Causal Inference: Counterfactuals

Bruce Rushing

University of California, Irvine

3 May 2021

Table of Contents

Table of Contents

What is a counterfactual?

Suppose I am traveling home and come to a fork in the road one night, with one direction leading to the freeway ($X = 1$) and the other to a surface street, Sepulveda Boulevard ($X = 0$). I take Seuplveda but end up home late by an hour $Y = 1$. What would have happened if I had taken the freeway?

What is a counterfactual?

Suppose I am traveling home and come to a fork in the road one night, with one direction leading to the freeway ($X = 1$) and the other to a surface street, Sepulveda Boulevard ($X = 0$). I take Seuplveda but end up home late by an hour $Y = 1$. What would have happened if I had taken the freeway?

- The subjunctive conditional, "if I had taken the freeway, I would have arrived home sooner" is called a *counterfactual* because the antecedent is unrealized or contrary-to-fact.

What is a counterfactual?

Suppose I am traveling home and come to a fork in the road one night, with one direction leading to the freeway ($X = 1$) and the other to a surface street, Sepulveda Boulevard ($X = 0$). I take Seuplveda but end up home late by an hour $Y = 1$. What would have happened if I had taken the freeway?

- The subjunctive conditional, "if I had taken the freeway, I would have arrived home sooner" is called a *counterfactual* because the antecedent is unrealized or contrary-to-fact.
- These are useful because they allow us to compare outcomes in identical conditions, differing only in the antecedent.

What is a counterfactual?

Suppose I am traveling home and come to a fork in the road one night, with one direction leading to the freeway ($X = 1$) and the other to a surface street, Sepulveda Boulevard ($X = 0$). I take Seuplveda but end up home late by an hour $Y = 1$. What would have happened if I had taken the freeway?

- The subjunctive conditional, "if I had taken the freeway, I would have arrived home sooner" is called a *counterfactual* because the antecedent is unrealized or contrary-to-fact.
- These are useful because they allow us to compare outcomes in identical conditions, differing only in the antecedent.
- Knowing the outcome of the decision is important because my prior conditional probability of the consequent given the antecedent ($\Pr(Y = y|X = 1)$) might be different from posterior conditional probability given my actual decision ($X = 0 \land Y = y'$).

do-expressions Not Sufficient

Suppose we wanted to estimate the effect of taking the freeway using *do*-expressions.

do-expressions Not Sufficient

Suppose we wanted to estimate the effect of taking the freeway using *do*-expressions.

- We would write this as:

$$\mathbb{E}[\textit{driving time} \mid do(\textit{freeway}), \textit{driving time} = 1 \textit{ hour}]$$

do-expressions Not Sufficient

Suppose we wanted to estimate the effect of taking the freeway using *do*-expressions.

- We would write this as:

$$\mathbb{E}[\textit{driving time} | \textit{do}(\textit{freeway}), \textit{driving time} = 1 \textit{ hour}]$$

- But if the *driving time* on both sides of the | are the same variable, we would not find what we want.

do-expressions Not Sufficient

Suppose we wanted to estimate the effect of taking the freeway using *do*-expressions.

- We would write this as:

$$\mathbb{E}[\textit{driving time} \mid \textit{do}(\textit{freeway}), \textit{driving time} = 1 \textit{ hour}]$$

- But if the *driving time* on both sides of the | are the same variable, we would not find what we want.
- We need to distinguish:
  1. Actual driving time.

do-expressions Not Sufficient

Suppose we wanted to estimate the effect of taking the freeway using *do*-expressions.

- We would write this as:

$$\mathbb{E}[\textit{driving time} | do(\textit{freeway}), \textit{driving time} = 1 \textit{ hour}]$$

- But if the *driving time* on both sides of the | are the same variable, we would not find what we want.
- We need to distinguish:
  1. Actual driving time.
  2. Hypothetical driving time under freeway conditions when actual surface driving time is known to be 1 hour.

do-expressions Not Sufficient

Suppose we wanted to estimate the effect of taking the freeway using *do*-expressions.

