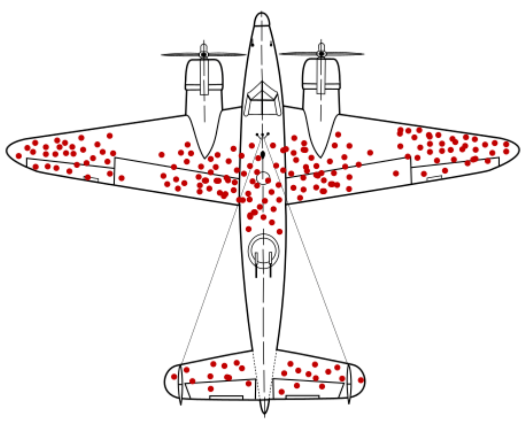
Causal Inference

Bias in Observational Data: Example

“During World War II, the statistician Abraham Wald took survivorship bias into his calculations when considering how to minimize bomber losses to enemy fire ....” (source Wikipedia)

Below is an image showing where planes were shot. These data were collected once the plane returned to airbase.

Q: Where will you reinforce the armor?



Bias in Observational Bias

1. **Survivor bias** (what we've just seen)

2. Self-reporting bias: After major elections, surveys typically find that 60–70% of respondents say they voted, but actual turnout records (which are public in many states) show only ~50–55% turnout.

3. **Post-treatment Bias**: Does a college education increase income?
   Suppose we adjust for “occupation” when estimating the effect of college on income. But occupation is partly determined by having gone to college.

4. **Attrition Bia**: Do online learning platforms improve performance?
   If struggling students are more likely to drop out of an online course, the final test score average may seem high — but it only reflects those who didn’t drop out.

5. **Reverse Causality**: Does stress cause lack of sleep?
   Maybe. But also: lack of sleep causes stress. If we just look at correlation, we might mistake the direction of causality.

6. **Collider Bias** (Berkson's Paradox): Is exercise unrelated to healthy eating?
   Among hospital patients, those who exercise tend to eat less healthily — but this is not true in the general population.
   Why? Hospital admission might depend on either bad diet or lack of exercise — conditioning on this collider creates a spurious relationship.

Paradoxes in Statistics

Simpson's paradox. The following table summarizes number of successes and fails for two treatments (A and B) for small and large kidney stones. Treatment A is an open surgical procedures, and Treatment B is a minimally-invasive procedure. A more successful treatment should yield higher success rate compare to its alternative.

	Small	**stones**	**Large**	**stones**
	success	fail	success	fail
Treatment A	81	6	192	71
Treatment B	234	36	55	25

We can calculate the success rates as follow Treatment A v.s. Treatment B).

- Success rate for small stones: 0.93 (81/87) > 0.87 (234/270)
- Success rate for large stones: 0.73 (192/263) > 0.69 (55/80)
- Overall success rate: 0.78 (273/350) < 0.83 (289/350)

Which treatment is more effective? Read more about Simpson's paradox and also Lord's paradox

Spurious Correlations from Observational Data

Today we rely heavily on quantitative methods to recover (conditional) correlation or causal relationship from data.

But feeding data to untrained methods might yield spurious correlation...

Pursuit of Causation

However, it is often the causal inference that matters in the real world. For instance, we might be interested in the following questions.

- What would happen to the patient if they received treatment A instead of B?
- What would happen to the unemployment rate if the U.S. government increased minimum wages?
- What would happen to the case number if a state took a different action in April?

In all these what-ifs, we notice that it is always comparing *outcomes* under *different* conditions for the *same* subject(s). In other words, **causal inference** is the comparison between *potential outcomes* under *treatment* and *control* for the *same* unit(s).

We will take a close look at the potential outcome framework

Applications of Potential Outcome

Using the potential outcome framework, we can describe the sources of bias and more:

- Propensity score methods

- Instrumental variable analysis
