# Homework 1 Answers

## Instructions

- This homework is due Wednesday, January 19 at 3pm EST.
- Please format your homework solutions using R Markdown. You are welcome to simply add your answers below each question.
  - If the question requires a figure, make sure you have informative title, axis labels, and legend if needed.
- Turn in both the .rmd file and the knitted .pdf file.
  - Knitting the .rmd file to a .pdf file should help ensure your code runs without errors, but double check the .pdf output is what you expected.

## Question 1 (Exercise 1 from GG Ch. 1)

In your own words, what is an experiment and how does it differ from an observational study?

Answer:

A randomized experiment is a study in which observations are allocated by chance to receive some type of treatment; in an observational (or non-experimental) study, treatments are not assigned randomly.

## Question 2 (Exercise 2 from GG Ch. 1)

Would you classify the study described in the following abstract as a field experiment, a naturally occuring experiment, a quasi-experiment, or none of the above? Why?

"This study seeks to estimate the health effects of sanitary drinking water among low-income villages in Guatemala. A random sample of all villages with fewer than 2,000 inhabitants was selected for analysis. Of the 250 villages sampled, 110 were found to have unsanitary drinking water. In these 110 villages, infant mortality rates were, on average, 25 deaths per 1,000 live births, as compared to 5 deaths per 1,000 live births in the 140 villages with sanitary drinking water. Unsanitary drinking water appears to be a major contributor to infant mortality."

Answer:

This study is a quasi-experiment. Although villages are sampled randomly, random assignment is not used to determine which villages receive sanitary drinking water (the treatment in this study). The lack of random assignment means that this study does not qualify as either an experiment or natural experiment, the latter being a special kind of experiment in which governments or other non-academic entity allocates treatments randomly.

## Question 3

Using DeclareDesign, declare a population with 100 people. Include age as a covariate. Assume `age` in our population is distributed `age` $\sim N(40, 10)$.

Next, imagine (i.e., declare) potential outcomes for each person in your sample. We are interested in the outcome of the person's salary. If a person is treated they attend a job training program, and if a person is untreated they do not attend the program. *If* everyone attended the job training program, assume the average salary for the population would be 65K. *If* everyone did *not* attend the job training program, assume the average salary for the population would be 55K. Also assume some heterogeneity in the population's salaries whether they attend the job training program or not. We'll assume a similar amount of heterogeneity in both cases, so in either condition, add noise distributed $N(0, 5K)$.

Finally, create your `design` object by adding these two declarations together.

Answer:

```
library(DeclareDesign)
```

```
## Loading required package: randomizr
```

```
## Loading required package: fabricatr
```

```
## Loading required package: estimatr
```

```
set.seed(13248)

N <- 100
treatment_mean <- 65
control_mean <- 55
population <- declare_population(N = N,
                                 age = rnorm(N, 40, 10),
                                 u_0 = rnorm(N, 0, 5),
                                 u_1 = rnorm(N, 0, 5))

potential_outcomes <- declare_potential_outcomes(Y ~ Z*(treatment_mean + u_1) +
                                                   (1-Z)*(control_mean + u_0))

design <- population + potential_outcomes
```

## Question 4

Using the `draw_data` function, draw one hypothetical population's potential outcomes using the provided code. Print the first ten lines of your dataset. Then, answer the following questions.

```
single_draw <- draw_data(design)
single_draw[1:10, ]
```

```
##     ID      age        u_0          u_1    Y_Z_0    Y_Z_1
## 1  001 26.60539  5.2277918  -0.02213325 60.22779 64.97787
## 2  002 47.11757  5.6573460  -1.99139872 60.65735 63.00860
## 3  003 51.71674 -3.9693810  -4.29616009 51.03062 60.70384
## 4  004 42.12265 -8.4521390  -1.11319963 46.54786 63.88680
## 5  005 46.23664 -0.6239806   0.69413815 54.37602 65.69414
## 6  006 27.38906 -4.3586224   1.26765822 50.64138 66.26766
## 7  007 46.50530 -1.2952470   2.65476425 53.70475 67.65476
## 8  008 34.52227 -3.4247263   3.61010573 51.57527 68.61011
## 9  009 28.94785 -2.2054755 -10.46380628 52.79452 54.53619
## 10 010 48.74396  4.0774393  -8.17378707 59.07744 56.82621
```

### 4a

Create a column for the causal effect on unit i. Call it `tau`.

