midterm

ZiqianHe

Question 1

1. Define and calculate the appropriate measure for the effect of the treatment and interpret it.

A=1: assigned new treatment

A=0: assigned standard treatment

Y=1: disease prevented

Y=0: disease not prevented

$$ACE = E[Y_1] - E[Y_0] = \frac{12}{20} - \frac{5}{20} = \frac{7}{20}$$

new treatment is better than standard treatment on average. The average causal effect is

$$\frac{7}{20}$$

2. Write the formula for the appropriate measure for this study, calculate it and interpret it.

$$ACE = E[Y|A_1] - E[Y|A_0] = \frac{5}{10} - \frac{4}{10} = \frac{1}{10}$$

Difference in observed group means is

$$\frac{1}{10}$$

3. Compare this estimate with what you obtained from Question 1.

- In question2, the difference in observed group means is 0.1, which is smaller than the average causal effect in question1.
- We need the assignment mechanism contain individualistic, probabilistic, unconfounded, known and controlled to ensure the difference in observed group means equal to ACE. In study 1, we don't know the assignment mechanism, the study didn't match all the mechanism. The difference between question 1 and question 2 might due to potential confounders for example maybe those who got standard treatment may be healthier.

4. Explain how this type of data might arise in (a) an observational study and (b) a randomized controlled trial.

- `a)` observational study: decide eligibility criteria(including inclusion and exclusion criteria), select 10 patients from each group record. Get all data together (covatiates, treatment, ourcomes) the treatment is not randomized and it will be the case that individuals select to take the active treatment based on their health condition.

- b)randomized controlled: the design phase is done before access to outcomes and analysis. Decide eligibility criteria(including inclusion and exclusion criteria). Select 20 patients based on eligibility criteria, assign 10 of them to new treatment and 10 of them to standard treatment randomly and record.

5. Can you rule out a particular study design given what you observe?

- In a randomized experiment, the assignment mechanism is regular (unconfounded, individualistic, probabilistic, controlled) by design. So the association between A and Y is the same as the average causal effect. While the observational study may not hold all the mechanism. So we need to rule out this.
- While if the sample size is too small maybe the association between A and Y do not equal to ACE because covariate. In this situation we could not rule out observaltional study.

6. Please describe your process in words and provide a table showing your study (i.e., your observed data) under this assignment mechanism.

In block randomizarion, units can be partitioned into strta with similar condition. The treatment assignment is randomized with in each block to ensure the unconfounded and probalistics. In reality we will stratify by baseline covatiates such as gender, education... but here we know the counterfactual outcomes.

I want to partition the units into 4 different blocks that each blocks units have similar health situation base on the truth and do randomization to assign treatment within each blocks to ensure that `every unit has a positive probability of being in either treatment` and `assihnment does not depend on the potential outcome`.

- `outcome all 1(seems healthy)` block in total 4 people and random assign to treatment (0 0 1 1)
- `outcome all 0(seems unhealthy)` block in total 7 people and random assign to treatment (0 1 1 1 0 0 0)
- `outcome Y1=0, Y0=1(seems uneffect)` block in total 1 people and random assign to treatment (1)
- `outcome Y1=1, Y0=0(seems effect)` block in total 8 people and random assign to treatment (0 1 0 1 0 1 1 0)

So my table is

```
##     A Y|A=1 Y|A=0  type
##  1: 0     .     1     H
##  2: 0     .     1     H
##  3: 1     1     .     H
##  4: 1     1     .     H
##  5: 0     .     0   unh
##  6: 1     0     .   unh
##  7: 1     0     .   unh
##  8: 1     0     .   unh
##  9: 0     .     0   unh
## 10: 0     .     0   unh
## 11: 0     .     0   unh
## 12: 1     0     . uneff
## 13: 0     .     0   Eff
## 14: 1     1     .   Eff
## 15: 0     .     0   Eff
## 16: 1     1     .   Eff
```

