



Welcome to this session

Skills Bootcamp:

Causal Inference

The session will start shortly...

Questions? Drop them in the chat.
We'll have dedicated moderators
answering questions.



Skills Bootcamp Data Science Housekeeping

- The use of disrespectful language is prohibited in the questions, this is a supportive, learning environment for all - please engage accordingly. **(Fundamental British Values: Mutual Respect and Tolerance)**
- No question is daft or silly - **ask them!**
- There are **Q&A sessions** midway and at the end of the session, should you wish to ask any follow-up questions. We will be answering questions as the session progresses as well.
- If you have any questions outside of this lecture, or that are not answered during this lecture, please do submit these for upcoming Academic Sessions. You can submit these questions here: **Questions**

Skills Bootcamp Data Science Housekeeping

- For all **non-academic questions**, please submit a query: www.hyperiondev.com/support
- Report a safeguarding incident: www.hyperiondev.com/safeguardreporting
- We would love your feedback on lectures: [Feedback on Lectures.](#)
- Find all the lecture **content** in your [Lecture Backpack](#) on GitHub.
- If you are hearing impaired, kindly use your computer's function through Google chrome to enable captions.

Safeguarding & Welfare

We are committed to all our students and staff feeling safe and happy; we want to make sure there is always someone you can turn to if you are worried about anything.

If you are feeling upset or unsafe, are worried about a friend, student or family member, or you feel like something isn't right, speak to our safeguarding team:



Ian Wyles
Designated Safeguarding
Lead



Simone Botes



Nurhaan Snyman



Rafiq Manan



Ronald Munodawafa



Tevin Pitts

Scan to report a
safeguarding concern



or email the Designated
Safeguarding Lead:
Ian Wyles
safeguarding@hyperiondev.com

Skills Bootcamp Progression Overview

✓ Criterion 1 - Initial Requirements

Specific achievements **within the first two weeks** of the program.

To meet this criterion, students need to, by no later than **01 December 2024 (C11)** or **22 December 2024 (C12)**:

- **Guided Learning Hours (GLH):** Attend a **minimum of 7-8 GLH per week** (lectures, workshops, or mentor calls) for a total minimum of **15 GLH**.
- **Task Completion:** Successfully complete the **first 4 of the assigned tasks**.

✓ Criterion 2 - Mid-Course Progress

Progress through the successful completion of tasks **within the first half** of the program.

To meet this criterion, students should, by no later than **12 January 2025 (C11)** or **02 February 2025 (C12)**:

- **Guided Learning Hours (GLH):** Complete at least **60 GLH**.
- **Task Completion :** Successfully complete the **first 13 of the assigned tasks**.

Skills Bootcamp Progression Overview

✓ Criterion 3 – End-Course Progress

Showcasing students' progress nearing the completion of the course.

To meet this criterion, students should:

- **Guided Learning Hours (GLH):** Complete the **total minimum required GLH**, by the **support end date**.
- **Task Completion : Complete all mandatory tasks**, including any necessary resubmissions, by the end of the bootcamp, **09 March 2025 (C11)** or **30 March 2025 (C12)**.

✓ Criterion 4 - Employability

Demonstrating progress to find employment.

To meet this criterion, students should:

- **Record an Interview Invite:** Students are required to record proof of invitation to an interview by **30 March 2025 (C11)** or **04 May 2025 (C12)**.
 - **South Holland Students** are required to proof and interview by **17 March 2025**.
- **Record a Final Job Outcome :** Within 12 weeks post-graduation, students are required to record a job outcome.



Question



What is your understanding of Causal Inference?

Learning Outcomes

- Discuss the Concept of Causal Inference
- Identify and Apply Causal Inference Methods
- Assess Causal Relationships in Data
- Explain challenges in causal inference

Lecture Overview

- Introduction to causal inference
- Causal graphs
- Key techniques for causal inference
- Real world applications
- Challenges in causal inference



Introduction to Causal Inference

Wikipedia Definition

- ❖ According to wikipedia, **causal inference** is the process of determining the independent, actual effect of a particular phenomenon that is a component of a larger system.

