CompSci 295, Causal Inference

Rina Dechter, UCI

Lecture 6: Counterfactuals

Primer, chapter 4, Causality Chapter 9, PCH paper

The Covid Poli

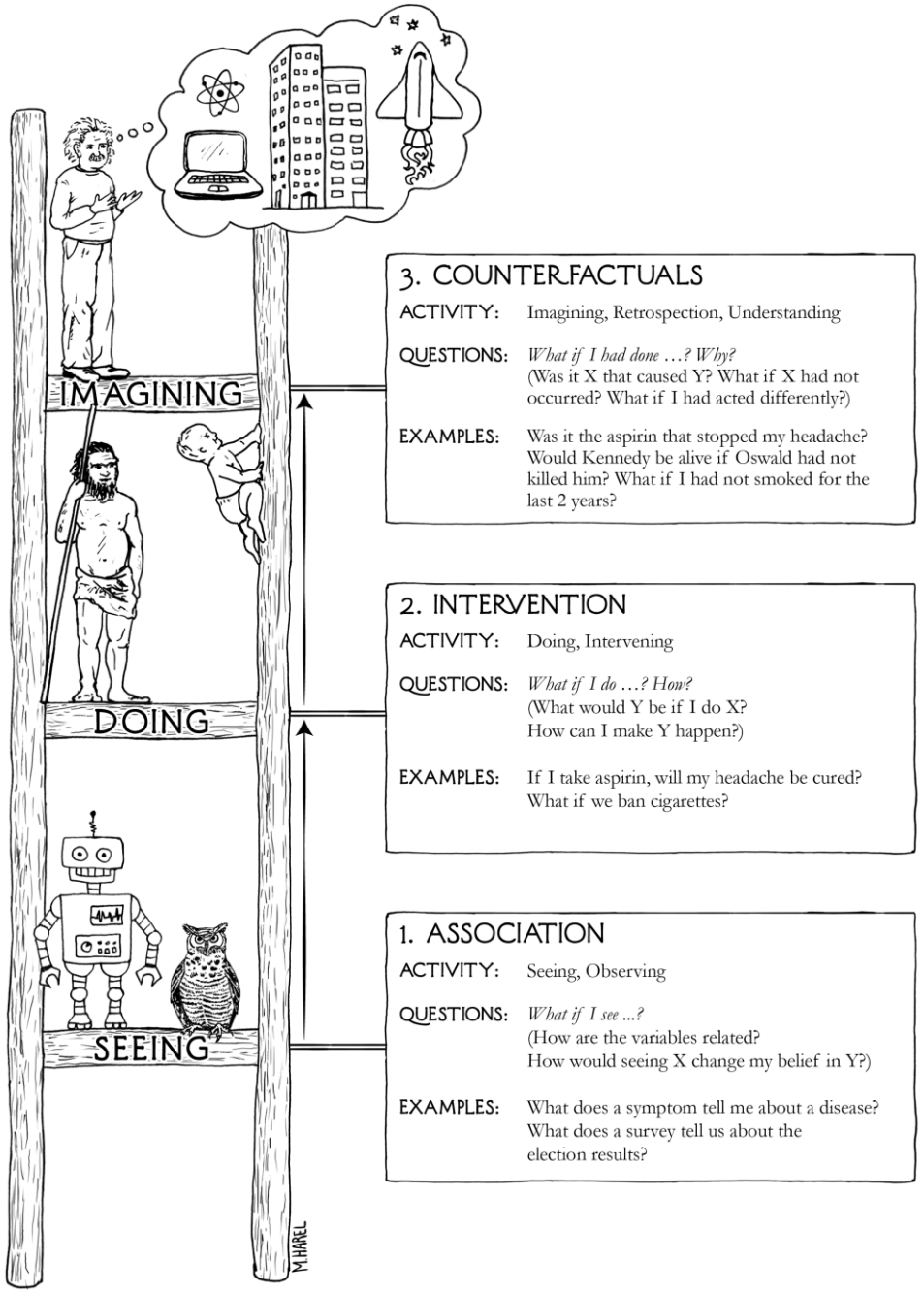
IF THE C.D.C. had recommended better masks from the beginning, how many people would have worn them and for how long? If the Biden administration had flooded stores with cheap rapid tests, would people have used them? If boosters had been pushed earlier, and more loudly, would the United States no longer trail peer nations in vaccinations?

Put differently: How much would getting our pandemic policies right have mattered?

It's easy to speak as if policy smoothly reshapes reality. I'm more guilty of that than most. But policy lies downstream of society. Mandates are not self-executing; to work, policies need to be followed, guidance needs to be believed. Public health is

Table of Contents

Ladder of Causation`



seeing, doing, and imagining.

- Most animals, learning machines are on the first rung, learning from association.
- Tool users, such as early humans, are on the second rung, if they act by planning and not merely by imitation. We can also use experiments to learn the effects of interventions, and presumably this is how babies acquire much of their causal knowledge.
- On the top rung, counterfactual learners can imagine worlds that do not exist and infer reasons for observed phenomena.

Counterfactuals subsumes the higher levels.

What are Counterfactuals

Common sense: “While driving home last night, I came to a fork in the road where I had to make a choice: to take the freeway (X = 1) or go on a surface street named Sepulveda Boulevard (X = 0). I took Sepulveda, only to find out that the traffic was bad and go. As I arrived home, an hour later, I said to myself: “Gee, I should have taken the freeway.”

Economy: Would a customer by the shoes online had the advertisement not been there?