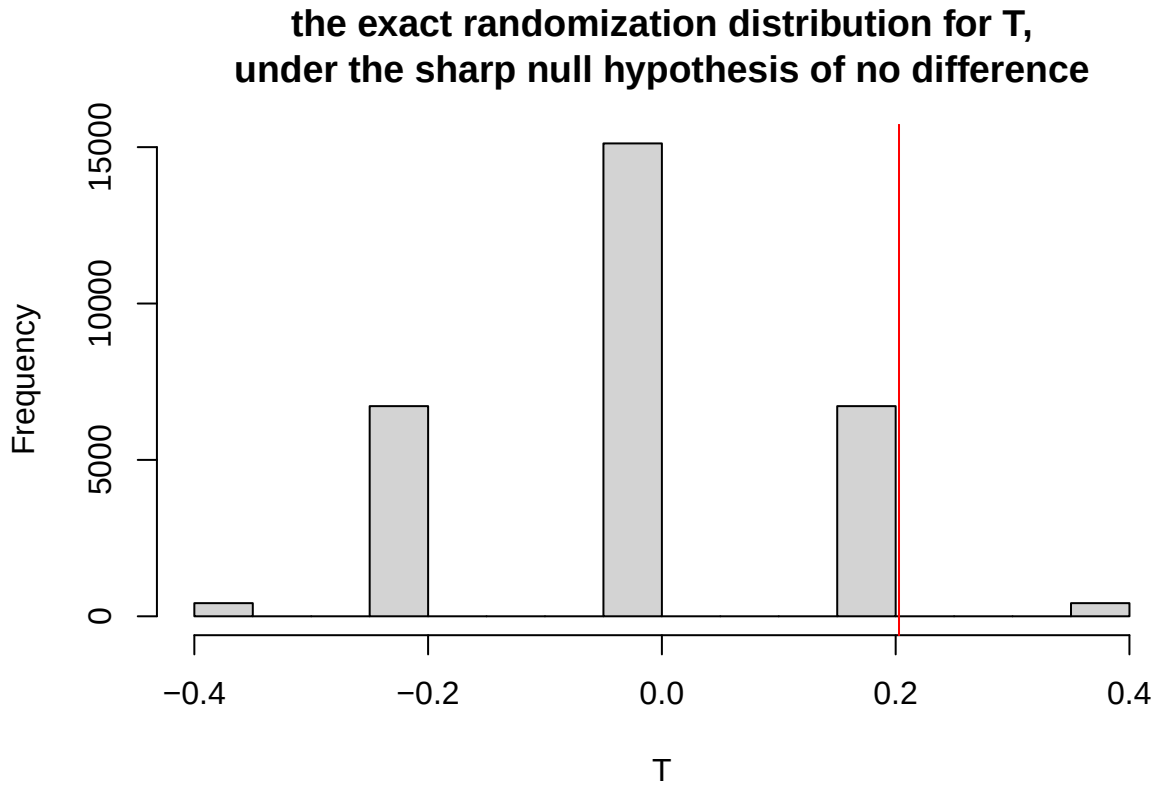
```
## 17: 0     .     0   Eff
## 18: 1     1     .   Eff
## 19: 1     1     .   Eff
## 20: 0     .     0   Eff
```

7. Write and test the sharp null hypothesis of no causal effect in your study (describe your process and plot the randomization distribution). Interpret your results.

`Sharp null hypothesis:` there is no treatment effect, so the vector of observed outcomes Y dose not change with different A. $H_0 : \tau_i = Y_{0i} - Y_{1i} = 0$

to calculate conveniently, I combine the two blocks with odd number of individuals(outcome all 0 and outcome Y1=0, Y0=1). So now I have three block: 4 people for Y1=1 and Y0=1, 8 people for Y1=0 and Y0=0 or Y1=0 and Y0=1, and 8 people for Y1=1 and Y0=0.

