Midterm

Xinyi Lin

10/24/2019

Question 1

Question 1

Individual	Y1	Y0	individualCE
1	1	0	1
2	1	1	0
3	1	0	1
4	0	0	0
5	1	1	0
6	1	1	0
7	1	0	1
8	0	0	0
9	1	0	1
10	0	0	0
11	0	1	-1
12	0	0	0
13	0	0	0
14	1	1	0
15	1	0	1
16	0	1	-1
17	1	0	1
18	1	0	1
19	0	0	0
20	1	0	1

$ACE = E[Y_1 - Y_0] = E[Y_1] - E[Y_0] = \frac{12}{20} - \frac{6}{20} = \frac{3}{10}$

Interpretation: As the average causal effect $\frac{3}{10} > 0$, treatment 1(new treatment) is better than treatment 0(standard treatment) on average.

Question 2

$E[Y|A = 1] - E[Y|A = 0] = \frac{5}{10} - \frac{3}{10} = \frac{2}{10}$

Question 3

Compared to what we get in question 1, the estimator in question 2 is smaller. The value calcualted in question 1 is average causal effect while the value calculated in question 2 is association.

In question 1, as we know all the potential outcomes, each individual is assigned to both new treatment and standard treatment, so the probability a unit is assigned to different treatments are 1 and doesn't influence by covariates or potential outcomes of the other units, so the assignment mechanism is individualistic, probabilistic, unconfounded, know and controlled. However, in study 1, we don't know the assignment mechanism, so the study does not have individualistic, probabilistic, unconfounded, know and controlled properties. The difference between question 1 and question 2 might due to potential confounders.

Question 4

a) Assume some patients have treatment1(new treatment), some patients have treatment0(standard treatment). In an obervational study, What the study might do is first, decide eligibility criteria(including inclusion and exclusion criteria) and then, select 10 patients from treatment1 group, record their outcomes and select 10 patients from treatment0 group, record their outcomes.

b) In a randomized controlled trial, what the study might do is first, decide eligibility criteria(including inclusion and exclusion criteria). Then, select 20 patients based on eligibility criteria, randomly assign 10 of them to treatment1 group and 10 of them to treatment0 group and record outcomes of these 20 patients.

The key different between an observational study and a randomizad controlled trial is whether the study contains a randomization process.

Question 5

If we assume in a randomized controlled trial, assumptions of individualistic, probabilistic, unconfounded, know and controlled are acheived and the association between A and Y is the same as the average causal effect of A on Y, then a randomized controlled trial can be ruled out. Beacuse what I calculate in question 2 is not the same as what I calculate in question 1.

If we assume that even in a randomized controlled trial, there might be some baseline covariates that cannot be balanced and might cause difference of the association between A and Y and the ACF, then we cannot rule out a particular study design.

Question 6

With a block randomized design, the experimenter divides subjects into subgroups called blocks, such that the variability within blocks is less than the variability between blocksr with respect to one or more covariates. Then, subjects within each block are randomly assigned to treatment conditions.

As we do not have covariate informationssume, I assume first 10 of these 20 individuals are male, others are female and except for gender, they are similar in other covariates like age, health status. In order to do randomization using a block randomization method, I will

1)divide these 20 individuals into female group and male group and give them ID from 1-10 sepretaly.

2)use computational software to randomly draw 5 numbers between 1-10 without replacement from each group.

3)corresponding 5 individuals in each group will be assigned to treatment1 and rest will be assigned to standard treatment.