- We would write this as:

$$\mathbb{E}[\mathit{driving\ time}|\mathit{do}(\mathit{freeway}), \mathit{driving\ time} = 1\ \mathit{hour}]$$

- But if the *driving time* on both sides of the | are the same variable, we would not find what we want.
- We need to distinguish:
  1. Actual driving time.
  2. Hypothetical driving time under freeway conditions when actual surface driving time is known to be 1 hour.
- The *do*-operator only gives us $\Pr(\mathit{driving\ time}|\mathit{do}(\mathit{freeway}))$ and $\Pr(\mathit{driving\ time}|\mathit{do}(\mathit{Suplveda}))$.

Getting Around the Impasse

The way around is to discriminate the consequent variables based on their antecedent variables:

Getting Around the Impasse

The way around is to discriminate the consequent variables based on their antecedent variables:

- Recall that $X = 0$ means we took Sepulveda Blvd and $X = 1$ means we took the freeway.

Getting Around the Impasse

The way around is to discriminate the consequent variables based on their antecedent variables:

- Recall that $X = 0$ means we took Sepulveda Blvd and $X = 1$ means we took the freeway.
- Denote the value of our driving time Y when we we take Sepulveda as $Y_{X=0}$ and when we take the freeway as $Y_{X=1}$. Then what we want to estimate is:
$$\mathbb{E}[Y_{X=1}|X = 0, Y = 1]$$

Getting Around the Impasse

The way around is to discriminate the consequent variables based on their antecedent variables:

- Recall that $X = 0$ means we took Sepulveda Blvd and $X = 1$ means we took the freeway.
- Denote the value of our driving time Y when we we take Sepulveda as $Y_{X=0}$ and when we take the freeway as $Y_{X=1}$. Then what we want to estimate is:
$$\mathbb{E}[Y_{X=1}|X = 0, Y = 1]$$
- Another way to think of $Y_{X=1}$ is the value of Y conditional on the intervention of $do(X = 1)$. So $\mathbb{E}[Y|do(X = 1)] = \mathbb{E}[Y_{X=1}]$.

Getting Around the Impasse

The way around is to discriminate the consequent variables based on their antecedent variables:

- Recall that $X = 0$ means we took Sepulveda Blvd and $X = 1$ means we took the freeway.
- Denote the value of our driving time Y when we we take Sepulveda as $Y_{X=0}$ and when we take the freeway as $Y_{X=1}$. Then what we want to estimate is:

$$\mathbb{E}[Y_{X=1}|X = 0, Y = 1]$$

- Another way to think of $Y_{X=1}$ is the value of Y conditional on the intervention of $do(X = 1)$. So $\mathbb{E}[Y|do(X = 1)] = \mathbb{E}[Y_{X=1}]$.
- Notation: we also write $Y_{X=x}$ as Y_x.

Getting Around the Impasse

- The difference between the counterfactual case and intervention case is that the counterfactual involves expressions that apply to "different worlds."

Getting Around the Impasse

- The difference between the counterfactual case and intervention case is that the counterfactual involves expressions that apply to "different worlds."
- $\mathbb{E}[Y_{X=1}|X=0, Y=1]$ involves the expression $X=0$, which by definition is a different world from $Y_{X=1}$.

Getting Around the Impasse

- The difference between the counterfactual case and intervention case is that the counterfactual involves expressions that apply to "different worlds."
- $\mathbb{E}[Y_{X=1}|X = 0, Y = 1]$ involves the expression $X = 0$, which by definition is a different world from $Y_{X=1}$.
- Essentially, we ask what the drive time would be in a world where $do(X = 1)$ given that in our actual world, $X = 0$ and $Y = 1$.

Getting Around the Impasse

- The difference between the counterfactual case and intervention case is that the counterfactual involves expressions that apply to "different worlds."
- $\mathbb{E}[Y_{X=1}|X=0, Y=1]$ involves the expression $X=0$, which by definition is a different world from $Y_{X=1}$.
- Essentially, we ask what the drive time would be in a world where $do(X=1)$ given that in our actual world, $X=0$ and $Y=1$.
- But in the case of $\mathbb{E}[Y|do(X=x)]$, we estimate the drive time across in a specific world where $X=x$, irrespective to any other world.

