Midterm

Yuki Joyama

Question 1

(1) We know all the potential outcomes, so the causal effect of the treatment for an individual can be calculated by $Y_1 - Y_0$ (Y_1: the potential outcome of a disease if an individual is assigned new treatment, Y_0: the potential outcome of a disease if an individual is assigned standard treatment). The average causal effect is $E[Y_1 - Y_0] = 0.3$. On average, the new treatment prevents disease in 30% more individuals compared to the standard treatment.

(2) Under consistency, SUTVA, randomizeation and positivity assumption, $ACE = E[Y_1 - Y_0] = E[Y_1 \mid A = 1] - E[Y_0 \mid A = 0] = E[Y \mid A = 1] - E[Y \mid A = 0]$
Given the table, $E[Y \mid A = 1] = \frac{1+0+1+0+0+0+0+1+1+0}{10} = 0.4$ and $E[Y \mid A = 0] = \frac{0+0+1+0+0+0+0+1+0+0}{10} = 0.2$. So $ACE = 0.4 - 0.2 = 0.2$.
This suggest that the new treatment has 20% higher likelihood of disease prevention compared to standard treatment.

(3) The observed effect in (2) is smaller than in (1). The counterfactuals in (2) are unknown, and knowing (1), it appears that some individuals who would have shown a preventive effect with the new treatment were assigned to the standard treatment group in (2). This potentially attenuates the effect observed in (2).

(4)

(a) Observational study
In an observational study, the intervention is not assigned randomly and the exposure is likely to be influenced by various factors. Moreover, we may observe different number of individuals in each exposure groups by the nature of the data collection process. This can leads to a biased estimates of the outcome effect.

(b) Randomized controlled trial
In RCT, the randomization of treatment (intervention) assignment leads to expected balance on both observed and unobserved covariates in each group. This can minimize confounding and allows for an unbiased estimate of treatment effect under the necessary assumptions.

(5) Given the data, we can rule out the crossover trial because each individual receive only one of the treatment.

(6) Knowing the true potential outcomes under both treatments, I first divide the participants into three groups, G=1, G=2, and G=3, with G=1 being individuals whose outcomes are not affected by treatment assignment, G=2 being individuals whose outcomes are positively affected by the new treatment, and G=3 being individuals whose outcomes are negatively affected by the new treatment. I then use R to randomize the treatment within each group based on the complete randomization method, with the size of the new treatment group equal to the standard treatment. Below is a table showing the results of this assignment mechanism.

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

(7)

```
# set up the data frame based on the table
df <- data.frame(
  Individual = 1:20,
  A = c(1, 1, 1, 0, 1, 0, 0, 0, 0, 0, 1, 1, 0, 1, 0, 0, 1, 0, 1, 1),
  Y = c(1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 1, 0, 0, 1),
  G = c(2, 1, 2, 1, 1, 1, 2, 1, 2, 1, 3, 1, 1, 1, 2, 3, 2, 2, 1, 2)
)

Yb_obs1 = df |>
  filter(A == 1) |>
  summarize(mean_Y_obs = mean(Y)) |>
  pull(mean_Y_obs)
Yb_obs0 = df |>
  filter(A == 0) |>
  summarize(mean_Y_obs = mean(Y)) |>
  pull(mean_Y_obs)

# calculate t_obs
T_obs = Yb_obs1 - Yb_obs0

# possible number of treatment assignment
A_num = choose(10, 5)*choose(8, 4)*choose(2, 1)

# record possible treatment assignment in each blosck as a matrix
g1 = chooseMatrix(10,5)
g2 = chooseMatrix(8,4)
g3 = chooseMatrix(2,1)

# create a function to generate the combinations
```

```r
gen_comb <- function(g1, g2, g3) {
  # Get the number of rows for each matrix
  num_rows_g1 <- nrow(g1)
  num_rows_g2 <- nrow(g2)
  num_rows_g3 <- nrow(g3)

  # initialize an empty matrix to store results
  result <- matrix(nrow = 0, ncol = ncol(g1) + ncol(g2) + ncol(g3))

  # loop through all rows of g1, g2, and g3
  for (i in 1:num_rows_g1) {
    for (j in 1:num_rows_g2) {
      for (k in 1:num_rows_g3) {
        # concatenate rows from g1, g2, and g3
        new_row <- c(g1[i, ], g2[j, ], g3[k, ])
        # append the new row to the result matrix
        result <- rbind(result, new_row)
      }
    }
  }
  return(result)
}

# call the function to generate the matrix
A <- gen_comb(g1, g2, g3)

# create a placeholder for the reordered matrix based on the table for individual 1-20
new_A <- matrix(nrow = nrow(A), ncol = length(df$G))

# loop through df$G and assign the columns in the desired order
g1_cols <- 1:10      # columns for g=1
g2_cols <- 11:18     # columns for g=2
g3_cols <- 19:20     # columns for g=3

for (i in 1:length(df$G)) {
  if (df$G[i] == 1) {
    new_A[, i] <- A[, g1_cols[1]]  # take the first column for g=1
    g1_cols <- g1_cols[-1]         # remove the used column from g1
  } else if (df$G[i] == 2) {
    new_A[, i] <- A[, g2_cols[1]]  # take the first column for g=2
    g2_cols <- g2_cols[-1]         # remove the used column from g2
  } else if (df$G[i] == 3) {
    new_A[, i] <- A[, g3_cols[1]]  # take the first column for g=3
    g3_cols <- g3_cols[-1]         # remove the used column from g3
  }
}
```

