Midterm

Ngoc Duong

10/28/2020

Question 1

1.1

a) A=1: assigned new treatment

A=0: assigned standard treatment

Y=1: disease prevented

Y=0: disease not prevented

On the population level, the average causal effect of treatment on the outcome is defined as:

$$ACE = E[Y_1] - E[Y_0]$$

which is the average difference between the two counterfactuals: outcomes if these 20 people were given the new treatment and outcomes if the same people were given the standard treatment.

individual	Y1	Y0	Y1-Y0
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

$$\Rightarrow ACE = E[Y_1] - E[Y_0] = \frac{12}{20} - \frac{6}{20} = \frac{3}{10} = 0.3$$

Interpretation:

Since $E[Y_1 = 1] > E[Y_0 = 1]$, or ACE > 0, the treatment has a causal effect on the outcome Y, and the new treatment is better than the standard treatment on average.

Meanwhile, $Y_1 - Y_0$ is the treatment effect on the outcome for each individual.

1.2

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

Given we only observe one counterfactual outcome for each individual, we can only calculate the associational effect/risk difference. There are 10 people assigned to the new treatment (5 among whom have disease prevented), and 10 people assigned to the standard treatment (3 among whom have disease prevented):

$$E[Y|A = 1] - E[Y|A = 0] = \frac{5}{10} - \frac{3}{10} = \frac{2}{10} = 0.2$$

1.3

Under this assignment, the difference in observed group means/the associational effect is 0.2, which is smaller than the average causal effect.

We can expect some variability in the associational effects depending on how the random assignment is made. We need the assignment mechanism to be individualistic, probabilistic, unconfounded, and controlled for the associational effect to be equal to ACE. Since we don't know about the assignment mechanism from study 1, it's hard to say whether the mechanism satisfies all the necessary assumptions. Here, it doesn't seem to be the case.

A possible assignment that can violate the assumptions and might explain the difference is that the healthier response type (individuals who will have the disease prevented regardless) are more likely to be assigned to the standard treatment and vice versa (less healthy more likely to be assigned to new treatment).

1.4

a) In observational study, such data might be collected together including the treatments and other covariates. There might be inclusion/exclusion criteria, and from a larger qualified cohort of individuals assigned to either the new or standard treatment, 10 individuals might have been randomly sampled from each group. Regardless, we don't know about things like whether the treatment assignment was independent of the potential outcomes, whether why an individual received new treatment is because they are weaker/because they requested, etc.

b) In a randomized control trial, we might also have inclusion/exclusion criteria. Qualified individuals would be been recruited if they meet the criteria. The difference is that, once we have obtained 20 subjects, they are given the treatments in a controlled fashion before observing outcome (10 received new treatment and the remaining receive standard treatment randomly) . In other words, treatment assignment was randomized independently of the outcome, while observational studies may have treatment assisgnment depending on potential outcomes.

1.5

For randomized control trials, typically the assignment mechanism is unconfounded, individualistic, probabilistic, and controlled by design. Therefore, we can expect the associational effect to be the same as the average causal effect. Many well-designed and large RCTs should satisfy the above assignment mechanisms. On the other hand, in oservational studies, the above assumptions are hard to justify/test, and likely will not hold, so we cannot rule out observational study.

However, given a relatively small sample size and under simple randomization, we may find that by chance, some assignments might violate the positivity or exchangability assumption in the sample. Therefore, the observed table can also potentially come from a randomized trial. We cannot rule it out entirely, but the chance the observed data comes from a completely randomized RCT might be lower.

1.6

For block randomization, we can partition the individuals into strata that share similar aspects. The treatment assignment can be randomized within each block. This will ensure balance for covariate of interest and reduce the risk of confoundedness by chance for small sample seen in simple randomization.

Given that I know the truth, I will create 4 blocks for 4 response types (specifics in Appendix), and then perform simple randomization within these 4 blocks. In reality, since we don't know both counterfactual outcomes, we might want to stratify by relevant baseline covariates (age, sex, race, health profile, etc.) if they are available.

