P8122 Midterm yz4184

Yunlin Zhou

2022-10-21

Question 1

1)

ACE

$$ACE = E[Y_1] - E[Y_0] = \frac{12}{20} - \frac{5}{20} = \frac{7}{20} = 0.35$$

The new treatment is better on average since the ACE is larger than 0.

2)

$$E[Y|A = 1] - E[Y|A = 0] = \frac{5}{10} - \frac{4}{10} = \frac{1}{10} = 0.1$$

Difference in observed group means is 0.1.

From the association parameter above, we can conclude that the new treatment is better on average.

3)

The results from question 1 and 2 are not same.

- In the scenario in question 1, the assignment mechanism is regular and known and controlled. $E[Y_a]$ are population quantities that are computed by taking an average of potential outcome among all individuals is the population.
- In the scenario in question 2, the assignment mechanism was not known or controlled in advance.The result was only defined post selection. $E[Y|A = a]$ is computes by taking on average of observed outcomes on in the subset of the population with A=a.
- The result in question 2 in smaller, which make the effect of the new treatment seem smaller. The possible reason is that those who received the standard treatment might be healthier. And those who received the new treatment might be weaker. From the truth in question 1, we know that the disease would be prevented in both treatments for subject 3 5 11 16 . In the data in question2, we can see that subject 3 5 11 are assigned to the control group. Similarly, many of those whose disease would not be prevented in both treatments are assigned to treatment group. Thus, the result would be lower in real world data in this case.

4)

a) In an observational study, we typically get all data together (covariates, treatment, outcomes), and the assignment mechanism is not known or controlled. It will typically be the case that individuals select or are selected to take the active treatment based on their underlying health condition. So the data might arise like the one in question 2.

b) In a randomized controlled trial, the assignment mechanism is regular and known and controlled. Also, randomization enforces the assumption of unconfoundedness or exchangeability marginally across covariants. So the data might arise more like the result in theory.

5)

- There is a significant difference between the result in question 1 and 2. It is possible that the experiment is not a randomized trial. But we do not have enoght evidence for this assumption.
- This study might not be a Bernoulli randomized experiment. Because the treatment group size and the control group size are equally assigned. But the group a size in Bernoulli randomized experiment is likely to be unequally assigned.

6)

From the "truth" we know that: for some weak patients, the disease would not be prevented in both treatments; for some strong patients, the disease would be prevented in both treatments. And for the normal patients, the effects would be different due to different treatments. So the health status of patients is a covariate and we will stratify the patients due to their health status. The process is below:

- filter the patients.
  - weak: Y1 = Y0 = 0
  - normal: Y1 =! Y0
  - strong: Y1 = Y0 = 1
- Completely randomize the units with *sample_frac* function within each block.
- Combine the "real-world" dataset with different health status data.

```
individual <- c(1:20)
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)

df = cbind(individual,Y1,Y0)%>%
  as.data.frame()%>%
  mutate(status = ifelse(Y1 == 0 & Y0 == 0, "weak",
                         ifelse(Y1 ==1 & Y0 == 1, "strong","normal")))%>%
  mutate(status = as.factor(status))

df_weak = df %>%
  filter(status == "weak")

df_normal = df %>%
  filter(status == "normal")
```

```
df_strong = df %>%
  filter(status == "strong")
```

```
set.seed(8122)

norm_y1 <- sample_frac(df_normal,0.5)%>% mutate(Y0 = NA)
norm_y0 <- df_normal %>%
  filter(!(individual %in% (norm_y1$individual))) %>%
  mutate(Y1 = NA)
norm_rw = rbind(norm_y1,norm_y0) %>%
  arrange(individual)
knitr::kable(norm_rw)
```

individual	Y1	Y0	status
1	1	NA	normal
2	1	NA	normal
7	1	NA	normal
8	NA	0	normal
10	0	NA	normal
13	NA	0	normal
17	NA	0	normal
18	NA	0	normal
20	NA	0	normal

```
weak_y1 <- sample_frac(df_weak,0.5)%>% mutate(Y0 = NA)
weak_y0 <- df_weak %>%
  filter(!(individual %in% (weak_y1$individual))) %>%
  mutate(Y1 = NA)
weak_rw = rbind(weak_y1,weak_y0)%>% arrange(individual)
knitr::kable(weak_rw)
```

