# Causal Inference
# Problem Set 1

Due Wednesday April 4th

Please 1) Write up your answers to all questions (including computational questions and any related graphics), 2) include your `R` code *with comments*, in your write-up, and 3) email your write up with an easy-to-recognize file names (e.g., `ps1_grimmer.pdf`). Please ensure that all of these are completed *before* class begins. For any problems that require calculations, please show your work. You are encouraged to work in groups, but you should write up the problem set alone.

## Problem 1

Review of unbiasedness and consistency.

(a) Consider a vector $\boldsymbol{y}$ of length $N$, where $y_i \in \mathbb{R}$ and are fixed numbers (*not* random variables, or degenerate random variables) $\forall\ i$.
Now consider a situation in which $y_i$s are being randomly sampled from $\boldsymbol{y}$ (without replacement), and define a random vector $\boldsymbol{z}$, also of length $N$, where $z_i$ takes a value of 1 if $y_i$ is sampled and 0 otherwise. For this and all subsequent parts in Problem 1, assume a sample of size $n$.

Explain what the sample space of the random vector $\boldsymbol{z}$ is, and specify how many elements (i.e. possible outcomes) the sample space contains.

The sample space of $z$ is the set of all possible random sampling outcomes (in terms of which $y_i$ are selected from $y$) with samples of size $n$.
The function $z_i$ maps each unit to the set $\{0, 1\}$ depending upon whether or not it is contained in the sample for a given outcome.
Given some arbitrary ordering of the $y_i$s in the vector $y$, the sample space of $z$ for a given sample size $n$ contains every N-tuple with $n$ 1's and $N - n$ 0's. There are $\binom{N}{n}$ such N-tuples.

(b) Denote the mean of the $y_i$'s as $\mu_y$.
What is the expected value of $\sum_{i=1}^{N} \frac{z_i y_i}{n}$? Is it unbiased for $\mu_y$?

Let $n$ denote the size of our random sample. Then:

$$E\left(\sum_{i=1}^{N} z_i \frac{y_i}{n}\right) = \sum_{i=1}^{N} E(z_i)\frac{y_i}{n} = \sum_{i=1}^{N} \frac{n}{N}\frac{y_i}{n} = \sum_{i=1}^{N} \frac{y_i}{N} = \mu_y$$

Does your answer to change if we drop the assumption of random sampling? Why or why not?

Yes. Without random sampling, it is not the case that $E(z_i) = n/N \ \forall \ i$.

And if $c \neq 0$ is a constant, is $\frac{c}{n} + \sum_{i=1}^{N} \frac{z_i y_i}{n}$ an unbiased estimator of $\mu_y$? What is its expectation?

No, this is not an unbiased estimator.

$$E\left(\frac{c}{n} + \sum_{i=1}^{N} z_i \frac{y_i}{n}\right) = E\left(\frac{c}{n}\right) + E\left(\sum_{i=1}^{N} z_i \frac{y_i}{n}\right) = \frac{c}{n} + \mu_y$$

(c) Derive the probability limit of $\frac{c}{n} + \sum_{i=1}^{N} \frac{z_i y_i}{n}$. Be very explicit in your use of relevant theorems. Is this estimator consistent for $\mu_y$?

Since the population is finite (N) and we are sampling without replacement, we will use an alternative notion of consistency that has been defined in the survey sampling literature, which is that $\hat{\theta} = \theta$ if $n = N$

$$\begin{aligned}
&\underset{n \to N}{plim}\left(\frac{c}{n} + \sum_{i=1}^{N} z_i \frac{y_i}{n}\right) \\
=&\underset{n \to N}{plim}\left(\frac{c}{n}\right) + \underset{n \to N}{plim}\left(\sum_{i=1}^{N} z_i \frac{y_i}{n}\right) \text{ Algebraic property of probability limits.} \\
=&\frac{c}{N} + \mu_y \text{ by WLLN.}
\end{aligned}$$

No, the estimator is not consistent for $\mu_y$, given the set-up of sampling without replacement from a finite population.

(d) Show that $s^2 = \frac{1}{n-1} \sum z_i (y_i - \bar{y})^2$ is a biased estimator of $\sigma_y^2$, the variance of $y$.

The proof to show that $s^2$ is a biased estimator of the true variance $\sigma_y^2$ given the conditions introduced in this question follows the same logic and process as in part (e). See below.

(e) Is $\hat{\sigma}^2 = \frac{1}{n} \sum z_i (y_i - \bar{y})^2$ an unbiased estimator of $\sigma_y^2$?

