**Humphreys, Essex 1 August 2016**

**Causal Inference and Experimentation**

**1 Group work on potential outcomes**

**1.1** Consider an outcome *Y* that can be affected by two variables *X1* and *X2*. All variables are binary.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | *X1* is a necessary and sufficient condition for Y | *X1* is necessary but not sufficient | *X1* is sufficient but not necessary | *X1* sometimes causes Y but is neither necessary nor sufficient |
| Y(0,0) = | 0 | 0 | 0 | 0 |
| Y(1,0) = | 1 | 0 | 1 | 1 |
| Y(0,1) = | 0 | 0 | 1 | 1 |
| Y(1,1) = | 1 | 1 | 1 | 0 |

**1.2** Consider an outcome *Y* that can be affected by two variables *X1* and *X2* but say that *X2* can itself be affected by *X1*. Write down the potential outcomes for *Y1* and*X2* when:

|  |  |  |
| --- | --- | --- |
|  | X1 causes X2 and Y, but X1 does not cause Y though X2 | X1 causes X2 and X2 causes Y, but X1 does not cause Y |
| Y(0,0) = | 0 | 0 |
| Y(1,0) = | 1 | 0 |
| Y(0,1) = | 0 | 1 |
| Y(1,1) = | 1 | 0 |
| X2(0) = | 0 | 0 |
| X2(1) = | 1 | 1 |

**1.3** Consider a process with: Y(0,0) =0, Y(1,0) = 1, Y(0,1) =1, Y(1,1) =1.

Say X1=1, X2=1. Then Y = 1. **What caused *Y* = 1?**

Say X1=0, X2=0. Then Y = 0. **What caused *Y* = 0?**

Say *X1* = 1 with 10% probability, otherwise 0 and, independently, *X2* = 1 with 50% probability, otherwise 0. Then what is the average effect of *X1* on Y? What is the average effect of *X2* on Y? *Which cause has the biggest effect?*

**1.4** Discuss these possibly causal questions.

* *Janes’ paper was not cited because she is a woman.*Is this a causal claim? Can you write down two or three statements that capture some part of this claim idea but that have treatments that you could manipulate in principle?
* *The building of a new health clinic in 1990 led to 10% increase in the birthweight of newborns by 1992.*Is this a causal claim? If this is a statement about average causal effects, what is being averaged over?

**Humphreys, Essex 1 August 2016**

**Causal Inference and Experimentation**

**2 Group work on randomization**

**2.1** 100 students sign up to take part in an experiment. You want to measure the effect of immigration on social trust. Half your subjects are men and half are women and you believe gender is very predictive of social trust.

Your experiment involves varying whether a “native” or an “immigrant” facilitator instructs players in how to play a trust game. You have five native and five immigrant facilitators and you want them each to conduct one session with 10 subjects.

* You are free to assign both subjects and facilitators to sessions. Describe your optimal randomization strategy. Is it blocked? Clustered? Both?
* Say now that subjects have already signed up for sessions. You can only assign facilitators to sessions, but you have access to the subject lists before you do so. Describe your optimal randomization strategy. Is it blocked? Clustered? Both?

**2.2** You have access to a network of all friendship links in a classroom. You want to provide political information to a set of students and see how muc more likely it is that a student that you do not give information to receives the information if a friend is treated compared to the situation in which a friend of a friend is treated. So you want to be sure that some subjects have friends treated and some have only friends of friends treated. How would you assign treatment? How can you work out your treatment assignment probabilities?

**2.3** Say you sampled subject *A* with probability .6 and subject *B* with probability .4. Say you assigned each to treatment with probability .6. What weight should you put on *A* in your analysis if they end up in treatment? What if they end up in control? How about *B*?

