Causal Inference for Beginners

01 Causality

How Not to Lie with Statistics

Jiahao Zhu

Mars, International School of Business & Finance, Sun Yat-sen University

2023-09-19

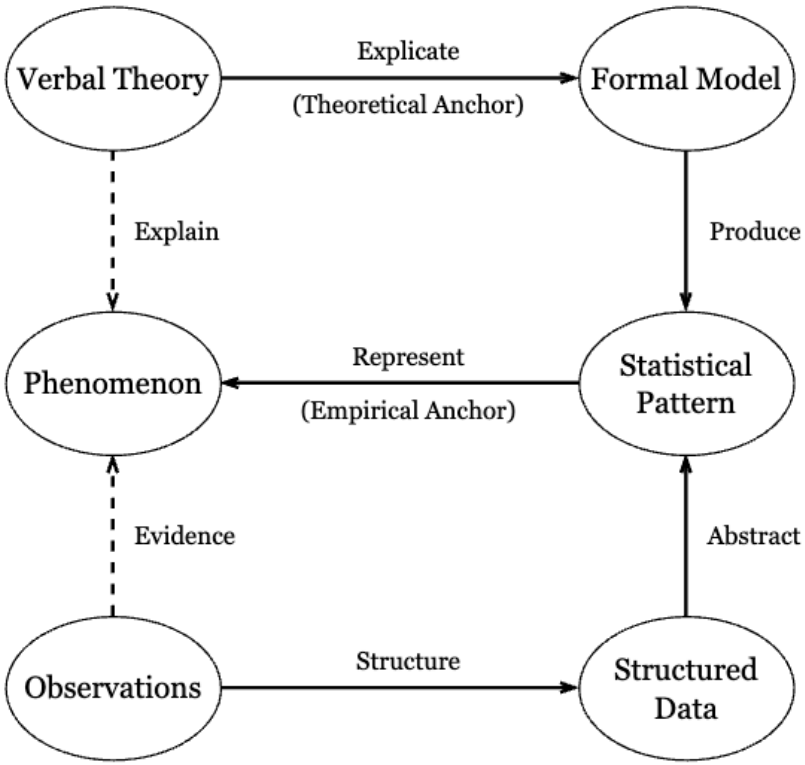
Outline

1. Review
2. Warm-up
3. Correlation vs. Causality
4. Experiments vs. Observations
5. Conclusion

Review

Productive explanation

Figure 2
The Productive Explanation Model.



Productive explanation framework (van Dongen et al., 2022)

- How can we infer *statistical patterns* from *structured data*?
- How can we construct a *verbal theory* or even a *formal model*?

