Homework1 Solution: Introduction to Causal Inference

1 (20'). Supposed that you have a random sample of size n from the population of interest. Answer the following questions that are designed to help you get familiar with potential outcomes. Try to keep your answers brief and your language precise. Throughout the problem, assume that the Stable Unit Treatment Value Assumption (SUTVA) holds.

a) Contrast the meaning of Y^0 with the meaning of Y

Y^0 is the potential outcome, representing what the outcome would have been if the treatment had not been applied.
Y is the actual outcome that we observe after treatment has been applied (or not).

b) Contrast the meaning of $E(Y^0)$ with the meaning of $E(Y|A=0)$

The first is the expected value of potential outcome if everyone in the population had not received the treatment.
The second is the expected value of observed outcome only among those who did not receive the treatment.

c) Contrast the meaning of $E(Y^0|A=1)$ with the meaning of $E(Y^0|A=0)$

The first is the expected value of potential outcome if the treatment had not been applied, among those in the treated group
The second is the expected value of potential outcome if the treatment had not been applied, among those who did not receive treatment.

d) Which of the following quantities can be identified from observed data, assuming SUTVA?
   a. $E(Y^0|A=1)$
   b. $E(Y^0)$
   c. $E(Y|A=0)$
   d. $E(Y^0|A=0)$

c and d. Note c and d are equal under SUTVA/consistency assumption, which states "observed outcome for those who didn't receive treatment is equal to their potential outcome if untreated"

e) Now, further assume that the units in this sample are randomly assigned to treatments, which means the assumption of exchangeability holds. Which of the above quantities can be identified from the observed data?

a)-d). Exchangeability assumption ensures the potential outcome for a subject would have the same expectation no matter its assignment of treatment. With both assumption of SUTVA and exchangeability, a-d are all identifiable and equal as a = d = $E(Y|A=0)$, which is equal to (b).

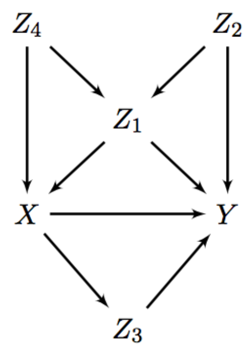
2 (10'). Under the complete randomization of treatment assignment, the average treatment effect is identified because the randomization guarantees statistical independence between the treatment indicator A and the observed outcome Y.

a) True

b) False **(it should be independence between the treatment indicator A and all potential outcomes, not observed outcome Y)**

Why? Justify your choice.

3 (20'). Consider the following DGA



a) Enumerate all paths from X to Y

   x→y;
   x→z3→y;
   x←z1→y;
   x←z4→z1→y;
   x←z1←z2→y;
   x←z4→z1←z2→y

b) In the path $X \leftarrow Z_4 \rightarrow Z_1 \leftarrow Z_2 \rightarrow Y$, what type of node is Z_1? Does conditioning on z1 block or unblock this path from X to Y?

   Z_1 is a Collider. Conditioning on Z_1 will unblock this path from X to Y

c) In the path $X \leftarrow Z_1 \rightarrow Y$. In this path, what type of node is Z_1? Does conditioning on Z_1 here block or unblock this path?

   Z_1 is a Confounder. Conditioning on Z_1 Block the path

d) Pearl's back-door criterion is a way to rule out confounding via conditioning, thus identifying the effect of one variable on another. For identifying the effect of any *X* on any *Y*, the backdoor criterion has two parts. First, the conditioning set C may not include any descendent of *X*. That means, anything that *X* affects cannot be conditioned on (c.f. post-treatment bias). Second, the conditioning set must block all back-door paths from X to Y. That is, by conditioning on our set C, we should break any back-door paths that may simultaneously generate confounded covariance in X and Y.

   a. Is $X \rightarrow Z_3 \rightarrow Y$ a back door path from X to Y? Why?

No, because it does not begin with a directed edge that points to the first variable X.

b. Based on your DAG, enumerate the minimum conditioning sets that satisfy the back door criterion for identifying the effect of X on Y?

{z1 and z4} or {z1 and z2}

4. Consider a randomized experiment with four observations, of which two units were randomly assigned to treatment via complete randomization. We use $A_i \in \{0, 1\}$ and Y_i to denote the treatment (1 for treatment and 0 for control) and the observed outcome for unit i, respectively.

a) (4') The table below shows the data observed from this experiment, augmented with columns for potential outcomes and the treatment effect for each unit. Fill in all the empty cells in the table based on the observed information, denoting unknown information with "?"

i	Y	A	Y^1	Y^0	τ
1	2	1	2	?	?
2	0	0	?	0	?
3	1	0	?	1	?
4	3	1	3	?	?

