$se^2)*(stratum_ests$N_g/N_all)^2)
se_stratum = sqrt(var_stratum)
se_stratum

## [1] 0.01002836

# Confidence interval (95%)
ci_95_stratum = c(point_stratum - qnorm(.975)*se_stratum,
                   point_stratum + qnorm(.975)*se_stratum)
ci_95_stratum

## [1] -0.036646815  0.002663651
```

We cans see htat overall there are very little varience between different voting station, as a result, we may not reject the null of no average treatment effect at the $\alpha = 0.05$ level

## Part E

In Table 4 of the paper, Hyde uses an estimator for the average treatment effect of a polling place receiving election monitors in round 1 on the incumbent’s vote share in round 1 *conditional* on the total number of votes cast in the election. Will this approach be unbiased for the average treatment effect of election monitors on the incumbent's vote share if we believe that one of the mechanisms through which election monitoring operates is by reducing the incidence of ballot-stuffing (inflating the number of "cast" votes in the election)? Why or why not?

I think it is because when reduing the ballot-stuffing, each people are more likely to vote onece instead of voting maybe multiple time intentionally or by mistake. As a result, it is more easy to make an objective result about the election.

## Problem 2

Consider an experiment with $N$ units. Each unit $i$ in the sample belongs to one of $G$ mutually exclusive strata. $G_i = g$ denotes that the $i$th unit belongs to stratum $g$. $N_g$ denotes the size of stratum $g$ and $N_{t,g}$ denotes the number of treated units in that stratum. Suppose that treatment is assigned via block-randomization. Within each stratum, $N_{t,g}$ units are randomly selected to receive treatment and the remainder receive control (complete randomization). Suppose also that the proportion of treated units in each stratum, $\frac{N_{t,g}}{N_g}$, varies depending on the stratum. After treatment is assigned, you record an outcome $Y_i$ for each unit in the sample. Assume consistency holds with respect to the potential outcomes:

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0)$$

Let $w(g) = P(D_i = 1 | G_i = g)$ denote the known (constant) probability that unit $i$ would receive treatment if it's stratum membership is $g$.

Instead of using the stratified difference-in-means estimator, your colleague suggests an alternative that assigns a weight to each unit and takes two weighted averages.

$$\hat{\tau}_w = \frac{1}{N} \sum_{i=1}^{N} \frac{D_i Y_i}{w(G_i)} - \frac{(1 - D_i) Y_i}{1 - w(G_i)}$$

Show that $\hat{\tau}_w$ is unbiased for the average treatment effect $\tau$:

$$\tau = E[Y_i(1) - Y_i(0)]$$

Hints:

- For a unit with $G_i = g$, the probability of receiving treatment is $\frac{N_{t,g}}{N_g}$
- Consider splitting up the sum from $i = 1$ to $N$ to a double-sum over each of the $G$ groups and over the units $i$ in group $G_i = g$ (since the groups are exhaustive and mutually exclusive, you can think of this as partitioning of the set of $i$ observations by their group membership).

$$\hat{\tau}_w = \frac{1}{N}\sum_{i=1}^{N}\frac{D_iY_i}{w(G_i)} - \frac{(1-D_i)Y_i}{1-w(G_i)} = \frac{1}{N^t}\sum_{(i:D_i=1)} Y_i\,(1)\,\frac{N_tg}{N_g} - \frac{1}{N_cg}Yi(0)\frac{N_c}{N_g}$$

$$= \frac{1}{N_g}\sum_{(i:D_i=1)} E\,[Y_i(1)] - E[Y_i(0)] = \frac{N\tau}{N} = \tau$$

## Problem 3

For this question you will need the `SierraLeone_data.dta` dataset (available on the course website) based on a field experiment conducted by Casey et al 2012. You will re-analyze this experiment using Fisher's randomization inference.

Aid organizations in developing countries spend billions of dollars every year promoting accountability, competence and inclusion of under-representated groups. Arguably the most popular strategy for these efforts has been community-driven development (CDD) programs. CDD projects are initatives that attempt to bolster local coordination and enhance political participation by providing financial grants for local public goods and small entrepise development.

Casey et al 2012 explore the effectiveness of a CDD program in post-conflict Sierra Leone. The researchers block-randomized treatment (access to the CDD program) at the village level. That is, within each block (here chiefdoms) consisting of $N_g$ villages, the researchers randomly assigned $N_{t,g}$ villages to receive treatment. Overall, $N_t = 116$ villages received the treatment out of a total of $N = 233$.

