# Problem set 4

Your name here

Due 10/22/2021 at 5pm

NOTE1*: Start with the file* `ps4_2021.Rmd` *(available from the github repository at https://github.com/UChicago-pol-methods/IntroQSS-F21/tree/main/assignments). Modify that file to include your answers. Make sure you can "knit" the file (e.g. in RStudio by clicking on the* `Knit` *button). Submit both the Rmd file and the knitted PDF via Canvas*

## Question 1

Consider a study that seeks to measure the effect of daily exercise on subjective mental health during finals week in December 2021 (as measured by a survey) among UChicago students.

1a) For a given student, what (in words) are the potential outcomes in this study?

1b) With reference to the Fundamental Problem of Causal Inference, explain why it is difficult if not impossible to measure the effect of this treatment on this outcome for any individual student.

1c) In observational studies, treatment may be related to the potential outcomes. Give one plausible account of how and why treatment may be related to the potential outcomes in this example.

## Question 2

The code below creates a fake dataset with a population of 1000 observations and two variables: `Y0` is $Y(0)$, the potential outcome with treatment set to 0, and `Y1` is $Y(1)$, the potential outcome with treatment set to 1. (Note that observing both potential outcomes is generally not possible; we can do it here because it's a fake dataset.)

```
set.seed(30500)
n <- 1000
dat <- tibble(Y0 = runif(n = n, min = 0, max = 1)) %>%
  mutate(Y1 = Y0)
```

2a) Compute the *individual treatment effect (ITE)* for each individual and plot an ECDF of the ITEs. (*Hint*: see ECDF code in lecture 3.2.)

2b) Suppose we choose as our estimand the average treatment effect (ATE). What is the ATE for this population?

2c) Add a treatment variable `D` that takes on the value 1 with probability `Y1` and 0 otherwise. (Hint: use the `rbinom()` function.)

2d) Compute the difference in means using this treatment variable and compare it to the ATE. Why is the difference in means a bad estimator for the ATE in this case?

2e) Create a new treatment variable `D_random` that is assigned at random, as if this were a randomized experiment.

2f) Compute the difference in means using this treatment variable and compare it to the ATE.

## Question 3

The code below creates another fake dataset with a population of 1000 observations and the same two variables, `Y0` and `Y1`.

```
dat <- tibble(Y0 = rnorm(n = n, mean = 0, sd = 1)) %>%
  mutate(Y1 = Y0 + rnorm(n = n, mean = .5, sd = .5))
```

3a) Compute the *individual treatment effect (ITE)* for each individual and plot an ECDF of the ITEs.

3b) Create a scatterplot of `Y1` (vertical axis) against `Y0` (horizontal axis).

3c) If this were a study of students, and `Y` were a measure of academic achievement (with `D` a study skills training session), how would you interpret a point at (2,2) on the previous plot? How about a point at (-1, 0)?

3d) Suppose we choose as our estimand the average treatment effect (ATE). What is the ATE for this population?

3e) Create a new variable `pr_treatment` that is `1 - exp(Y0)/(1 + exp(Y0))`. Plot `pr_treatment` (vertical axis) as a function of `Y0`.

3f) Again supposing `Y` is a measure of academic achievement and `D` a study skill training, why might the probability of treatment be related to `Y0` as in this hypothetical example?

3g) Add a treatment variable `D` that takes on the value 1 with probability `pr_treatment` and 0 otherwise. Hint: use the `rbinom()` function.

3h) Compute the difference in means using this treatment variable and compare it to the ATE. Why is the difference in means a bad estimator for the ATE in this case?

3i) Create a new treatment variable `D_random` that is assigned at random, as if this were a randomized experiment.

3j) Compute the difference in means using this treatment variable and compare it to the ATE.