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 : Causal identification

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?

One potential outcome for a given student is the amount of time it would take them to complete the pop quiz if they got a free lunch. The other potential outcomes is the amount of time it would take them to complete the pop quiz if they did not get a free lunch.

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).

Per Holland (1986), you could assume "temporal stability," i.e., potential outcomes do not change over time, and "causal transience," i.e., whether students get lunch or not one day does not affect their test times the next day. There are other assumptions that might also facilitate identification.

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.

You could assume, for example, that the effect is constant, and students who get a free lunch take the quiz 10 seconds faster than they would otherwise, and students who don't get a free lunch are 10 seconds slower than they would be if they did get a free lunch. This is not the only assumption that would allow you to identify the individual counterfactuals!

```
# test times of students who got a free lunch
# IF they had NOT gotten a free lunch
Y_given_free_lunch - 10
```

```
##  [1]  45  55  60  70  70  80  80  95 110 170
```

```
# test times of students who did not get a free lunch
# IF they HAD gotten a free lunch
Y_given_no_free_lunch + 10
```

```
##  [1]  65  65  75  80  85  90  90 100 100 115 130
```

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?

We could simply assume independence; whether or not students are enrolled in the free lunch program has nothing to do with their potential outcomes (i.e. the test times they would receive under each treatment condition).

```
mean(Y_given_free_lunch) - mean(Y_given_no_free_lunch)
```

```
## [1] 13.04545
```

Alternatively, we could assume constant effects, as proposed here for part 1b. If that were the case, the average treatment effect would be exactly 10.

1e)

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

We don't know very much about the current free lunch program, but the assumption of independence is probably not realistic. This assumption would be violated if, for example, students who got free lunches on average tend to take tests faster (or slower) than students who do not, regardless of whether they get free lunch or not.

Question 2 : Counterfactuals

Consider the below table.

2a) 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	

answer

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

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

```
tau_i <- c(1,0,0,1,-1,0,0,0,-1)
mean(tau_i)
```

```
## [1] 0
```

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

```
Y_0obs <- c(1,0,0,1)
Y_1obs <- c(1,1,0,1,1)
(hat_tau <- mean(Y_1obs) - mean(Y_0obs))
```

```
## [1] 0.3
```

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.

```
dat |>
  mutate(ITE = Y1 - Y0)
```

```
## # A tibble: 1,000 x 3
##       Y0    Y1   ITE
##    <dbl> <dbl> <dbl>
##  1 0.420 0.420     0
##  2 0.855 0.855     0
##  3 0.842 0.842     0
##  4 0.355 0.355     0
##  5 0.454 0.454     0
##  6 0.681 0.681     0
##  7 0.364 0.364     0
##  8 0.683 0.683     0
##  9 0.539 0.539     0
## 10 0.642 0.642     0
## # ... with 990 more rows
```

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

It's obvious that the ATE is zero from the setup, but we can compute it thus:

```
dat|>
  summarize(ate = mean(Y1) - mean(Y0))
```

```
## # A tibble: 1 x 1
##     ate
##   <dbl>
## 1     0
```

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

```
dat2 <- dat |>
  mutate(D = rbinom(n, size = 1, prob = Y1))
```

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?

```
dat2 |>
  summarize(mean(Y1[D == 1]) - mean(Y0[D==0]))
```

```
## # A tibble: 1 x 1
##   `mean(Y1[D == 1]) - mean(Y0[D == 0])`
##                                   <dbl>
## 1                                 0.361
```

The treament is related to the potential outcomes: here, subjects with higher $Y(1) = Y(0)$ have a higher probability of being treated. This means that the treated units would have a higher value of the outcome variable than the control units even if (as is true here) the treatment has no effect.

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

```
dat3 <- dat |>
  mutate(D_random = sample(x = rep(c(0,1), n/2)))
# here, we will have exactly balanced treatment and control assignments, but in
# a random order
```

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

```
dat3 |>
  summarize(mean(Y1[D_random == 1]) - mean(Y0[D_random==0]))
```

```
## # A tibble: 1 x 1
##   `mean(Y1[D_random == 1]) - mean(Y0[D_random == 0])`
##                                                 <dbl>
## 1                                           -0.00811
```

It should be close. It differs from the ATE only due to random variation in the treatment variable `D`.