Getting Around the Impasse

Unfortunately, we cannot reduce $\mathbb{E}[Y_{X=1}|X = 0, Y = 1]$ to a coherent *do*-expression.

Getting Around the Impasse

Unfortunately, we cannot reduce $\mathbb{E}[Y_{X=1}|X=0, Y=1]$ to a coherent *do*-expression.

- Note that

$$\mathbb{E}[Y_{X=1}|X=0, Y=1] \neq \mathbb{E}[Y|do(X=1)] \neq \mathbb{E}[Y|do(X=0)]$$

  Why?

Getting Around the Impasse

Unfortunately, we cannot reduce $\mathbb{E}[Y_{X=1}|X=0, Y=1]$ to a coherent *do*-expression.

- Note that

$$\mathbb{E}[Y_{X=1}|X=0, Y=1] \neq \mathbb{E}[Y|do(X=1)] \neq \mathbb{E}[Y|do(X=0)]$$

  Why? $X=0$ and $do(X=1)$ are incompatible.
- We could try to do a RCT involving other drivers. Why would this not work?

Getting Around the Impasse

Unfortunately, we cannot reduce $\mathbb{E}[Y_{X=1}|X=0, Y=1]$ to a coherent *do*-expression.

- Note that

$$\mathbb{E}[Y_{X=1}|X=0, Y=1] \neq \mathbb{E}[Y|do(X=1)] \neq \mathbb{E}[Y|do(X=0)]$$

  Why? $X=0$ and $do(X=1)$ are incompatible.
- We could try to do a RCT involving other drivers. Why would this not work?
  1. The conditions are not replicated between drivers and freeway and side road conditions.

Getting Around the Impasse

Unfortunately, we cannot reduce $\mathbb{E}[Y_{X=1}|X=0, Y=1]$ to a coherent *do*-expression.

- Note that

$$\mathbb{E}[Y_{X=1}|X=0, Y=1] \neq \mathbb{E}[Y|do(X=1)] \neq \mathbb{E}[Y|do(X=0)]$$

  Why? $X=0$ and $do(X=1)$ are incompatible.
- We could try to do a RCT involving other drivers. Why would this not work?
  1. The conditions are not replicated between drivers and freeway and side road conditions.
  2. At best, it would only be an approximation. Approximation is not a definition.

Table of Contents

Goal

We aim to show that using *do* expressions and SCMs, we can leverage our structural equations to define what counterfactuals stand for, how to read counterfactuals from a given model, and how probabilities of counterfactuals can be estimated when portions of models are unknown.

Structural Causal Models

Recall

A structural causal model $M = \langle V, U, \mathcal{F}, \Pr(u) \rangle$ where:

- V is a set of endogenous (observed) variables.
- U is a set of exogenous (unobserved) variables.
- $\mathcal{F}$ is a set of functions $f : D \mapsto V_i$ where $D \subseteq V \cup U$ and $V_i \in V$.
- $\Pr(u)$ is a probability distribution on U.

Structural Causal Models

Recall

A structural causal model $M = \langle V, U, \mathcal{F}, \Pr(u) \rangle$ where:

- V is a set of endogenous (observed) variables.
- U is a set of exogenous (unobserved) variables.
- $\mathcal{F}$ is a set of functions $f : D \mapsto V_i$ where $D \subseteq V \cup U$ and $V_i \in V$.
- $\Pr(u)$ is a probability distribution on U.

Deterministic Models

Call a model M fully specified or deterministic when we know both the functions $\mathcal{F}$ and for every member of U, we know their values.

Deterministic Models

Call a model M fully specified or deterministic when we know both the functions $\mathcal{F}$ and for every member of U, we know their values.

- An assignment $U = u$ uniquely determines the values of every $V_i \in V$.
- We can think of these assignments as identifying individuals in a population.

Deterministic Models

Call a model M fully specified or deterministic when we know both the functions $\mathcal{F}$ and for every member of U, we know their values.

- An assignment $U = u$ uniquely determines the values of every $V_i \in V$.
- We can think of these assignments as identifying individuals in a population.
- These assignments correspond to a "situation in nature".

