Homework 5

Lee Kennedy-Shaffer

Course Week 12

Do **all** parts.

Show your work and express all answers in lowest terms. You are welcome to use results and definitions from the posted document on the Moodle, but be sure to explain why a result or property can be used if it is not obvious (e.g., if certain conditions must be met, how do we know they are?).

Note: If you use R to get results, please also write the answers in context or make clear which number is which result.

Question 1: Cluster Randomized Trial Design

Suppose we are interested in designing a trial to test the effect of music education in high school on college graduation within 10 years of entering high school. We plan to conduct a randomized trial, randomizing entering high school students to being either required to take at least 2 years of music classes in high school (intervention arm) or no such requirement (control arm). Assume we will randomize 1:1, so half of the total enrollment is in the intervention arm and half in the control arm. Note that the control arm will not be prohibited from taking music classes; they will follow their school's standard requirements and options. In 10 years, we will follow up with these students and record whether they have graduated from college. We will compare the proportion in the intervention arm to the control arm. We plan to do a cluster-randomized trial, randomizing entire high schools to implement this requirement. We will use only one class year of students: those entering high school in 2022.

A. About 50% of U.S. high school graduates currently complete college within 6 years of completing high school. So we'll assume that in our control arm ($X = 0$), $E[Y|X = 0] = 0.50$. We want this study to have 80% power to detect an increase of 6 percentage points in the college graduation rate, with a significance level of $\alpha = 0.05$. If this were a standard, individually randomized RCT, how many people would we need to enroll in each arm? How many total?

B. Assume that each high school has 400 students in its entering class. Assume the intracluster correlation coefficient is $\rho = 0.005$. What is the design effect of the CRT?

C. Use the design effect to find the sample size per arm and total sample size required for the CRT. How many schools are required in each arm?

D. We can choose whether to enroll and collect outcome data on all students or a sample of students from each high school. Find the number of schools per arm and number of participants per arm we would need to enroll if we sample 50, 100, and 200 students per school (you can put your results directly into the first two columns of the table below E). Out of these designs and the one where we sample all 400 per school, which would you choose and why?

E. Recalculate the results from D if we assume a higher ICC of 0.01. Fill out the table below to summarize the results of D and E.

ρ:	0.005	0.005	0.01	0.01
m	Schools Per Arm	Students Per Arm	Schools Per Arm	Students Per Arm
50				
100				
200				
400				

F. Plot the number of schools required by the number of students per school for number of students per school ranging from 25 to 500, for both $\rho = 0.005$ and $\rho = 0.01$ (you can put them on the same plot or two different plots).

Question 2: Comparing Designs

For this question, consider the same question of interest as in Question 1 (and the same CRT design where appropriate).

A. Explain the estimand that we are actually going to estimate through the CRT in Question 1. How well or poorly does it match the original question of interest?

B. Name at least one advantage and disadvantage each comparing this CRT to a similar individually-randomized trial and a similar observational study (like the example school study from class). Be specific.

C. Imagine that New York implements a policy requiring all high schools to require at least 2 years of music education. Explain how we could use a quasi-experimental design (difference-in-differences or synthetic control) to assess the results. What comparators would you use? What would the advantages and disadvantages of this design be?

D. Rank the designs (CRT, individually randomized trial, observational study, quasi-experiment) in terms of how strong and interpretable you think the evidence they generate would be, and comment on the relative feasibility or ethics of them as well. Explain your rankings very briefly.

Question 3: Confounding Bias

A frequent talking point in professional sports today is whether money can buy championships. That is, do teams that spend more money on players perform better than teams that spend less money? We are going to conduct an observational study to see whether increased payroll (the amount spent on players in dollars) causes an increase in the winning percentage of a team (measured as a decimal between 0 and 1, denoted Y). To make it simpler (and because we don't think there's a linear relationship), we are going to categorize all teams into either high or low spenders ($X = 1$ for high spenders and $X = 0$ for low spenders).

A. Suppose we have 30 data points from baseball teams last year. There are 15 high-spending observations and 15 low-spending observations. Suppose, for now, that there is no effect of spending on winning percentage. Suppose that winning percentage for each observation is normally distributed with mean 0.5 and standard deviation 0.08 (so variance 0.0064): $Y \sim N(0.5, 0.0064)$. Find the expectation and variance of the estimator $\hat{\theta} = \bar{Y}_1 - \bar{Y}_0$, where $\bar{Y}_1$ is the average winning percentage among high-spending observations and $\bar{Y}_0$ is the same among low-spending observations. Assume that all observations are independent (not a realistic assumption if some of the observations are the same team in different years or different teams in the same year).

B. Continue to assume that there is no causal effect of spending on winning percentage. But, a team's performance the year before affects both their winning percentage and their payroll. We'll measure this confounder using L, an indicator of whether the team was good the year before ($L = 1$) or bad the year before ($L = 0$). The 30 observations can be split as follows: 15 have $L = 1$ and 15 have $L = 0$. Among those with $L = 1$, 3/5 have $X = 1$ and 2/5 have $X = 0$. Among those with $L = 0$, 2/5 have $X = 1$ and 3/5 have $X = 0$. So being good the year before increases your chances of being a high spender. Among those with $L = 1$, $Y|L = 1 \sim N(0.53, 0.005)$. *Note that stratification has reduced the variance too, like it did for RCTs.* Among those with $L = 0$, $Y|L = 0 \sim N(0.47, 0.005)$. So the effect of being good the year before is a bit less than a one-standard deviation difference in winning percentage. Find $E[Y|X = 1]$ and $E[Y|X = 0]$. *Hint: Use the law of total expectation*: $E[Y|X = 1] = E[Y|X = 1, L = 1]P[L = 1|X = 1] + E[Y|X = 1, L = 0]P[L = 0|X = 1]$, and similar for $X = 0$.