- Number of individuals in A=1 group: $N_1 = 10$
- Number of individuals in A=0 group: $N_0 = 10$
- Total number of individuals: $N = 20$

- Mean of the outcome variable for A=1 group: $\bar{Y}_1^{obs} = 0.6$
- Mean of the outcome variable for A=0 group: $\bar{Y}_0^{obs} = 0.1$
- $T_{obs} = \bar{Y}_1^{\text{obs}} - \bar{Y}_0^{\text{obs}} = 0.5$

Under the assignment mechanism in (6), there are $\binom{10}{5} \times \binom{8}{4} \times \binom{2}{1} = 35280$ possibilities for A.

The sharp null hypothesis:
$H_0 : Y_i^1 = Y_i^0 \quad$ for all i
where Y_i^1 is the potential outcome for individual i if they are assigned to $A = 1$, and Y_i^0 is the potential outcome for individual i if they are assigned to $A = 0$.

```
# create df that has the group assignment based on the first row of matrix A
df2 = df
df2$A = new_A[1,]

# calculate t under the first possibility of A, under the sharp null hypothesis
T_stat = mean(df2$Y[df2$A == 1]) - mean(df2$Y[df2$A == 0])
```

Under the sharp null hypothesis, the test statistic under the first row of matrix `new_A` is -0.1.