```
# Read in the data
sierraLeone <- read_dta("SierraLeone_data.dta")
```

The variables you will need are:

- `communitybank` - Whether the village has a community bank by the end of the experiment
- `treat_control` - The village's treatment status
- `chief_2004census` - The census area (chiefdom) at which block randomization was conducted
- `id_vill` - Unique village identifier

### Part A

Estimate the average treatment effect of the CDD program on the probability that a village has a community bank using a stratified difference-in-means estimator (the strata here being chiefdoms).

```
sierraLeone = subset(sierraLeone, is.na(sierraLeone$community_bank) == FALSE)
sierraLeone$stratum[sierraLeone$chief_2004census == "biriwa"] = 1
```

```
## Warning: Unknown or uninitialised column: `stratum`.
```

```
sierraLeone$stratum[sierraLeone$chief_2004census == "bombali shebora"] = 2
sierraLeone$stratum[sierraLeone$chief_2004census == "bothe town"] = 3
sierraLeone$stratum[sierraLeone$chief_2004census == "bum"] = 4
sierraLeone$stratum[sierraLeone$chief_2004census == "gbanti kamaranka"] = 5
sierraLeone$stratum[sierraLeone$chief_2004census == "gbendenbu ngowahun"] = 6
sierraLeone$stratum[sierraLeone$chief_2004census == "imperi"] = 7
sierraLeone$stratum[sierraLeone$chief_2004census == "jong"] = 8
sierraLeone$stratum[sierraLeone$chief_2004census == "makari gbanti"] = 9
sierraLeone$stratum[sierraLeone$chief_2004census == "nongoba bullom"] = 10
sierraLeone$stratum[sierraLeone$chief_2004census == "safroko limba"] = 11
sierraLeone$stratum[sierraLeone$chief_2004census == "sanda loko"] = 12
sierraLeone$stratum[sierraLeone$chief_2004census == "sanda tendaren"] = 13
sierraLeone$stratum[sierraLeone$chief_2004census == "sella limba"] = 14

stratum_ests = bind_rows(map(unique(sierraLeone$stratum), function(x)
  diff_in_means_stratum(sierraLeone$community_bank[sierraLeone$treat_control
== 1&sierraLeone$stratum == x],sierraLeone$community_bank[sierraLeone$treat_c
ontrol == 0&sierraLeone$stratum == x], x))) %>% select(stratum, est, se, N_g
= N)


N_all = sum(stratum_ests$N_g)
point_stratum = sum(stratum_ests$est*(stratum_ests$N_g/N_all))
point_stratum
```

```
## [1] 0.7108099
```

```
var_stratum = sum((stratum_ests$se^2)*(stratum_ests$N_g/N_all)^2)
se_stratum = sqrt(var_stratum)
se_stratum
```

```
## [1] 0.04506488
```

```
ci_95_stratum = c(point_stratum - qnorm(.975)*se_stratum,
                  point_stratum + qnorm(.975)*se_stratum)
ci_95_stratum
```

```
## [1] 0.6224844 0.7991354
```

```
stratum_ests
```

```
## # A tibble: 14 x 4
##    stratum   est     se   N_g
##      <dbl> <dbl>  <dbl> <int>
##  1       1 1      0        12
##  2      11 0.667 0.142     23
##  3      12 0.633 0.260     11
##  4       2 0.909 0.0909    23
##  5      13 0.2   0.2       10
##  6      14 0.333 0.279     12
##  7       6 0.833 0.0904    34
##  8       5 0.833 0.167     12
##  9       9 0.667 0.211     12
## 10       4 0.636 0.168     22
## 11       7 0.6   0.245     11
## 12       8 0.735 0.148     23
## 13      10 0.8   0.2       11
## 14       3 0.667 0.333      7
```

## Part B

Let's obtain a p-value under the sharp null of no treatment effect using randomization inference. We'll use the absolute value of the difference-in-means as our test statistic.

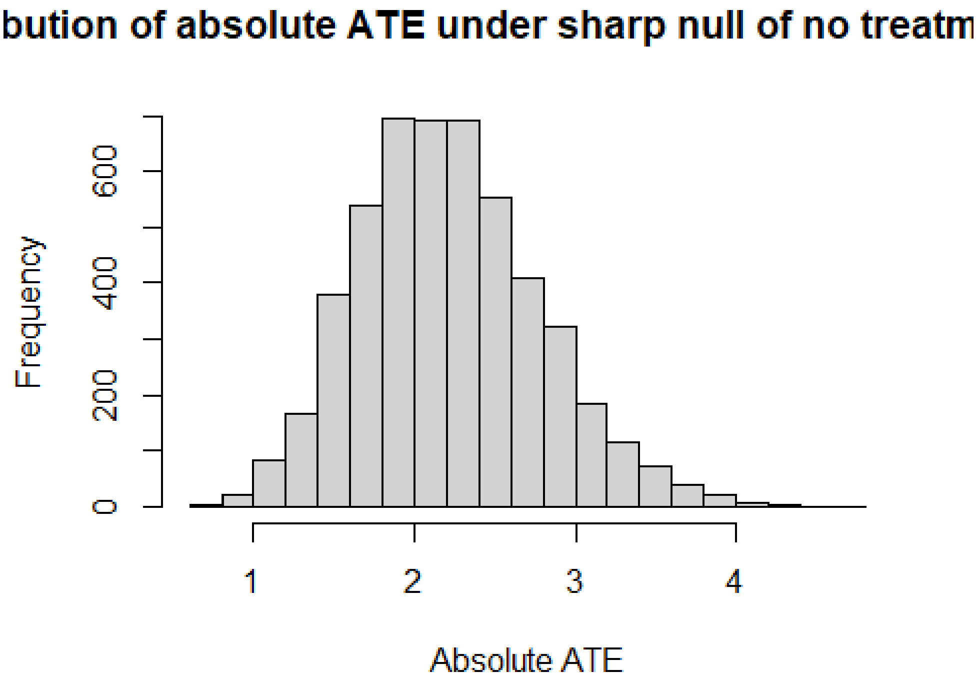
Assume $N_{t,g}$ is fixed for each block and that within each block, the researchers assigned treatment using complete randomization.

Approximate the randomization distribution of the absolute difference-in-means given the stratified randomization procedure (fixed $N_{t,g}$ for each stratum and complete randomization within each stratum). Use a simulation with 5000 total draws. Set your random seed to 10003 prior to the start of the simulation.

Make a histogram of your draws.