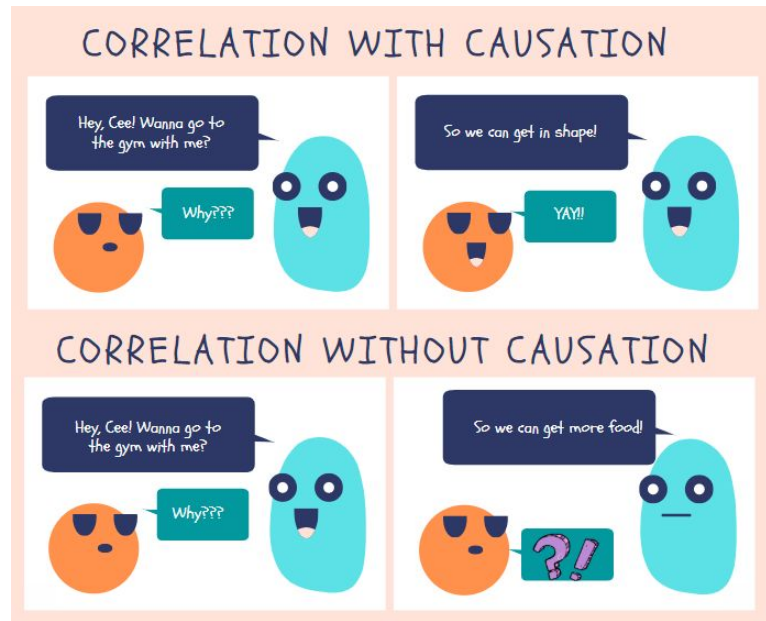
What is Causal Inference?

- ❖ Causal inference is the scientific process of determining cause-and-effect relationships between variables. It goes beyond correlation to establish whether one variable directly influences another. This distinction is crucial because, as the classic statement reminds us: "Correlation does not imply causation." For instance, just because ice cream sales and drowning incidents rise together does not mean one causes the other. Understanding this distinction is essential in fields ranging from social science and medicine to business and technology.
- ❖ Without a rigorous approach, it is easy to make incorrect causal claims. A common question is: "Can a regression analysis establish causation?" The short answer is no, unless additional steps are taken. This is where causal inference methods come into play.

Introduction to Causality

Correlation vs Causation

- ❖ One of the most fundamental misconceptions in data science is mistaking correlation for causation. Correlation simply measures the relationship between two variables, while causation implies that one variable directly influences the other.
- ❖ Example:
 - Ice cream sales and drowning incidents are correlated, but ice cream does not cause drowning. The underlying confounder is temperature. Hotter weather increases both ice cream sales and swimming activities.



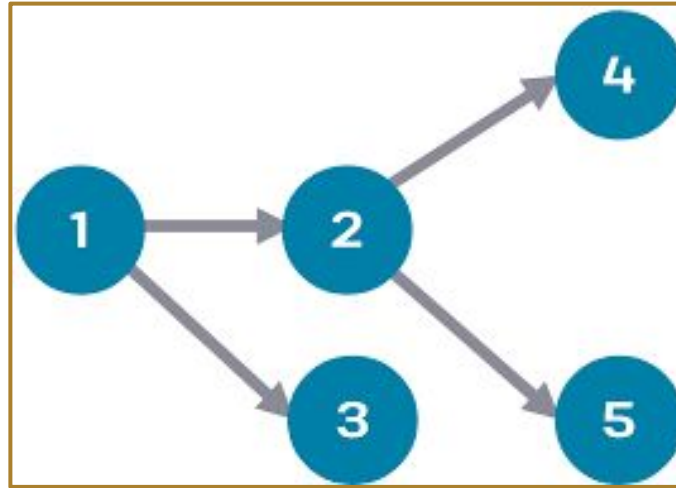
Why is Causal Inference important?

- ❖ In data-driven decision-making, understanding causality helps:
 - **Predict Outcomes More Accurately:** Helps in making better policy or business decisions.
 - **Avoid Spurious Relationships:** Identifies real causal effects rather than coincidental correlations.
 - **Optimize Interventions:** Determines what actions will yield the desired outcomes.

Causal Graphs: Directed Acyclic Graphs (DAGs)

Understanding DAGs

- ❖ A **Directed Acyclic Graph (DAG)** is a graphical representation of causal relationships between variables. Nodes represent variables, and directed edges (arrows) indicate causal effects.