individual	Y1	Y0	status
4	NA	0	weak
6	NA	0	weak
9	0	NA	weak
12	0	NA	weak
14	NA	0	weak
15	0	NA	weak
19	0	NA	weak

```
strong_y1 <- sample_frac(df_strong,0.5)%>% mutate(Y0 = NA)
strong_y0 <- df_strong %>%
  filter(!(individual %in% (strong_y1$individual))) %>%
  mutate(Y1 = NA)
strong_rw = rbind(strong_y1,strong_y0)%>% arrange(individual)
knitr::kable(strong_rw)
```

individual	Y1	Y0	status
3	NA	1	strong
5	1	NA	strong
11	NA	1	strong
16	1	NA	strong

So my final "real-world" data is :

```
df_rw = rbind(norm_rw,weak_rw, strong_rw) %>%
  arrange(individual)

knitr::kable(df_rw)
```

individual	Y1	Y0	status
1	1	NA	normal
2	1	NA	normal
3	NA	1	strong
4	NA	0	weak
5	1	NA	strong
6	NA	0	weak
7	1	NA	normal
8	NA	0	normal
9	0	NA	weak
10	0	NA	normal
11	NA	1	strong
12	0	NA	weak
13	NA	0	normal
14	NA	0	weak
15	0	NA	weak
16	1	NA	strong
17	NA	0	normal
18	NA	0	normal
19	0	NA	weak
20	NA	0	normal

7)

Sharp null hypothesis is that there is no treatment effect:

$$H_0 : \tau_i = Y_{1i} - Y_{0i} = 0$$

for all i

```
df_rw_new = df_rw %>%
pivot_longer(cols = c ("Y1","Y0"),names_to = "A",values_to = "Y", values_drop_na =T) %>%
  mutate(A = ifelse(A == "Y1", 1, 0))
Y = df_rw_new$Y
A = df_rw_new$A

T_stat <- mean(Y[A == 1]) - mean(Y[A == 0])
T_stat
```

```
## [1] 0.3
```

First, we build a new data set including assignment(A) and outcome(Y). We also calculated the sharp null t test which is 0.3.

```
Abold_1 = chooseMatrix(9, 4)
Abold_1 = t(Abold_1)
ncol(Abold_1)
```

```
## [1] 126
```

```
Abold_2 = chooseMatrix(7, 4)
Abold_2 = t(Abold_2)
ncol(Abold_2)
```

```
## [1] 35
```

```
Abold_3 = chooseMatrix(4, 2)
Abold_3 = t(Abold_3)
ncol(Abold_3)
```

```
## [1] 6
```

```
ncol(Abold_1)*ncol(Abold_2)*ncol(Abold_3)
```

```
## [1] 26460
```

```
Abold <- genperms(A, maxiter = 26460)
```

```
## Too many permutations to use exact method.
## Defaulting to approximate method.
## Increase maxiter to at least 184756 to perform exact estimation.
```

```
Abold <- genperms(A)
```