The number of total possible randomizations is $\binom{4}{2}\binom{8}{4}\binom{8}{4} = 29400$. Then obtain the exact randomization distribution and calculate the p-value as the proportion of test tatistic at least as extreme as my observed test



statistics.
`interpret` p-value is $0.0142857 < 0.05$, we reject the sharp null hypothesis and conclude that there is individual causal effect of treatment type on disease prevention.

8. Provide the point estimate and confidence interval inverting the hypothesis testing procedure you conducted in (7). Describe the procedure and interpret your results.

code in appendix

need to create a grid of possible sharp null hypotheses and calculate p-value for each sharp null and pick the value that is "least surprising" under the null as point estimate as well as confidence interval.

```
## [1] 0.1666998 0.6666722
```

The point estimate is 0.395 so those who have the new treatment have more cases of disease prevented than those who have standard treatment on average and 95 %confidence interval is [0.1666998,0.6666722]

9. Provide the point estimate and confidence interval of the marginal average causal effect in your study using Neyman's approach. Interpret your results.

code in appendix

```
## [1] 0.8769046
```

```
## [1] -0.0769046
```

Point estimate = $E[Y|A=1] - E[Y|A=0] = 0.4$

Sampling variance estimate = $\frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} = 0.044$

95% CI = $0.4 \pm 2.26 \times \sqrt{0.044} = [-0.0769046, 0.8769046]$

Therefore, the point estimate of the marginal causal effect is 0.4, means that who have the new treatment have more cases of disease prevented than those who have standard treatment on average, and the 95%CI is [-0.0769046, 0.8769046] using Neyman's approach. The p-valis $0.0451337 < 0.05$ so reject the null hypothesis and conclude that there is individual causal effect of treatment type on disease prevention.

10. Compare the estimates you obtained in (8) and (9) with what you computed in (1).

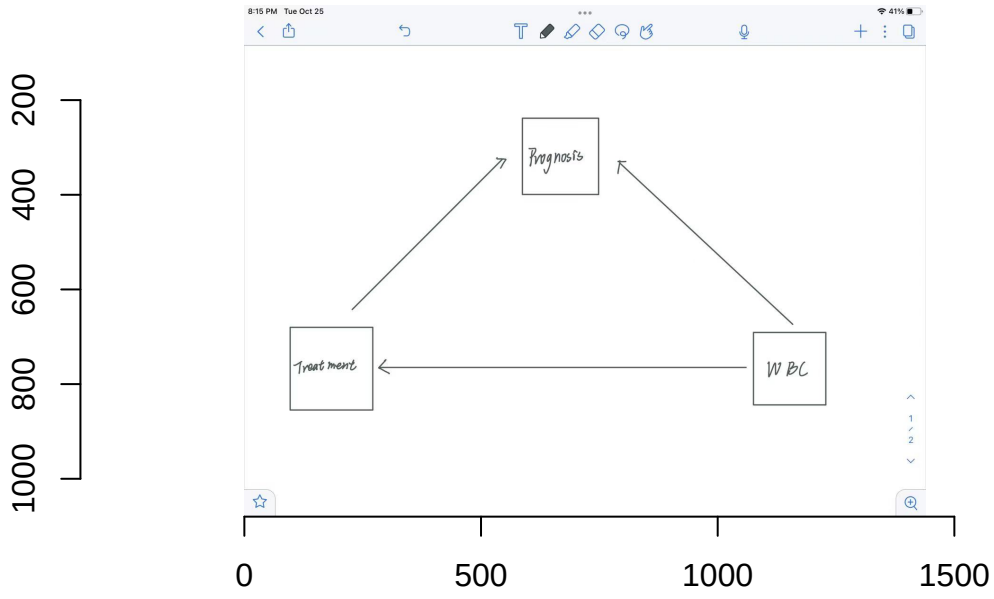
For (8) and (9), the estimate of the causal effect greater than to 0 so those who have the new treatment have more cases of disease prevented than those who have standard treatment on average. and through the p-value we reject the null hypothesis and conclude that there is treatment effect on average at 5% significance level.

