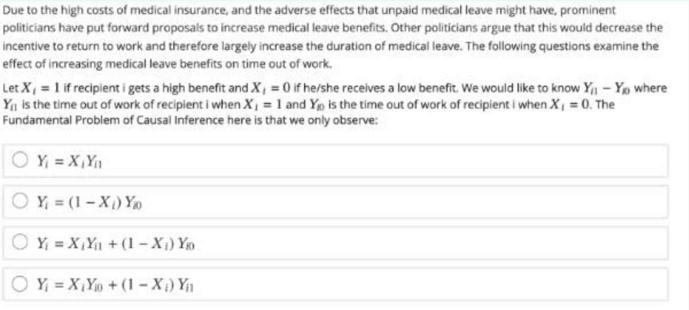
Causal Inference Equation



Due to the high costs of medical insurance, and the adverse effects that unpaid medical leave might have, prominent politicians have put forward proposals to increase medical leave benefits. Other politicians argue that this would decrease the incentive to return to work and therefore largely increase the duration of medical leave. The following questions examine the effect of increasing medical leave benefits on time out of work.

Let $X_i = 1$ if recipient i gets a high benefit and $X_i = 0$ if he/she receives a low benefit. We would like to know $Y_{i1} - Y_{i0}$ where Y_{i1} is the time out of work of recipient i when $X_i = 1$ and Y_{i0} is the time out of work of recipient i when $X_i = 0$. The Fundamental Problem of Causal Inference here is that we only observe:

○ $Y_i = X_i Y_{i1}$

○ $Y_i = (1 - X_i) Y_{i0}$

○ $Y_i = X_i Y_{i1} + (1 - X_i) Y_{i0}$

○ $Y_i = X_i Y_{i0} + (1 - X_i) Y_{i1}$

I have a problem related to Microeconomics Theory and Public Policy, as shown in the picture. Please help to solve.

To solve the problem, we need to consider the Fundamental Problem of Causal Inference: for each individual i, we can only observe one of the two potential outcomes Y_{i1} or Y_{i0}, depending on whether $X_i = 1$ or $X_i = 0$.

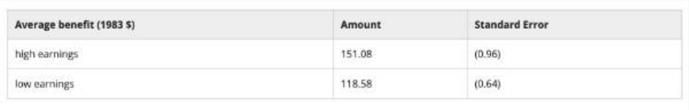
Explanation of the Options:

1. $Y_i = X_i Y_{i1}$: This only accounts for Y_{i1} when $X_i = 1$, but it does not consider Y_{i0} when $X_i = 0$. Hence, this is incomplete.
2. $Y_i = (1 - X_i) Y_{i0}$: This only accounts for Y_{i0} when $X_i = 0$, but it does not consider Y_{i1} when $X_i = 1$. Again, this is incomplete.
3. $Y_i = X_i Y_{i1} + (1 - X_i) Y_{i0}$: This equation correctly reflects the observed outcome Y_i, as it switches between Y_{i1} and Y_{i0} depending on the value of X_i. If $X_i = 1$, then $Y_i = Y_{i1}$; if $X_i = 0$, then $Y_i = Y_{i0}$.
4. $Y_i = X_i Y_{i0} + (1 - X_i) Y_{i1}$: This incorrectly assigns Y_{i0} to $X_i = 1$ and Y_{i1} to $X_i = 0$, which is inconsistent with the definition of Y_{i1} and Y_{i0}.

Correct Answer:

The correct equation is: $Y_i = X_i Y_{i1} + (1 - X_i) Y_{i0}$

This ensures that we observe the appropriate outcome depending on whether $X_i = 1$ or $X_i = 0$.



Average benefit (1983 $)	Amount	Standard Error
high earnings	151.08	(0.96)
low earnings	118.58	(0.64)

Meyer and co-authors have an administrative data set that allows them to see the maximum amount of weekly medical leave benefits that different workers are entitled to. They have aggregated the data in the picture showing the average medical leave benefits for two different groups of workers, high earners and low earners (standard errors in parentheses). You are their research assistant. You notice that the first group is granted substantially higher medical benefits than the second group. You also see that from the data set you can calculate the average weeks of sick leave that both groups take.

