Simpson's Paradox

Max Turgeon

DATA 2010–Tools and Techniques in Data Science

Israel COVID Vaccine Data

Age	Population (%)		Severe cases		Efficacy
	Not Vax %	Fully Vax %	Not Vax per 100k	Fully Vax per 100k	vs. severe disease
All ages	1,302,912 18.2%	5,634,634 78.7%	214 16.4	301 5.3	**67.5%**

Source: Jeffery Morris, https://www.covid-datascience.com/post/israeli-data-how-can-efficacy-vs-severe-disease-be-strong-when-60-of-hospitalized-are-vaccinated

Israel COVID Vaccine Data Redux

Age	Population (%)		Severe cases		Efficacy
	Not Vax %	Fully Vax %	Not Vax per 100k	Fully Vax per 100k	vs. severe disease
All ages	1,302,912 18.2%	5,634,634 78.7%	214 16.4	301 5.3	**67.5%**
<50	1,116,834 23.3%	3,501,118 73.0%	43 3.9	11 0.3	**91.8%**
>50	186,078 7.9%	2,170,563 90.4%	171 90.9	290 13.6	**85.2%**

Source: Jeffery Morris, https://www.covid-datascience.com/post/israeli-data-how-can-efficacy-vs-severe-disease-be-strong-when-60-of-hospitalized-are-vaccinated

What is going on?

- Vaccination and severity are not uniform over age groups.
    - Most unvaccinated people are < 50 years old.
    - Most severe events occur in people who are > 50 years old.
- When stratifying by age, we see good vaccine efficacy.

Simpson's Paradox

- **Simpson's Paradox** refers to situations where we see an effect (e.g. vaccine efficacy) in subgroups, but the effect disappears, or attenuates, in aggregate.
- This paradox can be resolved by careful *interpretation* and *modeling* of the data.
    - In the example above, to accurately model the effect of vaccines, we need to include age in our analysis.