`compare with 1`: Base on the truth, I designed the block randomized trial to ensure probabilistic, unconfounded, exchangeability, positivity, SUTVA/consistency and obtained the point estimate similar to ACE in question 1.

11. scientific question of interest for this observational study under Neyman's approach to inference

what is the difference between these average, the average causal effect, if all units were in new treatment versus all units were in standard treatment.

12. Represent the situation described in a DAG.



* `Treatment`: treatment that individuals have new treatment or standard treatment * `Prognosis`: outcome that individuals have diseased prevented or not * `WBC`: confounder that individuals have normal white blood cell

13. What does the DAG imply about the crude association? leaving L unadjusted will?

- Because the WBC is a confounder, the new treatment and the standard treatment are not comparable and didn't follow all the mechanism. So the crude association between the treatment A and ourcome Y will be different from the average causal effect.

- b)Bias in the estimate such that it overestimates the average causal effect (i.e., the estimate is larger in magnitude than the true causal effect).
  Because L=1, with normal WBC, individuals are more likely to assigned to new treatment while these people also more likely to have a better disease prognosis. So even if the effect of different treatment and outcome are the same, people in new treatment group are more likely to have better outcome and this will overestimates the ACE.

14. Compute the estimate and confidence interval for ACE using the g-formula for observational studies.

code in appendix

g-formula: $E(Y_a) = \sum_c E(Y|A = a, C = c)Pr(C = c)$

$E(Y_1) = E(Y|A=1, L=1)Pr(L=1) + E(Y|A=1, L=0)Pr(L=0) = 0.575.$

$E(Y_0) = E(Y|A=0, L=1)Pr(L=1) + E(Y|A=0, L=0)Pr(L=0) = 0.2.$

$E(Y_1) - E(Y_0) = 0.375.$

estimate of ACE is 0.375 so new treatment is better than standard treatment on average.

Sampling variance estimate = $\frac{S_1^2}{N_1} + \frac{S_0^2}{N_0}$

95% CI = $0.375 \pm z^* \times \sqrt{0.021}$

The confidence interval is 0.0713122 to 0.6786878

```
##                         V1              mean                se
## 1                 Observed               0.4 0.069372184627558
## 2             No Treatment 0.214285714285714 0.139531465245289
## 3                Treatment 0.585714285714286 0.118264845656253
## 4 Treatment - No Treatment 0.371428571428572  0.22157075028267
##                     ll                ul
## 1   0.264033016601123 0.535966983398877
## 2 -0.0591909323051546 0.487762360876583
## 3   0.353919447590841 0.817509123837731
## 4 -0.0628421191529803 0.805699262010124
```

- With boostrap, estimate of ACE is 0.371428571428572 so new treatment is better than standard treatment on average and new treatmet group have a better disease prognosis.
- The confidence interval is [-0.0628421191529803, 0.805699262010124] we don't have a positivity violation (given that Pr(C=1)>0 then we should see Pr(A=1|C=1)>0).

15. Compare your result with what you obtained in Question 1.

- The result is similar to question 1.
- Since we don't have positivity violatio and follow the assumption below.
- All confounders of the relationship between treatment A and Y be in the model and seems that L is the only confounder.
- No unobserved confounding assumption: L suffices to account for confounding, and therefore within levels of L, it is as if A were randomized.
- Consistency Assumption
- Positivity assumption: Every unit has some chance of being assigned to either treatment group, conditional on covariates.
- No measurement error

16. Do you find support for your hypothesis in the data? Please explain why or why not and, where applicable, show any additional calculations you performed to reach your conclusion.