**Humphreys, Essex 1 August 2016**

**Causal Inference and Experimentation**

**Group work on inference**

3.1 Consider the following data. Treatment is assigned in two blocks. One third of block 1 got treatment (randomly); one third of block two got treatment (randomly). The data are as below:

|  |  |  |
| --- | --- | --- |
| Block | Z | Y |
| 1 | 0 | 0 |
| 1 | 0 | 0 |
| 1 | 1 | 1 |
| 2 | 0 | 0 |
| 2 | 0 | 0 |
| 2 | 1 | 0 |
| 2 | 1 | 0 |
| 2 | 1 | 1 |
| 2 | 1 | 1 |

Can you estimate the ATE? How about the ATT? And the ATC?

How do these compare to a simple difference in means between treatment and control? What would you get if you did OLS and controlled for block?

**3.2** Two of four units are going to be assigned to treatment.A researcher sets up a design in which subjects can decide for themselves the probability with which they receive a treatment. Requested propensities are as below.

Can you :

1. List the set of admissible treatment allocations
2. Describe a scheme for allocating subjects to treatment
3. Calculate your estimate under each allocation
4. Assess whether your estimate will be biased or not.

|  |  |  |
| --- | --- | --- |
| Requested propensity | Y0 | Y1 |
| 0.2 | 0 | 1 |
| 0.4 | 0 | 2 |
| 0.6 | 1 | 3 |
| 0.8 | 1 | 4 |

[This is hard and long; so focus on the principles even if you do not have time to do all the calculations]

**3.3 Group work on hypothesis testing**

Say there are 20 students. One of them is randomly assigned to have an extra maths class. They take a test and are given grades. The student that got the extra class scored 73% which is higher than any of the other 19 students.

You are interested in calculating a p value for the hypothesis that the math class had no effect on grades. As a test statistic you use the difference in means between the treated student and the rest. What is the probability that the difference you see is so big under the null of no effect?

**3.4 Confidence Intervals**

Consider the same problem and say there were a gazillion students with grades spread evenly between 0 and 100. Say there was one treated student and this student had a grade of 95%. The difference in means estimate for the treatment effect is 45. Under the assumption of homogeneous treatment effects (and so allowing for the possibility that grades could go over 100 for treated units) can you construct a 90% confidence interval for this estimate.

**Group work on topics**

**4.1 Spillovers**

Imagine a study with three subjects. Each subjects potential outcomes are as follows:

* 0 if in control
* n if in treatment and n subjects are assigned to treatment

So for example if one unit is in treatment that unit has outcome Y=1, and the others have Y = 0; if all 3 are in treatment they all have Y=3.

1. Write down the potential outcomes for all possible assignments, including all and none assigned o treatment
2. Write down the difference in means calculation for each possible assignment
3. Define two estimands of interest.
4. Define a randomization scheme that will return an unbiased estimate of each estimand.
5. Define a randomization scheme that will return a biased estimate of each estimand.

**4.2 LATE**

Which of these problems could be addressed using instrumental variables. In each case what kinds of concern might you have about the IV strategy?

* Experimenters introduce a unconditional cash transfers into a set of villages in 2012 and use it to measure access school attendance in 2015. You come on the scene later and are interested in whether the transfer could have led to greater political participation in 2017.
* Experimenters introduce an unconditional cash transfers into a set of villages in 2012 and use it to measure access school attendance in 2015. You come on the scene later and are interested in whether the increased school attendance could have led to greater political participation in 2017.
* You want to understand the effects of attending a rally on subsequent support for a candidate. You send a random set of voters a flyer about an upcoming demonstration.
* You want to understand the effects of attending a rally on subsequent support for a candidate. You send a random set of voters a flyer about an upcoming demonstration but you find out later that your enumerators did not deliver the flyers in a bunch of areas.
* You want to understand the effects sending flyers about an upcoming demonstration but you find out later that your computer code used incomplete data when making assignments and so failed to assign treatment to a whole bunch of regions.
* You want to understand whether sending flyers increases participation because people actually go to the rallies or because people’s general level awareness of the election increases, whether or not they go.