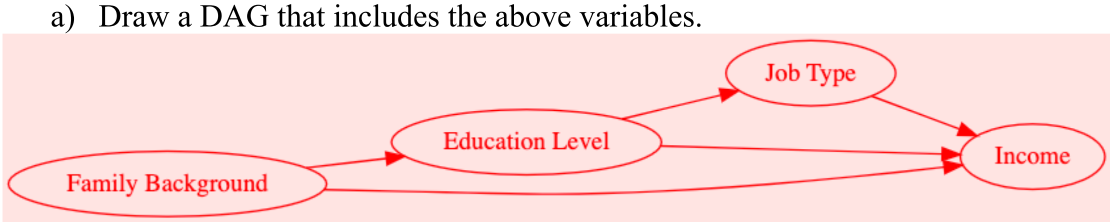
b) (6') Define the population average treatment effect for the treated (ATT) using the above notation and propose an unbiased estimator for this estimand. Then using the data in the table estimate this quantity.

$ATT = E(Y^1 - Y^0 | A = 1)$
$= E(Y|A = 1) - E(Y^0|A = 1)$ by consistency
$= E(Y|A = 1) - E(Y^0|A = 0)$ by exchangeability due to randomization
$= E(Y|A = 1) - E(Y|A = 0)$ by consistency

estimated by

$\frac{1}{n_t}\sum_{i=1}^{n_t} Y_i - \frac{1}{n-n_t}\sum_{i=1}^{n-n_t} Y_i = 2.5 - 0.5 = 2$

5. (10') Researchers are studying the effect of education level on income. They believe higher education leads to better job opportunities, which in turn increases income. Additionally, they suspect that family background influences both education and income.
   a) Draw a DAG that includes the above variables.



   b) Is Job Type a mediator or confounder in this DAG?

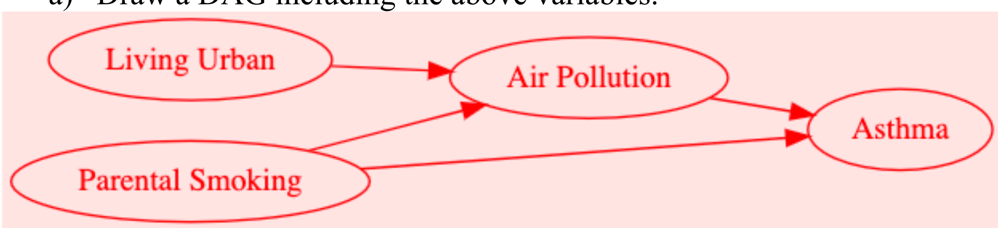
Job Type explains the relationship between an independent variable (in this case, education level) and a dependent variable (income). Job Type is a medicator since a mediator occurs when the independent variable affects the mediator, which in turn affects the dependent variable.

c) How would you use the DAG to assess the causal effect of Education Level on Income?

Family Background is identified as a confounder. To assess the causal effect accurately, we need to control for this confounder in the analysis.

6. (10') You are tasked with investigating the relationship between air pollution and asthma in children. Consider the following:
   - Parental Smoking increases the likelihood of both air pollution exposure and asthma in children.
   - Living in an urban area increases exposure to air pollution but is not directly related to asthma.

   a) Draw a DAG including the above variables.



b) What are the confounders you should adjust for?

Parental smoking is a confounder because it is associated with both:

- Increased likelihood of air pollution exposure (e.g., secondhand smoke contributing to indoor air pollution).
- Higher incidence of asthma in children (due to direct exposure to tobacco smoke).

c) Should you adjust for Urban Living? Explain why or why not based on your DAG.

You do not need to adjust for Urban Living, which increases exposure to air pollution but does not have a direct causal relationship with asthma. Thus, it is not a confounder in the context of the DAG you're considering.

Appendix: R codes for DAG

```
install.packages("DiagrammeR")
library(DiagrammeR)
grViz("
  digraph {
    graph [rankdir = LR, bgcolor = 'mistyrose']

    // Define nodes with red color
    F [label = 'Family Background', color = 'red', fontcolor = 'red']
    E [label = 'Education Level', color = 'red', fontcolor = 'red']
```

```
    J [label = 'Job Type', color = 'red', fontcolor = 'red']
    I [label = 'Income', color = 'red', fontcolor = 'red']

    // Define edges with red color
    F -> E [color = 'red'];
    F -> I [color = 'red'];
    E -> J [color = 'red'];
    J -> I [color = 'red'];
    E -> I [color = 'red'];
  }
")

grViz("
  digraph {
    graph [rankdir = LR, bgcolor = 'mistyrose']

    // Define nodes with red color
    F [label = 'Air Pollution', color = 'red', fontcolor = 'red']
    E [label = 'Asthma', color = 'red', fontcolor = 'red']
    J [label = 'Parental Smoking', color = 'red', fontcolor = 'red']
    I [label = 'Living Urban', color = 'red', fontcolor = 'red']

    // Define edges with red color
    F -> E [color = 'red'];
    J -> F [color = 'red'];
    J -> E [color = 'red'];
    I -> F [color = 'red'];
  }
")
```