Deterministic Models

Call a model M fully specified or deterministic when we know both the functions $\mathcal{F}$ and for every member of U, we know their values.

- An assignment $U = u$ uniquely determines the values of every $V_i \in V$.
- We can think of these assignments as identifying individuals in a population.
- These assignments correspond to a "situation in nature".
- For example, if $U = u$ are all of the identifying characteristics of an agricultural plot and Y is the yield of the plot in a season, then $Y(u)$ is the yield of the plot when $U = u$.

Deterministic Models

Call a model M fully specified or deterministic when we know both the functions $\mathcal{F}$ and for every member of U, we know their values.

- An assignment $U = u$ uniquely determines the values of every $V_i \in V$.
- We can think of these assignments as identifying individuals in a population.
- These assignments correspond to a "situation in nature".
- For example, if $U = u$ are all of the identifying characteristics of an agricultural plot and Y is the yield of the plot in a season, then $Y(u)$ is the yield of the plot when $U = u$.
- Consider "Y would be y had X been x, in situation $U = u$", denoted $Y_x(u) = y$, where X and Y are two variables in V.

Example of Deterministic Model

Example

Let $M = \langle \{X, Y\}, U, \mathcal{F} = \{f_X, f_Y\}, \Pr(u) \rangle$ where

$$f_X : X = aU \tag{1}$$

$$f_Y : Y = bX + U \tag{2}$$

Example of Deterministic Model

Example

Let $M = \langle \{X, Y\}, U, \mathcal{F} = \{f_X, f_Y\}, \Pr(u) \rangle$ where

$$f_X : X = aU \tag{1}$$

$$f_Y : Y = bX + U \tag{2}$$

To solve for $Y_X(u) = y$, we modify the model so that it becomes M_x where $\mathcal{F}$ is

$$f'_X : X = x \tag{3}$$

$$f_Y : Y = bX + U \tag{4}$$

Example of Deterministic Model

Example

Let $M = \langle \{X, Y\}, U, \mathcal{F} = \{f_X, f_Y\}, \Pr(u) \rangle$ where

$$f_X : X = aU \tag{1}$$

$$f_Y : Y = bX + U \tag{2}$$

To solve for $Y_X(u) = y$, we modify the model so that it becomes M_x where $\mathcal{F}$ is

$$f'_X : X = x \tag{3}$$

$$f_Y : Y = bX + U \tag{4}$$

and substitute in $U = u$ and solve for Y:

$$Y_X(u) = bx + u \tag{5}$$

Example of Deterministic Model

Example

What is the computed result for $X_y(u)$, i.e. what X would be had Y been y in situation $U = u$?

Example of Deterministic Model

Example

What is the computed result for $X_y(u)$, i.e. what X would be had Y been y in situation $U = u$? $\mathcal{F}$ is now

$$f_X = aU \tag{6}$$

$$f'_Y : Y = y \tag{7}$$

Example of Deterministic Model

Example

What is the computed result for $X_y(u)$, i.e. what X would be had Y been y in situation $U = u$? $\mathcal{F}$ is now

$$f_X = aU \tag{6}$$

$$f'_Y : Y = y \tag{7}$$

Substituting $U = u$ and solving for X, we have

$$X_y = au \tag{8}$$

which is just the observed value for X. This invariance is expected because a hypothetical change in the future should not affect the past.

SCM Counterfactuals

Each SCM encodes many possible counterfactuals. Suppose U can assume the values $1, 2, 3$ and $a = b = 1$. Then we have the following table of possible values for our various counterfactual models:

u	$X(u)$	$Y(u)$	$Y_1(u)$	$Y_2(u)$	$Y_3(u)$	$X_1(u)$	$X_2(u)$	$X_3(u)$
1	1	2	2	3	4	1	1	1
2	2	4	3	4	5	2	2	2
3	3	6	4	5	6	3	3	3