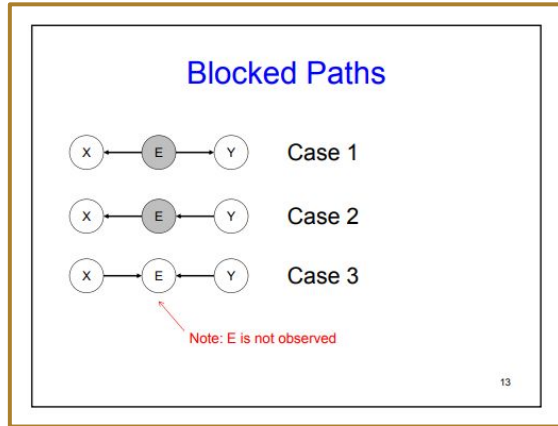


Example of a DAG

- ❖ Suppose we want to determine the effect of Exercise (X) on Weight Loss (Y).
- ❖ A potential confounder is Diet (Z).
- ❖ The DAG might look like:
 - $X \rightarrow Y$
 - $Z \rightarrow X$
 - $Z \rightarrow Y$
- ❖ Here, Diet (Z) affects both Exercise (X) and Weight Loss (Y), meaning failing to account for it may lead to incorrect conclusions about Exercise's effect on Weight Loss.

D-separation and Blocking Paths

- ❖ **D-separation** is a method to determine if two variables are independent given another variable.
- ❖ **Blocking paths:** If all backdoor paths (non-causal paths) between X and Y are blocked by controlling for confounders, we can estimate the causal effect accurately.

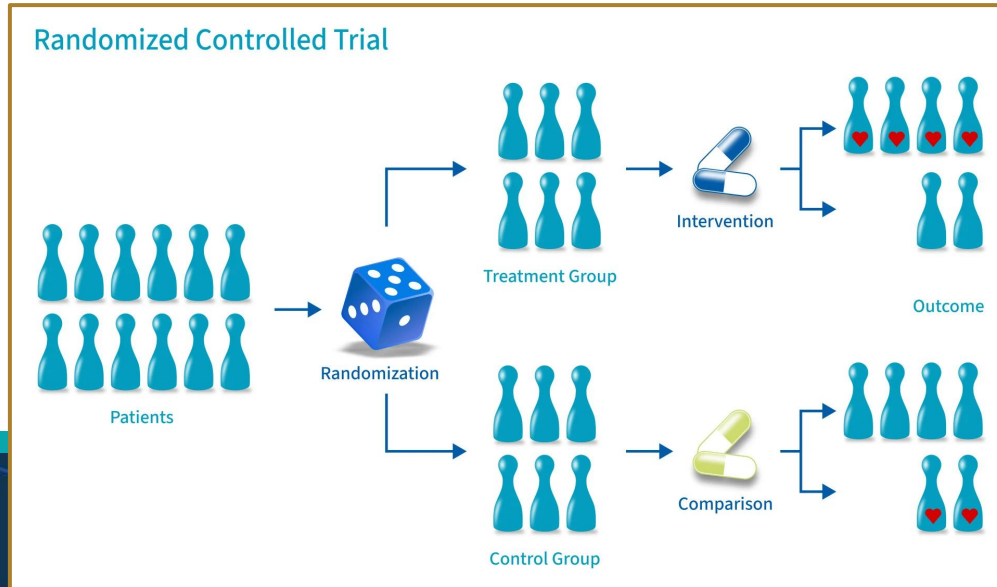




Key Techniques for Causal Inference

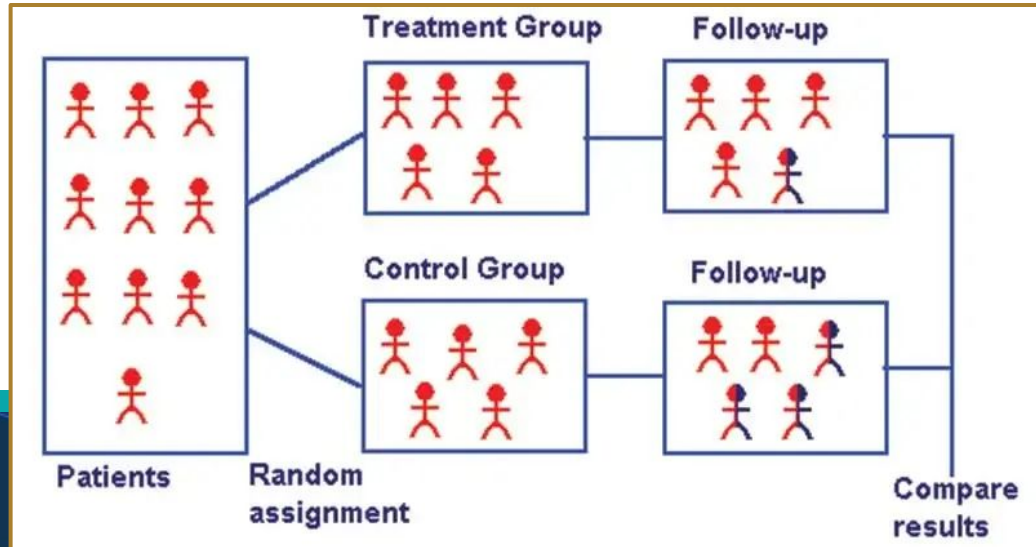
Randomized Controlled Trials(RCTs)