True or False? Using the data described above you can calculate the causal effect of the right to higher sick leave benefits on the amount of sick leave that people take. Note: numbers in parentheses are standard errors.

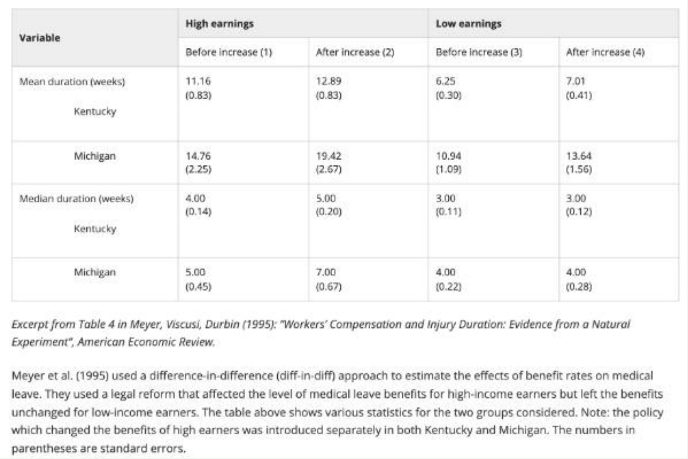
The statement is **False**.

Here's why:

The data provided shows the average medical leave benefits for high earners and low earners along with their respective standard errors. While this allows for the calculation of differences in the average benefits, it does not provide enough information to establish a causal effect of higher benefits on the amount of sick leave taken.

To calculate the causal effect, you would need a method to control for confounding variables that could influence the relationship between benefits and sick leave taken. This might include factors like worker health, job type, workplace policies, or other socioeconomic variables. Without a randomized controlled trial (RCT) or a quasi-experimental design (such as Difference-in-Differences or Instrumental Variables), the observed difference in sick leave could be due to other factors, not just the level of benefits.

Thus, using the provided data alone, you cannot definitively calculate the causal effect.



Variable	High earnings		Low earnings	
	Before increase (1)	After increase (2)	Before increase (3)	After increase (4)
Mean duration (weeks) Kentucky	11.16 (0.83)	12.89 (0.83)	6.25 (0.30)	7.01 (0.41)
Michigan	14.76 (2.25)	19.42 (2.67)	10.94 (1.09)	13.64 (1.56)
Median duration (weeks) Kentucky	4.00 (0.14)	5.00 (0.20)	3.00 (0.11)	3.00 (0.12)
Michigan	5.00 (0.45)	7.00 (0.67)	4.00 (0.22)	4.00 (0.28)

Excerpt from Table 4 in Meyer, Viscusi, Durbin (1995): "Workers' Compensation and Injury Duration: Evidence from a Natural Experiment", American Economic Review.

Meyer et al. (1995) used a difference-in-difference (diff-in-diff) approach to estimate the effects of benefit rates on medical leave. They used a legal reform that affected the level of medical leave benefits for high-income earners but left the benefits unchanged for low-income earners. The table above shows various statistics for the two groups considered. Note: the policy which changed the benefits of high earners was introduced separately in both Kentucky and Michigan. The numbers in parentheses are standard errors.

Using the information shown in the image, please help to compute the diff-in-diff estimates for “Median duration (weeks)” separately for Kentucky and Michigan.

To compute the difference-in-difference (diff-in-diff) estimates for the **median duration (weeks)** separately for Kentucky and Michigan, follow these steps:

Step 1: Understand the formula for diff-in-diff

The difference-in-difference estimate is calculated as:

$$\text{Diff-in-Diff} = (\text{Post-treatment High Earners} - \text{Pre-treatment High Earners}) - (\text{Post-treatment}$$

Where:

- **Post-treatment**: Refers to the "After increase" values.
- **Pre-treatment**: Refers to the "Before increase" values.