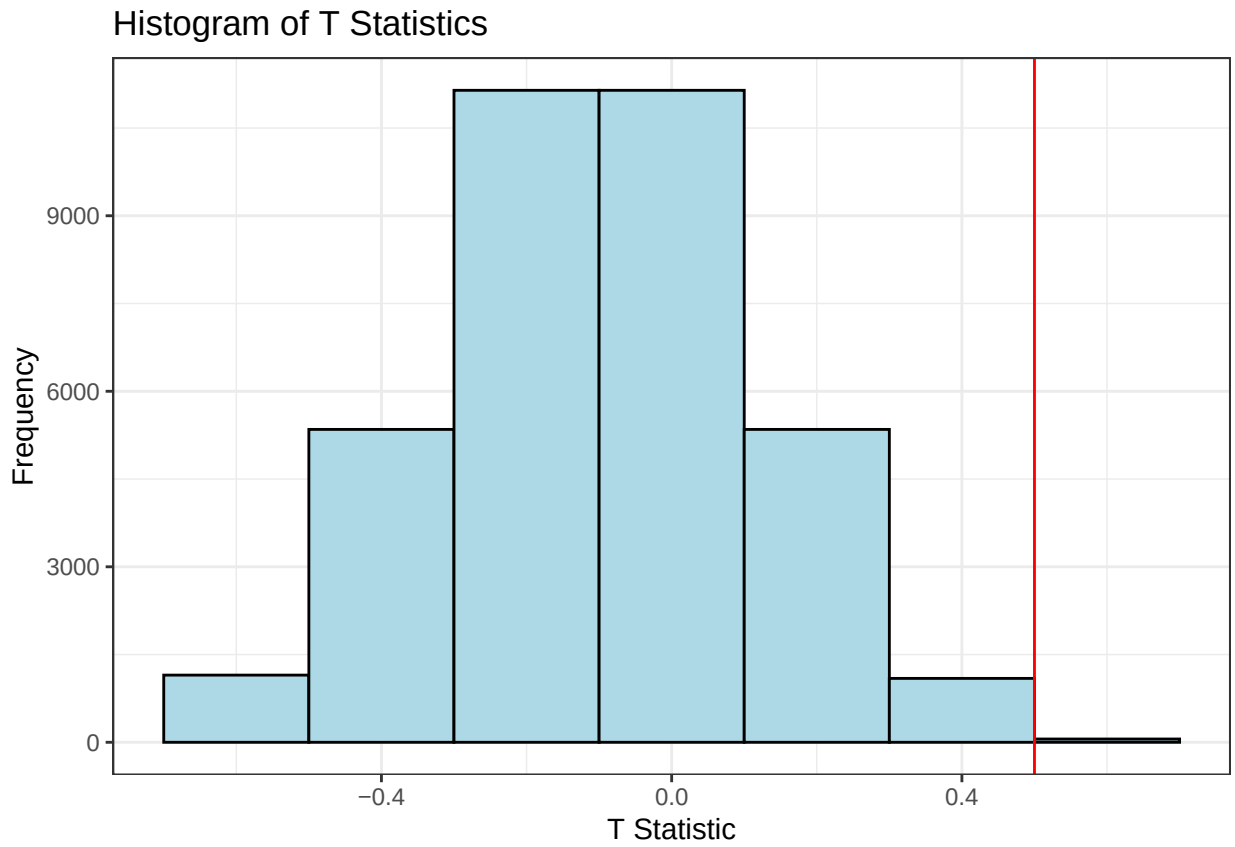
I will iterate this process for all the possibilities of matrix `new_A` to obtain the exact randomization distribution for T under the sharp null hypothesis.

```
# set up df to store T statistic values
rdist = rep(NA, times = A_num)

# iteration
for (i in 1:A_num) {
  df_ite = df2
  df_ite$A = new_A[i,]
  rdist[i] = mean(df_ite$Y[df_ite$A == 1]) - mean(df_ite$Y[df_ite$A == 0])
}
```

The T_{obs} is the red line in the plot.

```
# plot histogram
ggplot(data.frame(t_stat = rdist), aes(x = t_stat)) +
  geom_histogram(binwidth = 0.2, color = "black", fill = "lightblue") +
  geom_vline(aes(xintercept = T_obs), color = "red", size = 0.5) + # add T_obs line
  labs(title = "Histogram of T Statistics", x = "T Statistic", y = "Frequency")
```



Based on this distribution, we can obtain the exact p-value by the following formula:
$P(T(A,Y) \geq T_{obs} | Y_i^1 - Y_i^0 = 0) = \frac{\sum I(T(A,Y) \geq T_{obs})}{K}$ where $K = \binom{10}{5} \times \binom{8}{4} \times \binom{2}{1}$.

```
# calculate the exact p-value
p_val = sum(rdist >= T_obs) / length(rdist)
```

The exact p-value is 0.03253968.

Under $\alpha = 0.10$, the exact p-value < 0.10 suggests that the observed test statistic is unlikely to have occurred under the sharp null hypothesis. Therefore, we reject the null hypothesis. We can conclude that the new treatment compared to standard treatment has a causal effect on disease prevention.