Given response type is directly related to potential outcomes, this block design assures probabilistic and confounded assumptions because: 1) every unit has a positive probability of being in either treatment group for each level of the covariate/potential outcome (e.g. healthy response type has some chance of getting either old or new treatment), and 2) assignment mechanism does not depend on the potential outcome because they are within blocks/people within each blocks are exchangeable.

In terms of treatment assignment, choose a random number from a uniform distribution for each of the treatments (1 and 0) within a block, and then re-rank the random number in order, which gives a randomized treatment assignment.

First block (healthiest)

A	rand	rank
1	0.6469028	1
0	0.6185018	2
0	0.4768911	3
1	0.3942258	4

Second block (least healthy)

A	rand	rank
0	0.9620645	1
1	0.7103224	2
1	0.3896344	3
1	0.2461373	4
0	0.0913837	5
0	0.0109333	6

Third block (“conformers”)

A	rand	rank
1	0.8828502	1
1	0.7093150	2
0	0.6381935	3
0	0.6318950	4
1	0.4944254	5
1	0.3842035	6
0	0.2407255	7
0	0.1002037	8

Fourth block (“non-conformers”)

A	rand	rank
0	0.3721239	1
1	0.2655087	2

Therefore, the treatment assignment in first block is (1,0,0,1). The treatment assignment in second block is (0,1,1,1,0,0). The treatment in the third block is (1,1,0,0,1,1,0,0). And the treatment in fourth block is (0,1)

The table with my observed data is:

individual	type	A	$Y \mid A = 1$	$Y \mid A = 0$
1	C	1	1	.
2	H	1	1	.
3	C	0	.	0
4	LH	0	.	0
5	H	1	1	.
6	H	1	1	.

individual	type	A	Y | A = 1	Y | A = 0
7	C	0	.	0
8	LH	0	.	0
9	C	1	1	.
10	LH	0	.	0
11	NC	0	.	1
12	LH	1	0	.
13	LH	0	.	0
14	H	1	1	.
15	C	1	1	.
16	NC	1	0	.
17	C	0	.	0
18	C	0	.	0
19	LH	0	.	0
20	C	1	1	.