C. Find the expectation and bias of the estimator $\hat{\theta}$. Does it appear that high spending is helpful or harmful to a team?

D. To account for this confounding, we will conduct a stratified analysis, stratifying by L. Let $\hat{\theta}_1$ be the difference in average winning percentage between high- and low-spending teams **among teams with** $L = 1$ and let $\hat{\theta}_0$ be the difference in average winning percentage between high- and low-spending teams **among teams with** $L = 0$. The overall estimator will be $\hat{\theta}_{Strat} = \frac{\hat{\theta}_1 + \hat{\theta}_0}{2}$. Show that this estimator is unbiased.

E. Find the variance of the stratified estimator $\hat{\theta}_{Strat}$ from part D using the distributions given in part B. You can assume all observations/strata are independent. Find the MSE of this study.

F. Suppose we could get more data points, but without the ability to collect information on L. Then our bias would be the same as in part C (remember, larger sample size doesn't generally reduce bias). The unstratified estimator has the variance from part A, scaled by the inverse of the sample size. So if we double the sample size, the variance is multiplied by 1/2. Find the factor k by which we would have to increase the sample size so that the unstratified estimator $\hat{\theta}$ has the same or lower MSE compared to the stratified $\hat{\theta}_{Strat}$ from part E.

Question 4: Mathematical Analysis of Confounding

Generally, for an observational study, the magnitude of confounding bias depends on the confounder's association with the exposure and its association with the outcome. We can identify these relationships mathematically. Suppose we have a study of the effect of a binary exposure Z (that takes values 0 and 1 for unexposed and exposed individuals, respectively) and a continuous outcome Y. We are concerned about a continuous confounder X.

Assume that the outcome has the following relationship with the variables:

$$Y = \mu_0 + \theta Z + \beta X + \epsilon,$$

where μ_0, θ, β are constant parameters and $\epsilon \sim N(0, \sigma^2)$ is independent of any other RVs. Further, assume that:

$$X|Z = 1 \sim N(\nu_1, \sigma_2^2)$$
$$X|Z = 0 \sim N(\nu_0, \sigma_2^2)$$

When relevant, assume that we have n_1 exposed individuals and n_0 unexposed individuals in the study.

A. Which parameter represents the true causal effect of Z? Explain in a sentence or two.

B. What is the expected value of Y for an exposed individual ($E[Y|Z = 1]$)? An unexposed individual?

C. Let $\hat{\theta}_u$ be an estimator that takes the average outcome among exposed individuals and subtracts the average outcome among unexposed individuals. Under the assumptions of this problem, what will its expectation be? Is it unbiased for the true causal effect identified in part A? If not, find a formula for the bias and simplify as much as possible.

D. Identify any conditions by which this estimator would be unbiased. Describe these conditions in words as well as in mathematical formulae.

E. Let $\mu_0 = 10$, $\theta = 2$, $\beta = 0.7$, $\sigma^2 = 1$, $\nu_1 = 1.3$, $\nu_0 = 0.6$, and $\sigma_2^2 = 0.01$. Run 1000 simulations of a study with these conditions with $n_1 = n_2 = 30$. For each simulation, find $\hat{\theta}_u$ as defined above and $\hat{\theta}_r$, a regression estimator of θ that is the estimated coefficient of the exposure variable in a regression that adjusts for X. Find the empirical expectation and variance of each of these two estimators. Does it match results you had above? What is the bias of $\hat{\theta}_r$? Which estimator appears to be more efficient? **Some sample code is available at the end of the homework.**

Question 5: Quasi-Experimental Designs

Choose one randomized or observational study that we read for this class. Describe a quasi-experiment (difference-in-differences, synthetic control, or something else you've seen) that could address a similar question of interest. The policy change does not need to have actually happened, but it needs to be something that could happen. In one page or less, describe the relative advantages and disadvantages of the quasi-experiment you designed compared to the actual study. Consider the estimands they target, bias, variance, generalizability, or any other relevant properties.

Sample R Code for Question 4

The following function runs a single simulated trial by creating a data set of n0 unexposed observations and n1 exposed observations with the parameters mu0 (μ_0), theta (θ), beta (β), sigma2 (σ^2), nu1 (ν_1), nu0 (ν_0), and sigma22 (σ_2^2). It then generates random X and ϵ values according to their distributions and the inputted parameters and random Y values for each observation based on those X and ϵ values. The output is a vector with two values: theta_u gives the unadjusted estimator value ($\hat{\theta}_u$) for that simulated trial; theta_r gives the regression-adjusted estimator value ($\hat{\theta}_r$) for that simulated trial. You can run the simulation, get all 1000 simulated outputs, and work with them using similar code to past homeworks (the replicate and apply functions will be helpful).

```
OneSim <- function(mu0,theta,beta,sigma2,nu1,nu0,sigma22,n1,n0) {
  input <- tibble(Z=c(rep(0,n0),rep(1,n1))) %>%
    mutate(X=rnorm(n1+n0,mean=nu0+(nu1-nu0)*Z,sd=sqrt(sigma22)),
           epsilon=rnorm(n1+n0, mean=0, sd=sqrt(sigma2)),
           Y=mu0+theta*Z+beta*X+epsilon)
  reg <- lm(Y~Z+X, data=input)
  return(c(theta_u=mean(input %>% filter(Z==1) %>% pull(Y)) -
             mean(input %>% filter(Z==0) %>% pull(Y)),
           theta_r=reg$coefficients["Z"]))
}
```