# A tutorial on Causal Inference and its relevance in Astrophysics

Souhardya Sengupta

Department of Statistics, Harvard University

April 17, 2024

# Motivation

- A group of researchers claim that their treatment, when performed during the ages 1-2 of a child, results in much lesser chance of them developing diabetes later. The data presented by the researchers show that among the people who went through the treatment during their childhood, the incidence of diabetes is indeed very low.

# Motivation

- A group of researchers claim that their treatment, when performed during the ages 1-2 of a child, results in much lesser chance of them developing diabetes later. The data presented by the researchers show that among the people who went through the treatment during their childhood, the incidence of diabetes is indeed very low. WOULD YOU RECOMMEND THIS TREATMENT TO OTHERS?

# Motivation

- A group of researchers claim that their treatment, when performed during the ages 1-2 of a child, results in much lesser chance of them developing diabetes later. The data presented by the researchers show that among the people who went through the treatment during their childhood, the incidence of diabetes is indeed very low. WOULD YOU RECOMMEND THIS TREATMENT TO OTHERS?
- **Plot twist**: You find that most of children who get the treatment die before the age of 40.

# Motivation

- A group of researchers claim that their treatment, when performed during the ages 1-2 of a child, results in much lesser chance of them developing diabetes later. The data presented by the researchers show that among the people who went through the treatment during their childhood, the incidence of diabetes is indeed very low. WOULD YOU RECOMMEND THIS TREATMENT TO OTHERS?
- **Plot twist**: You find that most of children who get the treatment die before the age of 40.
- **Lessons**: We did observe that there is a high correlation between getting treated and having lower chances of diabetes. But that is not enough to guarantee that getting treated causes this.

# The Philosophy of Causality

- Logically, how to establish that a treatment caused an effect?

# The Philosophy of Causality

- Logically, how to establish that a treatment caused an effect?
- Whatever we observe only establishes correlation/association.

# The Philosophy of Causality

- Logically, how to establish that a treatment caused an effect?
- Whatever we observe only establishes correlation/association.
- **Key idea**: Ask the counterfactual question - What would have happened had the treatment not been administered?

# The Philosophy of Causality

- Logically, how to establish that a treatment caused an effect?
- Whatever we observe only establishes correlation/association.
- **Key idea**: Ask the counterfactual question - What would have happened had the treatment not been administered?
- Suppose you have $n$ subjects, you collect a response $Y_i$ and treatment status $T_i$, from each of the subject. We know that $Cor(Y_i, T_i)$ establish association between them.
- We need different quantities that establish causation.

# The Potential Outcomes Framework

- Assume we have $n$ subjects and for each one of them, we have a treatment status, $T_i \in \{0, 1\}$.
- We assume that there are two unobserved Potential Outcomes - $\{Y_i(0), Y_i(1)\}$ for the $i^{th}$ individual depending on whether they received the treatment or not.
- The administration of treatment picks one of the potential outcomes, which we observe, $Y_i$. We usually assume **Consistency:** $Y_i = Y_i(T_i)$.

# The Potential Outcomes Framework

- Assume we have $n$ subjects and for each one of them, we have a treatment status, $T_i \in \{0, 1\}$.
- We assume that there are two unobserved Potential Outcomes - $\{Y_i(0), Y_i(1)\}$ for the $i^{th}$ individual depending on whether they received the treatment or not.
- The administration of treatment picks one of the potential outcomes, which we observe, $Y_i$. We usually assume **Consistency:** $Y_i = Y_i(T_i)$.
- We define the Average Treatment Effect (ATE): $\tau = \mathbb{E}[Y_i(1) - Y_i(0)]$.
- The above is a causal quantity that includes an expectation over a counterfactual quantity - we do not observe both $Y_i(1)$ and $Y_i(0)$ together.
- We are interested in estimating $\tau$, testing $H_0 : \tau = 0$, etc.

# Randomized Control Trials and Estimation of ATE

# Randomized Control Trials and Estimation of ATE

- You are a researcher designing an experiment to estimate the ATE.

# Randomized Control Trials and Estimation of ATE

- You are a researcher designing an experiment to estimate the ATE.
- Ideal thing to do - For each subject, do an independent coin toss and decide whether you want to treat or not.

# Randomized Control Trials and Estimation of ATE

- You are a researcher designing an experiment to estimate the ATE.
- Ideal thing to do - For each subject, do an independent coin toss and decide whether you want to treat or not.
- Observe the treated potential outcome for the treated people and the un-treated (or control) potential outcome for the un-treated.
- Then,

$$\hat{\mathbb{E}}[Y_i(1)] = \frac{1}{\#\{i : T_i = 1\}} \sum_{i=1}^{n} T_i Y_i$$

$$\hat{\mathbb{E}}[Y_i(0)] = \frac{1}{\#\{i : T_i = 0\}} \sum_{i=1}^{n} (1 - T_i) Y_i$$

$$\implies \hat{\tau} = \frac{1}{\#\{i : T_i = 1\}} \sum_{i=1}^{n} T_i Y_i - \frac{1}{\#\{i : T_i = 0\}} \sum_{i=1}^{n} (1 - T_i) Y_i$$