1.7

Sharp null hypothesis

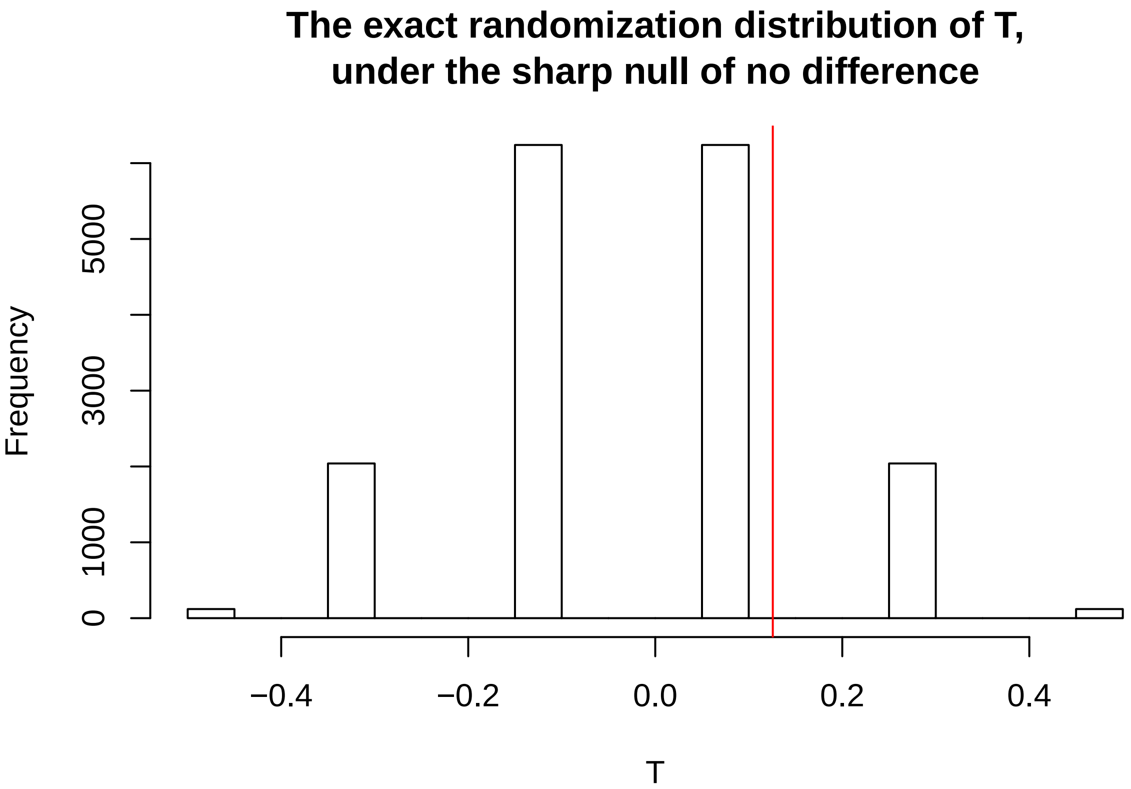
$$H_0 : \tau_i = Y_{1i} - Y_{0i} = 0$$

for all i

In other words, there is no individual-level treatment effect of treatment type on disease prevention for each observation.

Process:

For the observed vector of outcomes in **1.6**, I fix this outcome and permute the treatment assignments in each block of the 4 blocks. Then the number of total possible randomizations is $\binom{4}{2}\binom{8}{4}\binom{6}{3}\binom{2}{1} = 16800$. Then, we can calculate the statistic of interest under each randomziation scenario to obtain an exact randomization distribution. I used function "genperms" in package "ri" that allows for specification of blocks when doing permutations. Once I have obtained all test statistics under all possible randomzation scenarios, the p-value is calculated as the proportion of test statistic at least as extreme as my observed test statistic 0.3.



Interpretation: Since the exact p-value is $0.128 > 0.05$, we fail to reject the sharp null hypothesis and conclude that there is no individual causal effect of treatment type on disease prevention for all individuals in the sample.

1.8

Using Neyman's approach, the scientific question of interest would be: on average, what is the magnitude of the treatment effect on the outcome (e.g. in terms how many more/fewer people have disease prevented on average if all individuals got the new treatment versus the if they all got standard treatment), ie. we want to test the hypothesis of no mean difference $H_0 : \bar{Y}_1 = \bar{Y}_0$

1.9

Point estimate $= E[Y|A = 1] - E[Y|A = 0] = \frac{6}{10} - \frac{3}{10} = \frac{3}{10} = 0.3$

Sampling variance estimate $= \frac{S_1^2}{N_1} + \frac{S_0^2}{N_0} = \frac{0.267}{10} + \frac{0.233}{10} = 0.05$ (covariance term not included because technically we don't know the counterfactuals).

95% CI $= 0.3 \pm 2.26 \times \sqrt{0.05} = [-0.2058, 0.8058]$

Therefore, the point estimate of the marginal causal effect in my study is 0.3, and the 95%CI is [-0.2058,0.8058] using Neyman's approach.

Interpretation:

Point estimate $= 0.3$: those who have the new treatment have more cases of disease prevented than those who have standard treatment on average, and the average associational effect/difference is 0.3.

The true ACE will fall between -0.2058 and 0.8058 at least 95% of the times. Since the 95%CI contains the null value of 0, and the associated p-value is $0.1063 > 0.05$, we fail to reject the null hypothesis and conclude that there is no treatment effect on average at 5% significance level.