Causal Inference

Warm-up

Some questions

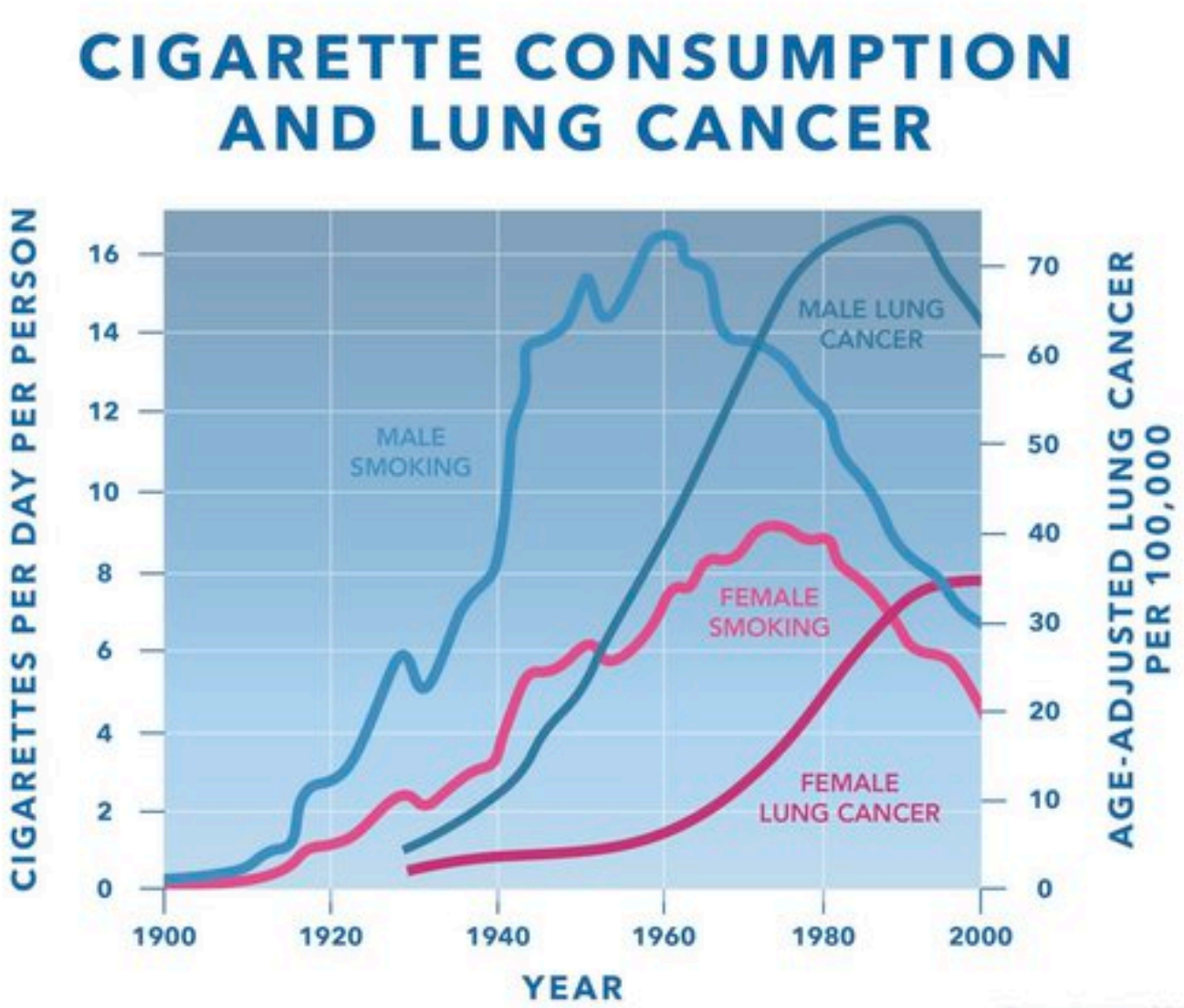
- Does smoking cause lung cancer?
- Does chocolate consumption increase cognitive ability?
- Does COVID-19 vaccines protect against hospitalization?
- How will the revenue of a company change if it increases the price of a product?
- How will the sales change if a company launches an ad?
- Do female graduate students applying for college have lower admission chances than male graduate students?

Warm-up

What will happen to Y if action X is performed?