The observed data I get from above process is as following:

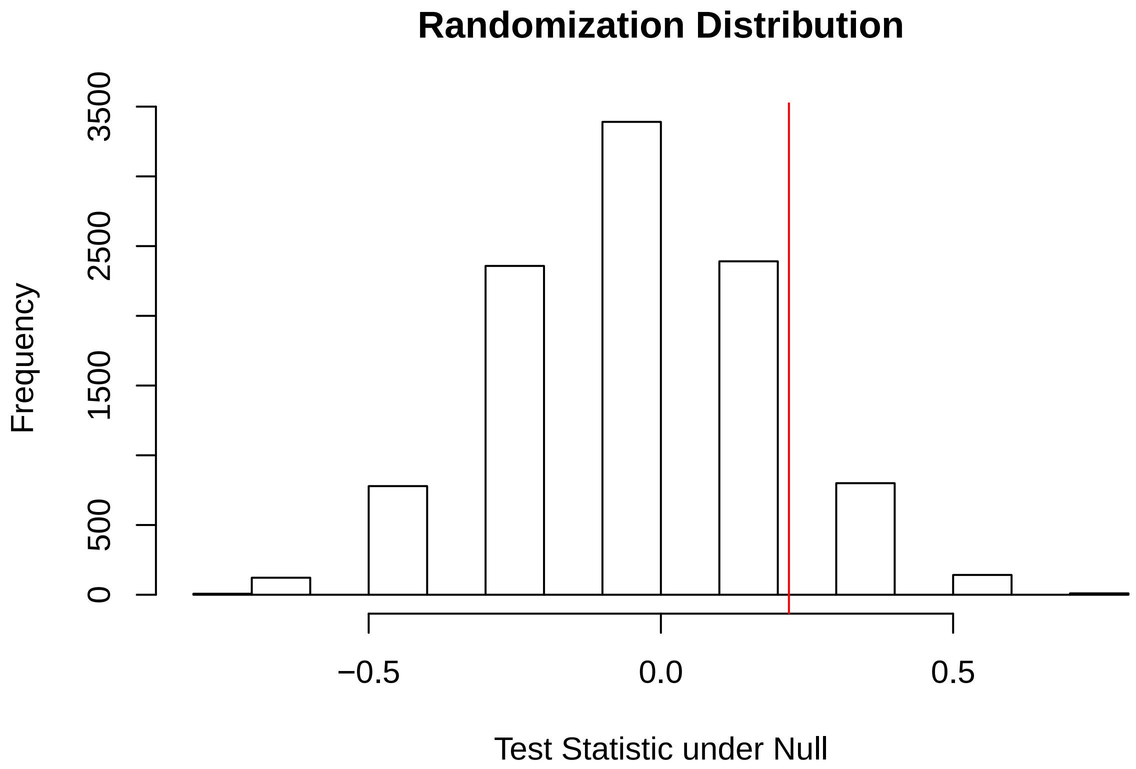
Individual	A	Y|A=1	Y|A=0
1	0	.	0
2	1	1	.
3	1	1	.
4	1	0	.
5	0	.	1
6	1	1	.
7	0	.	0
8	1	0	.
9	0	.	0
10	0	.	0
11	0	.	1
12	1	0	.
13	0	.	0
14	1	1	.
15	1	1	.
16	0	.	1
17	1	1	.
18	1	1	.
19	0	.	0
20	0	.	0

Question 7

The null hypothesis is $Y_{0i} - Y_{1i} = 0$. The test statistic is $T_i = \bar{Y}_1^{obs} - \bar{Y}_0^{obs}$