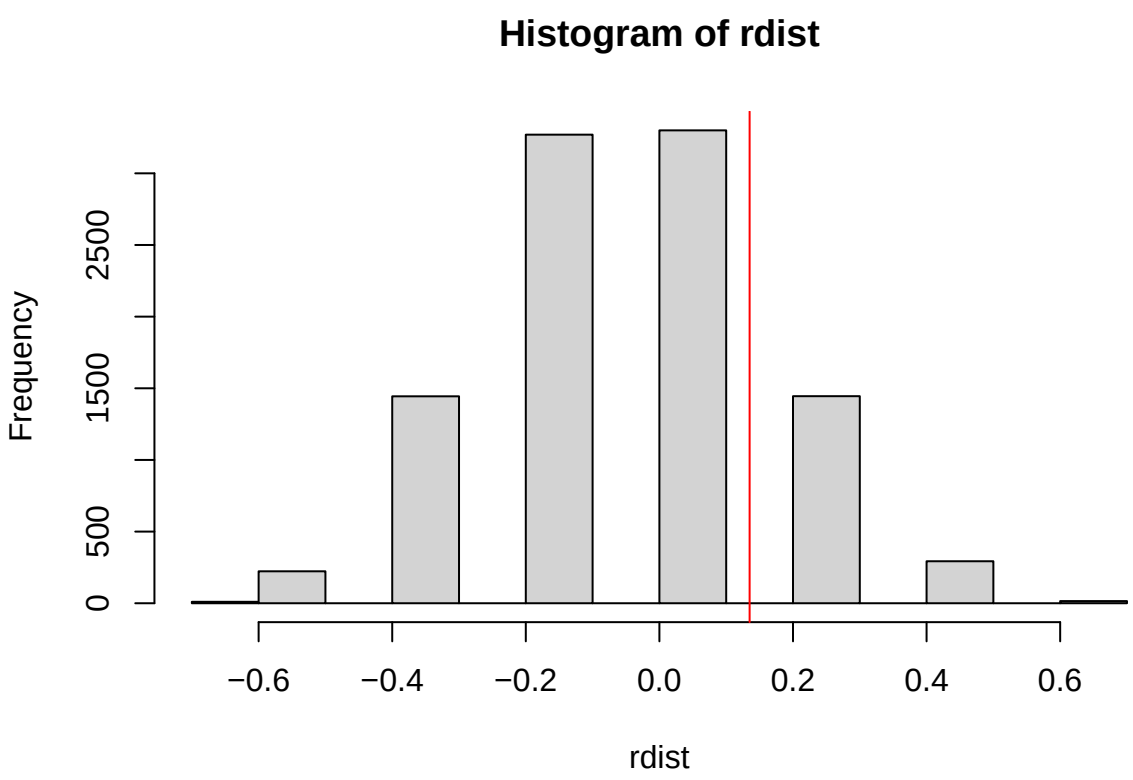
```
## Too many permutations to use exact method.
## Defaulting to approximate method.
## Increase maxiter to at least 184756 to perform exact estimation.
```

Then, we generate a matrix to show different possible assignment vectors. There are $\binom{9}{4}\binom{7}{4}\binom{4}{2}$ possibilities for A.

```
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])
}

pval <- mean(rdist >= T_stat)

quant <- quantile(rdist,probs = 1-pval)
hist(rdist)
abline(v = quant,col="red")
```



Finally, we use the bootstrap to generate the exact randomization distribution for T, under the sharp null hypothesis of no difference. Also, we calculated the p-value, and added the red line in the plot.

- Since the exact p-value is $0.1753 > 0.05$, we fail to reject the sharp null hypothesis that there is no difference so there is no treatment effect for all individuals in the sample.

8)