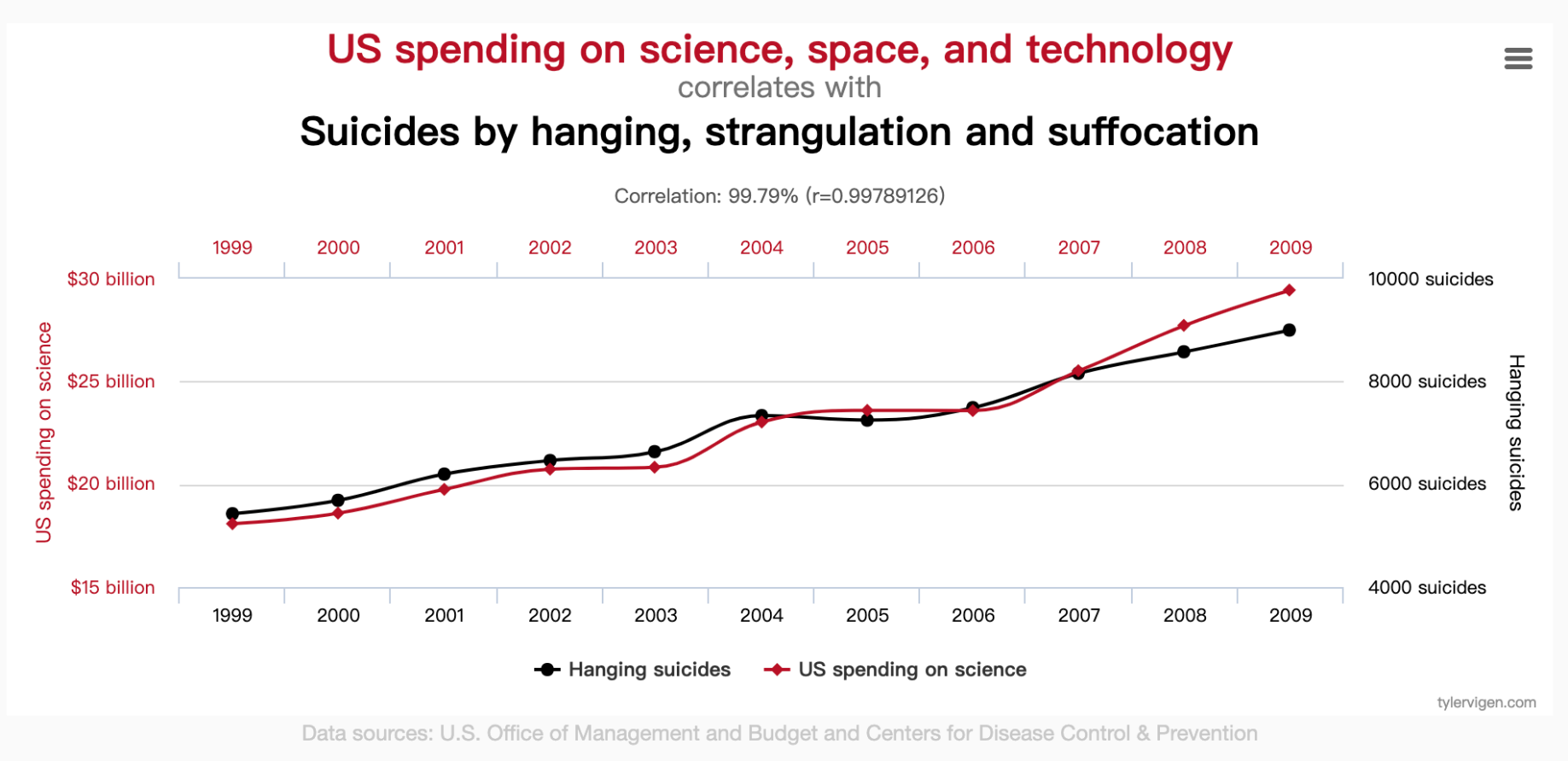
Correlation vs. Causality

Is correlation all you need?