Answer:

```
single_draw$tau <- single_draw$Y_Z_1 - single_draw$Y_Z_0
```

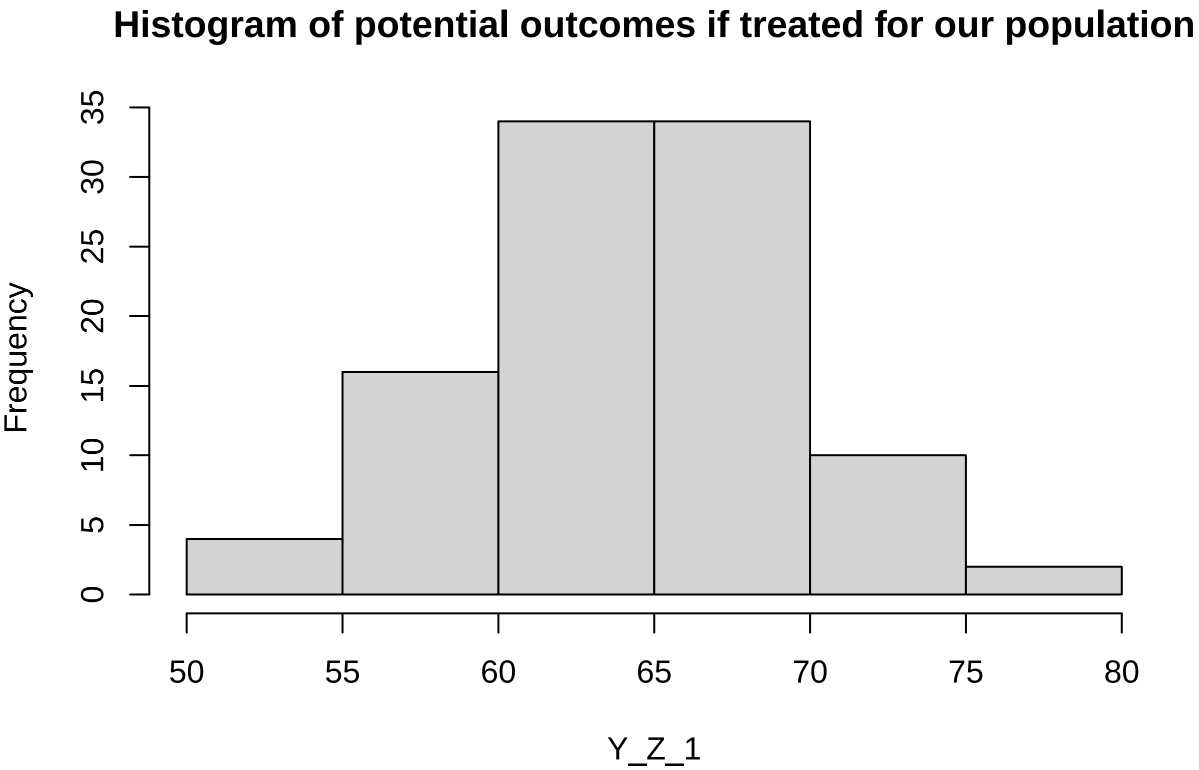
### 4b

Create three histograms. One for the potential outcomes if treated, one for the potential outcomes if not treated, and one for the individual causal effects of treatment. Comment on each.