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))
```

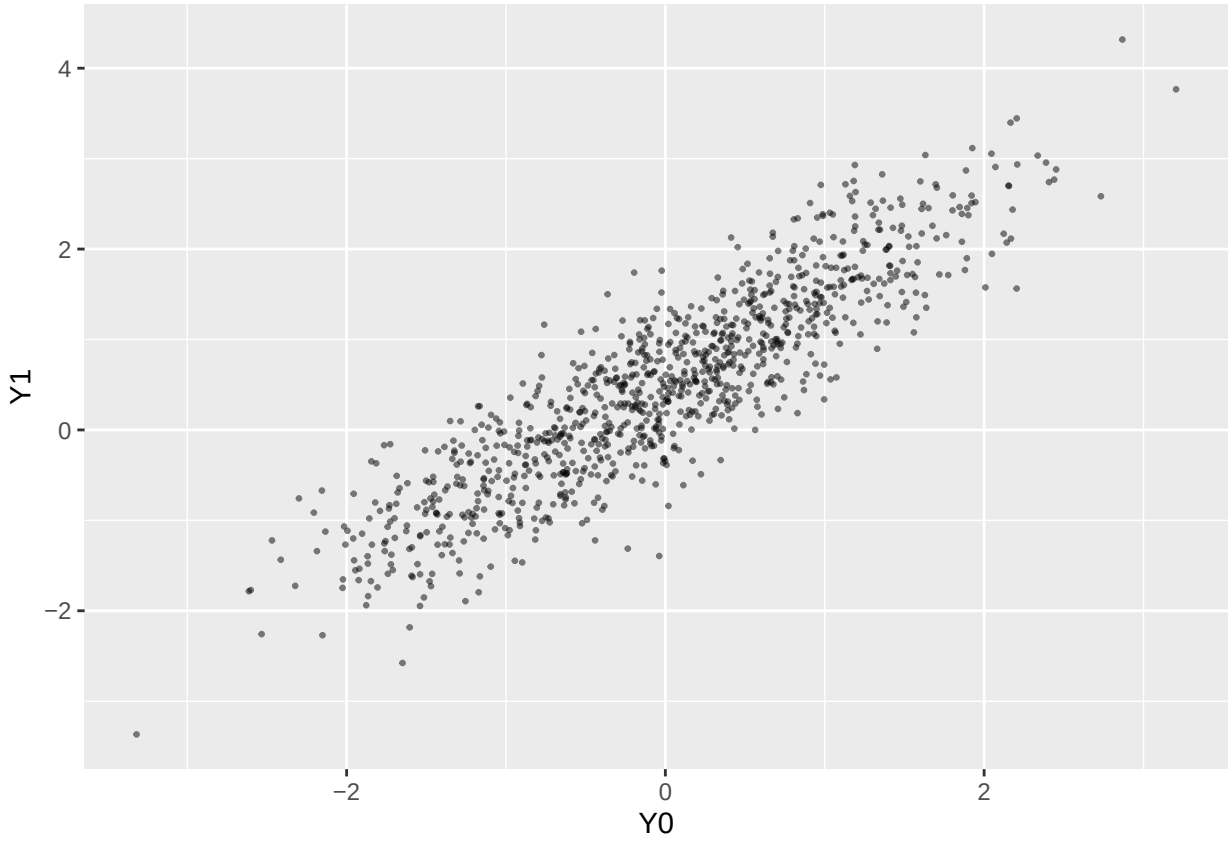
4a) Compute the *individual treatment effect (ITE)* for each individual.

```
dat |>
  mutate(ITE = Y1 - Y0)
```

```
## # A tibble: 1,000 x 3
##         Y0     Y1     ITE
##      <dbl>  <dbl>   <dbl>
##  1 -0.515   0.434  0.949
##  2  0.143   1.25   1.10
##  3  1.45    1.70   0.251
##  4 -1.54   -1.60  -0.0577
##  5  0.0957  0.202  0.106
##  6  0.778   1.24   0.465
##  7  2.34    3.03   0.697
##  8  1.27    2.05   0.777
##  9 -0.170   0.857  1.03
## 10 -0.0367  0.864  0.901
## # ... with 990 more rows
```

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