Correlation vs. Causality

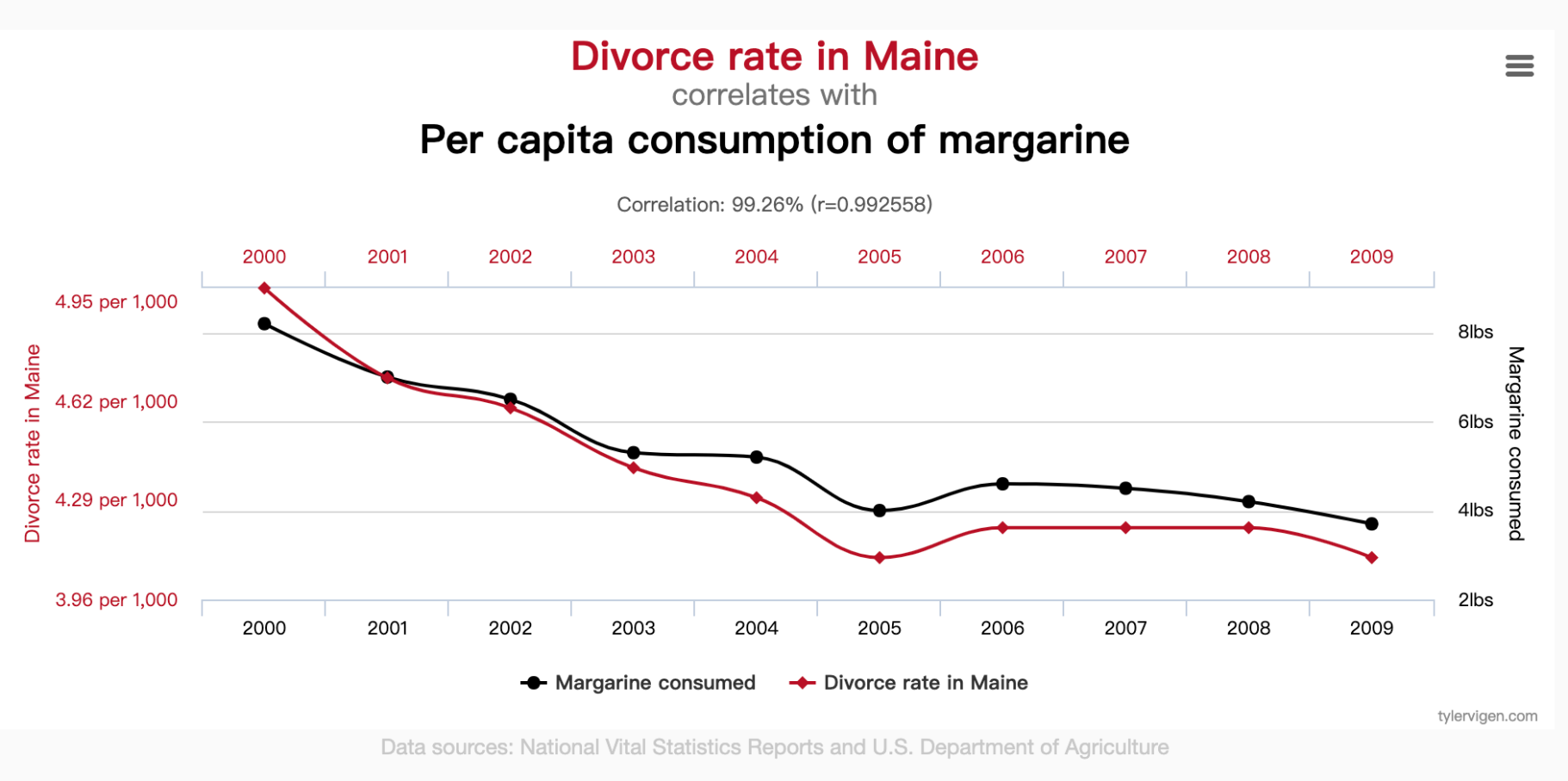
Is correlation all you need?



Does the US government kill people during technological development?

Correlation vs. Causality

Is correlation all you need?



Does eating margarine (人造黄油) make couples divorce?

Correlation vs. Causality

Is correlation all you need?

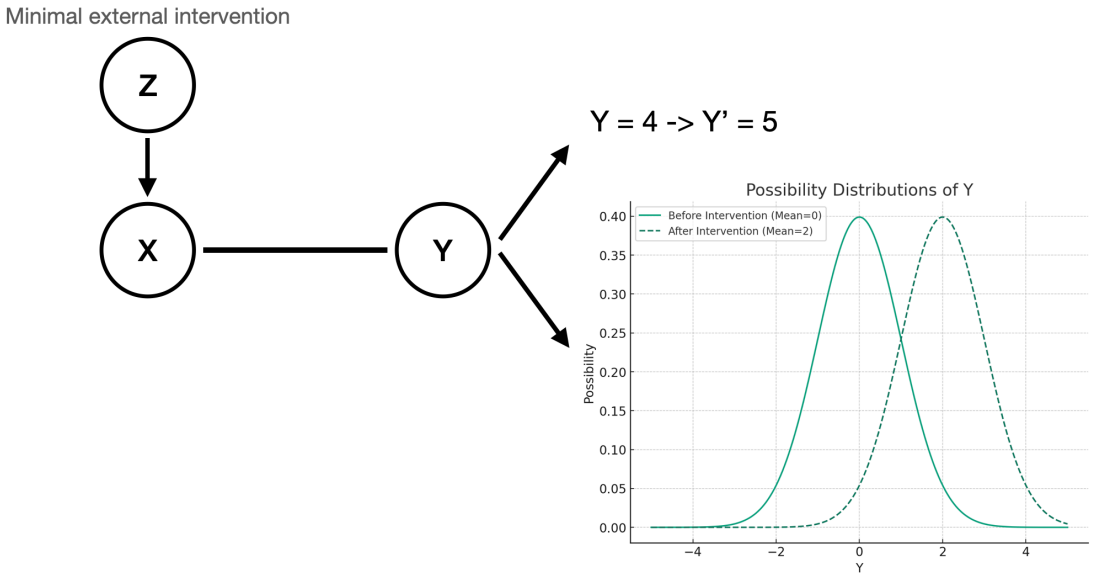
- We are surrounded by seemly reasonable but spurious causal claims with correlations (*https://tylervigen.com/spurious-correlations*)
  - rich people are happier
  - napping protects against Alzheimer's disease
  - drinking coffee decreases death risks
  - ...
- **Correlation is not causality**