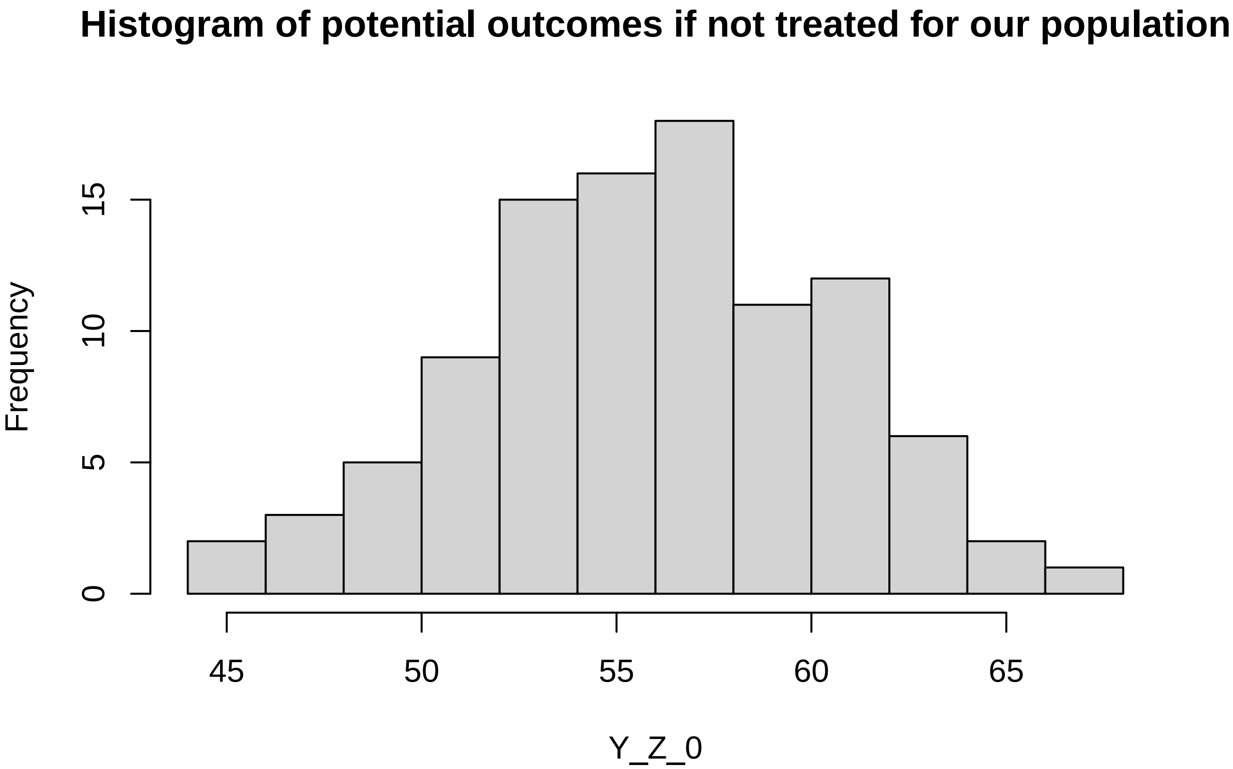
Answer:

The main point is that each is distributed roughly according to our simulation parameters above, including the noise we added because we don't expect treatment to "work" exactly the same on everyone. We see this variation in the potential outcomes, and thus, the the causal effects of treatment (which is simply the difference between the two potential outcomes.)