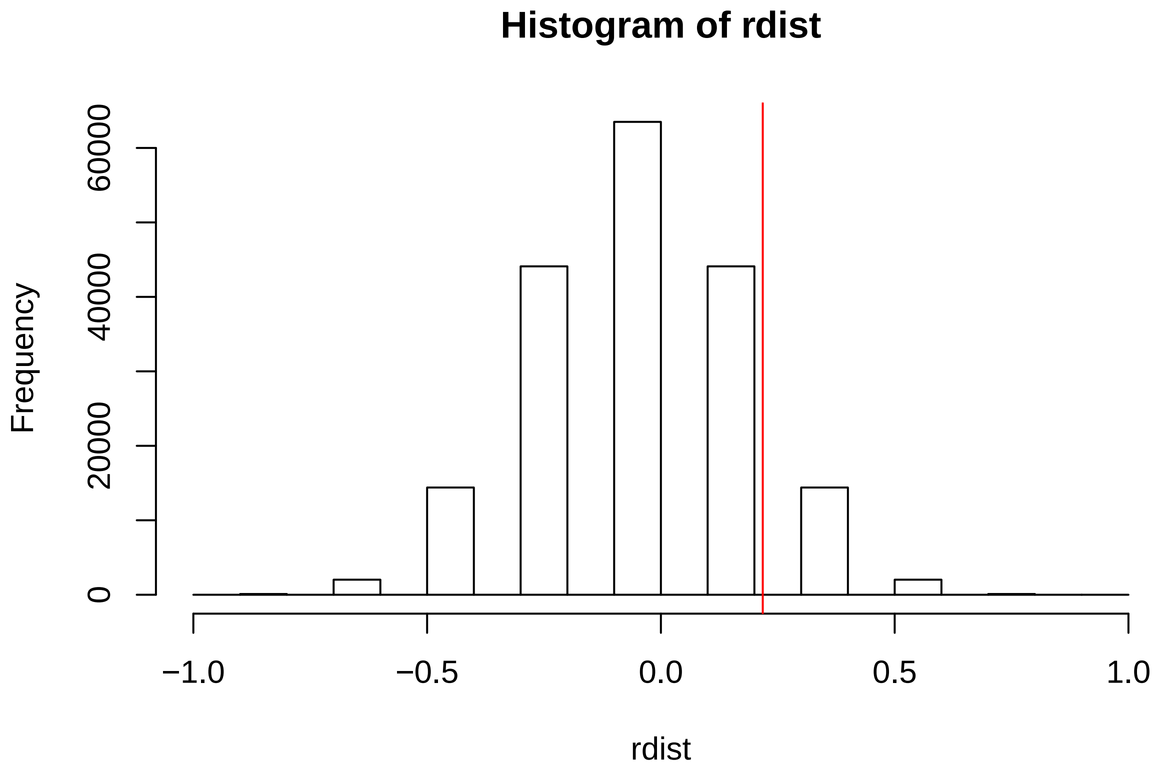
Simulation

The test statistic equals to 0.4. As a block randomization method is used, when generating sharp null distribution, the assignment mechanism should also follow a block randomization, which means for both the female group and male group, there should be 5 new treatment and 5 standard treatment sepretaly. I simulate 10000 assignment ways and keep the outcome of each individual the same. The distribution of sharp null hypothesis is as following, the red line represents the test statistic. In this way, the distribution is approximate distribution and the point estimator and 95% confidence interval are also approximate.



The p-value of this test statistics is 0.0952. As the p-value is larger than 0.05, we cannot reject the sharp null hypothesis and conclude that $Y_{0i} - Y_{1i} = 0$.

If we want to get the exact distribution and ignore the block randomization restriction when generating the sharp null distribution, then we should use the Fisher approach. The distribution of sharp null hypothesis is as following, the red line represent the test statistic.



The p-value of this test statistics is 0.0894477. As the p-value is larger than 0.05, we cannot reject the sharp null hypothesis and conclude that $Y_{0i} - Y_{1i} = 0$.

Question 8

By using Neyman's approach, the point estimate of the marginal average causal effect in my study is 0.4 and the 95% confidence interval is (-0.089, 0.889).

Interpretation: The estimator of the margianl average causal effect is 0.4. With 95% confidence, we can conclude that the true margianl average causal effect lies between -0.089 and 0.889.

This is a randomization experiment. Compare to what I get in question 1, which is the ture marginal average causal effect, the point estimator of the margianl average causal effect is bigger. However, the 95% confidence interval include what I get in question 1.

Using block randomization suppose to balance covariates and give us ture marginal average causal effect. But I stratify individuals based on covariates I assumed and it might not be the true covariates of these individuals. As a result, using block randomization does not make new treatment group and standard treatment group comparable and the point estimator is influenced by possible covariates.