1. Create a grid of possible sharp null hypotheses.
2. Calculate p-values for each sharp null.
3. For the point estimate: Pick the value that is "least surprising" under the null.
4. For the confidence interval: Find the range of the values that we would not reject under the null.

```
grid<-seq(-1,1, by=0.01)
p.ci<-rep(NA,length(grid))

rdist_1 <- rep(NA, times = ncol(Abold))
for (i in 1:length(grid)){
for (k in 1:ncol(Abold)) {
  A_tilde <- Abold[, k]
  rdist_1[k] <- mean(Y[A_tilde == 1]) - mean(Y[A_tilde == 0])+grid[i]
}
  p.ci[i]<-mean(rdist_1 >= T_stat)
}
```

```
cbind(p.ci,grid)
```

```
##          p.ci  grid
##   [1,] 0.0000 -1.00
##   [2,] 0.0000 -0.99
##   [3,] 0.0000 -0.98
##   [4,] 0.0000 -0.97
##   [5,] 0.0000 -0.96
##   [6,] 0.0000 -0.95
##   [7,] 0.0000 -0.94
##   [8,] 0.0000 -0.93
##   [9,] 0.0000 -0.92
##  [10,] 0.0000 -0.91
##  [11,] 0.0000 -0.90
##  [12,] 0.0000 -0.89
##  [13,] 0.0000 -0.88
##  [14,] 0.0000 -0.87
##  [15,] 0.0000 -0.86
##  [16,] 0.0000 -0.85
##  [17,] 0.0000 -0.84
##  [18,] 0.0000 -0.83
##  [19,] 0.0000 -0.82
##  [20,] 0.0000 -0.81
##  [21,] 0.0000 -0.80
##  [22,] 0.0000 -0.79
##  [23,] 0.0000 -0.78
##  [24,] 0.0000 -0.77
##  [25,] 0.0000 -0.76
##  [26,] 0.0000 -0.75
##  [27,] 0.0000 -0.74
##  [28,] 0.0000 -0.73
##  [29,] 0.0000 -0.72
##  [30,] 0.0000 -0.71
##  [31,] 0.0000 -0.70
##  [32,] 0.0000 -0.69
##  [33,] 0.0000 -0.68
##  [34,] 0.0000 -0.67
##  [35,] 0.0000 -0.66
##  [36,] 0.0000 -0.65
##  [37,] 0.0000 -0.64
##  [38,] 0.0000 -0.63
##  [39,] 0.0000 -0.62
##  [40,] 0.0000 -0.61
##  [41,] 0.0000 -0.60
##  [42,] 0.0000 -0.59
##  [43,] 0.0000 -0.58
##  [44,] 0.0000 -0.57
##  [45,] 0.0000 -0.56
##  [46,] 0.0000 -0.55
##  [47,] 0.0000 -0.54
##  [48,] 0.0000 -0.53
##  [49,] 0.0000 -0.52
##  [50,] 0.0000 -0.51
```

```
##  [51,] 0.0000 -0.50
##  [52,] 0.0000 -0.49
##  [53,] 0.0000 -0.48
##  [54,] 0.0000 -0.47
##  [55,] 0.0000 -0.46
##  [56,] 0.0000 -0.45
##  [57,] 0.0000 -0.44
##  [58,] 0.0000 -0.43
##  [59,] 0.0000 -0.42
##  [60,] 0.0000 -0.41
##  [61,] 0.0000 -0.40
##  [62,] 0.0015 -0.39
##  [63,] 0.0015 -0.38
##  [64,] 0.0015 -0.37
##  [65,] 0.0015 -0.36
##  [66,] 0.0015 -0.35
##  [67,] 0.0015 -0.34
##  [68,] 0.0015 -0.33
##  [69,] 0.0015 -0.32
##  [70,] 0.0015 -0.31
##  [71,] 0.0015 -0.30
##  [72,] 0.0015 -0.29
##  [73,] 0.0015 -0.28
##  [74,] 0.0015 -0.27
##  [75,] 0.0015 -0.26
##  [76,] 0.0015 -0.25
##  [77,] 0.0015 -0.24
##  [78,] 0.0015 -0.23
##  [79,] 0.0015 -0.22
##  [80,] 0.0015 -0.21
##  [81,] 0.0308 -0.20
##  [82,] 0.0308 -0.19
##  [83,] 0.0308 -0.18
##  [84,] 0.0308 -0.17
##  [85,] 0.0308 -0.16
##  [86,] 0.0308 -0.15
##  [87,] 0.0308 -0.14
##  [88,] 0.0308 -0.13
##  [89,] 0.0308 -0.12
##  [90,] 0.0308 -0.11
##  [91,] 0.0308 -0.10
##  [92,] 0.0308 -0.09
##  [93,] 0.0308 -0.08
##  [94,] 0.0308 -0.07
##  [95,] 0.0308 -0.06
##  [96,] 0.0308 -0.05
##  [97,] 0.0308 -0.04
##  [98,] 0.0308 -0.03
##  [99,] 0.0308 -0.02
## [100,] 0.0308 -0.01
## [101,] 0.1753  0.00
## [102,] 0.1753  0.01
## [103,] 0.1753  0.02
## [104,] 0.1753  0.03
```

