Problem set 4

Your name here

Due 10/25/2022 at 5pm

NOTE1: Start with the file `ps4_2022.Rmd` (available from the github repository at https://github.com/UChicago-pol-methods/IntroQSS-F22/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

Suppose you are interested in whether giving all school kids at the local elementary school a free lunch will affect how fast they complete a pop quiz in the afternoon. The (very small) school is our entire population of interest.

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

1b) You propose a study design. You will give all of the students free lunch one day. That afternoon, you will have all students take a pop quiz and time them. The next day, you plan to make sure no one gets a free lunch, and again you will have all of the students take a pop quiz in the afternoon. If you want to recover every student's individual treatment effect from your study, what assumptions could you make about the data? (your assumption does not need to be realistic).

1c) You're not satisfied with the design of your first study, so you try another study design. You record all of the students that are receiving free lunches already through a school program, and all of the students that do not currently receive a free lunch. That afternoon, you have all students take a pop quiz and time them. Test times in seconds are recorded below:

```
# test times of students who got a free lunch
Y_given_free_lunch <- c(55, 65, 70, 80, 80, 90, 90, 105, 120, 180)
# test times of students who did not get a free lunch
Y_given_no_free_lunch <- c(55, 55, 65, 70, 75, 80, 80, 90, 90, 105, 120)
```

If you want to identify every individual student's counterfactual test times under the other treatment, what assumptions could you make about the data? (your assumption does not need to be realistic). Under your assumptions, report test times for each group if they were in the other condition, e.g., what test times for students who got a free lunch WOULD be if they had NOT gotten a free lunch.

```
# your work here
```

1d) Suppose you only want to get the average treatment effect. What assumptions about the data could you make to identify the average treatment effect? Under these assumptions, how would you calculate the average treatment effect?

```
# your work here
```

1e) Was the assumption that you made in part 1c a realistic assumption? What is an example of how the assumption could be violated?

Question 2

2a) Consider the below table. Fill in the missing cells.

$Y_i(0)$	$Y_i(1)$	D_i	Y_i	τ_i
	1	1	1	1
1	1	0		
1		1		0
0	1	0		
1	0		0	
1	1	1		
	0	0	0	
1	1	1	1	0
	0		1	

2b) What is the true ATE for the population represented in the table?

```
# your work here
```

2c) If you take the difference in means based on *observed responses* in treatment and control, what is the estimated ATE?

```
# your work here
```

Question 3

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 data set) This data set represents our entire population of interest.

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

3a) Compute the *individual treatment effect (ITE)* for each individual.

```
# your work here
```

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

```
# your work here
```

3c) Add a treatment variable `D` that takes on the value 1 with probability `Y1` and 0 otherwise. (Hint: use the `rbinom()` function)

```
# your work here
```

3d) 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?

```
# your work here
```

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

```
# your work here
```

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

```
# your work here
```

Question 4

The code below creates another fake dataset with a population of 1000 observations and the same two variables, `Y0` and `Y1`. This data set again represents our entire population of interest.

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

4a) Compute the *individual treatment effect (ITE)* for each individual.

```
# your work here
```

4b) Create a scatterplot of `Y1` (vertical axis) against `Y0` (horizontal axis).

```
# your work here
```

4c) 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)?

4d) Suppose we choose as our estimand the average treatment effect (ATE). What is the ATE for this population?

```
# your work here
```

4e) Create a new variable `pr_treatment` that is `1 - exp(Y0)/(1 + exp(Y0))`. Plot `pr_treatment` (vertical axis) as a function of `Y0`.

```
# your work here
```

4f) 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?

4g) Add a treatment variable `D` that takes on the value 1 with probability `pr_treatment` and 0 otherwise. Hint: use the `rbinom()` function.

```
# your work here
```

4h) 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?

```
# your work here
```

4i) Create a new treatment variable `D_random` that is assigned at random, as if this were a randomized experiment.

```
# your work here
```

4j) Compute the difference in means using this treatment variable and compare it to the ATE.

```
# your work here
```