Politics: Had Hillery won the election had Comey not announced 10 days before election that FBI reopen The investigation into her email servers?

This kind of statement: an “if” statement in which the “if” portion is untrue or unrealized—is known as a **counterfactual**. The “if” portion of a counterfactual is called the hypothetical condition, or more often, the antecedent.

Require a new language beyond “do” or intervention

What are Counterfactuals

If the patient took an aspirin and the headache was cured, would the headache still be gone had they not taken the drug?

If an individual ended up getting a promotion, would this still be the case had they not earned a PhD? What if they had a different gender?

Rung 3 reasoning

do-Expressions Not Sufficient

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

- We would write this as:

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

- 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.
- The *do*-operator only gives us $\Pr(\textit{driving time} | do(\textit{freeway}))$ and $\Pr(\textit{driving time} | do(\textit{Suplveda}))$.

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:

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

- Another way to think of $Y_{X=1}$ is the value of Y conditional on the intervention of $do(X = 1)$. So $E[Y | do(X = 1)] = E[Y_{X=1}]$.
- 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."
- $E[Y_{X=1} | X = 0, Y = 1]$ 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 $E[Y | do(X = x)]$, we estimate the drive time across in a specific world where $X = x$, irrespective to any other world.

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 $M = (V, U, F, \Pr(u))$ where:

V is a set of endogenous (observed) variables.

U is a set of exogenous (unobserved) variables.

F is a set of functions $f : D \rightarrow V_i$ where $D \subseteq V \cup U$ and $V_i \in V$.

$\Pr(u)$ is a probability distribution on U.

Definition of Counterfactuals

- M is a structural causal model (V,U,F), exogenous variables U (latent) for which we know the potential domain values.
- U=u implies a single entity in the population (e.g., a person, a situation in Nature)
- X(u) is a characteristic at world (e.g., salary(joe))
- The counterfactual sentence: Y would be y had X been x in situation U=u denoted $Y_x(u) = y$, where Y and X are any two variables in V.
- "had X been x" can be thought of as an instruction to make a minimal modification in the current model so as to establish the antecedent condition $X = x$,

Example of Deterministic Model

Example

Let $M = (\{X, Y\}, U, F = \{f_X, f_Y\}, \Pr(u))$ 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 F is

$$f_X^I : X = x \tag{3}$$

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

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 $U = u$? F is now

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

$$f_Y^I : Y = y \tag{7}$$

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

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

$$X = aU$$

$$Y = bX + U$$

The Difference Between "Do" Operator and Counterfactuals

In this example we computed not merely the probability or expected value of Y under one intervention or another, but the actual value of Y under the hypothesized new condition $X = x$. For each situation $U = u$, we obtained a definite number, $Y_x(u)$, which stands for that hypothetical value of Y in that situation.

The do-operator, is only defined on probability distributions and, after deleting the factor $P(x_i|pa_i)$ always delivers probabilistic results such as $E[Y \mid do(x)]$.

the $do(x)$-operator captures the behavior of a population under intervention, whereas $Y_x(u)$ describes the behavior of a specific individual, $U = u$, under such 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

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

- 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?"

The Fundamental Law of Counterfactuals

In general, counterfactuals obey the following *consistency rule*:

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

If X is binary, then the consistency rule takes the convenient form:

$$Y = XY_1 + (1 - X)Y_0$$

which can be interpreted as follows: Y_1 is equal to the observed value of Y whenever X takes the value one. Symmetrically, Y_0 is equal to the observed value of Y whenever X is zero. All these constraints are automatically satisfied if we compute counterfactuals through Eq. (4.5).

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

From Population to Individual –Illustration in a Structural Equation Model (SEM)

$$X = U_X$$
$$H = a \cdot X + U_H$$
$$Y = b \cdot X + c \cdot H + U_Y$$
$$\sigma_{U_i U_j} = 0 \quad \text{for all } i, j \in \{X, H, Y\}$$

X = time in remedial program
H= the amount of homework
Y= student's score in exam

The value of each variable is the number of standard deviations above the mean the student falls
Students are assigned to the remedial sessions randomly.
Assume all U factors are independent and a = 0.5, b = 0.7, c = 0.4

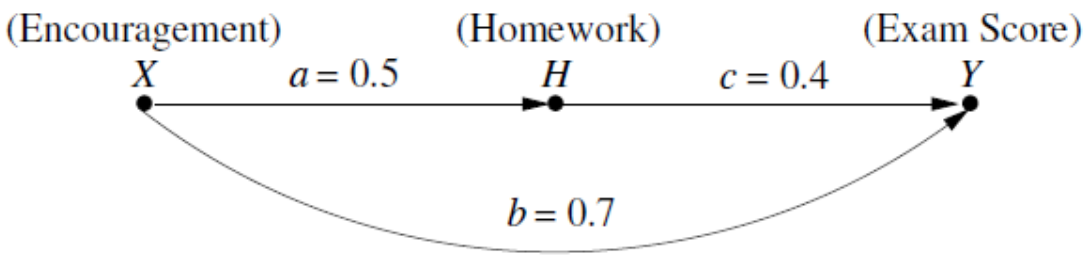
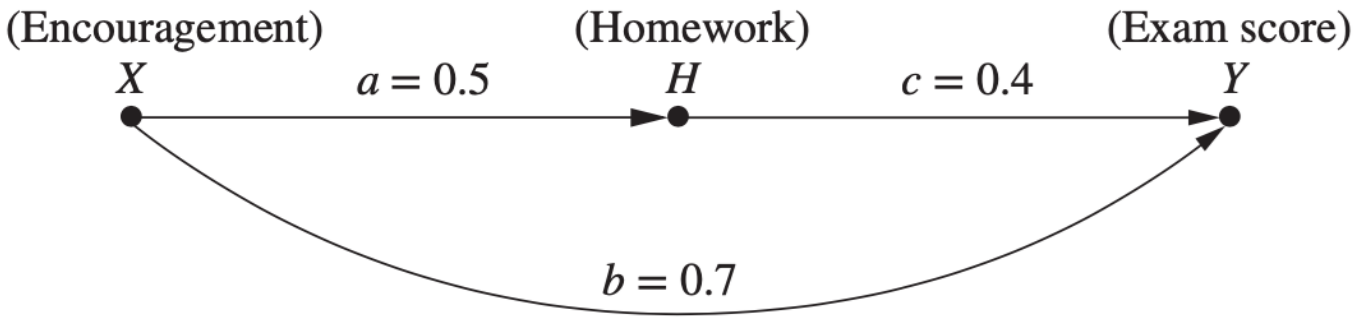