- ❖ The gold standard for causal inference is randomized experiments, where subjects are randomly assigned to treatment and control groups, ensuring that differences in outcomes are due to the intervention and not confounding factors.



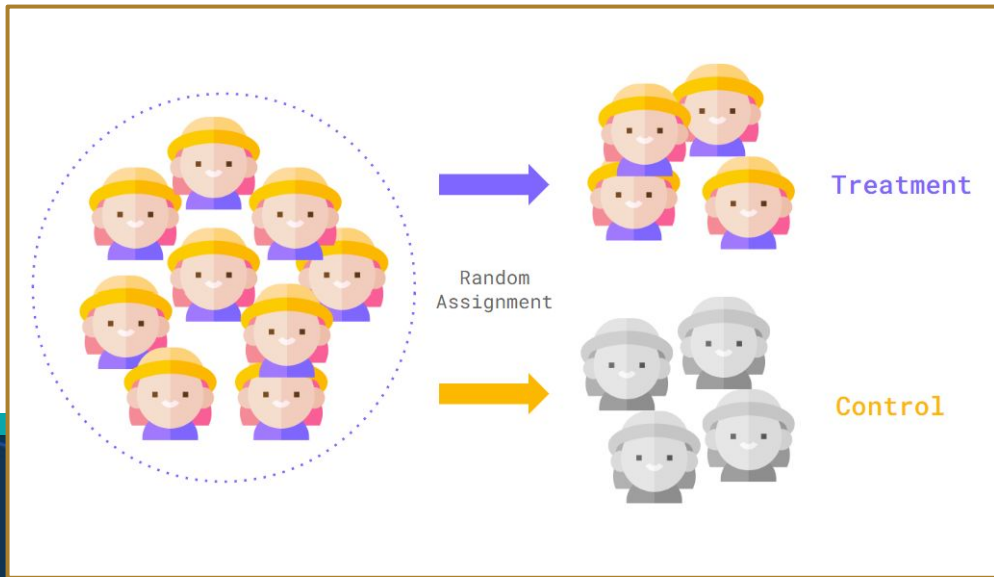
Example

- ❖ A pharmaceutical company tests a new drug.
- ❖ Patients are randomly assigned to receive either the drug or a placebo.
- ❖ Randomization ensures that differences in outcomes are due to the treatment rather than confounding factors.



Example 2

- ❖ A training program is introduced to improve student performance. Students are randomly assigned to receive the program (treatment group) or not (control group). By comparing their average scores, researchers can quantify the program's impact.