SCM Counterfactuals

Each SCM encodes many possible counterfactuals. Suppose U can assume the values $1, 2, 3$ and $a = b = 1$. Then we have the following table of possible values for our various counterfactual models:

u	$X(u)$	$Y(u)$	$Y_1(u)$	$Y_2(u)$	$Y_3(u)$	$X_1(u)$	$X_2(u)$	$X_3(u)$
1	1	2	2	3	4	1	1	1
2	2	4	3	4	5	2	2	2
3	3	6	4	5	6	3	3	3

We can compute each entry if we want. For example,
$Y_3(u) = b(3a) + 3 = (1)(3(1)) + 3 = 3 + 3 = 6.$

SCM Counterfactuals

Warning

We should not confuse counterfactuals with the *do*-operator. In the previous table, we computed not the expected value of Y under one intervention or another but the actual value of Y on the condition that $X = x$. The *do*-operator is only defined on probability distributions and so only can deliver $\mathbb{E}[Y|do(x)]$. This means it only applies to populations under interventions and not individuals. $Y_x(u)$ describes the behavior of a specific individual under those interventions.

Table of Contents

The Fundamental Law of Counterfactuals

Definition

Consider a structural model M and any arbitrary variables X and Y. Let M_x be the modified version of M with $X = x$. Then the counterfactual $Y_x(u)$ is

$$Y_x(u) = Y_{M_x}(u) \tag{4.5}$$

The Fundamental Law of Counterfactuals

Definition

Consider a structural model M and any arbitrary variables X and Y. Let M_x be the modified version of M with $X = x$. Then the counterfactual $Y_x(u)$ is

$$Y_x(u) = Y_{M_x}(u) \tag{4.5}$$

- We can think of this as the solution for Y in the surgically modified submodel M_x.

The Fundamental Law of Counterfactuals

Definition

Consider a structural model M and any arbitrary variables X and Y. Let M_x be the modified version of M with $X = x$. Then the counterfactual $Y_x(u)$ is

$$Y_x(u) = Y_{M_x}(u) \tag{4.5}$$

- We can think of this as the solution for Y in the surgically modified submodel M_x.
- This provides answer to such counterfactual questions as "what would Y had been if X had been x?"

Consistency Rule

All counterfactuals obey the following *consistency rule*:

$$\text{if } X = x, \text{ then } Y_x = Y \tag{4.6}$$

Consistency Rule

All counterfactuals obey the following *consistency rule*:

$$\text{if } X = x, \text{ then } Y_x = Y \tag{4.6}$$

Consider the previous example as found in this table:

u	$X(u)$	$Y(u)$	$Y_1(u)$	$Y_2(u)$	$Y_3(u)$	$X_1(u)$	$X_2(u)$	$X_3(u)$
1	1	2	2	3	4	1	1	1
2	2	4	3	4	5	2	2	2
3	3	6	4	5	6	3	3	3

Table of Contents

An Illustration

Example

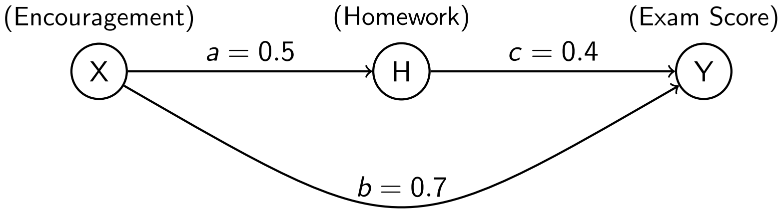
Consider the following model $M = \langle V, U, \mathcal{F}, P(u) \rangle$ where

- $V = \{X, H, Y\}$ where $X =$ time spent in after-school remedial program, $H =$ amount of homework, $Y =$ score on exam. Each variable is standardized (number of std dev. above mean).
- $U = \{U_X, U_H, U_Y\}$ where each U is independent of the others and $\sigma_{U_i U_j} = 0$ for all $i, j \in V$.
- $\mathcal{F} = \{f_X, f_H, f_Y\}$ where
  - $f_X : X = U_X$
  - $f_H : H = aX + U_H$
  - $f_Y : Y = bX + cH + U_Y$

  and $a = 0.5$, $b = 0.7$, $c = 0.4$.

An Illustration

The DAG for the previous model is given:



An Illustration

Example

Consider Joe, whose values we measure as $X = 0.5, H = 1, Y = 1.5$. What if we wanted to evaluate the query, given the evidence, of what his score would have been if he had doubled his study time?