code in appendix

- The ACE calculated in question 14 is 0.371428571428572 > 0,so new treatment have better treatment effect than standard treatment. So the hypothesis that the new treatment is better for disease prevention than the standard treatment is true.
- The probaility of being prescribed the new treatment among individuals with normal white blood cell is 0.6666667. The probaility of being prescribed the new treatment among individuals with abnormal white blood cell is 0.4. The formal probability is greater than the later one, so the hypothesis that

individuals with normal white blood cell (WBC) counts (L=1) are more likely to be prescribed the new treatment compared with individuals with abnormal WBC counts (L=0) is ture.

$$\frac{Pr(L=1, A=1)}{Pr(L=1)} > \frac{Pr(L=0, A=1)}{Pr(L=0)}$$

- The probability of having a better disease prognosis with normal white blood cell is 0.5333333. The probability of having a better prognosis with abnormal white blood cell is 0.32. The formal probability is greater than the later one, the hypothesis that individuals with normal white blood cell (WBC) counts (L=1) are more likely to have a better disease prognosis compared with individuals with abnormal WBC counts (L=0) is true.

$$\frac{Pr(L=1, Y=1)}{Pr(L=1)} > \frac{Pr(L=0, Y=1)}{Pr(L=0)}$$

17. Identify a variable or set of variables in the DAG that when conditioned on would close all back-door paths between A and Y.

Conditioning on L, U, F and B will close all back-door paths between A and Y.

18. What is the relationship between NUCA assumption of potential outcomes and a DAG?

Potential outcomes framework and DAGs help formalizing definition of causal effects, clarifying assumptions, and reason on whether such assumptions are met.

The assumption of no unobserved confounding(NUCA) essentially states that the observed C suffices to account for confounding, and therefore within levels of C, it is as if A were randomized (by nature).

For regression coefficients to have a causal interpretation we need both that 1) The linear regression to be correctly specified 2) All confounders of the relationship between treatment A and Y is in the model. So we need NUCA assumption of potential outcome and a correctly specified DAG.

19. Identify a variable in the DAG that when conditioned on would open a closed path from A to Y.

H

when conditioned on H would open a closed path from A to Y.

20. Conceptually, what is a collider and why is it problematic to adjust for a collider? Can you provide an example of a collider in the DAG?

H

Collider: A node on a path with both arrows on the path going into that node. Collider associated with the treatment and associated with the ourcome comditional on the treatment. Conditioning on the collider creates an association between outcome and treatment so will open a closed path. We need to block all back-door paths by conditioning to calculate causal effect. So adjusting for a collider is problematic.

Question 2

1. Identify the units, potential outcomes, treatment, and any observed covariates.

- `units`: each hospital in the survey.i = number of hospital, t = 1,2 where 1 means the baseline and 2 means after two years.
- `The potential outcomes`: the number of doctors from minority backgrounds that were promoted to leadership positions after 2 years.
- `The treatment`: whether being given workshop, A=1 means have workshop and A=0 means don't have workshop.
- `observed covariates`: whether have majority of white doctors in leadership positions at the baseline.opinions of hospital administrators towards the workshop(whether request to have workshop).

2. Define in words and with a mathematical formula the causal effect you would be interested in studying.

The causal effect of treatment on the outcome. Which means: is there difference between having workshop and do not have workshop, regarding the number of doctors from minority backgrounds that were promoted to leadership positions.

$$ACE = E(Y_1) - E(Y_0)$$

where 1 means have workshop. But in this question is an observational study so we could only get the association effect, means the difference in observation group.

3. How would you describe the study design in terms of the assignment mechanism? Is this study design appropriate to address your question?

- This study is individualistic but not probabilistic, unconfounded and controled.
- not probabilistic: cause hospital with majority of white doctors in leadership positions must assign to treatment that have the workshop so it's not probabilistic.
- not unconfounded: since the assignment depend on potential outcomes. Department want these hsopital have doctors from minority backgrounds are employed and have leadership positions so they give workshop to hospital that with majority of white doctors in leadership positions.
- not controled: this is an observational study
- individualistic: assignment of unit do not depend on the covariates or potential outcomes of other units

4. If you were to advise the New York State Department of Health, how would you suggest them to proceed (from design to analysis) in other to quantify the causal effect of interest?