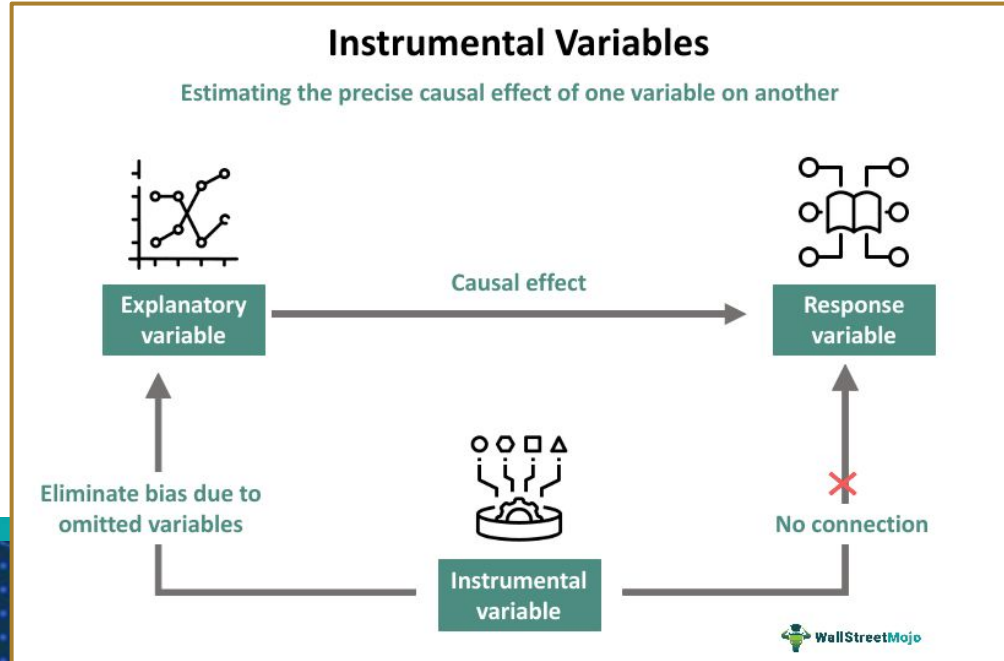


Limitations

- ❖ Expensive and time-consuming.
- ❖ Ethical concerns (e.g., withholding treatment from some patients).
- ❖ Not always feasible in real-world settings.
- ❖ RCTs are powerful but challenging to implement, especially in business settings where customer behavior cannot always be controlled.

Instrumental Variables (IV)

- ❖ When randomization is not possible, instrumental variables help estimate causal effects by introducing a variable that influences the treatment but not the outcome directly.



Instrumental Variables (IV)

❖ **Example:**

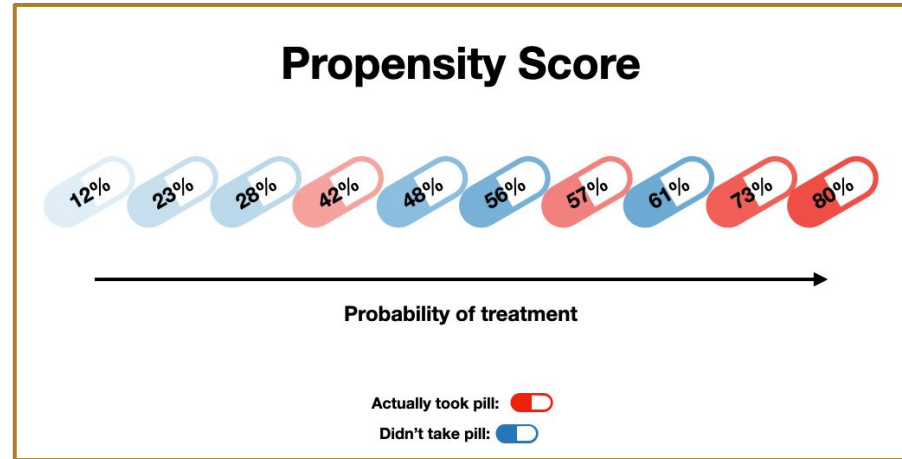
- To measure the effect of education on income, we can use proximity to a college as an instrument. It influences education but does not directly affect income beyond its impact on education.

❖ **Key Assumptions:**

- **Relevance:** The instrument affects the treatment.
- **Exogeneity:** The instrument is not related to the outcome, except through the treatment.

Propensity Score Matching (PSM)

- ❖ When randomization is not feasible, PSM attempts to match treated and untreated units based on their likelihood of receiving the treatment.



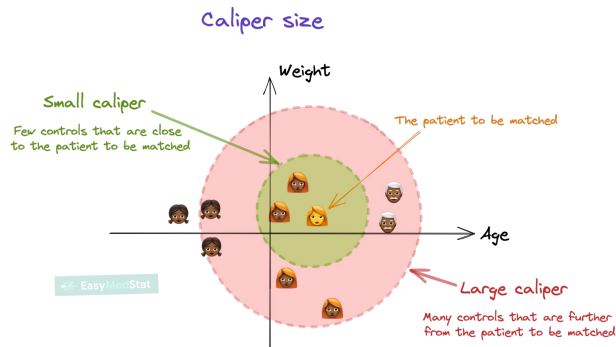
Propensity Score Matching (PSM)

❖ Example:

- In healthcare, comparing patients who received a new therapy with similar patients who did not.
- The propensity score is the probability of receiving treatment based on observed characteristics.