No, $\hat{\sigma}^2$ is not an unbiased estimator of $\sigma_y^2$. To see this, we can use our knowledge of the sampling set-up, notice that the notation corresponds to simple random sampling without replacement from a finite population, and rewrite the estimator.

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{N} z_i (y_i - \bar{y})^2 = \frac{1}{n} \sum_{i \in S} (y_i - \bar{y})^2$$

where $S$ corresponds to those units that were sampled (i.e. $i | z_i = 1$).

From there, we can recognize $\hat{\sigma}^2$ as the standard $\hat{\sigma}^2_n$, which we know to be biased as an estimator of $\sigma^2$ for BOTH sampling with and without replacement. In the case of sampling without replacement, the unbiased estimator is:

$$\hat{\sigma}^2_n \frac{n}{n-1} \frac{N-1}{N}$$

See handout for the full proof of this.

Thus:

$$\begin{aligned} E\left[\frac{1}{n}\sum_{i\in S}(y_i - \bar{y})^2\right] \\ = E\left[\hat{\sigma}^2_n\right] \\ = \sigma^2 \frac{n-1}{n}\frac{N}{N-1} \end{aligned}$$

which is clearly biased.

(f) [**Extra credit**] Derive the probability limit of $\hat{\sigma}^2 = \frac{1}{n}\sum z_i(y_i - \bar{y})^2$. Is $\hat{\sigma}^2$ a consistent estimator of $\sigma^2_y$?

$$\begin{aligned} \underset{n\to N}{plim}\ \hat{\sigma}^2 &= \underset{n\to N}{plim}\ \frac{1}{n}\sum_{i=1}^{N}\left[z_i(y_i - \bar{y})^2\right] \\ &= \frac{1}{N}\sum_{i=1}^{N}(y_i - \mu_y)^2 \\ &= \sigma^2_y \end{aligned}$$

# Problem 2

Review of distribution of the sample mean.

(a) Using `R`, draw 2000 observations from a non-normal distribution with finite mean and variance: specifically, use an exponential distribution with mean one. Call this vector $X$, and for the rest of the problem consider it fixed (that is, do not redraw it).

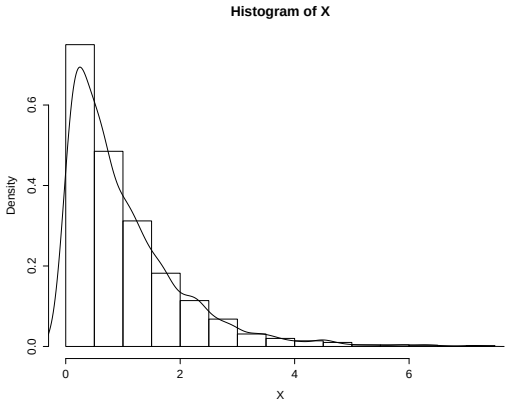
Example using draws from an exponential distribution, $X \sim exp(\lambda = 1)$

```
> set.seed(1)
```

```
> X <- rexp(2000, rate=1)
```

(b) Provide a histogram of $X$ and overlay a kernel density estimate.

```
> hist(X, prob=TRUE)
> lines(density(X))
```




(c) Draw 1000 random samples, each of size $n = 10$, from the original $X$ (do not redraw $X$; keep the original 2000 observations). In each of these 1000 samples, estimate the mean and standard error of the mean (*not* the standard deviation of the sample!). Use these mean and standard error estimates to estimate a 95% confidence interval for the mean in each of the 1000 samples (that is, 1000 confidence intervals). Use a normal approximation for the distribution of your mean estimator. Finally, calculate how often these confidence intervals contain the true mean—that is, the mean of the original $X$ vector. In other words: what proportion of the 1000 confidence intervals contain the mean of the original $X$ vector?

```
n <- 10
samp <- 1000

coverage <- function(n,samp){
  set.seed(1)
  mat <- matrix(NA, ncol=5, nrow=samp)
  cut <- abs(qt(.025,n-1))
  for (i in 1:samp){
    x <- sample(X,n,replace=TRUE)
    mat[i,1] <- mean(x) - 1.96*(sd(x)/sqrt(n))
    mat[i,2] <- mean(x) + 1.96*(sd(x)/sqrt(n))
    mat[i,3] <- mean(x)
    mat[i,4] <- mean(x) - cut*(sd(x)/sqrt(n))
    mat[i,5] <- mean(x) + cut*(sd(x)/sqrt(n))
  }
```

```
  within_norm <- matrix(NA, ncol=1, nrow=samp)
  within_t <- matrix(NA, ncol=1, nrow=samp)
  for (i in 1:samp){
    within_norm[i] <- if(mean(X)>=mat[i,1] & mean(X)<=mat[i,2]){1}else{0}
    within_t[i] <- if(mean(X)>=mat[i,4] & mean(X)<=mat[i,5]){1}else{0}
  }
  l <- list()
  l[1] <- mean(within_norm)
  l[[2]] <- mat[,3]
  l[[3]] <- mean(within_t)
  return(l)
}

coverage(n, samp)[[1]]
[1] 0.88
```