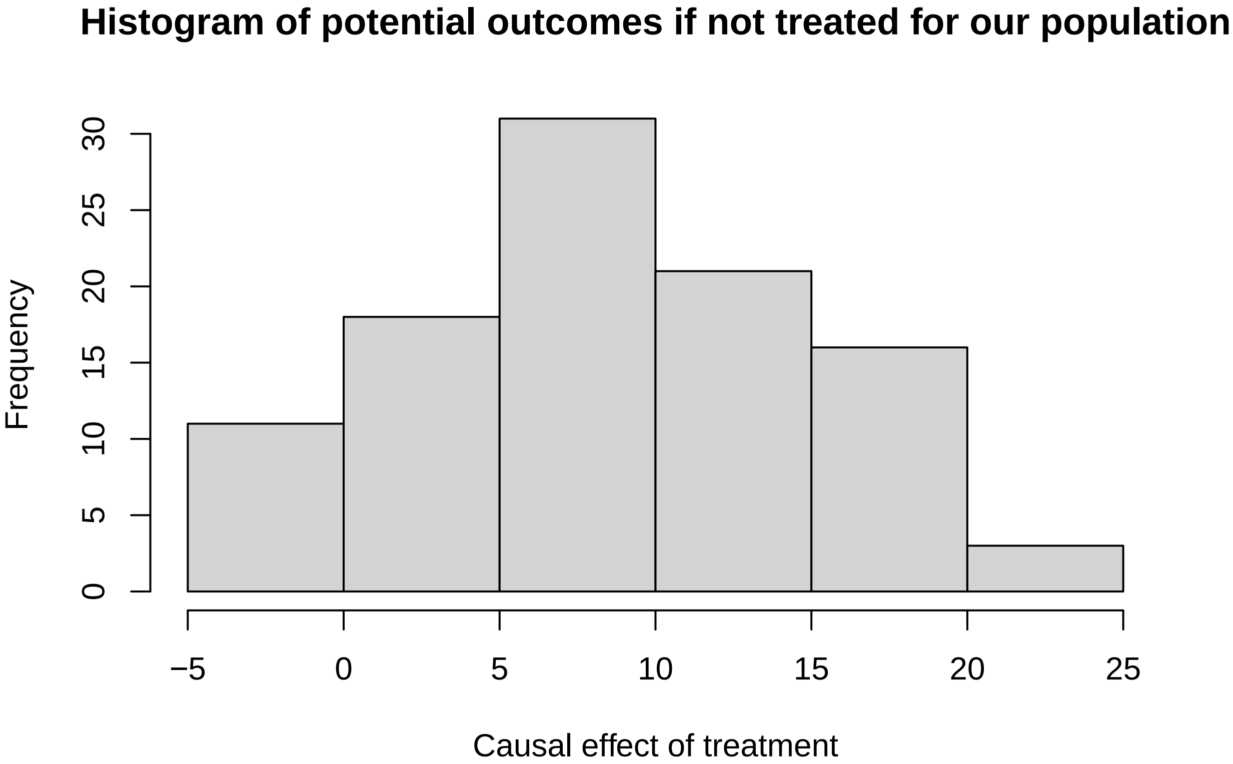
```
hist(single_draw$Y_Z_1,
     main = "Histogram of potential outcomes if treated for our population",
     "xlab" = "Y_Z_1")
```



```
hist(single_draw$Y_Z_0,
     main = "Histogram of potential outcomes if not treated for our population",
     "xlab" = "Y_Z_0")
```



```r
hist(single_draw$tau,
     main = "Histogram of potential outcomes if not treated for our population",
     "xlab" = "Causal effect of treatment")
```



### 4c

Describe in words the fundamental problem of causal inference in terms of the columns in our dataframe `single_draw`.

Answer:

For any given person, we can't observe both Y_Z_0 and Y_Z_1 so we can't observe `tau`.

### 4d

Calcuate the ATE.

Answer:

```
mean(single_draw$tau)
```

```
## [1] 8.558784
```

### 4e

Can we observe the ATE?

Answer:

No, we can't observe $\tau_i$ so we can't observe the *average* of these treatment effects either.