Figure 4.1: A model depicting the effect of Encouragement (X) on student's score

Assume Joe has X = 0.5,H = 1, and Y = 1.5.

What would Joe's score have been had he doubled his study time?

Population to Individuals



$$X = U_X$$

$$H = a \cdot X + U_H$$

$$Y = b \cdot X + c \cdot H + U_Y$$

$$\sigma_{U_i U_j} = 0 \quad \text{for all } i, j \in \{X, H, Y\}$$

- All variables have zero mean and unit variance
  - What kind of model is this?
  - Linear as coefficients of X, H, U's are constant in structural equations
- For an individual, X = 0.5, H = 1, Y = 1.5
  - What would exam score have been had they doubled their study time?
  - Y = ? when H = 2
  - U_X = ?, U_H = ?, U_Y = ?
  - U_X = **0.5**, U_H = 1 - 0.5·0.5 = **0.75**, U_Y = 1.5 - 0.7·0.5 - 0.4·1 = **0.75**
  - Now find $Y_{H=2}(U_X=0.5, U_H=0.75, U_Y=0.75)$
    - 0.7·0.5 + 0.4·2 + 0.75 = **1.9**

An Illustration

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 F to find the values of the members of U.

3 May 2021

$$X = U_X$$
$$0.5 = U_X$$

$$H = aX + U_H$$
$$1 = (0.5)(0.5) + U_H$$
$$1 = 0.25 + U_H$$
$$0.75 = U_H$$

$$Y = bX + cH + U_Y$$
$$1.5 = (0.7)(0.5) + (0.4)(1) + U_Y$$
$$1.5 = 0.35 + 0.4 + U_Y$$
$$1.5 = 0.75 + U_Y$$
$$0.75 = U_Y$$

Next, we simulate the action of doubling Joe's study time by replacing the structural equation for H with the constant $H = 2$. The modified model is depicted in Figure 4.2. Finally, we compute the value of Y in our modified model using the updated U values, giving

$$\begin{aligned} Y_{H=2}(U_X = 0.5, U_H = 0.75, U_Y = 0.75) \\ &= 0.5 \cdot 0.7 + 2.0 \cdot 0.4 + 0.75 \\ &= 1.90 \end{aligned}$$

We thus conclude that Joe's score, had he doubled his homework, would have been 1.9 instead of 1.5. This, according to our convention, would mean an increase to 1.9 standard deviations above the mean, instead of the current 1.5.

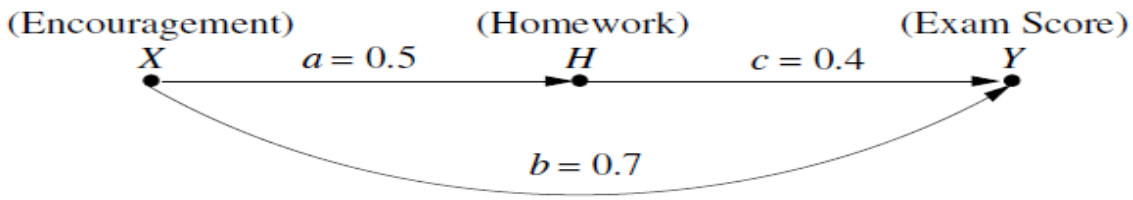


Figure 4.1: A model depicting the effect of Encouragement (X) on student's score

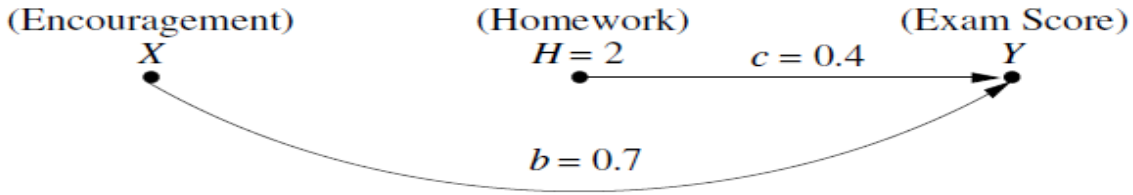


Figure 4.2: Answering a counterfactual question about a specific student's score, predicated on the assumption that homework would have increased to $H = 2$

Table of Contents

Outline

- Overview of last class:
  - Counterfactuals
  - Defining and computing counterfactuals.
  - The tree steps of computing counterfactuals (the deterministic case)
- Nondeterministic counterfactuals.
  - **The 3-steps**
  - Do operators are limited and Expressing do by counterfactuals
  - The graphical representation of counterfactuals

Three Steps for Computing Deterministic Counterfactuals

There is a three-step process for computing any deterministic counterfactual:

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

- **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 get M_x.

