Seminar 1 - Simpson's pardox

The Simpson's paradox occurs when the direction of association between two variables is reversed (or changes considerably) when subgroups of data are analyzed. The Simpson's paradox is described by De Grooth and Schervish (pages 654 and following) and in various online sources, such as

- Vaccine efficacy and vaccination on Isrealian COVID data
- Wikipedia
- Video explanation

Simpson's paradox is not actually a paradox but the occurrence of a puzzling result that occurs in some circumstances when data are aggregated.

- Describe the Simpson's paradox with two real-data examples, one that is based on two categorical variables and one where at least one numerical variable is considered.
- Simpson's paradox can be described also in probability terms:

  $P(I|A \cap B) > P(I|A \cap B^C)$

  $P(I|A^C \cap B) > P(I|A^C \cap B^C)$

  $P(I|B) < P(I| \cap B^C)$

  Define what I, A and B are in the example with two categorical variables that you choose.

  (Suppose that A and B are events such that $0 < P(A) < 1$ and $0 < P(B) < 1$.)
- Show that if $P(A|B) = P(A|B^C)$ then it is not possible for the three inequalities described above to hold.
- Show that if $P(B|A) = P(B|A^C)$ then it is not possible for the three inequalities described above to hold.
- Speculate how the occurrence of Simpson's paradox could prevented when an experimental study is planned (hint: think about how randomization of subjects could help).

- Write a short summary/outline of 1-2 pages of what you will present in class and submit it at least one day before the presentation. For example, summary should briefly present the examples that you choose and give some details on how you addressed the theoretical questions.
- Presentation date: **October 29th 2024. Presentation length: 25 minutes. Writ One day before the presentation**