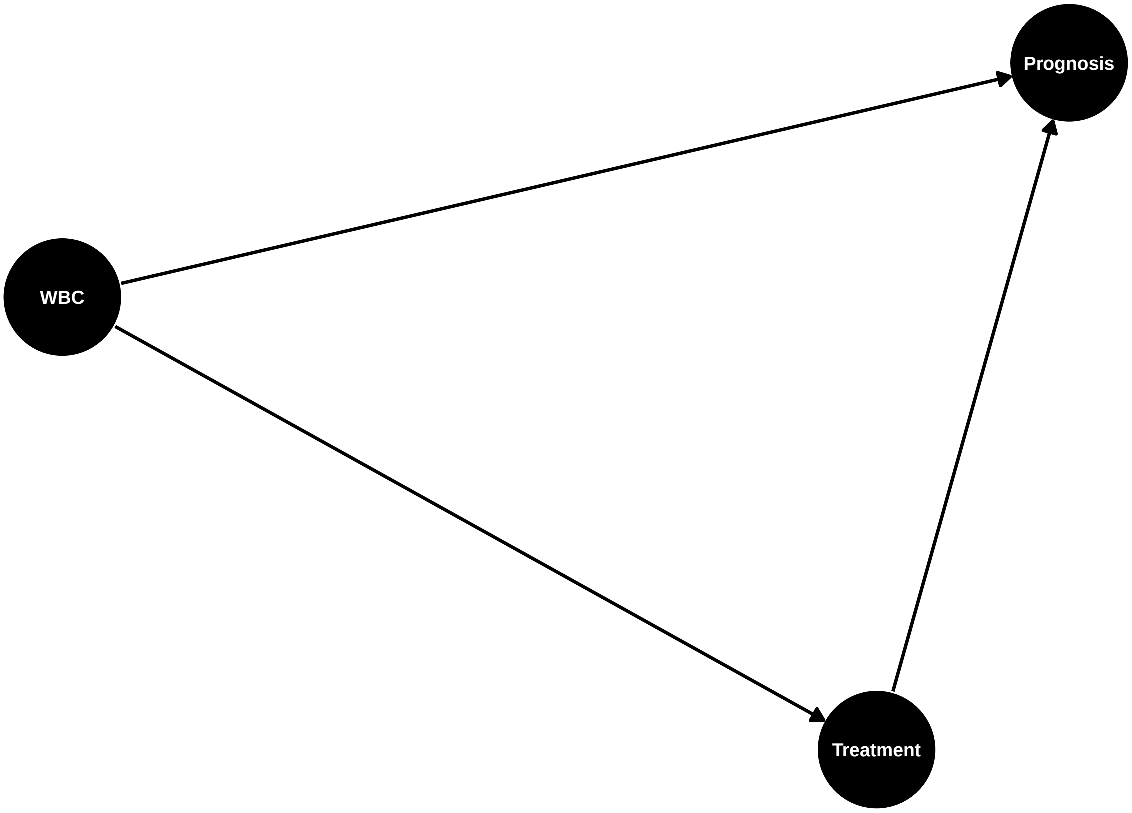
Question 9

Under Neyman's approach to inference, the scientific question is:

On average, how much better the effect on preventing disease of the new treatment is compared to the standard treatment(condition on WBC).

Question 10

Let `Treatment` indicates whether individuals will get new treatment or standard treatment and it is the treatment. Let `Prognosis` indicates whether individuals will be more likely to have diseased prevented and it is the outcome. Let `WBC` indicate whether individuals have normal blood cell and it is the confounder. The DAG representing the situation described is shown below.



Question 11

According to the DAG, we can find that in this observational study, whether individuals have normal white blood cell is a confounder. As this confounder is not controlled, individuals who had new treatment and individuals who had standard treatment is not comparable. Thus, the crude association between A and Y is different from the average causal effect of A on Y.

Leaving L unadjusted in analysis would lead to 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).

Beacuse for those with normal white blood cell, they are more likely to have disease prevented. As they are also more likely to be prescibed the new treatment, even if there is no effect difference between new treatment group and standard treatment group, outcome in new treatment group will be better. Thus, Bias in the estimate such that it overestimates the average causal effect.