Compare with 1.1: Since I know the truth, I designed the block randomized trial to ensure it is probabilistic and unconfounded. I obtained an observed average causal effect similar to that in question 1.1. Since we have exchangeability (from randomization), positivity (from probabilistic assignment), SUTVA/consistency (from the context of the question) hold, it follows that the observed effect here is equal to the ACE using all counterfactuals. However, in reality, we don't know the truth, although knowing and leveraging relevant covariates in block randomization may also lead to an unbiased estimator and/or a confidence interval that captures the true average causal effect.

Problem 2

2.1

The units are all school districts across the country. This is because they are the object of investigation, i.e. the object which the treatment (workshops) applies to and has an outcome of interest (number of female and male teachers hired).

Potential outcomes are number of male and female teachers employed if the school districts had been given workshops and if they had not been given workshops. This is because depending on the treatment received by the units, we can potentially observe either of these outcomes.

The treatment is workshop for school administrators on the benefits of diversity. This is because we expect the two outcomes to be different when the treatment was implemented versus when it was not.

Observed covariates include the (baseline) number of female and male teachers in the school districts (before intervention/workshops), and whether or not school administrators within the district request workshops. These are some characteristics inherent to the units that may affect the treatment assignment and/or the observed outcomes.

2.2

The assignment is observational, since the treatment (workshop) is not randomized but rather mostly administered to school districts based on their observed baseline characteristics (e.g. number of female teachers).

2.3

The assignment mechanism is not probabilistic, since if a school district currently only has female teachers, the probability of them receiving the workshop is 1 (they are required to take them).

2.4

The assignment mechanism is not unconfounded, since how the treatment is assigned is dependent on the potential outcomes given the covariates. Specifically, the treatment is assigned with a view to increasing male teacher hires (potential outcome), depending on the number of current female teachers in the district (observed covariate).

Problem 3

The confounders in this paper are: certain demographic variables – sex and age groups (through ≥ 85 years), and more importantly, the underlying health status/profiles between the vaccinated and unvaccinated senior groups which might cause preferential receipt of vaccination and overall better observed health-related outcome in the healthier group. Another confounder might be time.

How the authors adjust for confounders

In their primary model, the authors only include/adjust for sex and age group.

In their secondary model,the authors further adjust for a set of disease covariates (obtained from diagnosis codes) with a view to capturing the underlying health status confounder. These include: heart disease, lung disease, diabetes mellitus, renal disease, cancer, vasculitis and rehumatoligic disease, dementia, hypertension, lipid disorders, as well as indicators for medical utilization – hospitalization of pneumonia and indicator of ≥ 12 outpatient visits in previous year.

To account for differences/changes in the cohort by time, these covariates (except for sex) were updated every year and allowed to vary by time in the Cox PH model. New participants who meet eligibility criteria were also included every year. Single-year models were also analyzed and compared to the multi-year model.

Appendix

Question 1.6

Response types:

*Most healthy: individuals 2, 5, 6, 14.

*Least healthy: individuals 4, 8, 10, 12, 13, 19

*People who conform (get disease prevented from new treatemnt): individuals 1,3,7,9,15,17,18,20

