Quant II - Lab 1

Sylvan Zheng

The Science

Setup

- Download the file `thescience.tsv` from this week's lab folder on GitHub
- Move the file to a "lab 1" folder on your own computer
- Install the `tidyverse` and `here` R packages if you don't already have them

```r
knitr::opts_chunk$set(fig.width=4, fig.height=3)
library(tidyverse)
library(here)
df <- read_tsv(here('lab1/thescience.tsv'))
```

The data contains the following columns:

- **Potential Outcomes**: `y0` and `y1`
- **Observed Outcome**: `y`
- **Treatment**: `t`

```r
df %>% head
```

```
## # A tibble: 6 x 8
##      x1    x2      x3    y0    y1     t     y    t2
##   <dbl> <dbl>   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1 0.935 2.45   0.0717  3.34 4.84      1  4.84     1
## 2 0.824 3.90  -0.374   1.82 2.28      1  2.28     0
## 3 0.983 2.22   0.370   4.21 3.45      1  3.45     1
## 4 0.230 0.522  0.881   1.21 0.685     0  1.21     0
## 5 0.648 2.44   0.503   2.25 3.26      1  3.26     0
## 6 0.948 0.710  0.704   6.43 9.11      1  9.11     1
```

Lab Demonstration:

- How does the observed outcome `y` relate to the treatment `t` and the potential outcomes `y0` and `y1`?

```r
df %>% select(y0, y1, t, y) %>% tail
```

```
## # A tibble: 6 x 4
##      y0     y1     t      y
##   <dbl>  <dbl> <dbl>  <dbl>
## 1 1.74   2.33      0  1.74
## 2 1.92   2.73      1  2.73
## 3 1.38   0.432     1  0.432
## 4 1.66   1.97      0  1.66
## 5 0.744  0.470     1  0.470
## 6 1.17  -0.543     1 -0.543
```

- Calculate the difference in means between the treated and the untreated.

```r
mean(filter(df, t == 1)$y) - mean(filter(df, t == 0)$y)
```

```
## [1] 2.387075
```

- Calculate the true global treatment effect

```r
mean(df$y1 - df$y0)
```

```
## [1] 0.9567962
```

- Explain why they are different. Show this using R.

- What is the ATE vs the ATC and ATT? How would we calculate these from the science?

Now do the next part in pairs:

- You get to play omnipotent being! Create an alternate universe (ie, a new treatment assignment and new outcome variable) such that the difference in means between the treated and the untreated can be reliably estimated.
- Estimate the difference and means and compare it to the true effect.
- Are they different? Why/How?

Bias and Consistency of Estimators

Consider the following estimator for the population mean:

```r
my.estimator <- function(data) {
  data[[1]] + 5 / length(data)
}

N <- 5
mu <- 0
sigma <- 1

some.data <- rnorm(N, mu, sigma)
my.estimator(some.data)
```

```
## [1] 1.632246
```

```
some.data <- rnorm(N, mu, sigma)
my.estimator(some.data)
```

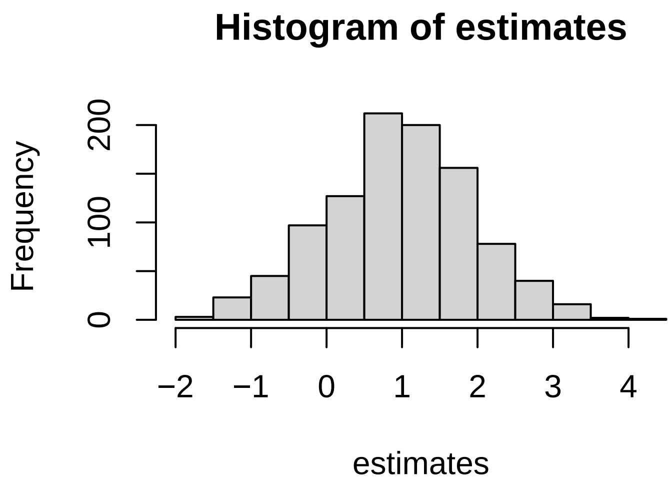
```
## [1] 0.244462
```

```
some.data <- rnorm(N, mu, sigma)
my.estimator(some.data)
```

```
## [1] 0.6601075
```

Is this estimator *biased*?

```
estimates <- c()
for(i in 1:1000) {
  some.data <- rnorm(N, mu, sigma)
  e <- my.estimator(some.data)
  estimates <- c(estimates, e)
}
hist(estimates)
```



```
mean(estimates)
```

```
## [1] 0.9745528
```

Is this estimator *consistent*?

Is this estimator *asymptotically biased*? How would we modify the above code to determine this?

Now compare for the following estimator:

```
my.estimator.2 <- function(data) {
  mean(data)
}
```

Write a function that pulls draws 1000 samples of size N from a normal distribution with mean `mu` and sd `sigma` and estimates the population mean using the above estimator.

Show a histogram or density plot of a few different values of N. Does the results suggest consistency? Unbiasedness? Asymptotic unbiasedness?

```r
my.estimator.2 <- function(data) {
  mean(data)
}
```

Write a function that pulls draws 1000 samples of size N from a normal distribution with mean `mu` and sd `sigma` and estimates the population mean using the above estimator.

Show a histogram or density plot of a few different values of N. Does the results suggest consistency? Unbiasedness? Asymptotic unbiasedness?