What is causality?

Correlation vs. Causality

Causality

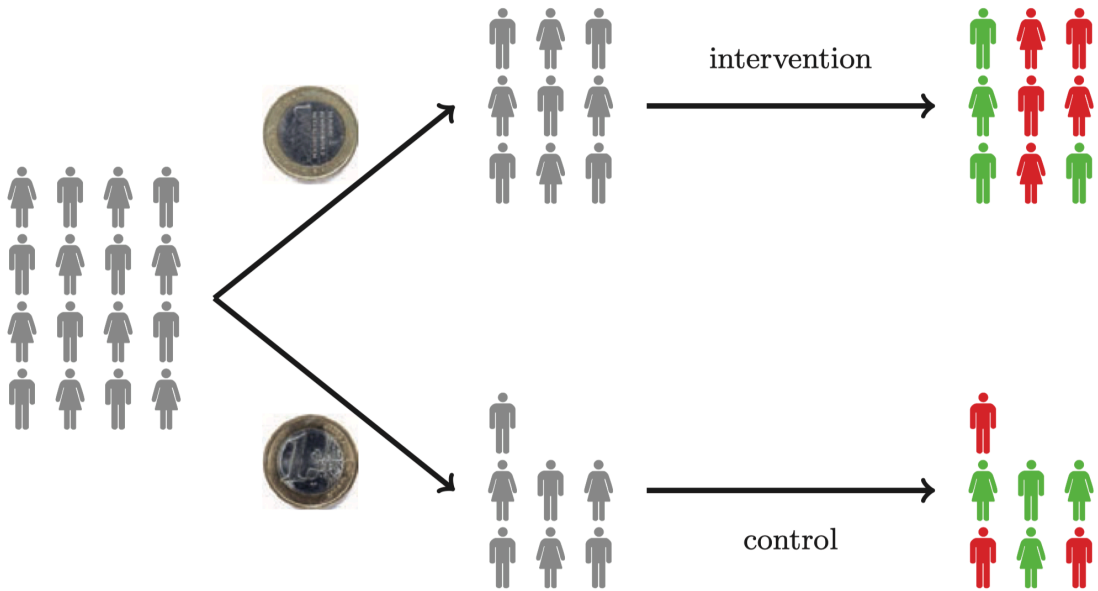
- Causality
  - caus(e) + al + ity
  - cause: make (something, especially something bad) happen (cf. Oxford Languages)
- Definition of causality in a *cause* manner
  - X cause Y
  - a minimal external intervention on the system that sets the value of X may change the possible values (or probability distribution) of Y



Experiments vs. Observations

Randomized Controlled Trials (RCTs)

- *Gold standard* of causal inference
- Limitations
  - money, time, effort, etc.
  - ethical problem
  - inclusion criteria



RCTs

Experiments vs. Observations

Simpson's paradox

- How do two COVID-19 vaccines (A and B) protect against hospitalization?