- **Prediction**: Use the modified model, M_x, and the value of U, to compute the value of Y , the consequence of the counterfactual.

In temporal metaphors, Step (i) explains the past (U) in light of the current evidence e;
Step (ii) bends the course of history (minimally) to comply with the hypothetical antecedent $X = x$; finally, Step (iii) predicts the future (Y) based on our new understanding of the past and our newly established condition, $X = x$.
This process will solve any deterministic counterfactual, enabled in structural models

Non-Deterministic Counterfactuals

- Counterfactuals can also be probabilistic, pertaining to a class of units within the population; for instance, in the after-school program example, we might want to know what would have happened if all students for whom $Y < 2$ had doubled their homework time.

- Nondeterminism enters causal models by assigning probabilities $P(U = u)$ over the exogenous variables U.

- The exogenous probability $P(U = u)$ induces a unique probability distribution on the endogenous variables V, $P(v)$, and we can compute not only the probability of any single counterfactual, $Y_x = y$, but also the joint distributions of all combinations of observed and counterfactual variables.

X = time in remedial program
H= the amount of homework
Y= student's score in exam

Nondeterminism

- In $E[Y_{X=x} | E = e]$, where $E = e$ is evidence
  - We allow $E = e$ to conflict with $X = x$ or Y, for example $E[Y_{X=x} | X = x', Y = y']$

1. **Abduction**: Update $P(U)$ by the evidence to obtain $P(U | E = e)$
2. **Action**: Modify M, by replacing $X = x$ in structural equations to obtain M_x
3. **Prediction**: Use M_x and $P(U | E = e)$ to compute $E[Y]$

- We can compute counterfactuals or give bounds without complete knowledge
  - Extraordinarily powerful, why?
  - <u>*Very rarely*</u> do we have complete knowledge of data *and* model
  - Counterfactual questions, like **probabilities of causation**, are often the most important questions in science and understanding

Revisiting earlier example; Adding P(U)

$$X = aU$$

$$Y = bX + U$$

$$P(U=1) = \frac{1}{2},\ P(U=2) = \frac{1}{3} \text{ and } P(U=3) = \frac{1}{6}.$$

Table 4.1 The values attained by $X(u), Y(u), Y_x(u)$, and $X_y(u)$ in the linear model of Eqs. (4.3) and (4.4)

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

For instance, we can compute the proportion of units for which Y would be 3 had X been 2, or $Y_2(u) = 3$. This occurs only in the first row when $U = 1$, and therefore $P(Y_2 = 3) = 1/2$. Similarly:
$P(Y\_1 = 4) = 1/6$, $P(Y\_1 = 3) = 1/3$, $P(Y\_2 > 3) = 1/2$

We can compute joint probability of any combination

$$P(Y_2 > 3,\ Y_1 < 4) = \frac{1}{3}$$

$$P(Y_1 < 4,\ Y - X > 1) = \frac{1}{3}$$

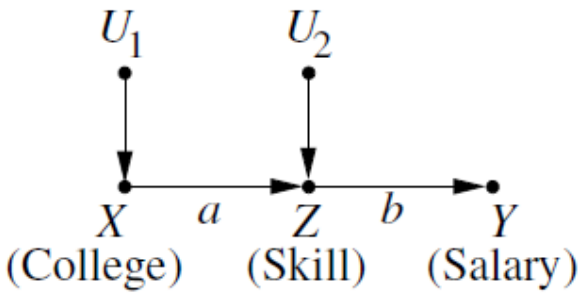
$$P(Y_1 < Y_2) = 1$$

Outline

- Overview of last class:
  - Counterfactuals
  - Defining and computing counterfactuals.
  - The tree steps of computing counterfactuals (the deterministic case)
- Nondeterministic counterfactuals.
  - The 3-steps
  - Do operators are limited and Expressing do by counterfactuals
  - The graphical representation of counterfactuals

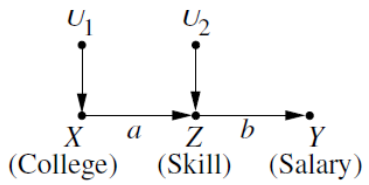
The Do Operator is Limited.

- Example model: $X = U_1 \quad Z = aX + U_2, \quad Y = bZ$
- X=1 has college education
- U_2 = professional experience
- Z = skill level
- Y = salary



The Do Operator is Limited.

$$X = U_1 \quad Z = aX + U_2, \quad Y = bZ$$



Let's compute $E[Y_{X=1}|Z = 1]$ = the expected salary of individuals with skill level $Z = 1$, had they received a college education.