# Randomized Control Trials and Estimation of ATE

- You are a researcher designing an experiment to estimate the ATE.
- Ideal thing to do - For each subject, do an independent coin toss and decide whether you want to treat or not.
- Observe the treated potential outcome for the treated people and the un-treated (or control) potential outcome for the un-treated.
- Then,

$$\hat{\mathbb{E}}[Y_i(1)] = \frac{1}{\#\{i : T_i = 1\}} \sum_{i=1}^{n} T_i Y_i$$

$$\hat{\mathbb{E}}[Y_i(0)] = \frac{1}{\#\{i : T_i = 0\}} \sum_{i=1}^{n} (1 - T_i) Y_i$$

$$\implies \hat{\tau} = \frac{1}{\#\{i : T_i = 1\}} \sum_{i=1}^{n} T_i Y_i - \frac{1}{\#\{i : T_i = 0\}} \sum_{i=1}^{n} (1 - T_i) Y_i$$

# Observational Studies : When things are not in our control

- RCTs are gold standard of scientific experiments as the scientist in completely under control of the treatment assignment.
- A more challenging, and often, more encountered situation is when the scientist observes data from an experiment done in retrospect. Such experiments are called *Observational Studies*.

# Observational Studies : When things are not in our control

- RCTs are gold standard of scientific experiments as the scientist in completely under control of the treatment assignment.
- A more challenging, and often, more encountered situation is when the scientist observes data from an experiment done in retrospect. Such experiments are called *Observational Studies*.
- Suppose this is the case and we observe iid data, $\{(Y_i, T_i)\}_{i=1}^n$. How should we estimate $\tau$?

# Observational Studies : When things are not in our control

- RCTs are gold standard of scientific experiments as the scientist in completely under control of the treatment assignment.
- A more challenging, and often, more encountered situation is when the scientist observes data from an experiment done in retrospect. Such experiments are called *Observational Studies*.
- Suppose this is the case and we observe iid data, $\{(Y_i, T_i)\}_{i=1}^n$. How should we estimate $\tau$?
- A simpler question: Is the following an unbiased estimator of $\mathbb{E}[Y_i(1)]$?

$$\hat{\tau}_1 = \frac{1}{\#\{i : T_i = 1\}} \sum_{i=1}^{n} T_i Y_i = \frac{1}{\#\{i : T_i = 1\}} \sum_{i=1}^{n} T_i Y_i(1).$$

# Observational Studies : When things are not in our control

- RCTs are gold standard of scientific experiments as the scientist in completely under control of the treatment assignment.
- A more challenging, and often, more encountered situation is when the scientist observes data from an experiment done in retrospect. Such experiments are called *Observational Studies*.
- Suppose this is the case and we observe iid data, $\{(Y_i, T_i)\}_{i=1}^n$. How should we estimate $\tau$?
- A simpler question: Is the following an unbiased estimator of $\mathbb{E}[Y_i(1)]$?

$$\hat{\tau}_1 = \frac{1}{\#\{i : T_i = 1\}} \sum_{i=1}^{n} T_i Y_i = \frac{1}{\#\{i : T_i = 1\}} \sum_{i=1}^{n} T_i Y_i(1).$$

- No! It estimates $\mathbb{E}[Y_i(1) \mid T_i = 1]$.

# Observational Studies : When things are not in our control

- RCTs are gold standard of scientific experiments as the scientist in completely under control of the treatment assignment.
- A more challenging, and often, more encountered situation is when the scientist observes data from an experiment done in retrospect. Such experiments are called *Observational Studies*.
- Suppose this is the case and we observe iid data, $\{(Y_i, T_i)\}_{i=1}^n$. How should we estimate $\tau$?
- A simpler question: Is the following an unbiased estimator of $\mathbb{E}[Y_i(1)]$?

$$\hat{\tau}_1 = \frac{1}{\#\{i : T_i = 1\}} \sum_{i=1}^{n} T_i Y_i = \frac{1}{\#\{i : T_i = 1\}} \sum_{i=1}^{n} T_i Y_i(1).$$

- No! It estimates $\mathbb{E}[Y_i(1) \mid T_i = 1]$.
- In RCT, we deliberately broke the association between the association between $\{Y_i(1), Y_i(0)\}$ and $T_i$, so that, $\mathbb{E}[Y_i(1) \mid T_i = 1] = \mathbb{E}[Y_i(1)]$.

# What went wrong?

- In general, $\hat{\tau}_1$ is only a good estimator under the assumption $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp T_i$, which in general is not the case for observational studies. What should we do now?

# What went wrong?

- In general, $\hat{\tau}_1$ is only a good estimator under the assumption $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp T_i$, which in general is not the case for observational studies. What should we do now?
- We assume that the association between $\{Y_i(1), Y_i(0)\}$ is due to a confounder - a set of covariates, $\boldsymbol{X}_i$, that influence both $\{Y_i(1), Y_i(0)\}$ and $T_i$. For example, rich people have access to better health-care facilities and hence have better chances of surviving a disease.