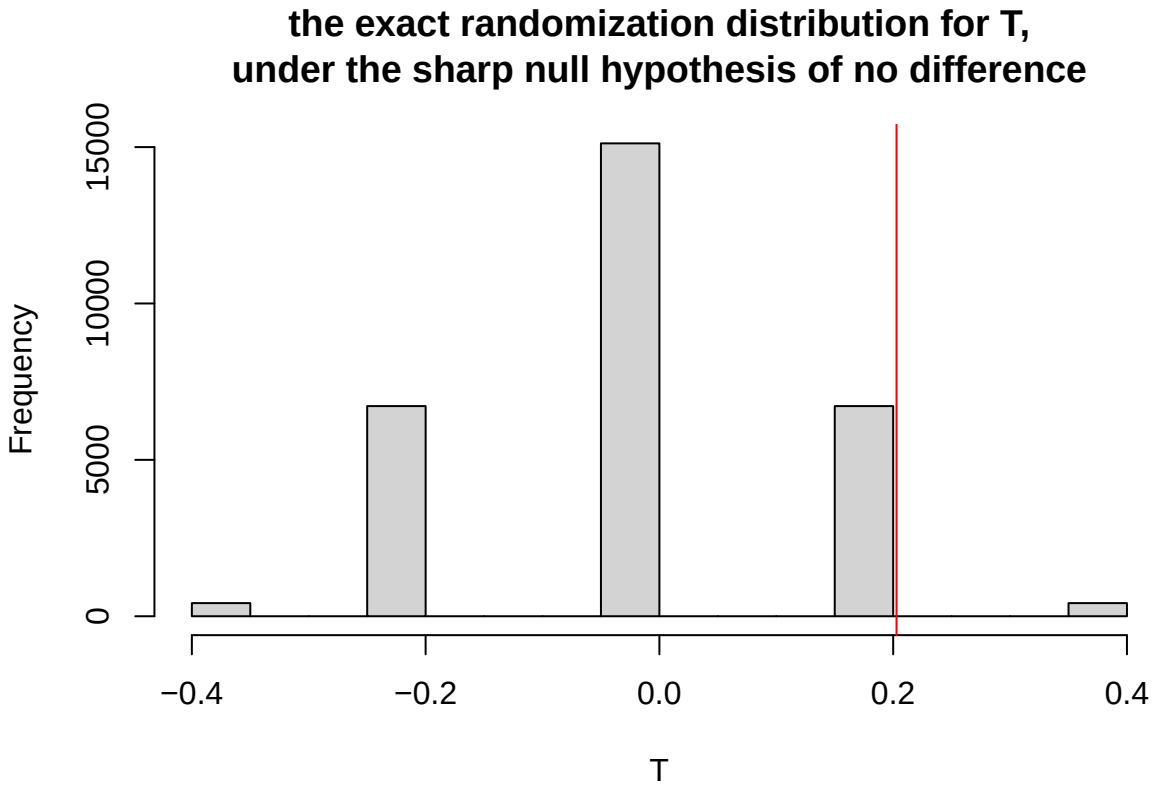
Need to match all the assumption. Decide eligibility criteria(including inclusion and exclusion criteria). Select hospitals based on eligibility criteria, assign them to having workshop and do not have workshop randomly and record. Or stratified. Hospitals are partitioned into blocks with the covatiates and are completelt randomized within each blocks.