```
## [105,] 0.1753  0.04
## [106,] 0.1753  0.05
## [107,] 0.1753  0.06
## [108,] 0.1753  0.07
## [109,] 0.1753  0.08
## [110,] 0.1753  0.09
## [111,] 0.1753  0.10
## [112,] 0.1753  0.11
## [113,] 0.1753  0.12
## [114,] 0.1753  0.13
## [115,] 0.1753  0.14
## [116,] 0.1753  0.15
## [117,] 0.1753  0.16
## [118,] 0.1753  0.17
## [119,] 0.1753  0.18
## [120,] 0.1753  0.19
## [121,] 0.5053  0.20
## [122,] 0.5053  0.21
## [123,] 0.5053  0.22
## [124,] 0.5053  0.23
## [125,] 0.5053  0.24
## [126,] 0.5053  0.25
## [127,] 0.5053  0.26
## [128,] 0.5053  0.27
## [129,] 0.5053  0.28
## [130,] 0.5053  0.29
## [131,] 0.5053  0.30
## [132,] 0.5053  0.31
## [133,] 0.5053  0.32
## [134,] 0.5053  0.33
## [135,] 0.5053  0.34
## [136,] 0.5053  0.35
## [137,] 0.5053  0.36
## [138,] 0.5053  0.37
## [139,] 0.5053  0.38
## [140,] 0.5053  0.39
## [141,] 0.8323  0.40
## [142,] 0.8323  0.41
## [143,] 0.8323  0.42
## [144,] 0.8323  0.43
## [145,] 0.8323  0.44
## [146,] 0.8323  0.45
## [147,] 0.8323  0.46
## [148,] 0.8323  0.47
## [149,] 0.8323  0.48
## [150,] 0.8323  0.49
## [151,] 0.8323  0.50
## [152,] 0.8323  0.51
## [153,] 0.8323  0.52
## [154,] 0.8323  0.53
## [155,] 0.8323  0.54
## [156,] 0.8323  0.55
## [157,] 0.8323  0.56
## [158,] 0.8323  0.57
```

```
## [159,] 0.8323  0.58
## [160,] 0.8323  0.59
## [161,] 0.9767  0.60
## [162,] 0.9767  0.61
## [163,] 0.9767  0.62
## [164,] 0.9767  0.63
## [165,] 0.9767  0.64
## [166,] 0.9767  0.65
## [167,] 0.9767  0.66
## [168,] 0.9767  0.67
## [169,] 0.9767  0.68
## [170,] 0.9767  0.69
## [171,] 0.9767  0.70
## [172,] 0.9767  0.71
## [173,] 0.9767  0.72
## [174,] 0.9767  0.73
## [175,] 0.9767  0.74
## [176,] 0.9767  0.75
## [177,] 0.9767  0.76
## [178,] 0.9767  0.77
## [179,] 0.9767  0.78
## [180,] 0.9767  0.79
## [181,] 0.9990  0.80
## [182,] 0.9990  0.81
## [183,] 0.9990  0.82
## [184,] 0.9990  0.83
## [185,] 0.9990  0.84
## [186,] 0.9990  0.85
## [187,] 0.9990  0.86
## [188,] 0.9990  0.87
## [189,] 0.9990  0.88
## [190,] 0.9990  0.89
## [191,] 0.9990  0.90
## [192,] 0.9990  0.91
## [193,] 0.9990  0.92
## [194,] 0.9990  0.93
## [195,] 0.9990  0.94
## [196,] 0.9990  0.95
## [197,] 0.9990  0.96
## [198,] 0.9990  0.97
## [199,] 0.9990  0.98
## [200,] 0.9990  0.99
## [201,] 1.0000  1.00
```