# What went wrong?

- In general, $\hat{\tau}_1$ is only a good estimator under the assumption $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp T_i$, which in general is not the case for observational studies. What should we do now?
- We assume that the association between $\{Y_i(1), Y_i(0)\}$ is due to a confounder - a set of covariates, $\boldsymbol{X}_i$, that influence both $\{Y_i(1), Y_i(0)\}$ and $T_i$. For example, rich people have access to better health-care facilities and hence have better chances of surviving a disease.
- We make the **Unconfoundedness** assumption, which states that $\boldsymbol{X}_i$ quantifies all systematic associations between $\{Y_i(1), Y_i(0)\}$ and $T_i$:

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp T_i \mid \boldsymbol{X}_i.$$

# What went wrong?

- In general, $\hat{\tau}_1$ is only a good estimator under the assumption $\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp T_i$, which in general is not the case for observational studies. What should we do now?
- We assume that the association between $\{Y_i(1), Y_i(0)\}$ is due to a confounder - a set of covariates, $\boldsymbol{X}_i$, that influence both $\{Y_i(1), Y_i(0)\}$ and $T_i$. For example, rich people have access to better health-care facilities and hence have better chances of surviving a disease.
- We make the **Unconfoundedness** assumption, which states that $\boldsymbol{X}_i$ quantifies all systematic associations between $\{Y_i(1), Y_i(0)\}$ and $T_i$:

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp T_i \mid \boldsymbol{X}_i.$$

- Does Unconfoundedness help us obtain an unbiased estimator of $\tau$.

# IPW estimators

- Let's revisit the problem of estimating $\mathbb{E}[Y_i(1)]$.
- Under the unconfoundedness assumption, we can define the propensity score:

$$\pi(\boldsymbol{x}) = \mathbb{P}(T_i = 1 \mid \boldsymbol{X}_i = \boldsymbol{x}).$$

# IPW estimators

- Let's revisit the problem of estimating $\mathbb{E}[Y_i(1)]$.
- Under the unconfoundedness assumption, we can define the propensity score:

$$\pi(\boldsymbol{x}) = \mathbb{P}(T_i = 1 \mid \boldsymbol{X}_i = \boldsymbol{x}).$$

- Assume **Positivity:** $0 < \pi(\boldsymbol{x}) < 1, \forall \boldsymbol{x}$.

# IPW estimators

- Let's revisit the problem of estimating $\mathbb{E}[Y_i(1)]$.
- Under the unconfoundedness assumption, we can define the propensity score:

$$\pi(\boldsymbol{x}) = \mathbb{P}(T_i = 1 \mid \boldsymbol{X}_i = \boldsymbol{x}).$$

- Assume **Positivity:** $0 < \pi(\boldsymbol{x}) < 1, \forall \boldsymbol{x}$.
- If we know $\pi(\boldsymbol{x})$, then we can define the Inverse Probability Weighted (IPW) estimator:

$$\hat{\tau}_{1,IPW} = \frac{1}{n} \sum_{i=1}^{n} \frac{T_i Y_i}{\pi(\boldsymbol{X}_i)}$$

# IPW estimators are unbiased

# IPW estimators are unbiased

- We have the following chain of equalities

$$
\begin{aligned}
&\mathbb{E}\left[\frac{T_i Y_i}{\pi(\boldsymbol{X}_i)}\right] = \mathbb{E}\left[\frac{T_i Y_i(1)}{\pi(\boldsymbol{X}_i)}\right] \\
=&\mathbb{E}\mathbb{E}\left[\frac{T_i Y_i(1)}{\pi(\boldsymbol{X}_i)} \mid \boldsymbol{X}_i\right] \\
=&\mathbb{E}\left[\frac{\mathbb{E}[Y_i(1) \mid \boldsymbol{X}_i]}{\pi(\boldsymbol{X}_i)}\mathbb{E}(T_i \mid \boldsymbol{X}_i)\right] \quad [\text{Unconfoundedness}] \\
=&\mathbb{E}\left[\frac{\mathbb{E}[Y_i(1) \mid \boldsymbol{X}_i]}{\pi(\boldsymbol{X}_i)}\pi(\boldsymbol{X}_i)\right] = \mathbb{E}[Y_i(1)],
\end{aligned}
$$

  so that, $\mathbb{E}[\hat{\tau}_{1,IPW}] = \mathbb{E}[Y_i(1)]$.

# IPW estimators are unbiased

- We have the following chain of equalities

$$\begin{aligned}
&\mathbb{E}\left[\frac{T_i Y_i}{\pi(\boldsymbol{X}_i)}\right] = \mathbb{E}\left[\frac{T_i Y_i(1)}{\pi(\boldsymbol{X}_i)}\right] \\
=&\mathbb{E}\mathbb{E}\left[\frac{T_i Y_i(1)}{\pi(\boldsymbol{X}_i)} \mid \boldsymbol{X}_i\right] \\
=&\mathbb{E}\left[\frac{\mathbb{E}[Y_i(1) \mid \boldsymbol{X}_i]}{\pi(\boldsymbol{X}_i)}\mathbb{E}(T_i \mid \boldsymbol{X}_i)\right] \quad [\text{Unconfoundedness}] \\
=&\mathbb{E}\left[\frac{\mathbb{E}[Y_i(1) \mid \boldsymbol{X}_i]}{\pi(\boldsymbol{X}_i)}\pi(\boldsymbol{X}_i)\right] = \mathbb{E}[Y_i(1)],
\end{aligned}$$

  so that, $\mathbb{E}[\hat{\tau}_{1,IPW}] = \mathbb{E}[Y_i(1)]$.
- The following is the IPW estimator of $\tau$:

$$\hat{\tau}_{IPW} = \frac{1}{n}\sum_{i=1}^{n}\frac{T_i Y_i}{\pi(\boldsymbol{X}_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-T_i)Y_i}{1-\pi(\boldsymbol{X}_i)}$$

# But we do not know $\pi(\boldsymbol{x})$!

- In general we won't know $\pi(\boldsymbol{x})$.