An Illustration

Example

Consider Joe, whose values we measure as $X = 0.5, H = 1, Y = 1.5$. What if we wanted to evaluate the query, given the evidence, of what his score would have been if he had doubled his study time?

- We use the evidence and the members of $\mathcal{F}$ to find the values of the members of U.

An Illustration

Example

Consider Joe, whose values we measure as $X = 0.5, H = 1, Y = 1.5$. What if we wanted to evaluate the query, given the evidence, of what his score would have been if he had doubled his study time?

- We use the evidence and the members of $\mathcal{F}$ to find the values of the members of U.

$$\begin{aligned} X &= U_X \\ 0.5 &= U_X \end{aligned}$$

An Illustration

Example

Consider Joe, whose values we measure as $X = 0.5, H = 1, Y = 1.5$. What if we wanted to evaluate the query, given the evidence, of what his score would have been if he had doubled his study time?

- We use the evidence and the members of $\mathcal{F}$ to find the values of the members of U.

$$\begin{aligned} X &= U_X & H &= aX + U_H \\ 0.5 &= U_X & 1 &= (0.5)(0.5) + U_H \\ & & 1 &= 0.25 + U_H \\ & & 0.75 &= U_H \end{aligned}$$

An Illustration

Example

Consider Joe, whose values we measure as $X = 0.5, H = 1, Y = 1.5$. What if we wanted to evaluate the query, given the evidence, of what his score would have been if he had doubled his study time?

- We use the evidence and the members of $\mathcal{F}$ to find the values of the members of U.

$$\begin{aligned} X &= U_X \\ 0.5 &= U_X \end{aligned} \qquad \begin{aligned} H &= aX + U_H \\ 1 &= (0.5)(0.5) + U_H \\ 1 &= 0.25 + U_H \\ 0.75 &= U_H \end{aligned} \qquad \begin{aligned} Y &= bX + cH + U_Y \\ 1.5 &= (0.7)(0.5) + (0.4)(1) + U_Y \\ 1.5 &= 0.35 + 0.4 + U_Y \\ 1.5 &= 0.75 + U_Y \\ 0.75 &= U_Y \end{aligned}$$

An Illustration

Example

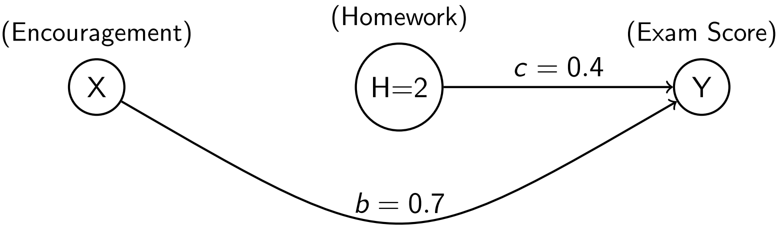
Consider Joe, whose values we measure as $X = 0.5, H = 1, Y = 1.5$. What if we wanted to evaluate the query, given the evidence, of what his score would have been if he had doubled his study time?

- We use the evidence and the members of $\mathcal{F}$ to find the values of the members of U.

$$\begin{aligned} X &= U_X \\ 0.5 &= U_X \end{aligned} \qquad \begin{aligned} H &= aX + U_H \\ 1 &= (0.5)(0.5) + U_H \\ 1 &= 0.25 + U_H \\ 0.75 &= U_H \end{aligned} \qquad \begin{aligned} Y &= bX + cH + U_Y \\ 1.5 &= (0.7)(0.5) + (0.4)(1) + U_Y \\ 1.5 &= 0.35 + 0.4 + U_Y \\ 1.5 &= 0.75 + U_Y \\ 0.75 &= U_Y \end{aligned}$$

- We can do this because the values of U are invariant to hypothetical interventions due to them being causally "upstream".