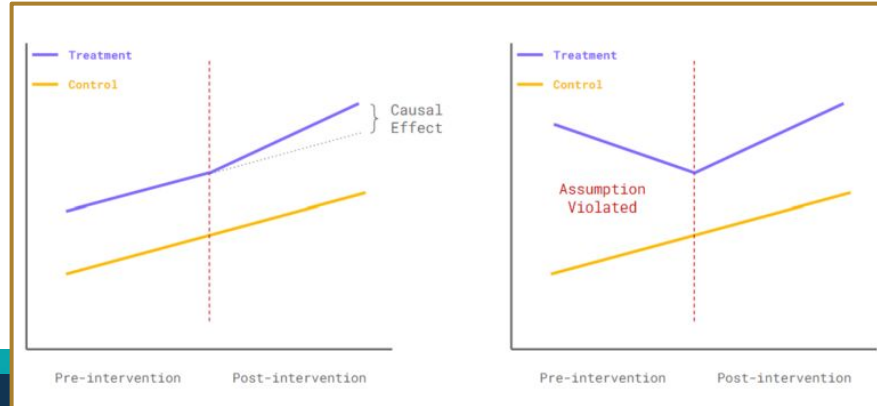
❖ Limitations:

- Can only account for observable confounders.
- Matching quality depends on data availability.



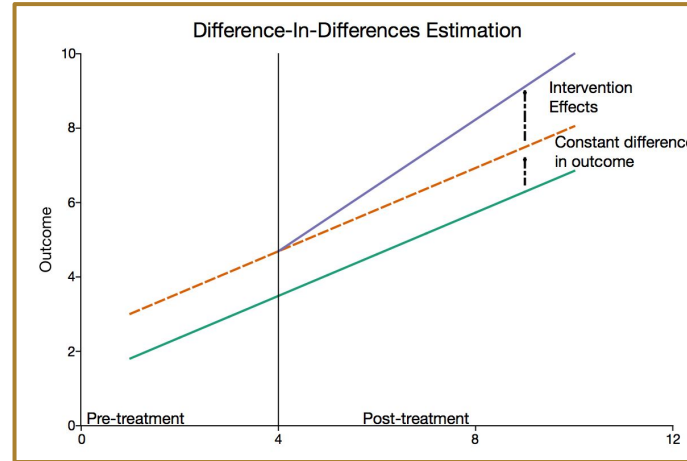
Difference-in-Differences (DiD)

- ❖ **Example:** Agricultural Productivity
 - A new farming technique is introduced in one region (treatment) but not another (control). By comparing the productivity change in both regions before and after implementation, researchers estimate the technique's impact.
- ❖ **Key Assumption:** Parallel Trends
 - DiD relies on the assumption that both groups would have followed similar trends in the absence of treatment. If violated, results can be biased.



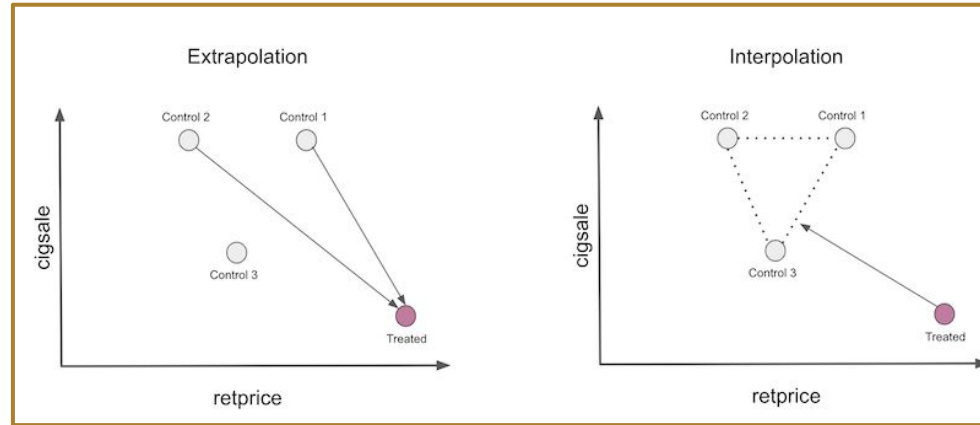
Difference-in-Differences (DiD)

- ❖ A quasi-experimental approach used when randomization is not possible. This method compares changes over time between a treatment and a control group.