# But we do not know $\pi(\boldsymbol{x})$!

- In general we won't know $\pi(\boldsymbol{x})$. We might try to use an estimate $\hat{\pi}(\boldsymbol{x})$, but how good the resulting estimator,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{T_i Y_i}{\hat{\pi}(\boldsymbol{X}_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-T_i)Y_i}{1-\hat{\pi}(\boldsymbol{X}_i)},$$

  is, would depend on many strong assumptions, which we have no way of verifying!
- What's a way out?

# But we do not know $\pi(\boldsymbol{x})$!

- In general we won't know $\pi(\boldsymbol{x})$. We might try to use an estimate $\hat{\pi}(\boldsymbol{x})$, but how good the resulting estimator,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{T_i Y_i}{\hat{\pi}(\boldsymbol{X}_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-T_i)Y_i}{1-\hat{\pi}(\boldsymbol{X}_i)},$$

  is, would depend on many strong assumptions, which we have no way of verifying!
- What's a way out?
- Propensity scores have this Balancing Property:

$$\mathbb{E}\left[\frac{T_i f(\boldsymbol{X}_i)}{\pi(\boldsymbol{X}_i)}\right] = \mathbb{E}[f(\boldsymbol{X}_i)], \forall \text{ bounded } f.$$

# But we do not know $\pi(\boldsymbol{x})$!

- In general we won't know $\pi(\boldsymbol{x})$. We might try to use an estimate $\hat{\pi}(\boldsymbol{x})$, but how good the resulting estimator,

$$\frac{1}{n}\sum_{i=1}^{n}\frac{T_i Y_i}{\hat{\pi}(\boldsymbol{X}_i)} - \frac{1}{n}\sum_{i=1}^{n}\frac{(1-T_i)Y_i}{1-\hat{\pi}(\boldsymbol{X}_i)},$$

  is, would depend on many strong assumptions, which we have no way of verifying!
- What's a way out?
- Propensity scores have this Balancing Property:

$$\mathbb{E}\left[\frac{T_i f(\boldsymbol{X}_i)}{\pi(\boldsymbol{X}_i)}\right] = \mathbb{E}[f(\boldsymbol{X}_i)], \forall \text{ bounded } f.$$

- Furthermore, IPW estimators belong to a class of weighing estimators: $\sum w_i Y_i T_i$, with $w_i = 1/(n\pi(\boldsymbol{X}_i))$.

# Weighing Estimators

- When the exact propensity score is unknown, people try to find weights that directly try to achieve the balancing property.

# Weighing Estimators

- When the exact propensity score is unknown, people try to find weights that directly try to achieve the balancing property.
- That is, for a class of functions, $\mathcal{M}$, they choose weights $\hat{w}_i$, such that,

$$\sup_{f \in \mathcal{M}} \left| \sum_{i=1}^{n} T_i \hat{w}_i f(\boldsymbol{X}_i) - \frac{1}{n} \sum_{i=1}^{n} f(\boldsymbol{X}_i) \right| < \delta,$$

and then use

$$\hat{\tau}_{1,\hat{\boldsymbol{w}}} = \sum_{i=1}^{n} \hat{w}_i Y_i T_i,$$

as an estimator for $\mathbb{E}[Y_i(1)]$.

# Weighing Estimators

- When the exact propensity score is unknown, people try to find weights that directly try to achieve the balancing property.
- That is, for a class of functions, $\mathcal{M}$, they choose weights $\hat{w}_i$, such that,

$$\sup_{f \in \mathcal{M}} \left| \sum_{i=1}^{n} T_i \hat{w}_i f(\boldsymbol{X}_i) - \frac{1}{n} \sum_{i=1}^{n} f(\boldsymbol{X}_i) \right| < \delta,$$

  and then use

$$\hat{\tau}_{1,\hat{\boldsymbol{w}}} = \sum_{i=1}^{n} \hat{w}_i Y_i T_i,$$

  as an estimator for $\mathbb{E}[Y_i(1)]$.
- In fact, if $m_1(\boldsymbol{x}) = \mathbb{E}[Y_i(1) \mid \boldsymbol{X}_i = \boldsymbol{x}] \in \mathcal{M}$, the bias in the above weighing estimator is controlled within $\delta$.

# Weighing Estimators

- When the exact propensity score is unknown, people try to find weights that directly try to achieve the balancing property.
- That is, for a class of functions, $\mathcal{M}$, they choose weights $\hat{w}_i$, such that,

$$\sup_{f \in \mathcal{M}} \left| \sum_{i=1}^{n} T_i \hat{w}_i f(\boldsymbol{X}_i) - \frac{1}{n} \sum_{i=1}^{n} f(\boldsymbol{X}_i) \right| < \delta,$$

  and then use

$$\hat{\tau}_{1,\hat{\boldsymbol{w}}} = \sum_{i=1}^{n} \hat{w}_i Y_i T_i,$$

  as an estimator for $\mathbb{E}[Y_i(1)]$.
- In fact, if $m_1(\boldsymbol{x}) = \mathbb{E}[Y_i(1) \mid \boldsymbol{X}_i = \boldsymbol{x}] \in \mathcal{M}$, the bias in the above weighing estimator is controlled within $\delta$.
- In general, people try to *balance* the first few moments by taking, $f(x) = x, x^2$, and so on.