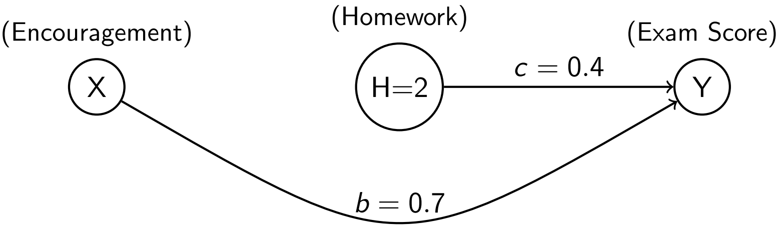
An Illustration



Example

Consider the hypothetical intervention where $H = 2$ (doubling Joe's study time). We now compute Y using the old values of U:

An Illustration



Example

Consider the hypothetical intervention where $H = 2$ (doubling Joe's study time). We now compute Y using the old values of U:

$$
\begin{aligned}
Y_{H=2}(U_X = 0.5, U_H = 0.75, U_Y = 0.75) &= aX + bH + U_Y \\
&= (0.7)(0.5) + (0.4)(2) + (0.75) \\
&= 1.90
\end{aligned}
$$

Table of Contents

Three Steps: Deterministic Counterfactuals

We can compute any deterministic counterfactual using the following three steps:

Three Steps: Deterministic Counterfactuals

We can compute any deterministic counterfactual using the following three steps:

1. Abduction: Use evidence $E = e$ to determine the value of U.

Three Steps: Deterministic Counterfactuals

We can compute any deterministic counterfactual using the following three steps:

1. Abduction: Use evidence $E = e$ to determine the value of U.
   - In Joe's example, we relied upon our structural equations to estimate each member of U using the given evidence.

Three Steps: Deterministic Counterfactuals

We can compute any deterministic counterfactual using the following three steps:

1. Abduction: Use evidence $E = e$ to determine the value of U.
   - In Joe's example, we relied upon our structural equations to estimate each member of U using the given evidence.
   - This is equivalent to us using those measurements to know the characteristics of an individual in our population.

Three Steps: Deterministic Counterfactuals

We can compute any deterministic counterfactual using the following three steps:

1. Abduction: Use evidence $E = e$ to determine the value of U.
   - In Joe's example, we relied upon our structural equations to estimate each member of U using the given evidence.
   - This is equivalent to us using those measurements to know the characteristics of an individual in our population.
2. Action: Modify the model, M, by removing the structural equations for the variables in X and replacing them with the appropriate functions $X = x$, to obtain the modified model, M_x.

Three Steps: Deterministic Counterfactuals

We can compute any deterministic counterfactual using the following three steps:

1. Abduction: Use evidence $E = e$ to determine the value of U.
   - In Joe's example, we relied upon our structural equations to estimate each member of U using the given evidence.
   - This is equivalent to us using those measurements to know the characteristics of an individual in our population.
2. Action: Modify the model, M, by removing the structural equations for the variables in X and replacing them with the appropriate functions $X = x$, to obtain the modified model, M_x.
   - We surgically alter our model to move to the hypothetical scenario we are wishing to estimate.

Three Steps: Deterministic Counterfactuals

We can compute any deterministic counterfactual using the following three steps:

1. Abduction: Use evidence $E = e$ to determine the value of U.
   - In Joe's example, we relied upon our structural equations to estimate each member of U using the given evidence.
   - This is equivalent to us using those measurements to know the characteristics of an individual in our population.
2. Action: Modify the model, M, by removing the structural equations for the variables in X and replacing them with the appropriate functions $X = x$, to obtain the modified model, M_x.
   - We surgically alter our model to move to the hypothetical scenario we are wishing to estimate.
   - In Joe's example, this is framed as intervening on M and setting $H = 2$.

Three Steps: Deterministic Counterfactuals

We can compute any deterministic counterfactual using the following three steps:

1. Abduction: Use evidence $E = e$ to determine the value of U.
   - In Joe's example, we relied upon our structural equations to estimate each member of U using the given evidence.
   - This is equivalent to us using those measurements to know the characteristics of an individual in our population.
2. Action: Modify the model, M, by removing the structural equations for the variables in X and replacing them with the appropriate functions $X = x$, to obtain the modified model, M_x.
   - We surgically alter our model to move to the hypothetical scenario we are wishing to estimate.
   - In Joe's example, this is framed as intervening on M and setting $H = 2$.
3. Prediction: Use the modified model, M_x, and the value of U to compute the value of Y, the consequence of the counterfactual.