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

ci = range(grid[which(0.05<p.ci & p.ci<0.95)])
```

- The point estimate is 0.295 and the confidence interval is 0, 0.59.
- Point estimate = 0.295: For the patients who received the new treatment are more likely to prevent the disease than the standard treatment on average since $0.295 > 0$.
- The true ACE would fall in the range 0, 0.59 at 5% significance level.

9)

$$\widehat{SACE} = \overline{Y_1^{obs}} - \overline{Y_0^{obs}} = \frac{\sum_{i=1}^{N} A_i Y_{1i}}{N_1} - \frac{\sum_{i=1}^{N}(1-A_i)Y_{0i}}{N_0} = \frac{5}{10} - \frac{2}{10} = \frac{3}{10} = 0.3$$

```
point_estimate_neyman = sum(Y[A==1])/10 - sum(Y[A==0])/10
```

- Point estimate = 0.3: For the patients who received the new treatment are more likely to prevent the disease than the standard treatment on average since $0.3 > 0$.

$$\widehat{var}(\widehat{SACE}) = \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} = 0.0456$$

```
t_crit = qt(0.975, 9)
var = var(Y[A==1])/10 + var(Y[A==0])/10
```

$$CI(low) = \widehat{SACE} - z^* \sqrt{\widehat{var}(\widehat{SACE})} = -0.1828291$$

$$CI(up) = \widehat{SACE} + z^* \sqrt{\widehat{var}(\widehat{SACE})} = 0.7828291$$

```
CI_low = point_estimate_neyman - t_crit*sqrt(var)
CI_up = point_estimate_neyman + t_crit*sqrt(var)
```

- The true ACE would fall in the range between -0.1828291 and 0.7828291at 5% significance level.

10)

The point estimates from question 8 and 9 very close to the ACE calculated from the "truth". The true ACE is in the range of the confidence interval.

- In Fisher's approach, we compare any test statistic to empirical randomization distribution under sharp null hypothesis. This is a design-based, assumption-free inference. And we derived a relatively accurate estimation using this approach.
- In Neyman's approach, we compare t-statistic to normal or t distribution under average null hypothesis. This approach considers random assignment and random sampling. However, this approach also relies on large N. Since we have a relatively small N, the estimation using this approach might have limitations.

11)

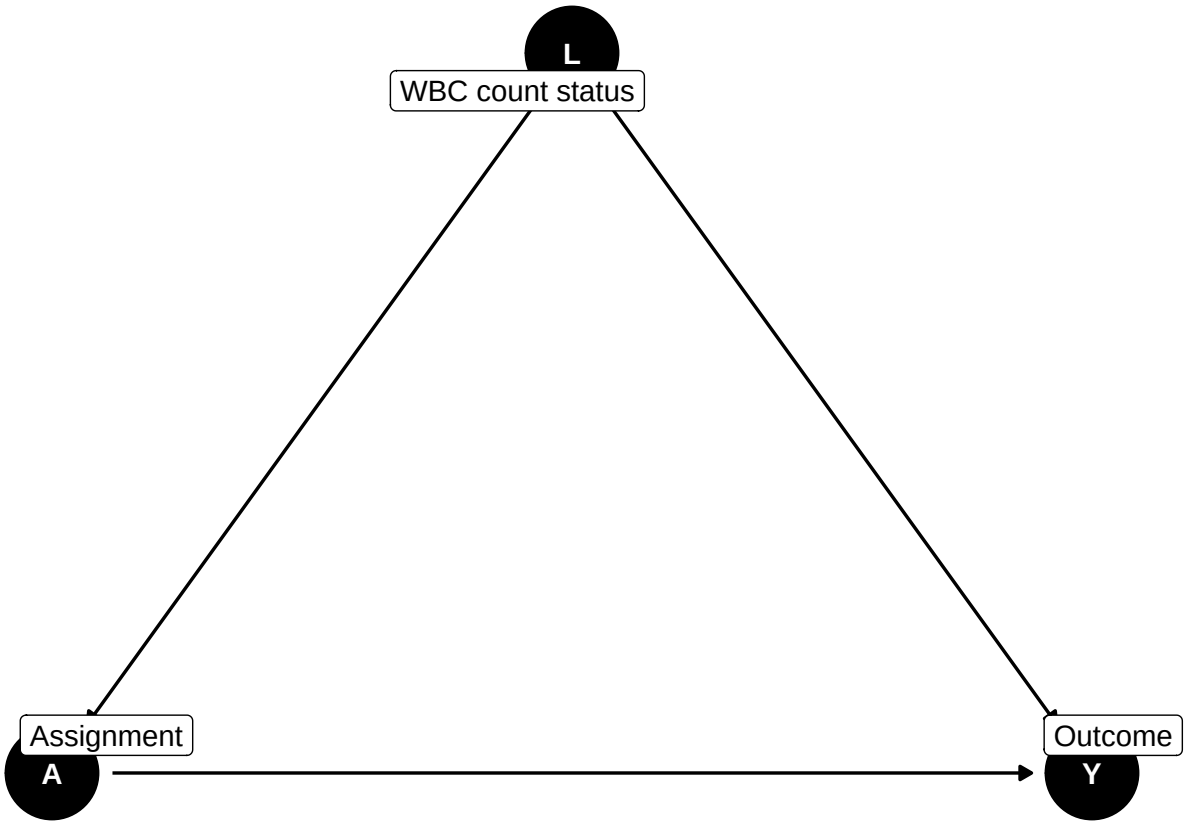
What is the difference between the average outcome if all units were exposed to the new treatment and the average outcome if all units were exposed to the standard treatment?

12)

- A: Assignment. Patients are assigned to either new treatment or standard treatment.
- Y: Outcome. Disease is prevented or not.
- L: White blood cell (WBC) counts. The status could be normal or abnormal.