# Outcome Regression estimators

# Outcome Regression estimators

- Another class of approach stems if we have access to the Outcome Regression functions,

$$m_1(\boldsymbol{x}) = \mathbb{E}[Y_i(1) \mid \boldsymbol{X}_i = \boldsymbol{x}]$$
$$\text{and } m_0(\boldsymbol{x}) = \mathbb{E}[Y_i(0) \mid \boldsymbol{X}_i = \boldsymbol{x}].$$

- Then an estimate of $\tau$ is given by,

$$\hat{\tau}_{OR} = \frac{1}{n}\sum_{i=1}^{n} m_1(\boldsymbol{X}_i) - \frac{1}{n}\sum_{i=1}^{n} m_0(\boldsymbol{X}_i).$$

# Outcome Regression estimators

- Another class of approach stems if we have access to the Outcome Regression functions,

$$m_1(\boldsymbol{x}) = \mathbb{E}[Y_i(1) \mid \boldsymbol{X}_i = \boldsymbol{x}]$$
$$\text{and } m_0(\boldsymbol{x}) = \mathbb{E}[Y_i(0) \mid \boldsymbol{X}_i = \boldsymbol{x}].$$

- Then an estimate of $\tau$ is given by,

$$\hat{\tau}_{OR} = \frac{1}{n}\sum_{i=1}^{n} m_1(\boldsymbol{X}_i) - \frac{1}{n}\sum_{i=1}^{n} m_0(\boldsymbol{X}_i).$$

- In general, we can use the $Y_i$'s to obtain estimates: $\hat{m}_0(\boldsymbol{x})$ and $\hat{m}_1(\boldsymbol{x})$, by training on the control and treatment groups - there can be several strategies here.
- Then, an estimate of $\mathbb{E}[Y_i(0) \mid T_i = 1]$ is given by,

$$\frac{1}{\#\{i : T_i = 1\}}\sum_{i=1}^{n} T_i \hat{m}_0(\boldsymbol{X}_i).$$

- There can be several strategies of estimating $\tau$ with these regression estimators:

$$\hat{\tau}_{reg} = \frac{1}{n}\sum_{i=1}^{n}(\hat{m}_1(\boldsymbol{X}_i) - \hat{m}_0(\boldsymbol{X}_i))$$

$$\hat{\tau}_{reg-imp} = \frac{1}{n}\sum_{i=1}^{n}\{T_i(Y_i - \hat{m}_0(\boldsymbol{X}_i) + (1 - T_i)(\hat{m}_1(\boldsymbol{X}_i) - Y_i)\}$$

- One can use a variety of machine learning algorithms for training these models.
- Can use for estimation of Conditional Average Treatment Effect (CATE):

$$\tau(\boldsymbol{x}) = \mathbb{E}[Y_i(1) - Y_i(0) \mid \boldsymbol{X}_i = \boldsymbol{x}].$$

# Sensitivity Analysis

- The full set of confounders might not be observed - this is generally the case! Thus unconfoundess no longer holds.

# Sensitivity Analysis

- The full set of confounders might not be observed - this is generally the case! Thus unconfoundess no longer holds.
- We assume that unconfoundedness holds if we include the unobserved confounder ($\boldsymbol{U}_i$):

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp T_i \mid (\boldsymbol{X}_i, \boldsymbol{U}_i)$$

- Under the above assumptions, our inference methods are not valid. There's not much we can do since $\boldsymbol{U}_i$'s are not observed.

# Sensitivity Analysis

- The full set of confounders might not be observed - this is generally the case! Thus unconfoundess no longer holds.
- We assume that unconfoundedness holds if we include the unobserved confounder ($\boldsymbol{U}_i$):

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp T_i \mid (\boldsymbol{X}_i, \boldsymbol{U}_i)$$

- Under the above assumptions, our inference methods are not valid. There's not much we can do since $\boldsymbol{U}_i$'s are not observed.
- However, we can try to assess what effect the unobserved confounder has on our method - Sensitivity Analysis.

# Sensitivity Analysis

- The full set of confounders might not be observed - this is generally the case! Thus unconfoundess no longer holds.
- We assume that unconfoundedness holds if we include the unobserved confounder ($\boldsymbol{U}_i$):

$$\{Y_i(1), Y_i(0)\} \perp\!\!\!\perp T_i \mid (\boldsymbol{X}_i, \boldsymbol{U}_i)$$

- Under the above assumptions, our inference methods are not valid. There's not much we can do since $\boldsymbol{U}_i$'s are not observed.
- However, we can try to assess what effect the unobserved confounder has on our method - Sensitivity Analysis.
- **A historical account:** Fisher once argued that the association between smoking and lung cancer is due to a common gene. Cornfield argued that if Fisher were right, then this gene should have had a very high association with propensity to smoke which is unrealistic.
- This sort of outlines the basis of argument for sensitivity analysis.