Probabilistic Counterfactuals

So far we have looked at estimating the values of our variables pertaining to a specific individual. What if we wanted to estimate the characteristics of a subset of our population in a contrary-to-fact situation?

- For example, suppose from our previous model we wanted to know if doubling homework for every student whose exam scores were two standard deviations below average would lead to improved scores?

Probabilistic Counterfactuals

So far we have looked at estimating the values of our variables pertaining to a specific individual. What if we wanted to estimate the characteristics of a subset of our population in a contrary-to-fact situation?

- For example, suppose from our previous model we wanted to know if doubling homework for every student whose exam scores were two standard deviations below average would lead to improved scores?
- We cannot use the *do*-operator for this because it applies to the population as a whole and not just a subset.

Probabilistic Counterfactuals

So far we have looked at estimating the values of our variables pertaining to a specific individual. What if we wanted to estimate the characteristics of a subset of our population in a contrary-to-fact situation?

- For example, suppose from our previous model we wanted to know if doubling homework for every student whose exam scores were two standard deviations below average would lead to improved scores?
- We cannot use the *do*-operator for this because it applies to the population as a whole and not just a subset.
- Or suppose we wanted to estimate our credence that Joe's score would have been $Y = y'$ if he had five more hours of encouragement, $X = X + 5$. In this case, we cannot uniquely determine the values of u for Joe. So we make do with $P(U = u)$.

Probabilistic Counterfactuals

- Our distribution $\Pr(U = u)$ induces a unique probability distribution on the endogenous variables V, $P(v)$, with which we can compute the probability of any counterfactual $Y_x = y$ along with the joint distributions of all observed and counterfactual variables.

Probabilistic Counterfactuals

- Our distribution $\Pr(U = u)$ induces a unique probability distribution on the endogenous variables V, $P(v)$, with which we can compute the probability of any counterfactual $Y_x = y$ along with the joint distributions of all observed and counterfactual variables.
- For example we can compute $\Pr(Y_x = y, Z_w = z, X = x')$, even though w or x' may conflict with x.

Probabilistic Counterfactuals

- Our distribution $\Pr(U = u)$ induces a unique probability distribution on the endogenous variables V, $P(v)$, with which we can compute the probability of any counterfactual $Y_x = y$ along with the joint distributions of all observed and counterfactual variables.
- For example we can compute $\Pr(Y_x = y, Z_w = z, X = x')$, even though w or x' may conflict with x.
- Typical query: "Given that we observe feature $E = e$ for a given individual, what would we expect the value of Y for that individual be if X had been x, i.e. $\mathbb{E}[Y_{X=x} | E = e]$?"

Three Steps: Probabilistic Counterfactuals

We have the same three steps as before, with modifications for uncertainty:

Three Steps: Probabilistic Counterfactuals

We have the same three steps as before, with modifications for uncertainty:

1. **Abduction**: Update $\Pr(U)$ by the evidence to obtain $\Pr(U|E = e)$.

Three Steps: Probabilistic Counterfactuals

We have the same three steps as before, with modifications for uncertainty:

1. **Abduction**: Update $\Pr(U)$ by the evidence to obtain $\Pr(U|E = e)$.
2. **Action**: Modify the model, M, by removing the structural equations for the variables in X and replacing them with the appropriate functions $X = x$, to obtain the modified model, M_x.

Three Steps: Probabilistic Counterfactuals

We have the same three steps as before, with modifications for uncertainty:

1. **Abduction**: Update $\Pr(U)$ by the evidence to obtain $\Pr(U|E = e)$.
2. **Action**: Modify the model, M, by removing the structural equations for the variables in X and replacing them with the appropriate functions $X = x$, to obtain the modified model, M_x.
3. **Prediction**: Use the modified model, M_x, and the updated probabilities over the U variables, $\Pr(U|E = e)$, to compute the expectation of Y, the consequence of the counterfactual.

End

Thank you!