(d) Repeat part (c) using samples of size of $n = 25, 50, 100, 200$. For each sample size, what proportion of the 1000 confidence intervals contain the true mean of $X$? What is going on here?

```
> coverage(10, samp)[[1]]
[1] 0.88
> coverage(25, samp)[[1]]
[1] 0.922
> coverage(50, samp)[[1]]
[1] 0.927
> coverage(100, samp)[[1]]
[1] 0.929
> coverage(200, samp)[[1]]
[1] 0.935
```

(e) Repeat parts (c) and (d), using the $t$ distribution (rather than the normal approximation) to select the critical values for the confidence intervals. Are there any meaningful changes? Why or why not?

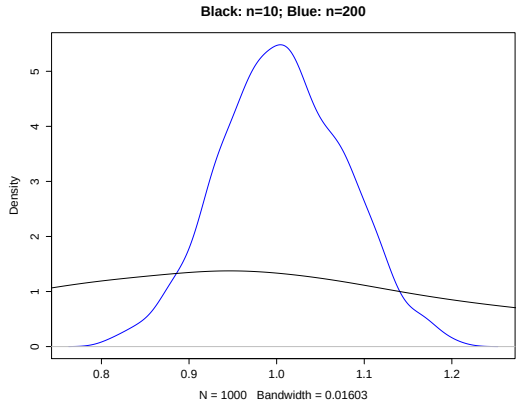
```
> coverage(10, samp)[[3]]
[1] 0.915
> coverage(25, samp)[[3]]
[1] 0.934
> coverage(50, samp)[[3]]
[1] 0.933
> coverage(100, samp)[[3]]
```

```
[1] 0.935
> coverage(200, samp)[[3]]
[1] 0.936
```

The confidence intervals have better coverage because using the t distribution is more conservative, but because the underlying data are not normally distributed, the coverage does not reach 95%, even for the larger sample sizes.

(f) Use a histogram or kernel density estimator to plot the distribution of the sample means themselves from above in the cases when $n = 10$ and when $n = 200$. Do they look different? Why?

```
> plot(density(coverage(200, samp)[[2]]), main="Black: n=10; Blue: n=200", col='blue')
> lines(density(coverage(10, samp)[[2]]))
```




(g) What important theorem from statistics explains the shape of the distribution of sample means when $n = 200$? Briefly describe what this theorem guarantees and why it is so important.

The Central Limit Theorem. Take a moment of silence to appreciate the power of this theorem. This says that, for iid $X_i$s with finite mean and finite, nonzero variance, scaled-up differences between the sample means and the population mean converge in distribution to a standard normal:

$$\sqrt{n}\frac{(\bar{X}_n - \mu)}{\sigma_X} \xrightarrow{d} \mathcal{N}(0, 1)$$

This implies that, as $n \to \infty$, the sample means themselves are approximately distributed normal around the population mean:

$$\bar{X}_n \sim \mathcal{N}\left(\mu, \frac{\sigma_X^2}{n}\right)$$

# Problem 3

Consider the model:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{e}$$

where $\mathbf{y}$ is an $N$x1 vector of realizations of an outcome variable, $\mathbf{X}$ is an $N$x2 matrix where the first column is a constant (a 1x$N$ vector of 1's) and the second column is a single variable $X$ (a 1x$N$ vector of realizations of $X$, each of which could be denoted $x_i$), $\mathbf{b}$ is a 2x1 vector of unknown parameters and $\mathbf{e}$ is an $N$x1 vector of errors. The variable $X$ takes on the value of 1 if a given observation receives an experimental treatment and 0 otherwise. There are $m$ treated observations and $N-m$ untreated observations.