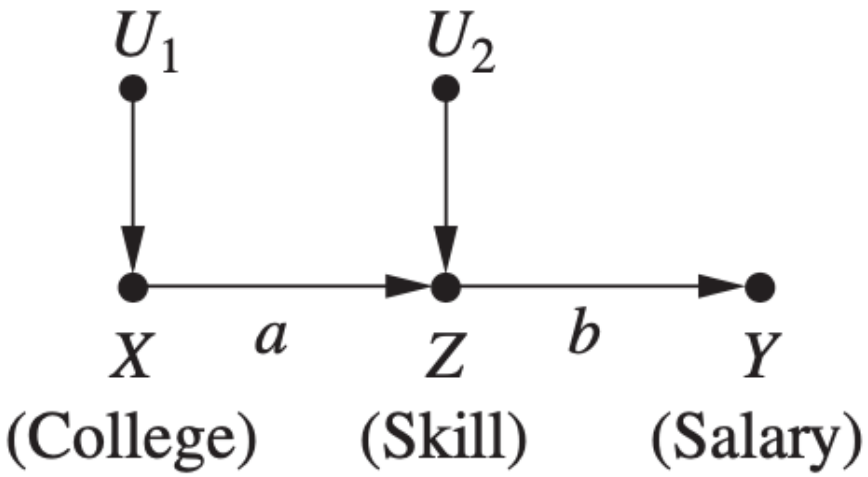
- $E[Y\ |do(X = 1), Z = 1]$ will not work: The do-expression stands for the expected salary of individuals who all finished college and have since acquired skill level $Z = 1$. The salaries of these individuals, as the graph shows, depend only on their skill, and are not affected by whether they obtained the skill through college or work experience.
- Conditioning on $Z = 1$, in this case, cuts off the effect of the intervention that we're interested in.

In contrast, some of those who currently have $Z = 1$ might not have gone to college and would have attained higher skill (and salary) had they gotten college education. Their salaries are of great interest to us, but they are not included in the do-expression.

Thus, in general, the do-expression $E[Y|do(X = 1), Z = 1] \neq E[Y_{X=1}|Z = 1]$

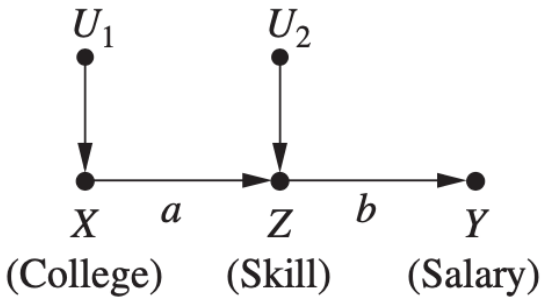
$E[Y\ |do(X = 1), Z = 1] = E[Y\ |do(X = 0), Z = 1]$, but $E[Y_{X=1}|Z = 1]$ is not equal to $E[Y_{X=0}|Z = 1]$;

Differences from do-expressions



- $E[Y|do(X = 1), Z = 1] = E[Y|do(X = 0), Z = 1]$?
  - Yes, Y only depends on Z
- $E[Y_{X=1}|Z = 1] = E[Y_{X=0}|Z = 1]$?
  - No, Z = 1 is a subset of the population, then we ask what would happen had they had X={0,1}
- Z = 1 is a post-intervention condition in the do-expression expectation
- Z = 1 is a pre-intervention condition in the counterfactual expectation
- What if we want a counterfactual with Z = 1 to be post-intervention?
  - $P(Y = y|do(X = 1), Z = 1) = P(Y = y, Z = 1|do(X = 1)) / P(Z = 1|do(X = 1) \Rightarrow$ $\mathbf{E[Y_{X=1}|Z_{X=1} = 1]}$
- Could conditioning on Z = 1 be pre-intervention?
  - Z could represent age and point to X, what happens to $E[Y_{X=1}|Z_{X=1} = 1]$?
  - Can simply drop the antecedent from Z: $E[Y_{X=1}|Z = 1]$

Counterfactual and do Calculations

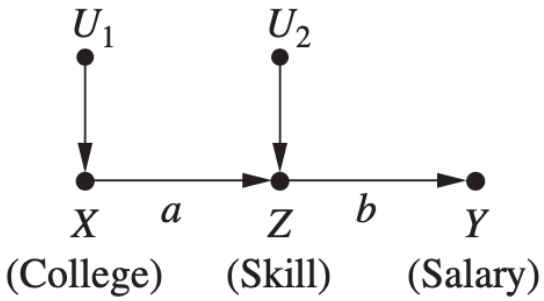


$$X = u_1 \quad Z = aX + u_2 \quad Y = bZ$$

u_1	u_2	$X(u)$	$Z(u)$	$Y(u)$	$Y_0(u)$	$Y_1(u)$	$Z_0(u)$	$Z_1(u)$
0	0	0	0	0	0	ab	0	a
0	1	0	1	b	b	$(a+1)b$	1	$a+1$
1	0	1	a	ab	0	ab	0	a
1	1	1	$a+1$	$(a+1)b$	b	$(a+1)b$	1	$a+1$

- $a \neq 0$, $a \neq 1$
- $E[Y_1|Z = 1] = ?$, $E[Y_0|Z = 1] = ?$
  - $E[Y_1|Z = 1] =$ **$(a + 1)\cdot b$**, $E[Y_0|Z = 1] =$ **b**
- $E[Y|do(X = 1), Z = 1] = ?$, $E[Y|do(X = 0), Z = 1] = ?$
  - $E[Y|do(X = 1), Z = 1] =$ **b**, $E[Y|do(X = 0), Z = 1] =$ **b** (anything suspicious?)
- $E[Y_1 - Y_0|Z = 1] = ?$
  - $a\cdot b$, note that $a\cdot b \neq 0$

