Homework 2 **Soc 723**
Research questions and estimands

Problem Set

In this exercise, you will simulate an intervention study with a predetermined average treatment effect. The goal is for you to understand the potential outcomes framework, and the properties of completely randomized experiments through simulation.

The setting for our hypothetical study is a class in which students take two quizzes. After quiz 1 but before quiz 2, the instructor randomly assigns half the class to attend an extra tutoring session. The other half of the class does not receive any additional help. Consider the half of the class that receives tutoring as the treated group. The goal is to estimate the effect of the extra tutoring session on average test scores for quiz 1.

1. Simulate a dataset consistent with the following assumptions:
   (a) The average treatment effect on all the students, δ, equals 5.
   (b) The population size, N, is 100.
   (c) Scores on quiz 1 follow a normal distribution with mean of 65 and standard deviation of 3.
   (d) The potential outcomes for quiz 2 should be linearly related to the pre-treatment quiz score, x. In particular, they should take the form:

$$y^0 = \beta_0 + \beta_1 x + u_0$$
$$y^1 = \beta_0 + \beta_1 x + \delta + u_1$$

   where the intercept $\beta_0 = 10$ and the slope $\beta_1 = 1.1$. Draw the errors u_0 and u_1 independently from normal distributions with mean 0 and standard deviations 1.
   (e) Assign half of the students to the treatment condition and the other half to the control.
2. Answer the following questions about the data-generating process:
   (a) What is your interpretation of δ?
   (b) How would you interpret the intercept for Y^0 and Y^1?
   (c) How would you interpret β_1?
3. Answer the following questions about the effect of the tutoring session on student performance:
   (a) Calculate the true sample average treatment effect ($SATE$) using potential outcomes and compare it to δ. Are they different?

(b) Using a difference in means in the observed outcome, estimate the sample average treatment effect $\widehat{SATE}$ and compare it to the true $SATE$ and δ. Are they different? Why?

(c) Repeat the treatment assignment 500 times, each time calculating the new $\widehat{SATE}$. Plot the distribution for $\widehat{SATE}$. What is the mean of this distribution? And the standard deviation?

(d) Using the data for the original experiment, estimate $\widehat{SATE}$ using two regressions, one including the quiz 1 scores and the other one without. Which estimate is closer to the true $SATE$?

Bonus question: Can you guess why? (If you can't that's okay, we did not discuss this in class, but give it a try).

(e) Repeat the treatment assignment 500 times again, but this time estimate the $\widehat{SATE}$ using the regression with the quiz 1 score as predictor. Plot the distribution of $\widehat{SATE}$ across this 500 iterations and calculate the mean and standard deviation. Compare it to the distribution of $\widehat{SATE}$ you got using the difference in means. Which is a better method to estimate $SATE$?