**Q1. Inventing causal assumptions**

The part of the table below that is not shaded is observed data. The exposure, A, is a program aimed at pregnant women to reduce risk factors for infant mortality. The outcome, Y, is the infant mortality observed in each exposure group.

The shaded part of the table are counterfactuals. represents the average value of Y that we would have observed had we taken everyone in the study and put them in this program. represents the average value of Y that we would have observed had we taken everyone in the study and not put them in this program.

|  |  |  |  |
| --- | --- | --- | --- |
| Observed exposure (A) | Observed infant mortality per 10,000 (Y) |  |  |
| 0 | 7.2 |  |  |
| 1 | 5.4 |  |  |
|  | Weighted average 🡪 |  |  |

1. Take the numbers from the observed data in the table (unshaded), and fill them into the light shaded cells. (Note: you will have to use the numbers more than once.)
2. Think through what assumption(s) you needed to make in order to plug that number in that cell.
3. Assume that half of all pregnant women are in A=0 and half are in A=1. Calculate the value for the dark shaded cells (the weighted average of the white cells above each). Take the difference between grey cells.
4. Assuming that the assumptions you came up with in part 2 are true, what can you say about the number you found in 3?

**Q2. Inventing partial identification**

A reviewer of your paper says that they think the exchangeability assumption is not likely to hold. They want to know if there are any conclusions you can come to without relying on it.

Can you invent partial identification? In other words, is there a weaker assumption than perfect exchangeability that allows you to say something about what the possible estimate could be?

I have included two tables in case you need more than one…

|  |  |  |  |
| --- | --- | --- | --- |
| Observed exposure (A) | Observed infant mortality per 10,000 (Y) |  |  |
| 0 | 7.2 |  |  |
| 1 | 5.4 |  |  |
|  | Weighted average 🡪 |  |  |

|  |  |  |  |
| --- | --- | --- | --- |
| Observed exposure (A) | Observed infant mortality per 10,000 (Y) |  |  |
| 0 | 7.2 |  |  |
| 1 | 5.4 |  |  |
|  | Weighted average 🡪 |  |  |

**Q3. Inventing triangulation**

You want to know whether increasing outdoor playing time among children decreases their need for corrective eyewear. You estimate an answer to this question two different ways. The first way is to use regression controlling for confounders. The second way is using the quasi-random introduction of outdoor play areas as an instrumental variable.

You would like to be able to say something about whether these estimates are biased or not by comparing whether or not the estimates agree with each other.

1. Assuming that both methods (regression and instrumental variable) are truly answering the same question. What is the only assumption you need to make about the bias of each method in order to detect bias? Assume, for the moment that random error does not exist (i.e. your estimates are perfectly precise).
2. If we now include random error (each estimate has a confidence interval around it), how does this complicate what we want to do in question a?
3. If we had two methods that we knew were biased in opposite directions, what benefits does this have for triangulation?
4. If the two methods, unbeknownst to us, are actually estimating different effects, what do we learn if the estimates disagree?
5. Here’s a tougher question. If we knew the instrumental variable got the right answer 90% of the time and the regression method go the right answer 60% of the time, how much confidence should we have in the instrumental variable estimate after we used triangulation and found that the estimates were different? How much confidence should we have in the regression estimate?
6. Repeat question e except assuming that the regression method is right 85% of the time.