- Assume the following setup where we have, binary treatment, binary outcomes and a binary confounder:

- Assume the following setup where we have, binary treatment, binary outcomes and a binary confounder:
  - Confounder $U \sim Ber(\pi)$
  - Treatment assignment: $\text{logit}\left(\mathbb{P}(Z = 1 \mid u)\right) = \gamma + \alpha u$.
  - Outcome model: $\text{logit}\left(\mathbb{P}(Y(z) = 1 \mid u)\right) = \beta_z + \delta_z u$
- The *sensitivity parameters*, $(\pi, \alpha, \delta_1, \delta_0)$ are unobserved.
- But for a fixed value of the sensitivity parameters, can obtain treatment effects.

- Assume the following setup where we have, binary treatment, binary outcomes and a binary confounder:
  - Confounder $U \sim Ber(\pi)$
  - Treatment assignment: $\text{logit}\left(\mathbb{P}(Z = 1 \mid u)\right) = \gamma + \alpha u$.
  - Outcome model: $\text{logit}\left(\mathbb{P}(Y(z) = 1 \mid u)\right) = \beta_z + \delta_z u$
- The *sensitivity parameters*, $(\pi, \alpha, \delta_1, \delta_0)$ are unobserved.
- But for a fixed value of the sensitivity parameters, can obtain treatment effects.
- Idea: Vary $\pi, \alpha, \delta_1, \delta_0$ over a grid of possible values and see how much they need to be varied for our inference to change significantly.

- Assume the following setup where we have, binary treatment, binary outcomes and a binary confounder:
  - Confounder $U \sim Ber(\pi)$
  - Treatment assignment: $\text{logit}\left(\mathbb{P}(Z = 1 \mid u)\right) = \gamma + \alpha u$.
  - Outcome model: $\text{logit}\left(\mathbb{P}(Y(z) = 1 \mid u)\right) = \beta_z + \delta_z u$
- The *sensitivity parameters*, $(\pi, \alpha, \delta_1, \delta_0)$ are unobserved.
- But for a fixed value of the sensitivity parameters, can obtain treatment effects.
- Idea: Vary $\pi, \alpha, \delta_1, \delta_0$ over a grid of possible values and see how much they need to be varied for our inference to change significantly.
- If there needs to be a drastic change in the sensitivity parameters to bring about this change in inference, our conclusions are pretty robust to the presence of un-measured confounders.

# Causal Estimands and Their identifiability

- In general, because causal estimands are stated in terms of the distribution of counter-factual quantities, not all of them can be *estimated*.

# Causal Estimands and Their identifiability

- In general, because causal estimands are stated in terms of the distribution of counter-factual quantities, not all of them can be *estimated*.
- We call a causal estimand identifiable, if one can find an unbiased estimator for them based on the *observed* data.

# Causal Estimands and Their identifiability

- In general, because causal estimands are stated in terms of the distribution of counter-factual quantities, not all of them can be *estimated*.
- We call a causal estimand identifiable, if one can find an unbiased estimator for them based on the *observed* data.
- For example, the following are some other identifiable causal quantities: Average Treatment Effect on the Treated (ATT): $\tau_{ATT} = \mathbb{E}[Y_i(1) - Y_i(0) \mid T_i = 1]$, Average Treatment Effect on the Control (ATC): $\tau_{ATC} = \mathbb{E}[Y_i(1) - Y_i(0) \mid T_i = 0]$, etc.
- An example of a quantity that is not identifiable: $\mathbb{E}[Y_i(1)Y_i(0)]$.

# Causal Estimands and Their identifiability

- In general, because causal estimands are stated in terms of the distribution of counter-factual quantities, not all of them can be *estimated*.
- We call a causal estimand identifiable, if one can find an unbiased estimator for them based on the *observed* data.
- For example, the following are some other identifiable causal quantities: Average Treatment Effect on the Treated (ATT): $\tau_{ATT} = \mathbb{E}[Y_i(1) - Y_i(0) \mid T_i = 1]$, Average Treatment Effect on the Control (ATC): $\tau_{ATC} = \mathbb{E}[Y_i(1) - Y_i(0) \mid T_i = 0]$, etc.
- An example of a quantity that is not identifiable: $\mathbb{E}[Y_i(1)Y_i(0)]$.
- Depending on the situation at hand, our causal estimand might be quite complicated and we impose a variety of assumptions on the potential outcomes to make the estimand identifiable (and hope these assumptions are feasible!), that is, writing it in terms of observable quantities.

# Difference-in-Difference Estimators

- An example where identification is tricky.