(8) To calculate 90% CI, I will first set the grid as follows:

```
grid = seq(-0.8, 0.8, by = 0.05)
```

Next, I will generate the randomization distribution for each hypothesized treatment effect.

```
p.ci = rep(NA, length(grid)) # initialize p-value vector
rdist = rep(NA, times = A_num) # initialize rdist for number of permutations

for (i in 1:length(grid)) {
  for (k in 1:A_num) {
    A_tilde <- new_A[k,] # get treatment assignment from new_A
    rdist[k] <- mean(df$Y[A_tilde == 1]) - mean(df$Y[A_tilde == 0]) + grid[i]
```

```
  }
  p.ci[i] <- mean(rdist >= T_stat)  # calculate p-value for each hypothesized
}

cbind(p.ci,grid)
```

```
##              p.ci  grid
##  [1,] 0.000000000 -0.80
##  [2,] 0.001587302 -0.75
##  [3,] 0.001587302 -0.70
##  [4,] 0.001587302 -0.65
##  [5,] 0.001587302 -0.60
##  [6,] 0.032539683 -0.55
##  [7,] 0.032539683 -0.50
##  [8,] 0.032539683 -0.45
##  [9,] 0.127579365 -0.40
## [10,] 0.184126984 -0.35
## [11,] 0.184126984 -0.30
## [12,] 0.184126984 -0.25
## [13,] 0.500000000 -0.20
## [14,] 0.500000000 -0.15
## [15,] 0.500000000 -0.10
## [16,] 0.500000000 -0.05
## [17,] 0.804960317  0.00
## [18,] 0.815873016  0.05
## [19,] 0.815873016  0.10
## [20,] 0.815873016  0.15
## [21,] 0.967460317  0.20
## [22,] 0.967460317  0.25
## [23,] 0.967460317  0.30
## [24,] 0.967460317  0.35
## [25,] 0.998412698  0.40
## [26,] 0.998412698  0.45
## [27,] 0.998412698  0.50
## [28,] 0.998412698  0.55
## [29,] 1.000000000  0.60
## [30,] 1.000000000  0.65
## [31,] 1.000000000  0.70
## [32,] 1.000000000  0.75
## [33,] 1.000000000  0.80
```

Given the output, the point estimate is the value with the highest p-value under the null. Therefore it is 0.60. Now, let's calculate the 90% CI.

```
prob <- rep(mean(df$A), length(df$A))  # probability of treatment assignment
perms <- t(new_A)  # transpose to get N-by-r matrix

# calculate the 90% confidence interval
ci_lower <- invert.ci(df$Y, df$A, prob, perms, 0.05)
ci_upper <- invert.ci(df$Y, df$A, prob, perms, 0.95)
```

90% CI: (0.2, 0.83)
The non-zero positive point estimate 0.60 indicates that the observed data suggests a positive effect of the

new treatment compared to the standard one. The 90% CI does not include 0, suggesting that we can reject the sharp null hypothesis at the 90% confidence level.

(9) First, I will calculate point estimate (SACE) and variance of the estimator.

```
Yb_obs1 <- mean(df$Y[df$A == 1])  # mean outcome for treatment group
Yb_obs0 <- mean(df$Y[df$A == 0])  # mean outcome for control group

# SACE point estimate
SACE <- Yb_obs1 - Yb_obs0

# variance
S1_sq <- var(df$Y[df$A == 1])  # new treatment group
S0_sq <- var(df$Y[df$A == 0])  # standard treatment group

N1 = 10
N0 = 10

# variance of the estimator
var_SACE = (S1_sq / N1) + (S0_sq / N0)
```

$\widehat{SACE} = 0.5$
$\hat{var}(\widehat{SACE}) = 0.037$

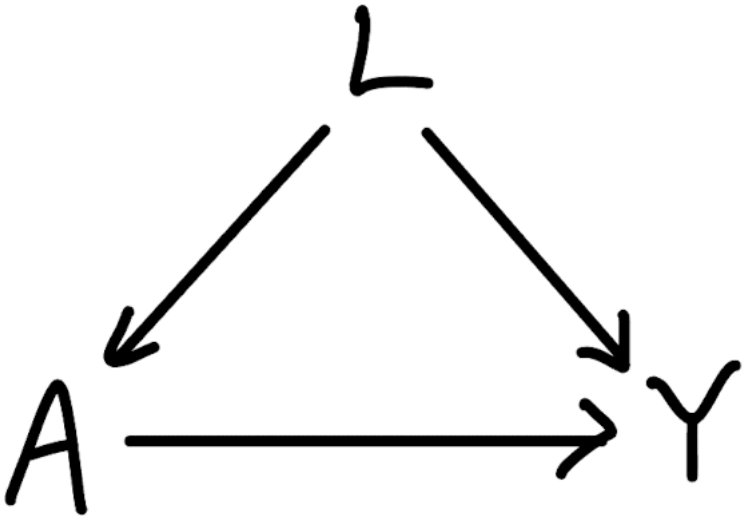
Confidence intervals can be calculated by $\widehat{SACE} \pm z\sqrt{\hat{var}(\widehat{SACE})}$. Therefore,

```
# 90% CI
z = qnorm(1 - 0.10 / 2)
ci_lower = SACE - z * sqrt(var_SACE)
ci_upper = SACE + z * sqrt(var_SACE)
```