Question 12

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

$E(Y_1) = E(Y|A = 1, L = 1)Pr(L = 1) + E(Y|A = 1, L = 0)Pr(L = 0) = 0.775$.

$E(Y_0) = E(Y|A = 0, L = 1)Pr(L = 1) + E(Y|A = 0, L = 0)Pr(L = 0) = 0.475.$

$ACE = E(Y_1) - E(Y_0) = 0.3.$

As ACE equals to 0.3 which is bigger than 0, it means new treatment have better treatment effect than standard treatment.

Question 13

What I get in question 13 is the same as what I get in question 1. This means L is the only confounder and using g-formula to adjust L makes two group comparable and helps us get the true average causal effect.

Question 14

According to the ACE calculated in question 12, we can find that new treatment have better treatment effect than standard treatment. This implys the hypothesis-the new treatment is better for disease prevention than standard treatment is true.

The probaility of being prescribed the new treatment among individuals with normal white blood cell is 0.667. The probaility of being prescribed the new treatment among individuals with abnormal white blood cell is 0.4. As the former probability is higher than the latter, the hypothesis that individuals with normal white blood cell (WBC) counts (L=1) are more likely to be prescribed the new treatment is ture.

The probability of having a better prognosis among individuals with normal white blood cell is 0.6666667. The probability of having a better prognosis among individuals with abnormal white blood cell is 0.4. As the former probability is higher than the latter, the hypothesis that individuals with normal white blood cell (WBC) counts (L=1) are more likely to have a better disease prognosis (i.e., more likely to have disease prevented) compared with individuals with abnormal WBC counts (L=0) is true.

Question 15

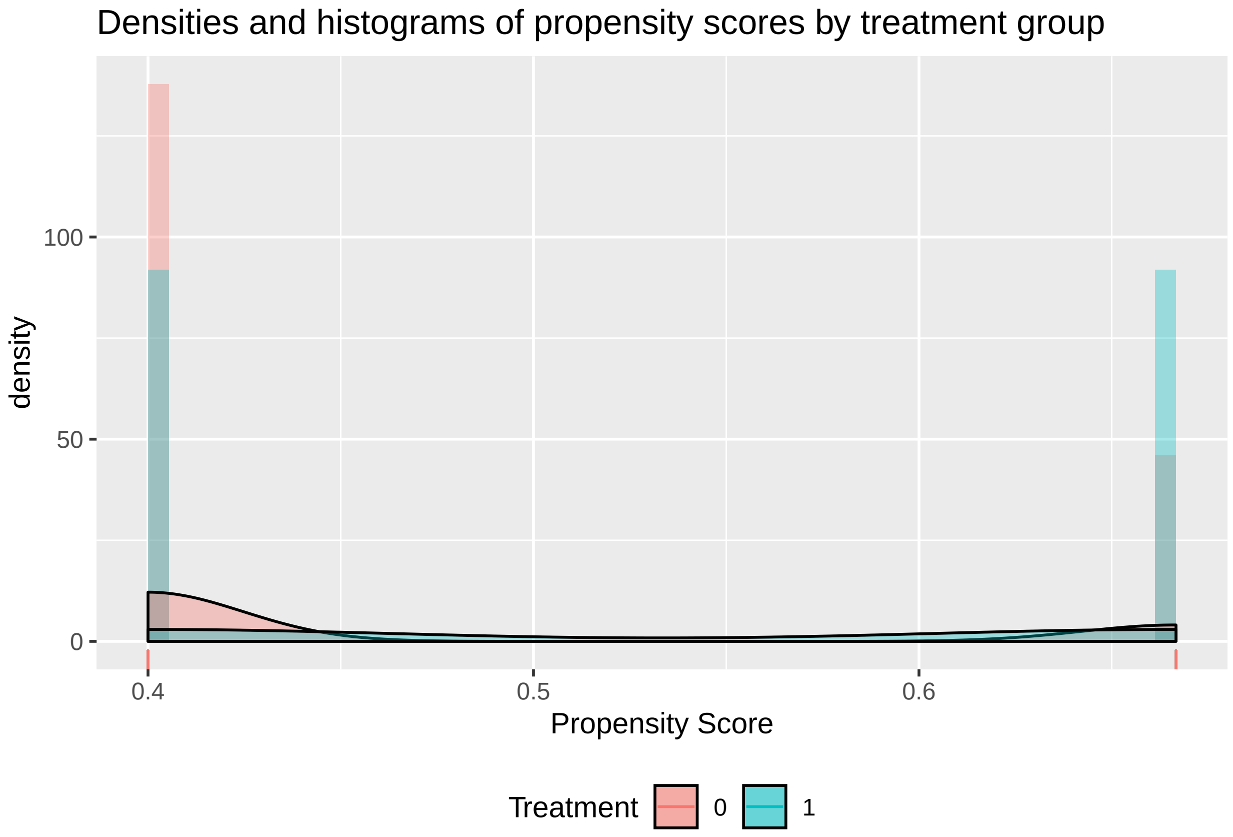
```
##
## Call:
## glm(formula = A ~ L, family = binomial, data = data1)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.48230  -1.01077  -0.05513   1.01382   1.35373
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -0.4055     0.4082  -0.993    0.321
## L1            1.0986     0.6831   1.608    0.108
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 55.452  on 39  degrees of freedom
## Residual deviance: 52.746  on 38  degrees of freedom
## AIC: 56.746
##
## Number of Fisher Scoring iterations: 4
```