Step 2: Extract values from the table

Kentucky (Median duration, weeks)

- High earners:
  - Pre-treatment: 4.00
  - Post-treatment: 5.00
- Low earners:
  - Pre-treatment: 3.00
  - Post-treatment: 3.00

Michigan (Median duration, weeks)

- High earners:
  - Pre-treatment: 5.00
  - Post-treatment: 7.00
- Low earners:
  - Pre-treatment: 4.00
  - Post-treatment: 4.00

Step 3: Compute the differences

Kentucky

- High earners' change: $5.00 - 4.00 = 1.00$
- Low earners' change: $3.00 - 3.00 = 0.00$

Michigan

- High earners' change: $7.00 - 5.00 = 2.00$
- Low earners' change: $4.00 - 4.00 = 0.00$

Step 4: Compute the diff-in-diff estimates

Kentucky

$$\text{Diff-in-Diff (Kentucky)} = (1.00) - (0.00) = 1.00$$

Michigan

$$\text{Diff-in-Diff (Michigan)} = (2.00) - (0.00) = 2.00$$

Final Results:

- **Kentucky diff-in-diff estimate**: 1.00
- **Michigan diff-in-diff estimate**: 2.00

Suppose that you learn that at the same time that the reform was passed, a new law was introduced to increase the cost of medical insurance for high-income earners. True or False? This would likely violate the parallel trend assumption of the diff-in-diff approach.

True.

The parallel trend assumption of the difference-in-difference (diff-in-diff) approach requires that, in the absence of the treatment (the reform in this case), the treated and control groups would have experienced similar trends over time. If a new law was introduced simultaneously, increasing the cost of medical insurance for high-income earners, this could independently affect their behavior (such as their decision to take medical leave), leading to a deviation in trends between the high-income and low-income earners unrelated to the reform.

This would violate the parallel trend assumption because the observed differences in outcomes (e.g., the median duration of medical leave) could be attributed not just to the reform but also to the new law affecting high-income earners, confounding the causal interpretation of the diff-in-diff estimates.

From the problems and solutions above, please extract the core concepts and knowledge required to solve such problems.
Please show all the relevant fancy terms. If any, please also show all the maths, and formulas.

Core Concepts and Knowledge Required to Solve the Problems

Here are the essential concepts, mathematical tools, and key terms needed to solve these types of problems:

---

1. Difference-in-Differences (Diff-in-Diff)

- **Definition**: A quasi-experimental econometric method used to estimate causal effects by comparing changes in outcomes over time between a treatment group and a control group.
- **Key Assumption**:

  - **Parallel Trends Assumption**: In the absence of treatment, the treatment and control groups would have followed the same trajectory over time.
- **Formula**:

$$\text{Diff-in-Diff Estimate} = (\text{Post-treatment Outcome}_{\text{Treatment}} - \text{Pre-treatment Outcome}_{\text{Treatme}}$$

  - Where:
    - Treatment: Group exposed to the intervention (e.g., high-income earners).
    - Control: Group not exposed to the intervention (e.g., low-income earners).

---

2. Causal Inference

- **Fundamental Problem of Causal Inference**: For each individual, we can only observe one of the two potential outcomes (treated or untreated). Thus, causal inference relies on assumptions and models to estimate the unobserved counterfactual.
- **Key Terms**:
  - **Potential Outcomes**:
    - Y_{i1}: Outcome for individual i under treatment.
    - Y_{i0}: Outcome for individual i under no treatment.
  - **Observed Outcome**:

$$Y_i = X_i Y_{i1} + (1 - X_i) Y_{i0}$$

    Where:
    - $X_i = 1$: Treated.
    - $X_i = 0$: Untreated.
- **Causal Effect**:

$$\text{Causal Effect} = Y_{i1} - Y_{i0}$$

  - Unobservable for a single individual due to the missing counterfactual.