90% CI: (0.19, 0.81)
The non-zero positive point estimate 0.5 indicates that on average, the new treatment increases the likelihood of disease prevention by 50% compared to the standard treatment. The 90% CI does not include 0, suggesting that we can reject the null hypothesis (there is no average new treatment effect) at the 90% confidence level.

(10) The estimate in (1) (0.3) is lower than the estimates obtained in both (8) (0.60) and (9) (0.5). The result from (9) suggests a stronger positive effect of the new treatment compared to (1). This discrepancy may arise from the small sample size of the study, as Neyman's inference relies on the Central Limit Theorem (CLT). The result from (8) focuses more on hypothesis testing rather than precise estimation. Despite the differences, both (8) and (9) point in the same direction, indicating a positive causal effect of the new treatment, consistent with the true estimate in (1).

(11) The scientific question would be "On average, how much does the new treatment affect disease prevention compared to the standard treatment, accounting for the potential confounding factor of normal or abnormal WBC count?"



(12)
A: Treatment (A = 1 new treatment, A = 0 standard treatment)
Y: Disease prevention (Y = 1 prevented, Y = 0 not prevented)
L: WBC count (L = 1 normal, L = 0 abnormal)
Representation:
$L \to A$: Individuals with normal WBC count (L = 1) are more likely to be prescribed the new treatment (A = 1)
$L \to Y$: Individuals with normal WBC count (L = 1) are more likely to have disease prevented (Y = 1)
$A \to Y$: The treatment (A) has a causal effect on disease prevention (Y)

(13) b. The causal estimate would appear to be larger in magnitude than the true causal effect. Those with normal WBC are more likely to receive the new treatment and also more likely to experience a better disease prognosis. This would exaggerate the apparent effect of the new treatment.

(14) g-formula for the observational study: $\sum_l E[Y|A=1, L=l]Pr(L=l) - \sum_l E[Y|A=0, L=l]Pr(L=l)$ under NUCA.
Given the table, I will calculate the estimate and 95% CI for ACE as follows:

```
# estimate
# probability of L
pl1 = 15/40
pl0 = 1 - pl1

# estimates of E[Y | A = a, L = l]
ya1l1 = (1+1+0+1+1+1+1+1+1+1)/10
ya1l0 = (0+1+1+1+1+0+0+1+0+1)/10
ya0l1 = (0+0+1+1+1)/5
ya0l0 = (0+0+0+0+1+1+1+1+0+0+0+0+1+0+1)/15

# estimate for ACE using g-formula
diff = (ya1l1*pl1+ya1l0*pl0) - (ya0l1*pl1+ya0l0*pl0)

# 95% CI
# critical value
z = qnorm(1 - 0.05 / 2)
```

```r
# size
na1l1 = 10
na1l0 = 10
na0l1 = 5
na0l0 = 15

# variance
va1l1 = var(c(1,1,0,1,1,1,1,1,1,1))
va1l0 = var(c(0,1,1,1,1,0,0,1,0,1))
va0l1 = var(c(0,0,1,1,1))
va0l0 = var(c(0,0,0,0,1,1,1,1,0,0,0,0,1,0,1))

# standard error
se = sqrt(va1l1/na1l1 + va1l0/na1l0 + va0l1/na0l1 + va0l0/na0l0)

# 95% CI
ci_lower = diff - z*se
ci_upper = diff + z*se
```

The estimate is 0.24 (95%CI: -0.42, 0.9). After adjusting for L, the point estimate indicates the positive effect of the new treatment compared to the standard treatment. However, the confidence interval for this estimate includes 0, implying that the effect is not statistically significant at the 95% confidence level.