The propensity score is shown above.

Question 16

```
##            Stratified by A
##             0          1          SMD
##   n         20         20
##   L = 1 (%)  5 (25.0)  10 (50.0)   0.535
```

The SMD which reflects covariate balance is shown above. As the SMD of L is larger than 0.2, the covariate does not balance well.



According to the plot, we can find that propensity score of treatment group 1 and treatment group 0 overlap well and we don't need to trim data.

Question 17

```
## 50%
## 0.4
```

```
##    subclass
##      0  1
##   0 15  5
##   1 10 10
```

As there are only 40 individuals in this study, I only choose 50% quantile and stratify them into 2 group. The break is 0.4 and numbers of individuals in different treatment group is shown as above.

By using propensity score stratification, we get the point estimator of the marginal average causal effect as 0.3 and the 95% confidence interval is (-0.02, 0.62).

Question 18

```
##
## Call:
## lm(formula = Y ~ A + L, data = data1)
##
## Residuals:
##    Min     1Q Median     3Q    Max
##   -0.9   -0.4    0.1    0.3    0.6
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   0.4000     0.1098   3.644 0.000818 ***
## A1            0.3000     0.1503   1.996 0.053317 .
## L1            0.2000     0.1552   1.289 0.205568
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4591 on 37 degrees of freedom
## Multiple R-squared:  0.168,  Adjusted R-squared:  0.123
## F-statistic: 3.736 on 2 and 37 DF,  p-value: 0.03329
```

By using linear probability model, we get the point estimator of the marginal average causal effect as 0.3 and the 95% confidence interval is (-0.005, 0.605).

Compared to what I get in question 17, point estimators of the marginal average causal effect are the same, while the confidence interval is slightly narrower. This means assumptions of linear probability model allows it to give a narrower 95% confidence interval and the linear model is correctly specified, consistency, SUTVA, exchangeability and positivity assumptions are meet.

Question 19

Following are two advantages of using propensity score methods to adjust for confounding:

1) It check propensity score of each individuals and in each subgroup, thus it allows two treatment group comparable in each subgroup and makes sure they are comparable.
2) When outcomes are binary, using linear probability model is not very appropriate beacuse outcomes in linear probability model are suppose to be continuous. This might causes bias. While using propensity score methods does not cause this problem.