---

3. Statistical Terms

- **Mean and Median**:
  - Mean: Average value, calculated as:

$$\text{Mean} = \frac{\sum_{i=1}^{N} Y_i}{N}$$

  - Median: The middle value in a sorted dataset.
- **Standard Error (SE)**:
  - Measures the precision of an estimated statistic (e.g., the mean or median).
  - Formula:

$$\text{SE} = \frac{\sigma}{\sqrt{n}}$$

    Where:

    - σ: Standard deviation.
    - n: Sample size.
- **Confidence Intervals**:
  - Represent the range within which the true parameter lies with a certain probability.
  - Formula:

$$\text{CI} = \hat{\theta} \pm z \cdot \text{SE}$$

    Where:
    - $\hat{\theta}$: Estimated parameter (e.g., mean or median).
    - z: Critical value from the standard normal distribution.

---

4. Bias in Causal Estimates

- **Confounding**:
  - Occurs when an extraneous variable affects both the treatment and the outcome, violating causal assumptions.
  - Example: Introduction of a new law (e.g., higher insurance costs for high earners) could confound the effect of the medical leave reform.
- **Violation of Parallel Trends**:
  - Happens if the treatment and control groups have different trends due to factors unrelated to the intervention.

---

5. Econometrics and Experimental Design

- **Quasi-Experiments**:
  - Research designs that exploit naturally occurring treatments or interventions.
  - Example: Policy changes affecting specific groups, as in the medical leave reform.
- **Natural Experiments**:
  - A type of quasi-experiment where the assignment of treatment approximates randomization.

---

6. Hypothesis Testing

- **Null Hypothesis**:
  - Assumes no causal effect of the treatment.
- **t-test**:
  - Used to compare means or differences between groups.
  - Formula:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

    Where:
    - $\bar{X}_1, \bar{X}_2$: Sample means of groups.
    - s_1^2, s_2^2: Sample variances.
    - n_1, n_2: Sample sizes.

---

7. Real-World Application

- **Policy Analysis**:
  - Example: Meyer et al.'s study of workers' compensation and injury duration using diff-in-diff.
- **Statistical Tools**:
  - Use data on pre- and post-treatment outcomes for both treatment and control groups.
  - Compute changes and compare trends using diff-in-diff.

Summary of Key Equations and Terms:

1. **Observed Outcome**:

$$Y_i = X_i Y_{i1} + (1 - X_i) Y_{i0}$$

2. **Causal Effect**:

$$Y_{i1} - Y_{i0}$$

3. **Diff-in-Diff Estimate**:

$$\text{Diff-in-Diff} = (\Delta_{\text{Treatment}}) - (\Delta_{\text{Control}})$$

4. **Standard Error**:

$$\text{SE} = \frac{\sigma}{\sqrt{n}}$$

By mastering these concepts and formulas, you will have the foundational tools to solve problems related to causal inference and policy evaluation.

To reinforce the understanding, please generate some similar problems and show the solutions.

Here are some similar problems to reinforce understanding of the concepts and formulas:

Problem 1: Diff-in-Diff with Unemployment Benefits

Scenario: A government introduced a policy in one state (treated state) that increased unemployment benefits, while another state (control state) remained unchanged. The table below shows the average duration of unemployment (in weeks) before and after the policy implementation.

Group	Before Policy (Mean)	After Policy (Mean)
Treated State	10 weeks	15 weeks
Control State	8 weeks	9 weeks

Question: Calculate the difference-in-difference (diff-in-diff) estimate of the policy effect on unemployment duration.

Solution:

1. Compute the changes in unemployment duration for both groups:
    - Treated state: $15 - 10 = 5$
    - Control state: $9 - 8 = 1$
2. Compute the diff-in-diff estimate:

$$\text{Diff-in-Diff} = (5) - (1) = 4$$

Answer: The diff-in-diff estimate is **4 weeks**.