(15) The ACE in (14) is lower than the ACE in (1). This is because, in (14), the potential confounder L (WBC count) was taken into account. By adjusting for L, we control for the bias introduced by the fact that individuals with normal WBC counts (who are more likely to receive the new treatment) also have a better prognosis. In (1), the confounder L was not accounted for, which may have led to an overestimation of the treatment effect due to this confounding factor.

(16) $Pr(A = 1|L = 1) = 0.67$
$Pr(A = 1|L = 0) = 0.4$
This supports the hypothesis of individuals with normal WBC counts (L = 1) to be more likely to be prescribed the new treatment.
$E[Y|A = 1, L = 1] - E[Y|A = 1, L = 0] = 0.3$
$E[Y|A = 0, L = 1] - E[Y|A = 0, L = 0] = 0.2$
When comparing WBC count status within each treatment group (as above), we can see that those with normal WBC counts (L = 1) to be more likely to have a better disease prognosis. Therefore, the data also supports the hypothesis.

(17) Conditioning on a set of variable {B, F} would close all the back-door paths between A and Y.

(18) NUCA: A is independent of Y_0 and Y_1 given the observed covariates.
If there are no unmeasured variables (e.g., U) that act as a confounder for both A and Y (i.e., no unmeasured variables are associated with A or Y conditional on A, and they do not lie on a causal pathway between A and Y), the DAG satisfies the NUCA.

(19) Conditioning on H would open a closed path from A to Y (e.g., $A \leftarrow H \leftarrow F \rightarrow Y$).

(20) Collider is a node on a path in a DAG where two arrows from different nodes both point into that node. Conditioning (adjusting) for a collider is problematic because it can create the association of its parents that has been d-separated. In other words, it can open a backdoor path between otherwise independent variables and create a non-causal association. The collider in the given DAG is H ($L \rightarrow H \leftarrow F$).

Question 2

(1)

- Units are the hospitals two years after the intervention (i.e., with or without workshops)
- Potential outcomes are the number of doctors from minority backgrounds promoted to leadership positions in each hospital two years later, either with or without receiving the workshop
- Treatment is the workshop focusing on the benefits of diversity in leadership (mandated due to majority of white doctors in leadership positions, or on request by hospital administrators)
- Observed covariates may include the current proportion of white doctors in leadership at baseline, hospital size, location, hospital type, insurance coverage, etc.

(2) I would be interested in examining the difference in the average number of physicians from minority backgrounds promoted to leadership positions between hospitals that took the diversity workshop and those that did not. In mathematical formula, this difference in effect can be described as:
$\sum_c E[Y|A=1,C=c]Pr(C=c) - \sum_c E[Y|A=0,C=c]Pr(C=c)$
where Y is the outcome, A is the treatment (1: receiving the workshop, 0: not receiving the workshop), and C is observed covariates or potential confounders.

(3) The study design is quasi-experimental because the workshop is assigned based on pre-existing conditions, rather than through randomized assignment. Specifically, hospitals with a higher proportion of white doctors in leadership positions automatically receive the workshop, while other hospitals receive it only upon request by hospital administrators. This non-randomized assignment introduces potential biases. This design is not ideal to address my question because the treatment assignment is influenced by pre-existing conditions (e.g., the racial composition of leadership, hospital budget, or administrators' race, which could affect their decision to request the workshop). As a result, the design may violate key assumptions such as positivity, randomization, and NUCA, which can complicate the interpretation of causal effects.

(4) A randomized controlled trial would be the ideal study design, where hospitals are randomly assigned to receive the workshop or not. Given potential discrepancies in the baseline characteristics of hospitals, I would recommend implementing block randomization, using the proportion of white physicians in leadership as a binary variable (>95% or $ $95%). Following this, the g-formula can be applied to estimate the causal effect between the treatment and control groups. If more hospital characteristics are available, propensity score matching would be appropriate to better balance the treatment and control groups. The causal effect could then be investigated using linear regression, incorporating all potential confounders into the model for a more precise estimation.