Question 20

In order to let the ACE recovered from linear regression model regressing Y on A and L and the ACE recovered using the g-formula be the same, following assumptions should be achived:

1)consistency, SUTVA, exchangeability and positivity assumptions are meet;

2)there is no interaction between treatment and covariates(the linear model is correctly specified);

3)the outcome is continous.

Question 21

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

Question 22

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). Only with NUCA, DAG can be regard as 'truth'.

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. These mean we need NUCA assumption of potential outcome and a correctly specified DAG.

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

Question 23

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

Question 24

Collider is a node on a path with both arrows on the path going into that node. Conditioning on the collider creates an association between outcome and treatment, thus open a closed path. But when we are calculating causal effect, we want all backdoor paths being blocked. So adjusting for a collider is problematic.

H in this DAG is a collider.

Question 2

Question 1

The units is each school district. The potential outcomes is the number of male and female teachers employed. The treatment is whether being given workshop. Obseved covariates include baseline number of male and female teachers employed in each school district and opinions of school administrators towards the workshop(whether school administrators request to have workshop).

Question 2

This assignment is observational. Beacuse the assignment is decided by what the Department observes. The Department does not randomly assign whether take a workshop to these school districts, the assignment is dependent with covariates.

Question 3

The assignment mechanism is not probabilistic, because for schools districts that currently only has female teachers, they can only be assigned to the group that is given the workshop.

Question 4

The assignment mechanism is confounded, because the department want schools districts with only female teachers hire both male and female teachers. That's why the workshop is given at all schools districts that currently have only female teachers. The assignment mechanism depends on potential outcome, thus it is confounded.