Appendix

question 1

```r
Y1 = c(1,1,1,0,1,0,1,1,0,0,1,0,1,0,0,1,1,1,0,1)
Y0 = c(0,0,1,0,1,0,0,0,0,1,1,0,0,0,0,1,0,0,0,0)
individual = c(1:20)
dt = data.table(individual,Y1,Y0)

## 1.6
## outcome all 1(healthy)
n = nrow(dt[Y1==1&Y0==1]) #4
rbinom(n, 1, 0.5)

## outcome all 0(unhealthy)
n = nrow(dt[Y1==0&Y0==0]) #7
rbinom(n, 1, 0.5)

## outcome Y1=1, Y0=0(effect)
n = nrow(dt[Y1==1&Y0==0]) #8
rbinom(n, 1, 0.5)

## outcome Y1=0, Y0=1(noneffect)
n = nrow(dt[Y1==0&Y0==1]) #1
rbinom(n, 1, 0.5)

A = c(0,1,0,0,0,1,0,1,1,1,1,1,0,0,0,1,1,1,0,0)
dt = data.table(A,Y0,Y1) %>%
  mutate("Y|A=1" = ifelse(A==1,Y1,"."),
         "Y|A=0" = ifelse(A==0,Y0,".")) %>%
  mutate(type = ifelse(Y1 == 1 & Y0 ==1 ,"H",
                               ifelse(Y1 == 1 & Y0 == 0, "Eff",
                                      ifelse(Y1 == 0 & Y0 == 0, "unh", "uneff")))) %>%
    select(-Y0,-Y1)
dt$type = factor(dt$type,levels= c("H","unh", "uneff","Eff"))
dt=dt %>% arrange(type)
dt
```

```r
## 1.7
A = dt$A
Y_obs= as.integer(ifelse(dt$`Y|A=1` == ".", dt$`Y|A=0`,dt$`Y|A=1`))
T_obs=mean(Y_obs[A==1])-mean(Y_obs[A==0])
block <- c(rep(1,4), rep(2,8),rep(3,8))
Abold <- genperms(A, blockvar=block, maxiter = 29400)
rdist <- rep(NA, times = ncol(Abold))
for (i in 1:ncol(Abold)) {
  A_tilde <- Abold[, i]
  rdist[i] <- mean(Y_obs[A_tilde == 1]) - mean(Y_obs[A_tilde == 0])
}


pval <- mean(rdist >= T_obs)
quant <- quantile(rdist,probs = 1-pval)
hist(rdist,xlab="T",main="the exact randomization distribution for T,\nunder the sharp null hypothesis
abline(v = quant,col="red")
```




```r
## 1.8
grid<-seq(-1,1, by=0.01)
p.ci<-rep(NA,length(grid))
rdist2 <- rep(NA, times = ncol(Abold))
for (i in 1:length(grid)){
for (k in 1:ncol(Abold)) {
  A_tilde <- Abold[, k]
  rdist2[k] <- mean(Y_obs[A_tilde == 1]) - mean(Y_obs[A_tilde == 0])+grid[i]
}
  p.ci[i]<-mean(rdist2 >= T_obs)
}

estimate = mean(grid[which(abs(p.ci - 0.5) == min(abs(p.ci - 0.5)))])


perms.ci <- Abold ## all possible permutations of assignment to treatment
probs.ci <- genprobexact(A) ## assuming complete randomization

c(invert.ci(Y_obs,A,probs.ci,perms.ci,0.025),invert.ci(Y_obs,A,probs.ci,perms.ci,0.975))
```

```r
## 1.9
t_crit = qt(0.975, 9)
var = var(Y_obs[A==1])/10 + var(Y_obs[A==0])/10
T_obs + t_crit*sqrt(var)
T_obs - t_crit*sqrt(var)
pt(0.4/sqrt(var), df = 9, lower.tail = FALSE)
```

```r
## 1.14
Y = c(1,0,1,0,1,1,0,1,1,1,0,1,1,0,0,0,0,0,0,0,1,0,1,1,0,0,0,0,0,1,0,1,0,0)
A = c(1,0,0,0,1,1,0,0,0,1,0,1,1,0,0,1,1,1,1,1,0,1,1,1,0,0,0,1,0,1,0,0,1,1,0,1,0,0)
L = c(1,1,0,0,1,0,1,0,0,1,0,1,1,0,0,1,0,0,0,1,1,0,0,0,0,0,0,0,0,0,0,1,1,0,0,1,1,1,0,0,0)
p_l1 = sum(L)/40