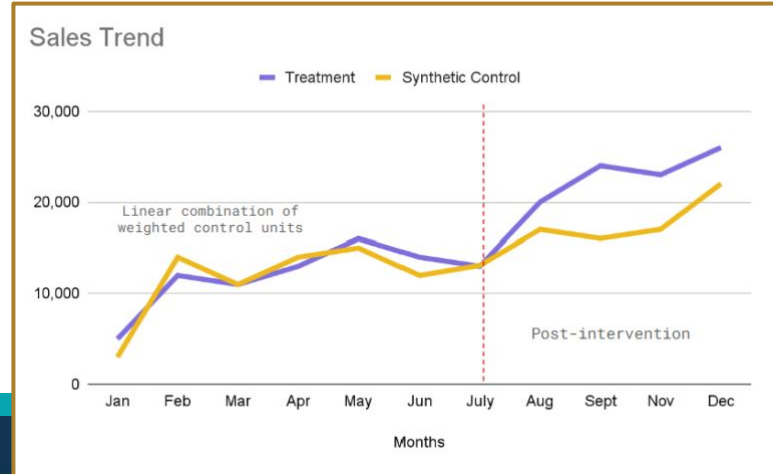
Synthetic Control (SC)

- ❖ SC is used when a single treated unit (e.g., a state or a company) undergoes an intervention. Instead of using a traditional control group, SC creates a weighted combination of multiple control units to serve as a synthetic counterfactual.



Synthetic Control (SC)

- ❖ **Example:** Marketing Campaign Impact
 - A company runs a major marketing campaign in Region A. To assess the impact, researchers construct a synthetic version of Region A using data from other similar regions. The difference in sales between actual and synthetic Region A estimates the campaign's true effect.
- ❖ **Limitations**
 - SC requires rich historical data and careful selection of control units to ensure validity.



Real-World Applications

Healthcare: Effectiveness of Medical Treatments

- ❖ Example: Determining if a new drug reduces heart attacks.
- ❖ Approach: Use RCTs or instrumental variables if randomization is infeasible.

Policy Evaluation: Minimum Wage Effects on Employment

- ❖ Example: Does increasing the minimum wage reduce employment?
- ❖ Approach: Use difference-in-differences (DiD) by comparing employment trends before and after a policy change in treated and untreated regions.

Marketing: Ad Campaign Effectiveness

- ❖ Example: Measuring the impact of an online ad campaign on sales.
- ❖ Approach: Use A/B testing (randomized experiments) or synthetic control methods to estimate the ad's true effect.



Challenges in Causal Inference

Confounding Variables

- ❖ Confounders are variables that affect both the treatment and the outcome, leading to biased causal estimates.
- ❖ Solution: Use DAGs to identify and control for confounders via matching, regression, or instrumental variables.

Omitted Variable Bias

- ❖ Occurs when an important variable is left out of the analysis, leading to incorrect causal conclusions.
- ❖ **Example:** Studying the effect of education on income without considering cognitive ability.
- ❖ **Solution:** Collect better data and use proxy variables if direct measurement is unavailable.

Sample Selection Bias

- ❖ When the sample is not representative of the entire population, results may not generalize.
- ❖ **Example:** Studying the impact of an executive MBA program on salary by only looking at enrolled students (ignoring those who couldn't afford or qualify).
- ❖ **Solution:** Use techniques like Heckman correction or apply reweighting methods to adjust for selection bias.