Appendix

```
knitr::opts_chunk$set(echo = FALSE)
library(ri)
library(ggdag)
library(tidyverse)
library(tableone)
library(personalized)
Individual = 1:20
Y1 = c(1,1,1,0,1,1,1,0,1,0,0,0,0,1,1,0,1,1,0,1)
Y0 = c(0,1,0,0,1,1,0,0,0,0,1,0,0,1,0,1,0,0,0,0)
data1 = data.frame(Individual, Y1, Y0)
data1 %>%
  mutate(individualCE = Y1-Y0) %>%
  knitr::kable()
# Question 7
A = c(0,1,1,1,0,1,0,1,0,0,0,1,0,1,1,0,1,1,0,0)
Y = c(0,1,1,0,1,1,0,0,0,0,1,0,0,1,1,1,1,1,0,0)
T.obs = mean(Y[A == 1]) - mean(Y[A == 0])
set.seed(123)
n.sim = 10000 #number of simulations
T.sim = rep(NA, n.sim)
for (i in 1:n.sim) {
  A1 = A[c(1:10)]
  A2 = A[c(11:20)]
  A.sim = c(sample(A1), sample(A2))
  T.sim[i] = mean(Y[A.sim==1], na.rm=TRUE) -  mean(Y[A.sim==0], na.rm=TRUE)
}
hist(T.sim, xlab = "Test Statistic under Null", main="Randomization Distribution")
pval <- mean(T.sim >= T.obs)
quant <- quantile(T.sim,probs = 1-pval)
abline(v = quant,col="red")
set.seed(123)
Abold <- genperms(A,maxiter = choose(20,10))
rdist <- rep(NA, times = ncol(Abold))
for (i in 1:ncol(Abold)) {
  A_tilde <- Abold[, i]
  rdist[i] <- mean(Y[A_tilde == 1]) - mean(Y[A_tilde == 0])
}
# p-value
pval <- mean(rdist >= T.obs)
quant <- quantile(rdist,probs = 1-pval)
hist(rdist)
abline(v = quant,col="red")
# Question 8
t = qt(0.975,9)
var = var(Y[A == 1])/10 + var(Y[A == 0])/10
lower_q = T.obs - t*sqrt(var)
higer_q = T.obs + t*sqrt(var)
# Question 10
tidy_ggdag <- dagify(Prognosis ~ Treatment + WBC,
             Treatment ~ WBC,
             exposure = "Treatment",
             outcome = "Prognosis") %>% tidy_dagitty()
ggdag(tidy_ggdag, node_size = 20, text_size = 2.5) + theme_dag()
```

```
# Question 12
Y = c(1,0,0,0,1,1,0,0,0,0,1,1,1,1,1,1,1,1,1,1,1,1,0,0,1,0,0,1,0,0,1,1,0,1,1,1,1,1,0,1)
A = c(1,0,0,0,1,1,0,0,0,1,0,1,1,0,0,1,1,1,1,1,0,1,1,1,0,0,0,1,0,1,0,1,0,0,1,1,0,1,0,0)
L = c(1,1,0,0,1,0,1,0,0,1,0,1,1,0,0,1,0,0,0,1,1,0,0,0,0,0,0,0,0,0,1,1,0,0,1,1,1,0,0,0)
pl1 = sum(L)/40
pl0 = 1-pl1
meanY1 = mean(Y[A==1&L==1])*pl1 + mean(Y[A==1&L==0])*pl0
meanY0 = mean(Y[A==0&L==1])*pl1 + mean(Y[A==0&L==0])*pl0
# Question 14
h21 = mean(A[L==1])
h20 = mean(A[L==0])
h31 = mean(Y[L==1])
h32 = mean(Y[L==0])
# Question 15
data1 = data.frame(Y = Y, A = A, L = L) %>%
  mutate(A = as.factor(A),
         L = as.factor(L))
ps.model<-glm(A~L,data=data1, family = binomial)
summary(ps.model)
# Question 16
vars = c("L")
cov_bal <- CreateTableOne(vars = vars, strata = "A", data = data1, test = FALSE)
print(cov_bal, smd = TRUE)
ps <- predict(ps.model, type="response")
x = data1
prop.func <- function(x, trt){
  # fit propensity score model
  propens.model <- glm(A~L,data=data1, family = binomial)
  pi.x <- predict(propens.model, type = "response")
  pi.x
}

check.overlap(x = x,
              trt = data1$A,
              type = "both",
              propensity.func = prop.func)
# Question 17
subclass.breaks = quantile(ps, c(.5)) # bins (initial try - modify as needed)
subclass.breaks
subclass = ps
subclass = as.numeric(ps>subclass.breaks[1])
subclass[which(ps>subclass.breaks[1])]<- 1
table(data1$A, subclass)
ACE0 <- mean(data1$Y[which(subclass==0 & data1$A==1)])-mean(data1$Y[which(subclass==0 & data1$A==0)])
ACE1 <- mean(data1$Y[which(subclass==1 & data1$A==1)])-mean(data1$Y[which(subclass==1 & data1$A==0)])

ace <- (nrow(data1[which(subclass==0),])/nrow(data1))*ACE0+(nrow(data1[which(subclass==1),])/nrow(data1

v01 <- var(data1$Y[which(subclass==0 & data1$A==1)])
v00 <- var(data1$Y[which(subclass==0 & data1$A==0)])
v11 <- var(data1$Y[which(subclass==1 & data1$A==1)])
v10 <- var(data1$Y[which(subclass==1 & data1$A==0)])
```

```r
n0 <- nrow(data1[which(subclass==0),])
n1 <- nrow(data1[which(subclass==1),])

n01 <- nrow(data1[which(subclass==0& data1$A==1),])
n11 <- nrow(data1[which(subclass==1& data1$A==1),])
n00 <- nrow(data1[which(subclass==0& data1$A==0),])
n10 <- nrow(data1[which(subclass==1& data1$A==0),])

varace <-(n1)^2/nrow(data1)^2*((v11/n11)+(v10/n10))+(n0)^2/nrow(data1)^2*((v01/n01)+(v00/n00))

sdace<-sqrt(varace)

CIL=ace-sdace*2
CIU=ace+sdace*2
# Question 18
lm.model = lm(Y~A+L, data = data1)
summary(lm.model)
t = qt(0.975, 37)
CIL = 0.3-t*0.1503
CIU = 0.3+t*0.1503
```