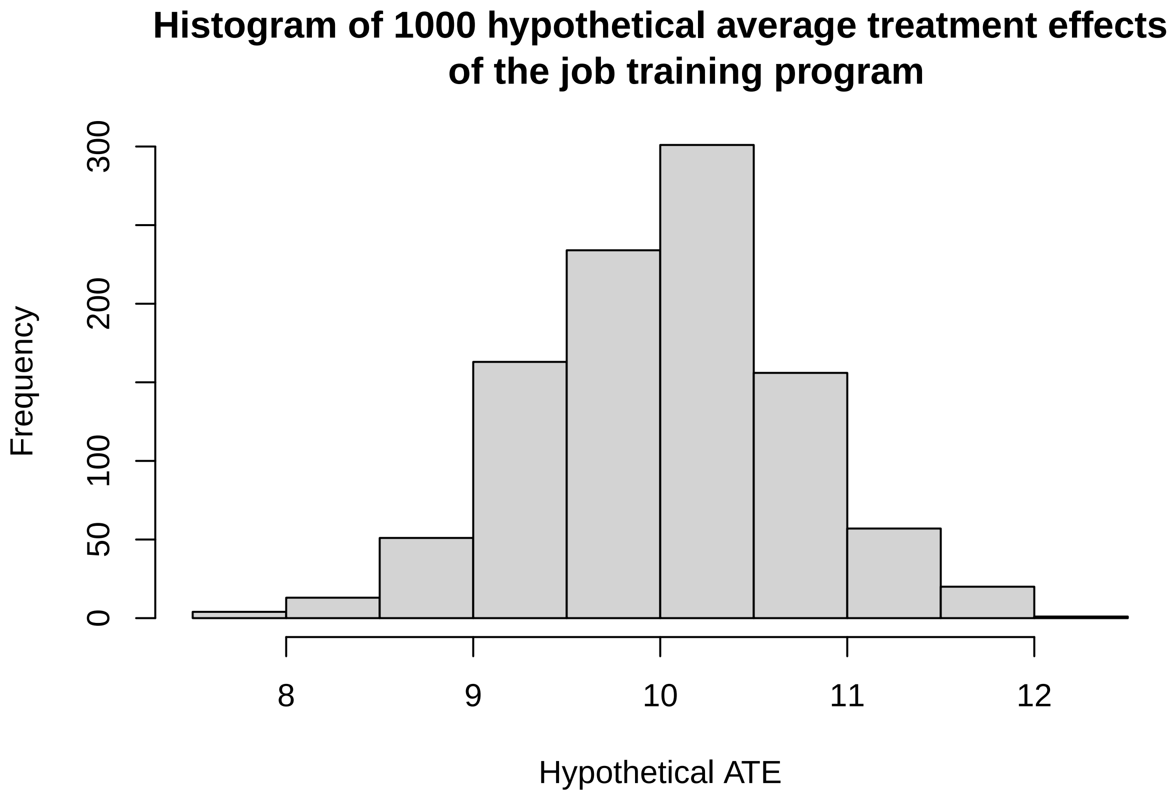
## Question 5

Now, instead of examining *one* hypothetical study like we did in Question 4, let's simulate many studies. Specifically, draw 1000 hypothetical studies, calculate the ATE for each, and plot a histogram of the ATE's. Comment on what you see in the histogram. (Note: there are many ways to accomplish this in code.)

Answer:

We confirm in the histogram our simulation parameters. Across many different studies the ATE may vary because the individual potential outcomes & causal effects have noise, but on average, the ATE will be 10K increase in salary.

```
hyp_ates <- rep(NA, 1000)
for(i in 1:1000){
  single_draw <- draw_data(design)
  single_draw$tau <- single_draw$Y_Z_1 - single_draw$Y_Z_0
  hyp_ates[i] <- mean(single_draw$tau)
}
hist(hyp_ates, main = "Histogram of 1000 hypothetical average treatment effects
     of the job training program",
     xlab = "Hypothetical ATE")
```


Histogram of 1000 hypothetical average treatment effects of the job training program
Frequency
0
50
100
200
300
8
9
10
11
12
Hypothetical ATE