```
coord_dag <- list(
  x = c(L = 1, A = 0, Y = 2),
  y = c(L = 1, A = 0, Y = 0))

dagify(Y ~ A,
       A ~ L,
       Y ~ L,
       coords = coord_dag,
       labels = c("Y" = "Outcome",
                  "A" = "Assignment",
                  "L" = "WBC count status")) %>%
  ggdag(use_labels = "label") + theme_void()
```



13)

The covariate WBC counts L is associates with both the assignment A and the outcome Y.

choice: b)

Explanation for choice:

- From the question we know that : individuals with normal white blood cell (WBC) counts (L=1) are more likely to be prescribed the new treatment and also more likely to have a better disease prognosis.
- Because there are more patients who have normal WBC counts (L = 1) in the new treatment group (A = 1) than in the other group (A = 0), one would have expected to find a higher disease prevention rate in the group A = 1 even under the null hypothesis of no effect of treatment A on Y. The effect estimate will be biased upwards in the absence of adjustment for L.

14)

$$ACE = \sum_c E[Y|A=1, C=c]Pr(C=c) - \sum_c E[Y|A=0, C=c]Pr(C=c)$$

```
df_40 = readxl::read_xlsx("40-patients.xlsx",col_names = c("individual","Y1","Y0","A","L")) %>%
  as.data.frame() %>%
  mutate(Y1 = as.numeric(Y1),
         Y0 = as.numeric(Y0),
         A = as.numeric(A),
         L = as.numeric(L)) %>%
  suppressWarnings()
df_40 = df_40[-1,]
```

```
df_40_new = df_40 %>%
pivot_longer(cols = c ("Y1","Y0"), values_to = "Y")
```

```
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 ~ as.factor(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(
 round(mean(d.onesample$predicted_meanY[d.onesample$interv == -1]),
3),
 round(mean(d.onesample$predicted_meanY[d.onesample$interv == 0]) ,
3),
 round(mean(d.onesample$predicted_meanY[d.onesample$interv == 1]) ,
3),
 round(mean(d.onesample$predicted_meanY[d.onesample$interv == 1]) -
 mean(d.onesample$predicted_meanY[d.onesample$interv == 0]),3)
 ))
}

# bootstrap
results <- boot(data = df_40_new,
 statistic = standardization,
 R = 5)
# generating confidence intervals
se <- c(round(sd(results$t[, 1]),3), #interv=-1
 round(sd(results$t[, 2]),3), #interv=0
 round(sd(results$t[, 3]),3), #interv=1
 round(sd(results$t[, 4]),3)) #interv=1-interv=0
mean <- results$t0
ll <- round(mean - qnorm(0.975) * se,3)
ul <- round(mean + qnorm(0.975) * se,3)
bootstrap <-
 data.frame(cbind(
 c(
 "Observed",
 "No Treatment",
 "Treatment",
 "Treatment - No Treatment"
 ),
 mean,
 se,
 ll,
 ul
 ))
knitr::kable(bootstrap)
```

V1	mean	se	ll	ul
Observed	0.4	0.076	0.251	0.549
No Treatment	0.214	0.112	-0.006	0.434
Treatment	0.586	0.076	0.437	0.735
Treatment - No Treatment	0.371	0.13	0.116	0.626

- For the patients who received the new treatment are more likely to prevent the disease than the standard treatment on average since $0.371 > 0$.
- The true ACE would fall in the range between 0.116 and 0.626 at 5% significance level.

15)

The result of question 14 is approximately the same as the result in question 1.

- Confounding adjustment: NUCA holds in this question and therefore within levels of L, it is as if A were randomized. The g-formula methods exploits conditional exchangeability in subsets defined by L to estimate the causal effect of A on Y in the entire population or in any subset of the population.
- Under the assumption of conditional exchangeability given L, g-methods simulate the A-Y association in the population if backdoor paths involving the measured variables L did not exist; the simulated A-Y association can then be entirely attributed to the effect of A on Y .

16)