# Difference-in-Difference Estimators

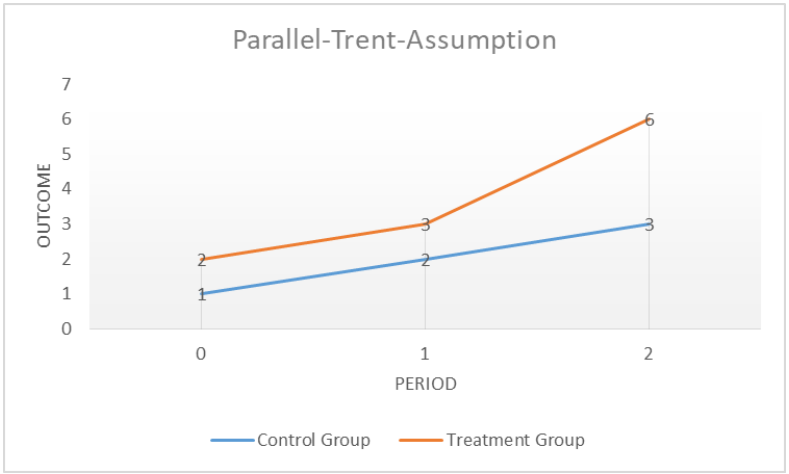
- An example where identification is tricky.
- Problem: We have two groups of people,
  - Interested in treatment effect on the first group.
  - Two time periods: pre-intervention ($t_1$) and post-intervention ($t_2$). Only first group is treated.
  - Estimand of interest: $\mathbb{E}[Y_{1i}(1, t_2) - Y_{1i}(0, t_2)]$.
  - Obstacle: Observe only $Y_{1i}(0, t_1) = Y_{1i}(t_1)$ and $Y_{1i}(1, t_2) = Y_{1i}(t_2)$.

- Use the Parallel Trends assumption:

$$\mathbb{E}[\underbrace{Y_{2i}(0,t_2)}_{=Y_{2i}(t_2)} - Y_{1i}(0,t_2)] = \mathbb{E}[\underbrace{Y_{2i}(0,t_1)}_{=Y_{2i}(t_1)} - Y_{1i}(0,t_1)]$$

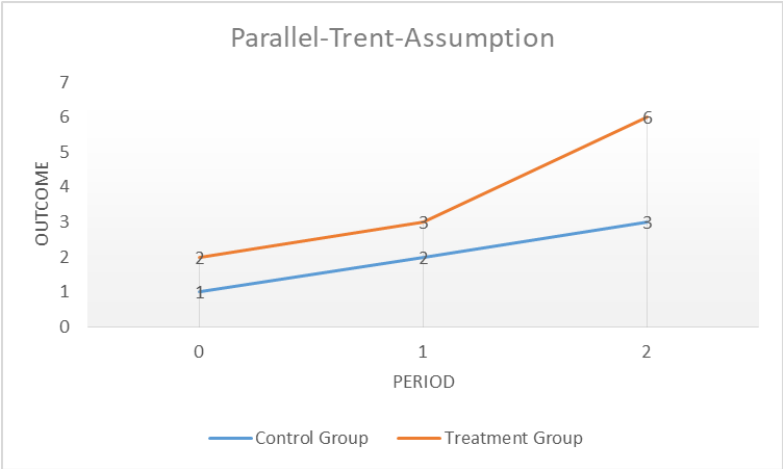
- Use the Parallel Trends assumption:

$$\mathbb{E}[\underbrace{Y_{2i}(0,t_2)}_{=Y_{2i}(t_2)} - Y_{1i}(0,t_2)] = \mathbb{E}[\underbrace{Y_{2i}(0,t_1)}_{=Y_{2i}(t_1)} - Y_{1i}(0,t_1)]$$



- Use the Parallel Trends assumption:

$$\mathbb{E}[\underbrace{Y_{2i}(0,t_2)}_{=Y_{2i}(t_2)} - Y_{1i}(0,t_2)] = \mathbb{E}[\underbrace{Y_{2i}(0,t_1)}_{=Y_{2i}(t_1)} - Y_{1i}(0,t_1)]$$




- Then,

$$\begin{aligned}
&\mathbb{E}[Y_{1i}(1,t_2) - Y_{1i}(0,t_2)] \\
=&\mathbb{E}[Y_{1i}(1,t_2) - Y_{2i}(0,t_2)] - \mathbb{E}[Y_{1i}(0,t_2) - Y_{2i}(0,t_2)] \\
=&\mathbb{E}[Y_{1i}(1,t_2) - Y_{2i}(0,t_2)] - \mathbb{E}[Y_{1i}(0,t_1) - Y_{2i}(0,t_1)] \\
=&\mathbb{E}[Y_{1i}(t_2) - Y_{2i}(t_2) - Y_{1i}(t_1) + Y_{2i}(t_1)].
\end{aligned}$$

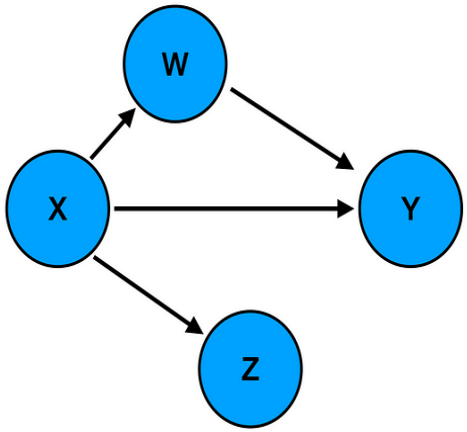
# Causal Discovery

- Usually, in a scientific experiment, we have a system of variables with all kinds of complex interactions.
- Often, it is of interest to identify which set of variables case an effect on others.
- A great way of representing such relations is via a Directed Acyclic Graph (DAG):

# Causal Discovery

- Usually, in a scientific experiment, we have a system of variables with all kinds of complex interactions.
- Often, it is of interest to identify which set of variables case an effect on others.
- A great way of representing such relations is via a Directed Acyclic Graph (DAG):