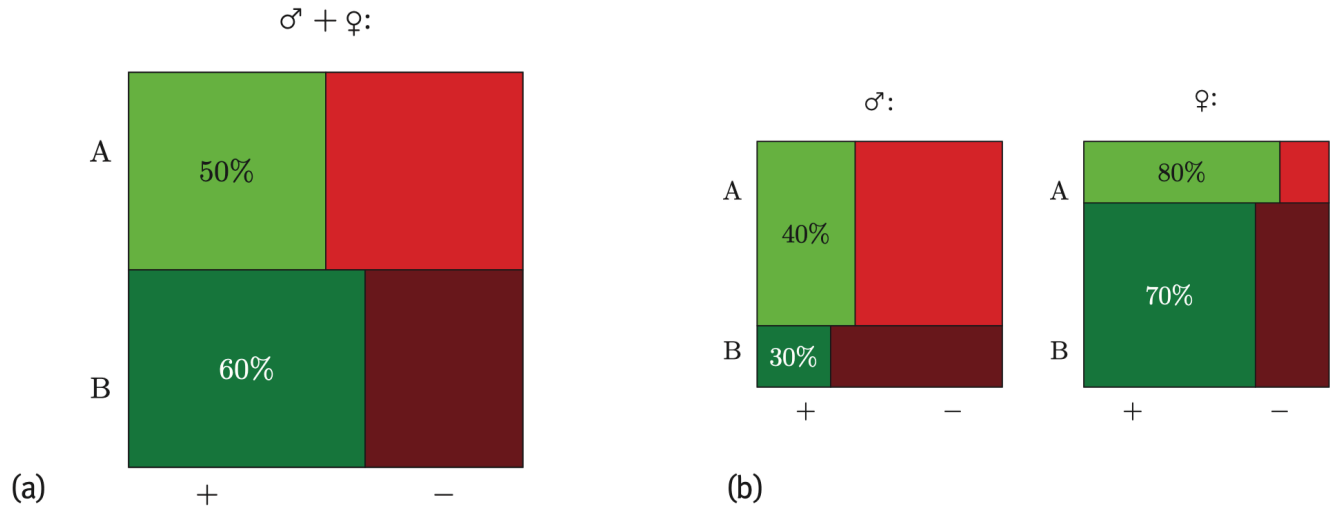
	♂ + ♀		♂		♀	
	$Y=+$	$Y=-$	$Y=+$	$Y=-$	$Y=+$	$Y=-$
$X=\mathrm{A}$	2500	2500	1500	2250	1000	250
$X=\mathrm{B}$	3000	2000	375	875	2625	1125

- Results
  - all sample: $hosp\%(A) = 0.5$; $hosp\%(B) = 0.4$ (B better than A)
  - male sub-sample: $hosp\%(A) = 0.6$; $hosp\%(B) = 0.7$ (A better than B)
  - female sub-sample: $hosp\%(A) = 0.2$; $hosp\%(B) = 0.3$ (A better than B)
- Wired conclusion
  - if we don't know a patient's gender, use vaccine A
  - if we do know a patient's gender, use vaccine B

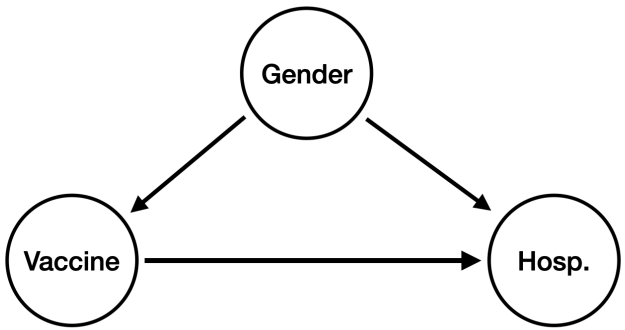
Experiments vs. Observations

Simpson's paradox

- Look deep into the data



- Problems: unbalanced sample



the potential data generative process

Conclusion

- Causal inference and assumption link theory and observation
- Correlation is not causality
- Causality is defined as an intervention on X will change Y
- Two famous ways to infer causality (experiments and observations)
  - experiments is good but may be limited
  - observation can deceive the unwary
- Causal inference roots in the assumption of how data is generated

References

- Mooij, J. M. (2022). Causality: from data to science.

Thanks for listening!

Q&A