```
pr_L1 = nrow(df_40[df_40$L == 1,])/nrow(df_40)
pr_L0 = nrow(df_40[df_40$L == 0,])/nrow(df_40)
E_A1_L1 = mean(df_40$Y1[df_40$A == 1 & df_40$L == 1])
E_A1_L0 = mean(df_40$Y1[df_40$A == 1 & df_40$L == 0])
E_A0_L1 = mean(df_40$Y0[df_40$A == 0 & df_40$L == 1])
E_A0_L0 = mean(df_40$Y0[df_40$A == 0 & df_40$L == 0])
```

```
E_A1_L1 * pr_L1 - E_A0_L1 * pr_L1
```

```
## [1] 0.1875
```

```
E_A1_L0 * pr_L0 - E_A0_L0 * pr_L0
```

```
## [1] 0.1875
```

We can calculate the causal effect within each subset using the dataset in study 3 with g-formula.

As we can see from above, the causal effect is same between the subset (L=1) and the subset (L=0). So we can say that conditional exchangeability is in 2 strata defined by L.

17)

Controlling for B and F and L suffices.

18)

- NUCA can be achieved only if we are able to measure all common causes of A and Y. In DAG, NUCA holds if all paths between A and Y are blocked after conditioning on L.

19)

H

20)

- Collider: A node on a path with both arrows on the path going into that node.
- Conditioning on the collider L creates an association between Y and A, so that while A and Y are marginally independent, they are conditionally dependent given L. This is known as collider selection bias.
- Example: H is a collider in the path A - L - H - F - Y.

Question 2

1.

- Units: The hospitals in New York state which are given the workshop.
- Potential outcomes: The number of doctors from minority backgrounds that were promoted to leadership positions after given workshop is larger or not.
- Treatment: The workshop for hospital administrators that focuses on the benefits of diversity in leadership.
- Observed Covariates: The ratio of white doctors in leadership positions (striking majority, majority and other).

2.

Since we don't know if we have unobserved confounding, we cannot assume that within strata the treatment groups are comparable.

We are interested in the conditional average causal effects given covariate C (The ratio of white doctors in leadership positions):

$$ACE = E[Y_1|C = c] - E[Y_0|C = c] = E[Y|A = 1, C = c] - E[Y|A = 0, C = c]$$

We are also interested in the marginal average causal effects given covariate C (The ratio of white doctors in leadership positions):

$$ACE = \sum_c E[Y|A = 1, C = c]Pr(C = c) - \sum_c E[Y|A = 0, C = c]Pr(C = c)$$

3.

According to the question, we know that the workshop is assigned to the hospitals with majority of white doctors in leadership positions or requested by hospital administrators. Thus we know the assignment is not randomized and there are unobserved covariates(like time effect, the willing of administrators to change the situation). So, within strata of the known confounding variables, the treatment groups are not comparable. We cannot calculate the ACE in that case.

4.

My suggestions are follow:

1. Before the assignment, we need a more thorough investigation on the baseline covariate, including the ratio of white doctors in leadership positions, the ratio of white doctors in the hospital, how much the administrators want to change the situation, etc.
2. After we have all the baseline covarites, we can use block randomization method to assign the workshop. For some hospitals, we do not assign them workshop and we use those hospitals as control group.
3. When analyzing the data, we can calculate both conditional and marginal ACE of new promoted doctor in minority backgrounds.