Problem Set 1 – Solutions

Research Design for Causal Inference
Due: April 7, 2015

Part I – Key topics

While doing the reading and problem set this week, focus on understanding the following key topics, which we will likely discuss in class:

- Potential outcomes (including notation).
- Average treatment effects.
- Expected value (including notation).
- Random assignment.
- Unbiased estimator.
- Excludability.
- Non-interference.

Part II – Gerber & Green

Complete the following exercises from FEDAI:

Chapter 1: Exercise 1(b).

> Unobserved heterogeneity refers to all the subtle, unseen (and perhaps unseeable) ways in which the units exposed to the one condition may vary from those exposed to another condition (experimentally or not). In interpreting correlations between any outcome of interest and exposure to a particular condition, the presence of unobserved heterogeneity across the units in the different conditions may confound whatever observed statistical relationships exist.

Chapter 2: Exercise 1.

(a) This refers to the potential outcome for the i^{th} unit under the control condition.

(b) This refers to the potential outcome for the i^{th} unit under the control condition conditional on the unit being treated under some hypothetical allocation of treatment. Using d in the conditional part of the expression (as opposed to D) would indicate that the unit was actually exposed to treatment.

(c) The first expression refers to the (unconditional) potential outcome for the i^{th} unit under the control condition. The second expression refers to the potential outcome for the i^{th} unit under the control condition conditional on the unit actually having been exposed to the control condition.

(d) Both expressions refer to the potential outcome for the i^{th} unit under the control condition. The first refers to this outcome conditional on the unit being hypothetically exposed to treatment; the second refers to this outcome conditional on the unit being hypothetically exposed to control.

(e) The first expression refers to the expected value for a unit under the control condition. The second expression refers to the expected value for a unit under the control condition conditional on the unit having been (hypothetically) exposed to the treatment condition.

(f) When D_i is randomly assigned, the two components of the "selection bias term" are equal in expectation and thus cancel out.

Chapter 2: Exercise 2. Please complete your calculations using the statistical software of your choice.

I'll complete this using R. First, I'll generate the schedule of potential outcomes, $Y_i(0)$ and $Y_i(1)$, as well as the difference between them, $Y_i(1) - Y_i(1)$, as vectors. Then I'll calculate the different expected values. Remember: the expected value of a random variable is your best guess for the value of a random draw from all the potential values that the variable might take. This is the same as the arithmetic mean.

```
y.0 <- c(10, 15, 20, 20, 10, 15, 15)
y.1 <- c(15, 15, 30, 15, 20, 15, 30)
y.diff <- y.0 - y.1

# Now calculate the expected values:

mean(y.0) - mean(y.1)

## [1] -5

mean(y.diff)

## [1] -5
```

Chapter 2: Exercise 5

(a) Strengths & weaknesses. For each method:

Method 1: Two key strengths of the coin flip approach are that assignment will be truly random and neither the subject nor the experimenter will have any knowledge of treatment assignment ahead of time. A weakness is that this assignment mechanism may result in an unbalanced design (more units in either treatment or control). This method also assumes a fair coin.

Method 2: Strengths of this approach are that it randomly assigns treatment and ensures a balanced design. One weakness is that the experimenter will know the schedule of treatment assignments ahead of time and this may introduce bias into the administration of each condition to each subject.

Method 3: The strengths of this approach are that it randomly assigns treatment, ensures a balanced design, and prevents foreknowledge of the schedule of treatment assignments ahead of time by the experimenter. A weakness is that the experimenter will know the final subject's assignment before it is administered, potentially introducing bias.

(b) Changing the number of subjects from 6 to 600: The concern about Method 1 producing an unbalanced design goes away because of the larger number of coin flips.

(c) Expected values of D_i under Methods 1 and 3: The expected value is 45 minutes under both methods.

Chapter 2: Exercise 7.

(a) An unbiased estimator has an expected value equal to the true value of the estimated quantity. In other words, over many repeated estimates, the mean value of the distribution of an unbiased estimator will be equal to the true value.

(b) The allocation procedure (randomly assigning two villages to treatment) does not produce biased estimates. Over many repeated iterations, the expected value of the estimated average treatment effect generated across all the potential permutations of treatment allocation would be equal to the true average treatment effect.

(c) Convenience-based treatment assignment is prone to bias because there may be an (observed or unobserved) relationship between the reasons why the treatment assignment is convenient and the outcome of interest. More formally, random assignment of treatment forces:

$$(D_i \perp\!\!\!\perp Y_i) \,|\, \mathbb{X}_i$$

You can read this expression: "treatment assignment for any unit, D_i, is statistically independent of that unit's outcome, Y_i, conditional on all covariate values for that unit, $\mathbb{X}_i$." Don't worry if this doesn't make perfect sense yet, we'll talk about it a few times in the coming weeks.

Part III – "No causation without manipulation"?

It is often said that within the potential outcomes framework effects can only be understood empirically in relation to causal variables that have been manipulated in some way.

(a) Explain why this notion of *manipulation* matters so much.

> Manipulation matters because *the possibility of exposure* to more than one condition (e.g., treatment vs. control), makes statistical estimation of the effect of a cause possible. Within the notation used by Gerber & Green, the existence of potential outcomes $Y_i(1)$ and $Y_i(0)$ for some unit i depends critically on whether each unit had some probability greater than 0 and less than 1 of being in either condition. This possibility of exposure makes the potential outcomes viable (think about it: they are not actually "potential outcomes" otherwise!). Statistically, the fact that each unit i might be exposed to treatment or control makes it possible to calculate estimates of the treatment effect τ_i using $E[(Y_i(1) - Y_i(0)]$.

(b) What are the implications of this for research into relatively stable attributes of people such as phenotype (skin color) or gender?

> Stable attributes of experimental units that cannot be manipulated cannot be analyzed as causes within the potential outcomes framework. For any unit i, the attribute could not plausibly have assumed any value other than the observed one. Potential outcomes under counterfactual conditions are therefore implausible for these kinds of attributes. (See Holland, 1986 for further details).