Exercise: Understanding Potential Outcomes

1. Generate data on a population of 1,000 people. Specifically, create a variable (a vector) that randomly assigns these people to be male or female (50:50). *Hint: In R, try rbinom and in Stata, try rbinomial.*

```
N <- 1000
x <- rbinom(N,1,0.5)
```

2. Now we are going to simulate the potential outcomes - a measure of attitudes - *under control* (y_0) for our population. Create another variable of random normally-distributed values with mean of 5 and standard deviation of 1.

```
y0 <- rnorm(N,5,1)
```

3. One problem with observational data is that potential outcomes are often correlated with other variables such as gender. Adjust your value of y_0 to add 1 (one) for all units who are male.

```
y0 <- y0 + x
```

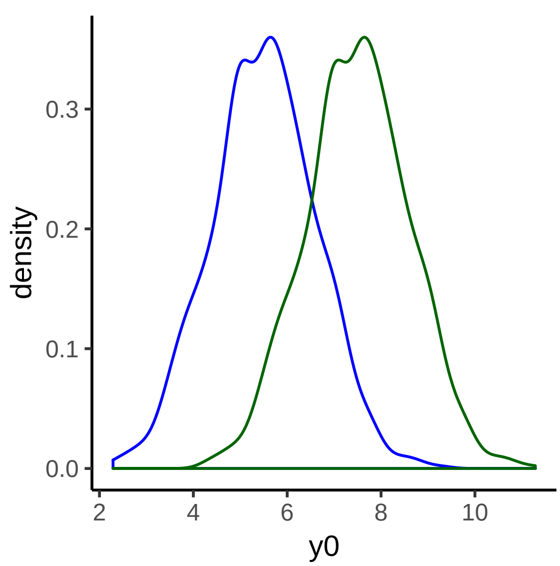
4. Now simulate potential outcomes *under treatment* (y_1) for all units. Define a *constant* treatment effect of $c = 2$ and create another vector $y_1 = y_0 + c$.

```
c <-2
y1 <- y0 + c
```

5. To compare our two sets of potential outcomes, plot two density charts on the same figure - one for y_0 and one for y_1.

```
data <- tibble(x,y0,y1)

data %>% ggplot() +
  geom_density(aes(x=y0), col="blue") +
  geom_density(aes(x=y1),col="dark green") +
  theme_classic()
```



6. Next, let us assume a specific **Treatment Assignment Mechanism** where men are more likely than women to receive treatment. First, create a temporary 'latent' variable which consists of two components added together: (i) $0.5 * x$ (i.e. half the value of the gender variable), and (ii) a random uniform value

between zero and one. Finally, create a new vector D which is equal to one when the latent variable is larger than 0.75, and zero otherwise.

```
data <- data %>% mutate(rnd=runif(N,0,1),
                        D=ifelse(0.5*x+rnd>0.75,1,0))
```

7. To show that gender and treatment - maybe we can think of treatment as low income and high income - are related, calculate the correlation between x and D.

```
cor(data$x,data$D)
```

```
## [1] 0.4879532
```

8. What is the average of the *real* indvidual treatment effects based on the potential outcomes, $E(y_1 - y_0)$?

```
Actual_causal_effect <- data %>%
  summarize(Actual_ATE=mean(y1-y0))
Actual_causal_effect
```

```
## # A tibble: 1 x 1
##   Actual_ATE
##        <dbl>
## 1          2
```

9. The Fundamental Problem of Causal Inference is that we *cannot* calculate (8.) above. Instead, we only observe one value: y_{obs}. Create a new variable y_{obs} which equals y_1 if $D = 1$ but which equals y_0 if $D = 0$.

```
data <- data %>% mutate(y_obs=case_when(D==1~y1,
                                        D==0~y0))
```

10. Based on the observable data, run the basic regression of treatment (D) on observable outcomes (y_{obs}). Interpret the result. Is this an accurate estimate of the treatment effect that we assumed at the start?

```
data %>% lm(y_obs~D,data=.) %>% stargazer(keep.stat=c("n"), header=F)
```

Table 1:

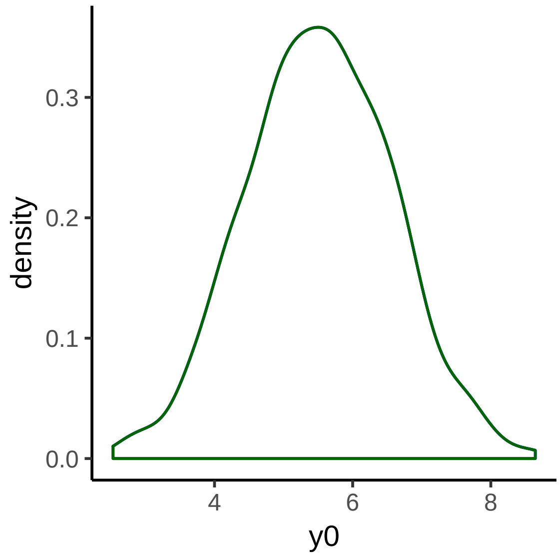
	Dependent variable:
	y_obs
D	2.586***
	(0.067)
Constant	5.219***
	(0.047)
Observations	1,000
Note:	*p<0.1; **p<0.05; ***p<0.01