```r
diff_in_means_stratum_abs <- function(treated, control, stratumname="all"){
  # Point Estimate of absolute value of difference
  point <- abs(sum(treated)-sum(control))

  # Standard Error
  se <- sqrt(var(treated)/length(treated) + var(control)/length(control))

  # Asymptotic 95% CI
  ci_95 <- c(point - qnorm(.975)*se,
             point + qnorm(.975)*se)

  # P-value
  pval <- 2*pnorm(-abs(point/se))

  # Return as a data frame
  output <- data.frame(stratum = stratumname, meanTreated = mean(treated), me
anControl = mean(control), est = point, se = se, ci95Lower = ci_95[1], ci95Up
per = ci_95[2], pvalue = pval, N= length(treated) + length(control))

  return(as_tibble(output))

}
```

```
set.seed(10003)
Niter = 5000
null_diff = rep(NA, Niter)
for (i in 1:Niter){
  sierraLeone$treat_control_permutation <- sample(sierraLeone$treat_control)

  stratum_ests <- bind_rows(map(unique(sierraLeone$stratum), function(x)
    diff_in_means_stratum_abs(sierraLeone$community_bank[sierraLeone$treat_co
ntrol_permutation == 1&sierraLeone$stratum == x],
                        sierraLeone$community_bank[sierraLeone$treat_control_
permutation == 0&sierraLeone$stratum == x], x))) %>%
    select(stratum, est, se, N_g = N)

  N_all <- sum(stratum_ests$N_g)
  point_stratum[i] <- sum(stratum_ests$est*(stratum_ests$N_g/N_all))
}
hist(abs(point_stratum), xlab="Absolute ATE", main = "Distribution of absolut
e ATE under sharp null of no treatment effect")
```



## Part C

Calculate a p-value for the observed test statistic under the sharp null of no treatment effects using your randomization distribution from Part B. Interpret your results. Would you reject the null at the $\alpha = .05$ level?

```
stratum_results <- sierraLeone %>% group_by(stratum)  %>% summarize(Ng =n(),
groups = "community_bank")

## `summarise()` ungrouping output (override with `.groups` argument)

#stratum_results
stratum_results$effect <- sapply(stratum_results$stratum, function(x)
mean(sierraLeone$stratum[sierraLeone$treat_control_permutation == 1&sierraLeo
ne$stratum == x]) -
mean(sierraLeone$stratum[sierraLeone$treat_control_permutation == 0&sierraLeo
ne$stratum == x]) )


effect_strat <- sum(stratum_results$effect*(stratum_results$Ng/sum(stratum_re
sults$Ng)))

mean(point_stratum)

## [1] 2.226332

abs(effect_strat)

## [1] 0

mean(point_stratum) - abs(effect_strat)

## [1] 2.226332
```

We get a p-value of 2.226332 – it is very unlikely that we would observe the difference between the two treatments that’s as extreme as we did under the sharp null of no treatment effect. We’d reject the null at α = .05.