Counterfactual and do Calculations



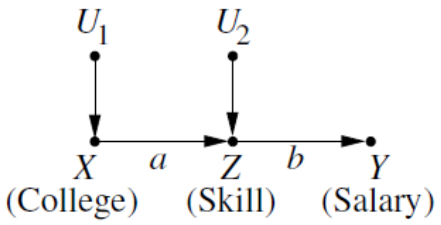
$$X = u_1 \quad Z = aX + u_2 \quad Y = bZ$$

u_1	u_2	$X(u)$	$Z(u)$	$Y(u)$	$Y_0(u)$	$Y_1(u)$	$Z_0(u)$	$Z_1(u)$
0	0	0	0	0	0	ab	0	a
0	1	0	1	b	b	$(a+1)b$	1	$a+1$
1	0	1	a	ab	0	ab	0	a
1	1	1	$a+1$	$(a+1)b$	b	$(a+1)b$	1	$a+1$

- a = 1, what changes about $E[Y_0|Z = 1]$?
  - Z = 1 happens when u_1=0 and u_2=1 and when u_1=1 and u_2=0
  - u_1 and u_2 are independent, so $P(u_1 = m, u_2 = n) = P(u_1 = m)\cdot P(u_2 = n)$
  - $E[Y_0|Z = 1] = b\cdot P(u_1 = 0)\cdot P(u_2 = 1)/[P(u_1 = 0)\cdot P(u_2 = 1) + P(u_1 = 1)\cdot P(u_2 = 0)]$
- $E[Y_1|Z = 1] = ?$
  - $2b\cdot P(u_1 = 0)\cdot P(u_2 = 1)/[P(u_1 = 0)\cdot P(u_2 = 1) + P(u_1 = 1)\cdot P(u_2 = 0)] +$
    $b\cdot(1 - P(u_1 = 0)\cdot P(u_2 = 1)/[P(u_1 = 0)\cdot P(u_2 = 1) + P(u_1 = 1)\cdot P(u_2 = 0)])$
    $= b\cdot(1 + P(u_1 = 0)\cdot P(u_2 = 1)/[P(u_1 = 0)\cdot P(u_2 = 1) + P(u_1 = 1)\cdot P(u_2 = 0)])$

Example of expectation of counterfactuals

X = u1, Z = aX+u2, Y = bZ



The table depicts the counterfactuals associated with the model for X. We replace the equation $X = u$ with the appropriate constant (zero or one) and solving for Y and Z.

Table 4.2 The values attained by $X(u), Y(u), Z(u), Y_0(u), Y_1(u), Z_0(u)$, and $Z_1(u)$ in the model of Eq. (4.7)

		$X=u_1$	$Z = aX + u_2$	$Y = bZ$				
u_1	u_2	$X(u)$	$Z(u)$	$Y(u)$	$Y_0(u)$	$Y_1(u)$	$Z_0(u)$	$Z_1(u)$
0	0	0	0	0	0	ab	0	a
0	1	0	1	b	b	$(a+1)b$	1	$a+1$
1	0	1	a	ab	0	ab	0	a
1	1	1	$a+1$	$(a+1)b$	b	$(a+1)b$	1	$a+1$

Using the table we can show:

$$E[Y_1|Z=1] = (a+1)b \tag{4.9}$$

$$E[Y_0|Z=1] = b \tag{4.10}$$

$$E[Y|do(X=1), Z=1] = b \tag{4.11}$$

$$E[Y|do(X=0), Z=1] = b \tag{4.12}$$

Despite the fact that Z separates X from Y in the graph we find that X has an effect on Y for those units falling under Z = 1: $E[Y_1 - Y_0|Z = 1] = ab \neq 0$

While the salary of those who have acquired skill level Z = 1 depends only on their skill, not on X, the salary of those who are currently at Z = 1 would have been different had they had a different past.

Outline

- Overview of last class:
  - Counterfactuals
  - Defining and computing counterfactuals.
  - The tree steps of computing counterfactuals (the deterministic case)
- Nondeterministic counterfactuals.
  - The 3-steps
  - Do operators are limited and Expressing do by counterfactuals
  - The graphical representation of counterfactuals

The Graphical Representation of Counterfactuals

Can we see counterfactual in our causal model's graph?
Yes. Based on the fundamental law of counterfactuals

$$Y_x(u) = Y_{M_x}(u)$$

If we modify model M to obtain the submodel M_x, then the outcome variable Y in the modified model is the counterfactual Y_x of the original model. Since modification calls for removing all arrows entering the variable X, the node associated with the Y variable serves as a surrogate for Y_x