Problem 2: Violation of Parallel Trends Assumption

Scenario: In the previous problem, suppose a new law was introduced in the treated state at the same time as the policy, requiring companies to offer higher wages to unemployed workers during job training.

Question: Would this violate the parallel trends assumption? Why or why not?

Solution:

Yes, this would violate the parallel trends assumption. The parallel trends assumption requires that, in the absence of the treatment, the treated and control groups would have experienced the same trend in unemployment duration over time. The new law independently affects the unemployment duration in the treated state, introducing a confounding factor that biases the diff-in-diff estimate.

Answer: **Yes**, it would violate the parallel trends assumption.

Problem 3: Observed Outcomes and Potential Outcomes

Scenario: Consider an experiment where workers are either given additional paid leave ($X_i = 1$) or not ($X_i = 0$). If a worker is treated, they take 12 weeks of leave ($Y_{i1} = 12$); otherwise, they take 8 weeks of leave ($Y_{i0} = 8$).

Question: Write the observed outcome Y_i as a function of X_i, Y_{i1}, and Y_{i0}.

Solution:

The observed outcome Y_i is:

$$Y_i = X_i Y_{i1} + (1 - X_i) Y_{i0}$$

Substitute $Y_{i1} = 12$ and $Y_{i0} = 8$:

- If $X_i = 1$: $Y_i = 1 \cdot 12 + (1 - 1) \cdot 8 = 12$
- If $X_i = 0$: $Y_i = 0 \cdot 12 + (1 - 0) \cdot 8 = 8$

Answer:

$$Y_i = X_i \cdot 12 + (1 - X_i) \cdot 8$$

Problem 4: Confidence Interval for Diff-in-Diff

Scenario: For the unemployment benefits example (Problem 1), the following standard errors are provided:

- Treated state change: $\mathrm{SE}_{\text{treated}} = 0.5$
- Control state change: $\mathrm{SE}_{\text{control}} = 0.4$

Question: Compute the 95% confidence interval for the diff-in-diff estimate.

Solution:

1. Compute the standard error for the diff-in-diff estimate:

$$\mathrm{SE}_{\text{diff-in-diff}} = \sqrt{\mathrm{SE}_{\text{treated}}^2 + \mathrm{SE}_{\text{control}}^2}$$

   Substituting the values:

$$\mathrm{SE}_{\text{diff-in-diff}} = \sqrt{0.5^2 + 0.4^2} = \sqrt{0.25 + 0.16} = \sqrt{0.41} \approx 0.64$$

2. Compute the 95% confidence interval:

$$\mathrm{CI} = \text{Diff-in-Diff Estimate} \pm z \cdot \mathrm{SE}_{\text{diff-in-diff}}$$

   With $z = 1.96$, and diff-in-diff estimate $= 4$:

$$\mathrm{CI} = 4 \pm 1.96 \cdot 0.64 = 4 \pm 1.25$$

$$\mathrm{CI} = [2.75, 5.25]$$

Answer: The 95% confidence interval is **[2.75, 5.25]**.

Problem 5: Causal Effect vs. Correlation

Scenario: In a study on medical leave benefits, researchers find that workers in the treated group take more leave than the control group after the reform. However, they also notice that workers in the treated group have more generous workplace policies.

Question: Does this finding indicate a causal effect of the reform? Why or why not?

Solution:

No, this does not indicate a causal effect. The observed difference could be due to the more generous workplace policies rather than the reform itself. Without controlling for confounding variables (e.g., workplace policies), the observed difference cannot be attributed solely to the reform. This highlights the importance of the parallel trends assumption and isolating the treatment effect.

Answer: **No**, the finding does not indicate a causal effect due to potential confounding variables.

These problems cover the core concepts, including diff-in-diff, potential outcomes, observed outcomes, standard errors, confidence intervals, and the importance of the parallel trends assumption. They reinforce both the theoretical understanding and practical application of these methods.