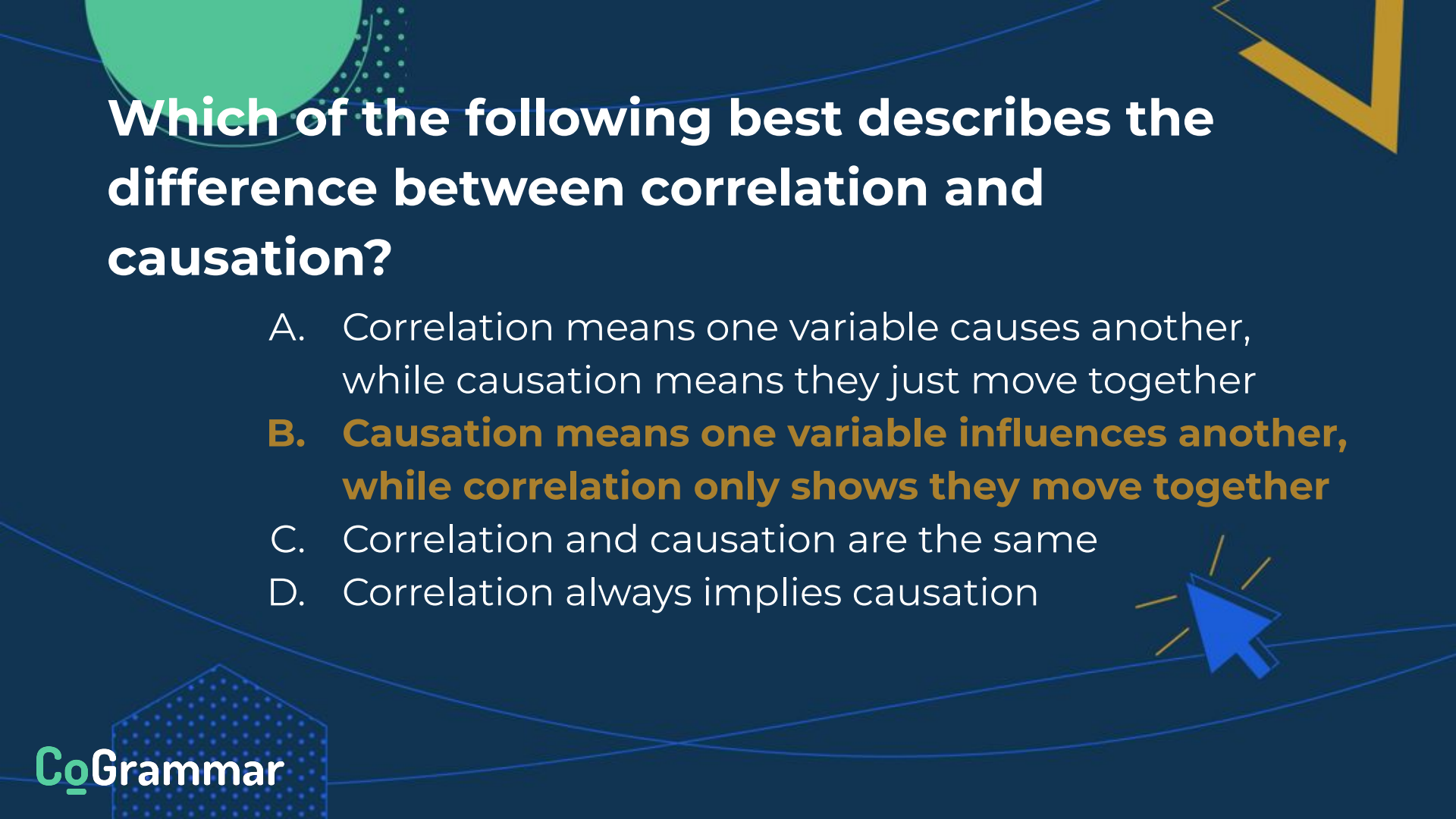
Polls Assessment



Which of the following best describes the difference between correlation and causation?

- A. Correlation means one variable causes another, while causation means they just move together
- B. Causation means one variable influences another, while correlation only shows they move together
- C. Correlation and causation are the same
- D. Correlation always implies causation





Which of the following best describes the difference between correlation and causation?

- A. Correlation means one variable causes another, while causation means they just move together
- B. Causation means one variable influences another, while correlation only shows they move together**
- C. Correlation and causation are the same
- D. Correlation always implies causation



Which of the following methods is most commonly used to determine causality?

- A. Randomized controlled trials (RCTs)
- B. Observing two variables move together
- C. Running a simple regression analysis
- D. Using only historical data to infer causality



Which of the following methods is most commonly used to determine causality?

- A. **Randomized controlled trials (RCTs)**
- B. Observing two variables move together
- C. Running a simple regression analysis
- D. Using only historical data to infer causality



Which technique helps eliminate confounding variables when estimating causal effects?

- A. Randomization
- B. Taking larger sample sizes
- C. Observing trends over time
- D. Increasing the number of variables in a model



Which technique helps eliminate confounding variables when estimating causal effects?

- A. Randomization**
- B. Taking larger sample sizes
- C. Observing trends over time
- D. Increasing the number of variables in a model

Questions and Answers



Thank you for attending



CoGrammar



Department
for Education