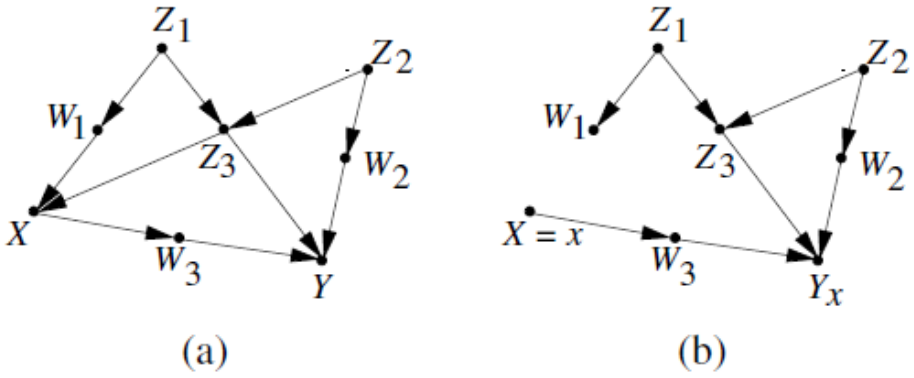
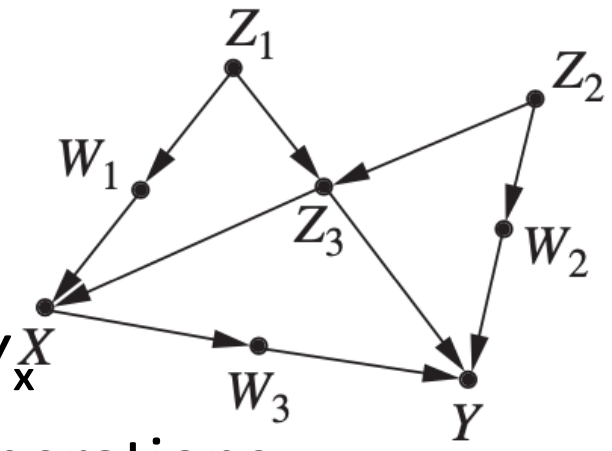
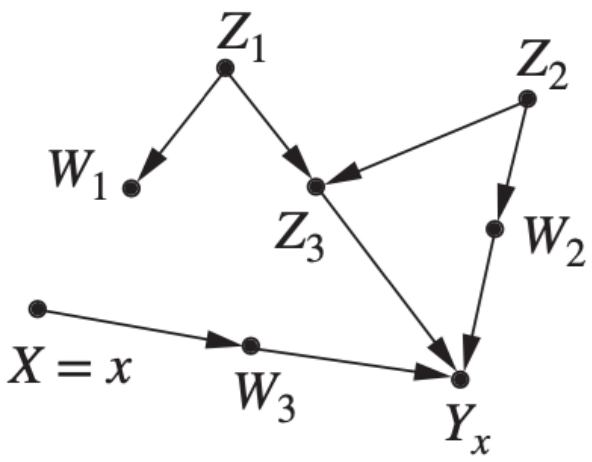


Figure 4.4: Illustrating the graphical reading of counterfactuals. (a) The original model. (b) The modified model M_x in which the node labeled Y_x represents the potential outcome Y predicated on $X = x$

Counterfactual Graphics





- We can visualize counterfactual Y_x
- Just like with interventional *do* operations
  - Remove arrows going into X
  - This new model is M_x
  - Y is now Y_x
  - Remember that conditioning Y_x on W_3 is a pre-interventional condition
- In M_x, which variables cause Y to vary?
  - Not shown are U_3 (error term for W_3) and U_Y (error term for Y)
  - Z_3, W_2, U_3, and U_Y
- How can we simply remove effect of arrows going into X?
  - This is how we can hold X constant
  - Condition on variables satisfying the backdoor criterion

The Graphical Representation of Counterfactuals

- When we ask about the statistical properties of Y_x, we need to examine what would cause Y_x to vary. Statistical variations of Yx are therefore governed by all exogenous variables capable of influencing Y when X is held constant at X=x, that is, when the arrows entering X are removed.
- The set of variables capable of transmitting variations to Y are the parents of Y , (observed and unobserved) as well as parents of nodes on the pathways between X and Y.
- For example, in the figure these parents are {Z_3,W_2,U_3,U_Y }, (U_Y and U_3, the error terms of Y and $W3$, are not shown in the diagram). Any set of variables that blocks a path to these parents also blocks that path to Yx, yield a conditional independence for Y_x. In particular, if we have a set Z that satisfies the backdoor criterion in M, that set also blocks all paths between X and those parents, and consequently, it renders X and Y_x independent for every Z = z.

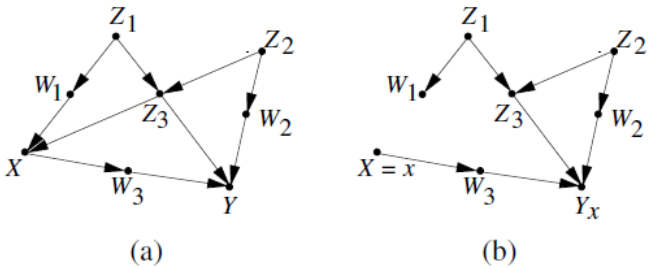


Figure 4.4: Illustrating the graphical reading of counterfactuals. (a) The original model. (b) The modified model M_x in which the node labeled Y_x represents the potential outcome Y predicated on $X = x$