11. Re-run all your code above but this time with $c = 0$ so we are assuming **NO** treatment effect. Run the regression in (10.) again - what is the result?

```
data_no_effect <- tibble(x=rbinom(N,1,0.5),
                         y0=x+rnorm(N,5,1),
                         y1=y0+0,
                         rnd=runif(N,0,1),
                         D=ifelse(0.5*x+rnd>0.75,1,0)) %>%
```

```r
  mutate(y_obs=case_when(D==1~y1,
                         D==0~y0))

data_no_effect %>% ggplot() +
  geom_density(aes(x=y0), col="blue") +
  geom_density(aes(x=y1),col="dark green") +
  theme_classic()
```



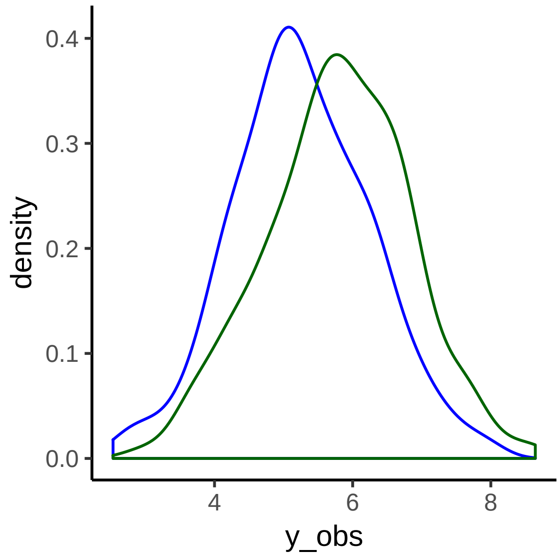

```r
data_no_effect %>% lm(y_obs~D,data=.) %>% stargazer(keep.stat=c("n"), header=F)
```

Table 2:

	Dependent variable:
	y_obs
D	0.520***
	(0.065)
Constant	5.259***
	(0.046)
Observations	1,000
Note:	*p<0.1; **p<0.05; ***p<0.01

12. To see why, let's plot two density charts on the same figure - one for the distribution of observable y_{obs} for the treated group ($y_{obs}|D == 1$) and one for the distribution of observable y_{obs} for the control group ($y_{obs}|D == 0$).

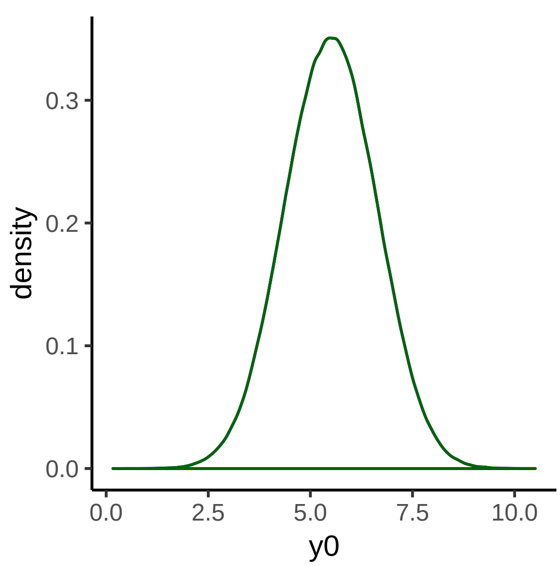
```r
data_no_effect %>% ggplot() +
  geom_density(data=data_no_effect %>% filter(D==0),aes(x=y_obs), col="blue") +
  geom_density(data=data_no_effect %>% filter(D==1),aes(x=y_obs),col="dark green") +
  theme_classic()
```



13. Run your code again for $c = 0$, but this time assume a larger population of $N = 1,000,000$. Does that solve the problem?

```
N <- 1000000
data_large_N <- tibble(x=rbinom(N,1,0.5),
                       y0=x+rnorm(N,5,1),
                       y1=y0+0,
                       rnd=runif(N,0,1),
                       D=ifelse(0.5*x+rnd>0.75,1,0)) %>%
  mutate(y_obs=case_when(D==1~y1,
                         D==0~y0))

data_large_N %>% ggplot() +
  geom_density(aes(x=y0), col="blue") +
  geom_density(aes(x=y1),col="dark green") +
  theme_classic()
```



```r
data_large_N %>% lm(y_obs~D,data=.) %>% stargazer(keep.stat=c("n"), header=F)
```

Table 3:

	Dependent variable:
	y_obs
D	0.503^{***}
	(0.002)
Constant	5.250^{***}
	(0.002)
Observations	1,000,000
Note:	*p<0.1; **p<0.05; ***p<0.01

14. For $c = 0$, run the regression of treatment on observable outcomes, but this time controlling for gender.

```r
data_no_effect %>% lm(y_obs~D + x,data=.) %>% stargazer(keep.stat=c("n"), header=F)
```

Table 4:

	Dependent variable:
	y_obs
D	0.029
	(0.069)
x	0.958^{***}
	(0.069)
Constant	5.026^{***}
	(0.045)
Observations	1,000
Note:	*p<0.1; **p<0.05; ***p<0.01

Stata Code

set obs 1000
gen x=rbinomial(1,0.5)
gen y0=rnormal(5,1)
replace y0=y0+x
gen y1=y0+2
kdensity y0, addplot(kdensity y1)
gen rnd=0.5*x+runiform(0,1)
gen D=0
replace D=1 if rnd>0.75
correlate x D
gen real_TE=y1-y0

```
mean real_TE
gen y_obs=y0
replace y_obs=y1 if D==1
regress y_obs D
regress y_obs D x
```