Hypothesis Testing
&
Causal Inference

Mentor: Kenny Zhang
Mentee: Serena Sun

What is Causal Inference?

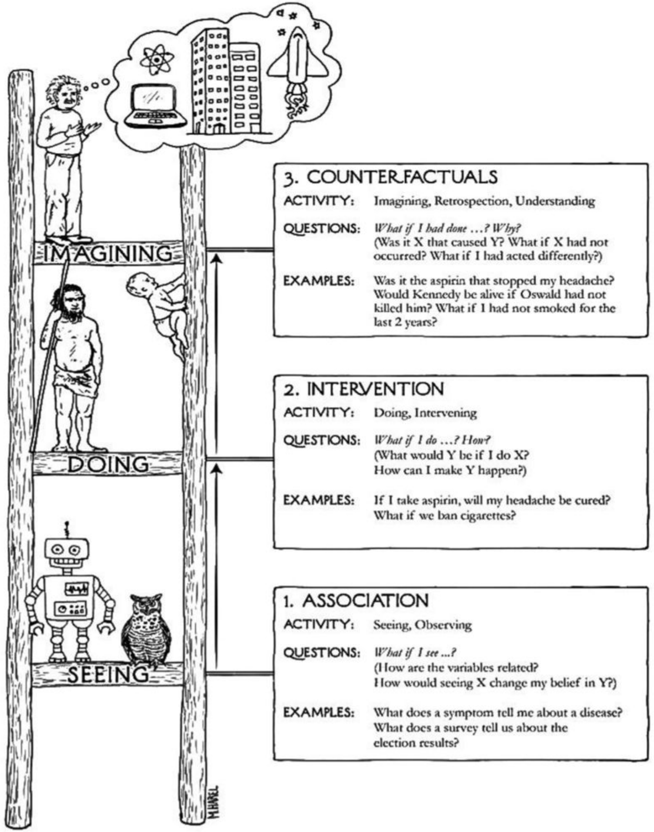
Causal inference refers to the process of **drawing conclusions about cause-and-effect relationships** between variables or events based on observed data.

Inferring the effects of any treatment/policy/Intervention/etc.

Real-life Questions:

- **Health Care**: Does a new drug treatment lead to a reduction in patient mortality rates?
- **Education**: What is the causal impact of a new teaching method or curriculum on student academic performance?
- **Economics:** Does foreign aid lead to economic growth and poverty reduction in developing countries?
- …

The Ladder of Causation



Imagining, Understanding

Doing, Intervening

Seeing, Observing

Why We Care?

Causal inference is crucial in tech, particularly in the context of **A/B testing**.

Scenario: A social media platform wants to increase user engagement with its new feature. (ins story)

A/B Testing: The platform **randomly** assigns users to two groups:

Group A(control group), which sees the traditional feed posts only

Group B(experimental group), which sees both feed posts and the new Story feature.

Metrics Selection: Key engagement metrics

views, likes, shares, and time spent.

Outcome: Users in Group B, exposed to the Story feature, exhibit significantly higher engagement metrics compared to those in Group A.

The platform concludes that the **Story feature** causally increases user engagement and decides to roll it out to all users.

Takeaway:

Through causal inference in A/B testing, tech companies can objectively **evaluate the impact** of new features and **make data-driven decisions** to enhance user experiences and drive platform growth.

Recap

A/B Testing: The platform **randomly** assigns users to two groups

Confounding Variables: Confounding variables are factors that are correlated with both the treatment (new feature) and the outcome, potentially leading to biased results.

Confounding variable
User Demographics
Device Type
Content Relevance
User Behavior Patterns
…

Independent variable
Presence or absence of the new feature

Dependent variable
Number of views
Number of (likes, comments, shares)
Time spent viewing the story
Click-through rate (if the feature includes a call-to-action)
…

Randomized Controlled Trials

“RCTs allows us to estimate the causal effect X $\longrightarrow$ Y without falling prey to confounders bias.”

- "The Book of Why" by Judea Pearl

Two benefits:

1. Eliminates confounder bias (it asks Nature the right answer)
2. Enables the researcher to quantify his uncertainty

Potential Outcome

Consider a study with n experimental units indexed by i = 1, . . . , n.

As a starting point, we focus on a treatment with two levels: 1 for the treatment and 0 for the control.

For each unit i, the outcome of interest Y has two versions: **Yi(1) and Yi(0)**, which are potential outcomes under the hypothetical interventions 1 and 0.

Table 1.1. ***Example of Potential Outcomes and Causal Effect with One Unit***

Unit	Potential Outcomes		Causal Effect
	Y(Aspirin)	*Y*(No Aspirin)	
You	No Headache	Headache	Improvement due to Aspirin

Assumption 2.1 (no interference) *Unit i's potential outcomes do not depend on other units' treatments. This is sometimes called the no-interference assumption.*

Assumption 2.2 (consistency) *There are no other versions of the treatment. Equivalently, we require that the treatment levels be well-defined, or have no ambiguity at least for the outcome of interest. This is sometimes called the consistency assumption.*[1]

Under SUTVA, Rubin (2005) called the $n \times 2$ matrix of potential outcomes the Science Table:

i	$Y_i(1)$	$Y_i(0)$
1	$Y_1(1)$	$Y_1(0)$
2	$Y_2(1)$	$Y_2(0)$
$\vdots$	$\vdots$	$\vdots$
n	$Y_n(1)$	$Y_n(0)$

Causal effects are functions of the Science Table. Inferring individual causal effects

$$\tau_i = Y_i(1) - Y_i(0), \quad (i = 1, \ldots, n)$$

is fundamentally challenging because we can only observe either $Y_i(1)$ or $Y_i(0)$

for each unit i, that is, we can observe only half of the Science Table. As a starting point, most parts of the book focus on the average causal effect (ACE):

$$\begin{aligned} \tau &= n^{-1} \sum_{i=1}^{n} \{Y_i(1) - Y_i(0)\} \\ &= n^{-1} \sum_{i=1}^{n} Y_i(1) - n^{-1} \sum_{i=1}^{n} Y_i(0). \end{aligned}$$

Thanks For Listening

Any Questions?