**Directed Acyclic Graphs (DAGs)**




**Structural Equation Models (SEMs)**

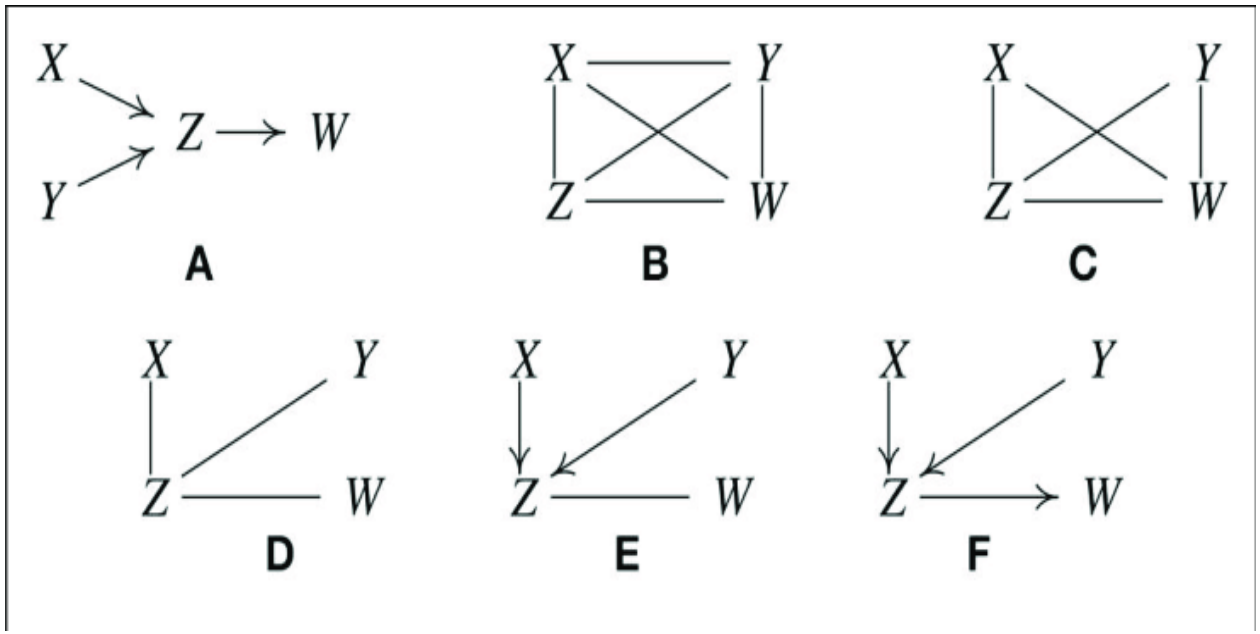
$$W := f_1(X)$$

$$Z := f_2(X)$$

$$Y := f_3(X, W)$$

# Structural Causal Modelling (SCM)

- SCM refers to the task of recovering this simple structure.
- Peter and Clark (PC) algorithm:

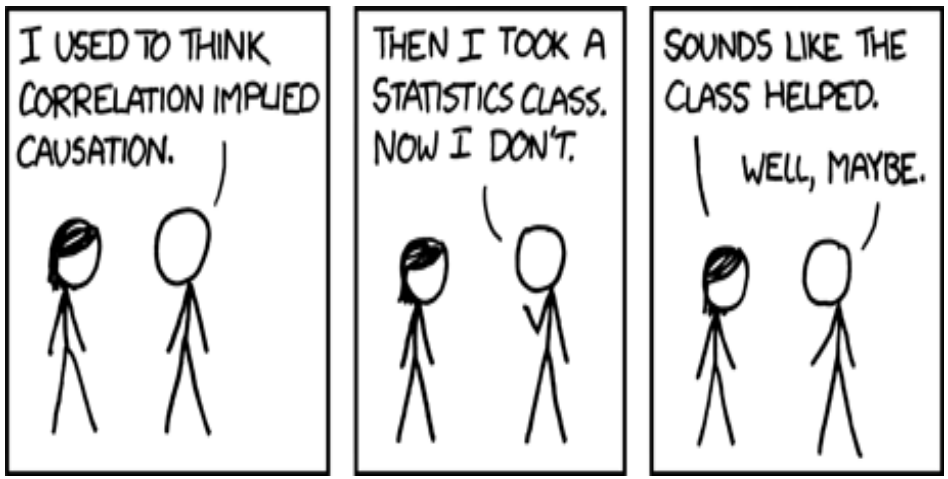


# Conclusion

- In this talk we explored some basic concepts of statistical causal thinking.
- Very immediate relevance with many scientific questions - people develop various frameworks to accommodate these settings.
- Closely related - Missing Data Analysis.
- A very recent development I am very excited about - Use of Conformal Inference in Causal Inference - Makes inference on Individual Treatment Effects (ITE)'s possible!

# Conclusion

- In this talk we explored some basic concepts of statistical causal thinking.
- Very immediate relevance with many scientific questions - people develop various frameworks to accommodate these settings.
- Closely related - Missing Data Analysis.
- A very recent development I am very excited about - Use of Conformal Inference in Causal Inference - Makes inference on Individual Treatment Effects (ITE)'s possible!




**Thank You!**