```
dat |>
  ggplot(aes(x = Y0, y = Y1)) +
  geom_point(alpha = .5, size = .5)
```



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)?

A point at (2,2) corresponds to a student who is totally unaffected by the treatment and would do well whether or not she attends the study skills session. A point at (-1,0) corresponds to a weaker student who does 1 point better when she attends the session than when she does not.

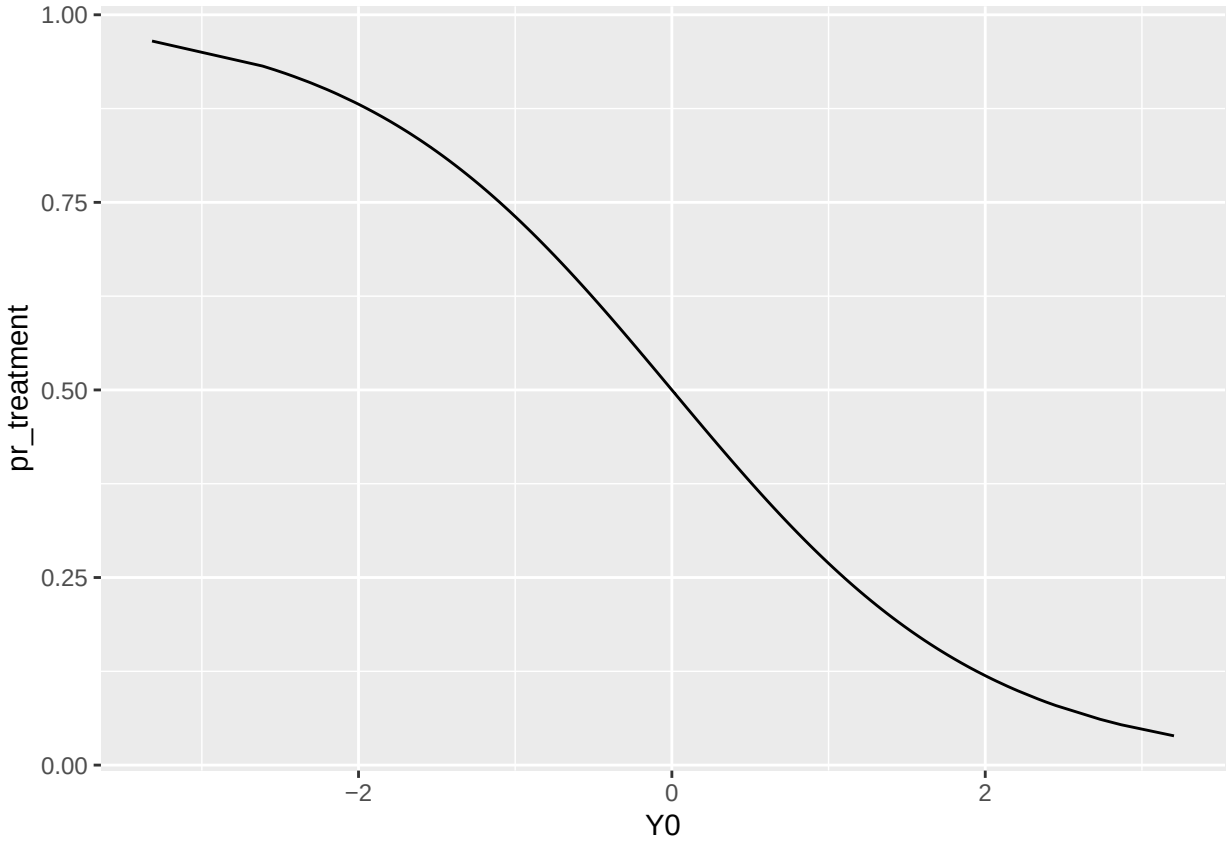
4d) Suppose we choose as our estimand the average treatment effect (ATE). What is the ATE for this population?

```
dat |>
  summarize(mean(Y1) - mean(Y0))
```

```
## # A tibble: 1 x 1
##   `mean(Y1) - mean(Y0)`
##                   <dbl>
## 1                 0.505
```

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`.

```
dat2 <- dat |>
  mutate(pr_treatment = 1 - exp(Y0)/(1 + exp(Y0)))
dat2 |>
  ggplot(aes(x = Y0, y = pr_treatment)) +
  geom_line()
```



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?

It could be that the weaker students are more likely to seek out support. Or perhaps they are more likely to be asked or required to attend.

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

```
dat3 <- dat2 |>
  mutate(D = rbinom(n, size = 1, prob = pr_treatment))
```

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?

```
dat3 |>
  summarize(mean(Y1[D == 1]) - mean(Y0[D==0]))
```

```
## # A tibble: 1 x 1
##   `mean(Y1[D == 1]) - mean(Y0[D == 0])`
##                                   <dbl>
## 1                                -0.282
```

Again, the treatment is related to the potential outcomes. Here the students who undertake the training would

have done worse than those who didn't even if none of them had undertaken the training. In this case it makes it seem like the training makes students do worse even though it actually makes them do .5 better on average.

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

```
dat3 <- dat2 |>
  mutate(D_random = sample(x = rep(c(0,1), n/2)))
```

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

```
dat3 |>
  summarize(mean(Y1[D_random == 1]) - mean(Y0[D_random==0]))
```

```
## # A tibble: 1 x 1
##   `mean(Y1[D_random == 1]) - mean(Y0[D_random == 0])`
##                                                 <dbl>
## 1                                               0.500
```

This is very close to the target of .5; any difference you get reflects random variation in the treatment assignment.