Exercises *Simpson's Paradox*

Joris Mooij

November 17, 2015

1. Simpson's Paradox

You are investigating the effectiveness of a drug against a deadly disease. You are given access to data collected by health insurance companies about their customers. You divide the diseased customers into two groups: those that took the drug (treated patients), and those that didn't take the drug (untreated patients). Some of the customers recovered, others unfortunately didn't. The reasons why some patients were treated and others were not, are unknown to you. You find the following numbers:

	Recovery	No recovery	Total	Recovery rate
Drug	20	20	40	...%
No drug	16	24	40	...%
Total	36	44	80	

1a. Calculate the recovery rates (in %) for both treated patients and untreated patients.
1b. If you were diseased, would you take the drug, or not?

Upon closer inspection of the data, you notice something peculiar when you group patients according to gender:

Males	Recovery	No recovery	Total	Recovery rate
Drug	18	12	30	...%
No drug	7	3	10	...%
Total	25	15	40	

Females	Recovery	No recovery	Total	Recovery rate
Drug	2	8	10	...%
No drug	9	21	30	...%
Total	11	29	40	

2a. Calculate the recovery rates (in %) for both the treated and the untreated patients, for both subpopulations (males and females).
2b. In light of these numbers, would you now take the drug, or not?

3. What would be your advice to a diseased patient with unknown gender?

This phenomenon is known as Simpson's paradox. A lot has been written about this paradox, but it dissolves once you recognize that you should not make the mistake of interpreting correlations as causations, as we'll see later today.