p_l0 = 1-p_l1
meanY1 = mean(Y[A==1&L==1])*p_l1 + mean(Y[A==1&L==0])*p_l0
meanY0 = mean(Y[A==0&L==1])*p_l1 + mean(Y[A==0&L==0])*p_l0

z = qt(0.975,19)
var = var(Y[A==1])/20 + var(Y[A == 0])/20
lo = meanY1-meanY0 - z*sqrt(var)
hi = meanY1-meanY0 + z*sqrt(var)



dt<- data.table(Y,A,L)

#boostrap
dt$interv <- -1 # 1st copy: equal to original one

interv0 <- dt # 2nd copy: treatment set to 0, outcome to missing
interv0$interv <- 0
interv0$A <- 0
interv0$Y <- NA

interv1 <- dt # 3rd copy: treatment set to 1, outcome to missing
interv1$interv <- 1
interv1$A <- 1
interv1$Y <- NA

onesample <- rbind(dt, interv0, interv1) # combining datasets

standardization <- function(data, indices) {
  # create a dataset with 3 copies of each subject
  d <- data[indices, ] # 1st copy: equal to original one`
  d$interv <- -1
  d0 <- d # 2nd copy: treatment set to 0, outcome to missing
  d0$interv <- 0
  d0$A <- 0
  d0$Y <- NA
  d1 <- d # 3rd copy: treatment set to 1, outcome to missing
  d1$interv <- 1
  d1$A <- 1
  d1$Y <- NA
  d.onesample <- rbind(d, d0, d1) # combining datasets

  # linear model to estimate mean outcome conditional on treatment and confounders
  # parameters are estimated using original observations only (interv= -1)
  # parameter estimates are used to predict mean outcome for observations with set
  # treatment (interv=0 and interv=1)
  fit <- glm(
```

```
    Y ~ A + as.factor(L),
    data = d.onesample
  )

  d.onesample$predicted_meanY <- predict(fit, d.onesample)

  # estimate mean outcome in each of the groups interv=-1, interv=0, and interv=1
  return(c(
    mean(d.onesample$predicted_meanY[d.onesample$interv == -1]),
    mean(d.onesample$predicted_meanY[d.onesample$interv == 0]),
    mean(d.onesample$predicted_meanY[d.onesample$interv == 1]),
    mean(d.onesample$predicted_meanY[d.onesample$interv == 1]) -
      mean(d.onesample$predicted_meanY[d.onesample$interv == 0])
  ))
}

results <- boot(data = dt,
                statistic = standardization,
                R = 5)

# generating confidence intervals
se <- c(sd(results$t[, 1]),
        sd(results$t[, 2]),
        sd(results$t[, 3]),
        sd(results$t[, 4]))
mean <- results$t0
ll <- mean - qnorm(0.975) * se
ul <- mean + qnorm(0.975) * se

bootstrap <-
  data.frame(cbind(
    c(
      "Observed",
      "No Treatment",
      "Treatment",
      "Treatment - No Treatment"
    ),
    mean,
    se,
    ll,
    ul
  ))

bootstrap
```

```
## 1.16
p_A1_L1 = nrow(dt[A==1&L==1])/sum(L)
p_A1_L0 = nrow(dt[A==1&L==0])/(40-sum(L))
p_Y1_L1 = nrow(dt[Y==1&L==1])/sum(L)
p_Y1_L0 = nrow(dt[Y==1&L==0])/(40-sum(L))
```