Prove that the second element of the vector $\hat{\mathbf{b}} = (\mathbf{X'X})^{-1}\mathbf{X'y}$ is equivalent to the difference between the mean value of $\mathbf{y}$ for treated observations and the mean value of $\mathbf{y}$ for untreated observations. Recall that for any nonsingular 2x2 matrix:

$$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

$$\mathbf{A}^{-1} = \frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$$

The familiar OLS estimator for $\hat{\beta}$ is $(X'X)^{-1}X'y$. Taking this step-by-step:

$$X'X = \begin{bmatrix} 1 & 1 & \dots 1 \\ x_1 & x_2 & \dots x_N \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ . & . \\ . & . \\ . & . \\ 1 & x_N \end{bmatrix} = \begin{bmatrix} N & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}$$

Using the formula for the inverse of a 2x2 matrix,

$$(X'X)^{-1} = \frac{1}{N\sum x_i^2 - \sum x_i \sum x_i} * \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & N \end{bmatrix}$$

$$X'y = \begin{bmatrix} 1 & 1 & \dots 1 \\ x_1 & x_2 & \dots x_N \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ . \\ . \\ . \\ y_N \end{bmatrix} = \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix}$$

Now let $m$ be the number of units in the treatment group and $(N-m)$ be the number of units in the untreated group. Also note that because $X$ only takes on the values 0 and 1 then $\sum x_i = \sum x_i^2 = m$.

$$\begin{aligned}
(X'X)^{-1}X'y &= \frac{1}{N\sum x_i^2 - \sum x_i \sum x_i} * \begin{bmatrix} \sum x_i^2 & -\sum x_i \\ -\sum x_i & N \end{bmatrix} \begin{bmatrix} \sum y_i \\ \sum x_i y_i \end{bmatrix} \\
&= \frac{1}{Nm - m^2} * \begin{bmatrix} m\sum y_i - m\sum x_i y_i \\ N\sum x_i y_i - m\sum y_i \end{bmatrix} \\
&= \frac{1}{m(N-m)} * \begin{bmatrix} m\sum y_i - m(\sum y_i | x_i = 1) \\ N(\sum y_i | x_i = 1) - m\sum y_i \end{bmatrix} \\
&= \begin{bmatrix} [m\sum y_i - m(\sum y_i | x_i = 1)]/m(N-m) \\ [N(\sum x_i y_i) - m\sum y_i]/m(N-m) \end{bmatrix} \\
&= \begin{bmatrix} [\sum y_i - (\sum y_i | x_i = 1)]/(N-m) \\ N(\sum y_i | x_i = 1)/m(N-m) - (m\sum y_i)/m(N-m) \end{bmatrix} \\
&= \begin{bmatrix} [\sum y_i | x_i = 0]/(N-m) \\ [N(\sum y_i | x_i = 1) - (m((\sum y_i | x_i = 1) + (\sum y_i | x_i = 0)))]/m(N-m) \end{bmatrix} \\
&= \begin{bmatrix} \bar{y}_{control} \\ [(N-m)(\sum y_i | x_i = 1) - (m(\sum y_i | x_i = 0))]/m(N-m) \end{bmatrix} \\
&= \begin{bmatrix} \bar{y}_{control} \\ ((\sum y_i | x_i = 1)/m) - ((\sum y_i | x_i = 0)/(N-m)) \end{bmatrix} \\
&= \begin{bmatrix} \hat{\beta}_0 = \bar{y}_{control} \\ \hat{\beta}_1 = \bar{y}_{treated} - \bar{y}_{control} \end{bmatrix}
\end{aligned}$$

# Problem 4

To reinforce the intuition behind the potential outcomes framework, consider the fictional data set "POdata.csv." In these fictional data, we observe an outcome for each unit both under treatment and under control (which, again, is usually impossible in the real world).

(a) Define individual level treatment effects and explain the fundamental problem of causal inference.

Individual-level treatment effects are $\tau_i = Y_{1i} - Y_{0i}$. The fundamental problem of causal inference is that we cannot observe both $Y_{1i}$ and $Y_{0i}$, and therefore we cannot usually calculate $\tau_i$ (this fictional data set is an exception).

(b) Define the Average Treatment Effect (ATE) and calculate the ATE in these data.

$\tau_{ATE} = E[Y_{1i} - Y_{0i}] = E[\tau_i]$

```
> set.seed(1)
> data <- read.csv("POdata.csv")
> ATE <- mean(data$Treat) - mean(data$Control)
[1] 0.4376972
```

(c) Plot the distribution of the individual treatment effects. Does the treatment seem to have an effect? How well is it captured by the ATE?

```
> plot(density(POdata$Treat-POdata$Control))
```