*People who are non-conformers (disease prevented from standard treatment): individuals 11, 16

```r
knitr::opts_chunk$set(echo = TRUE)
library(ri)
library(tidyverse)
library(perm)
library(ggplot2)
#Question 1.1, table as given
tibble(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),
       'Y1-Y0' = Y1-Y0) %>% #individual-level treatment effect
  knitr::kable()
#Question 1.2 table with observed outcome as given
tibble(individual = 1:20,
       A = c(1,1,0,0,0,0,1,0,0,0,1,1,1,0,0,1,1,1,1,0),
       "Y|A=1" = c(1,1,".",".",".",".",1,".",".",".",0,0,0,".",".",0,1,1,0,"."),
       `Y|A=0` = c(".",".",0,0,1,1,".",0,0,0,".",".",".",1,0,".",".",".",".",0)) %>%
  knitr::kable()
## 1.6
#First block (healthiest)
set.seed(2020)
h = tibble(A = c(1,1,0,0), rand = runif(4,0,1), #random number generation
           rank = rank(desc(rand))) #rank random numbers to obtain randomized assignment mechanism
```

```r
h_arr = h %>% arrange(rank)
h_arr %>% knitr::kable()
#Second block (least healthy)
set.seed(13)
lh = tibble(A = c(1,1,1,0,0,0), rand = runif(6,0,1), #random number generation
            rank = rank(desc(rand))) #rank random numbers to obtain randomized assignment mechanism
lh_arr = lh %>% arrange(rank)
lh_arr %>% knitr::kable()
#Third block ("conformers")
set.seed(1313)
c = tibble(A = c(1,1,1,1,0,0,0,0), rand = runif(8,0,1),  #random number generation
           rank = rank(desc(rand))) #rank random numbers to obtain randomized assignment mechanism
c_arr = c %>% arrange(rank)
c_arr %>% knitr::kable()
#Fourth block ("non-conformers")
set.seed(1)
nc = tibble(A = c(1,0), rand = runif(2,0,1), #random number generation
            rank = rank(desc(rand))) #rank random numbers to obtain randomized assignment mechanism
nc_arr = nc %>% arrange(rank)
nc_arr %>% knitr::kable()
#observed outcome given randomized treatment assignments in each block
tibble(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)) %>%
  #define response types
    mutate(type = ifelse(Y1 == 1 & Y0 ==1 ,"H",
                         ifelse(Y1 == 1 & Y0 == 0, "C",
                                ifelse(Y1 == 0 & Y0 == 0, "LH", "NC")))) %>%
    mutate(A = c(c_arr$A, h_arr$A, lh_arr$A, nc_arr$A),
           `Y | A = 1` = ifelse(A == 1, Y1, "."),
           `Y | A = 0` = ifelse(A == 0, Y0, ".")) %>% select(-Y1, -Y0) %>%
    knitr::kable()
## 1.7
#treatment assignment as determined
A <- c(c_arr$A, h_arr$A, lh_arr$A, nc_arr$A)

#specify the blocks sizes
block <- c(rep(1,8), rep(2,4),rep(3,6),rep(4,2))

#use genperms with argument blockvar to specify blocks
Abold <- genperms(A, blockvar=block, maxiter = 16800)

#the observed outcome in 1.6
Y_obs = c(1,1,0,0,1,1,0,0,1,1,1,1,0,0,0,0,0,0,1,0)

#create a vector of empty spots/placeholders for each statistic under each randomization scenario
rdist <- rep(NA, times = ncol(Abold))

#run a for loop through each randomization scenario and calculate the corresponding test statistic
#store the statistic in rdist vector
for (i in 1:ncol(Abold)) {
  A_tilde <- Abold[, i]
  rdist[i] <- mean(Y_obs[A_tilde == 1]) - mean(Y_obs[A_tilde == 0])
```

```
}
#observed statistic
t_obs = mean(Y_obs[A == 1]) - mean(Y_obs[A == 0])
#p-value calculated as the proportion of statistics equal or more extreme than the observed
#statistic under all possible randomizations
pval <- mean(rdist >= t_obs)
quant <- quantile(rdist,probs = 1-pval) #get the quantile in the distribution of this pval

#exact randomization distribution
hist(rdist, xlab = "T", main = "The exact randomization distribution of T,\nunder the sharp null of no
#where the observed test statistic is located in the exact randomization distribution
abline(v = quant,col="red")
## 1.9
t_crit = qt(0.975, 9) #95% t-crit with  9 degrees of freedom

#caculate estimate of sampling variance
var = var(Y_obs[A==1])/10 + var(Y_obs[A==0])/10

#calculate the 95% CI for the estimate
t_obs + t_crit*sqrt(var)
t_obs - t_crit*sqrt(var)

#calculate the p-value associated with the estimate under parametric t-distribution
pt(0.3/sqrt(var), df = 9, lower.tail = FALSE)
```