Counterfactual Interpretation of Backdoor

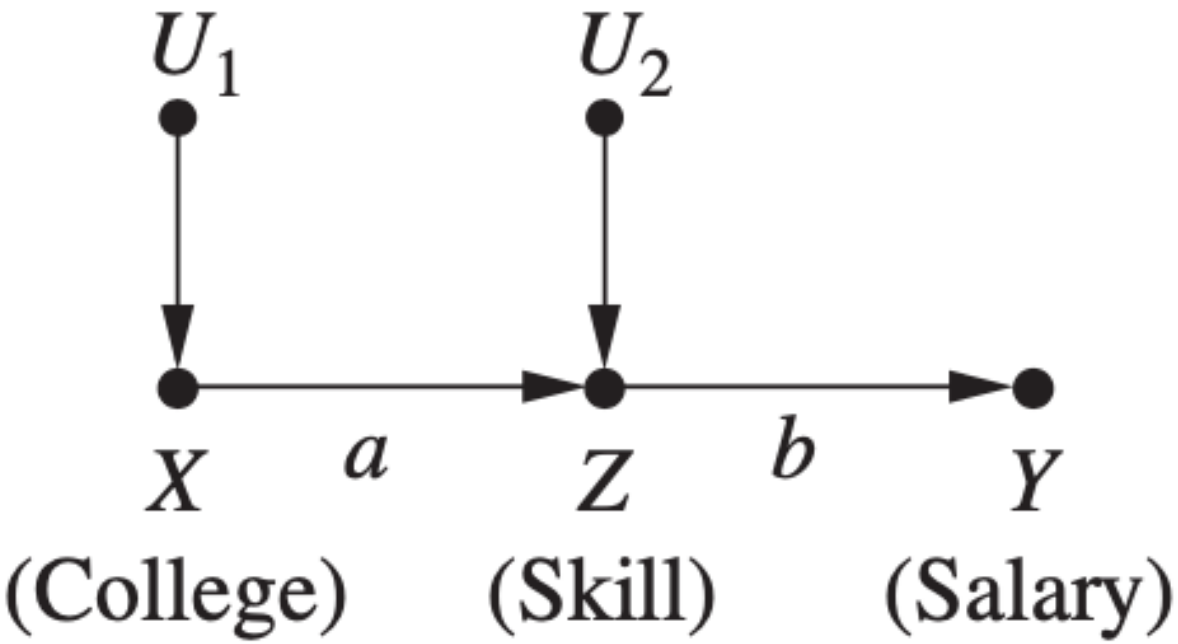
- If a set Z of variables satisfies the backdoor condition relative to (X, Y), then, for all x, the counterfactual Y_x is conditionally independent of X given Z

$$P(Y_x|X, Z) = P(Y_x|Z)$$

- How can we calculate $P(y_x)$ from data?
  - $P(y_x) = \sum_z P(y_x|Z = z)\cdot P(Z = z)$ -- law of total probability
  - $= \sum_z P(y_x|x, Z = z)\cdot P(Z = z)$ -- above theorem
  - $= \sum_z P(y|x, Z = z)\cdot P(Z = z)$ -- consistency rule
- What does this equation look like?
  - Backdoor adjustment formula!

Counterfactual Independence

- Does effect of education on salary (Y_x) depend on education, given skill z?
  - $Y_x \perp\!\!\!\perp X \mid Z$? Or $E[Y_x|X, Z] = E[Y_x|Z]$?
- But we know $E[Y|X, Z] = E[Y|Z]$, why?
  - Z blocks path $X \rightarrow Y$
- Is Y_x different?
  - Yes
  - Remove arrows into X
  - $Y \rightarrow Y_x$, which variables cause Y_x to vary?
  - $\{U_2\}$, U_2 is important, is $X \perp\!\!\!\perp U_2$?
  - Not when we condition on Z
  - $E[Y_x|X, Z] \neq E[Y_x|Z]$
- What does this mean?
  - Education matters in estimating Y_x



Outline

- Overview of last class:
  - Counterfactuals
  - Defining and computing counterfactuals.
  - The tree steps of computing counterfactuals (the deterministic case)
- Nondeterministic counterfactuals.
  - The 3-steps
  - Do operators are limited and Expressing do by counterfactuals
  - The graphical representation of counterfactuals
  - Counterfactuals in Experimental Settings
  - Practical use of counterfactuals

Counterfactual in Experimental Settings

So we can answer counterfactual question from a fully specified structural model.
But what to do when a model is not available, and we have only a finite sample of observed individuals?

Let's consider again the "encouragement design" model in which we analyzed the behavior of an individual named Joe. Assume that the experimenter observes a set of 10 individuals, with Joe being participant 1. Each, characterized by a distinct vector $U_i = (U_X, U_H, U_Y)$, as shown in the first 3 columns

$$X = U_X$$
$$H = a \cdot X + U_H$$
$$Y = b \cdot X + c \cdot H + U_Y$$
$$\sigma_{U_i U_j} = 0 \quad \text{for all } i, j \in \{X, H, Y\}$$

Table 4.3 Potential and observed outcomes predicted by the structural model of Figure 4.1 units were selected at random, with each U_i uniformly distributed over $[0, 1]$

	Participant characteristics			Observed behavior			Predicted potential outcomes				
Participant	U_X	U_H	U_Y	X	Y	H	Y_0	Y_1	H_0	H_1	$Y_{00}\ldots$
1	0.5	0.75	0.75	0.5	1.50	1.0	1.05	1.95	0.75	1.25	0.75
2	0.3	0.1	0.4	0.3	0.71	0.25	0.44	1.34	0.1	0.6	0.4
3	0.5	0.9	0.2	0.5	1.01	1.15	0.56	1.46	0.9	1.4	0.2
4	0.6	0.5	0.3	0.6	1.04	0.8	0.50	1.40	0.5	1.0	0.3
5	0.5	0.8	0.9	0.5	1.67	1.05	1.22	2.12	0.8	1.3	0.9
6	0.7	0.9	0.3	0.7	1.29	1.25	0.66	1.56	0.9	1.4	0.3
7	0.2	0.3	0.8	0.2	1.10	0.4	0.92	1.82	0.3	0.8	0.8
8	0.4	0.6	0.2	0.4	0.80	0.8	0.44	1.34	0.6	1.1	0.2
9	0.6	0.4	0.3	0.6	1.00	0.7	0.46	1.36	0.4	0.9	0.3
10	0.3	0.8	0.3	0.3	0.89	0.95	0.62	1.52	0.8	1.3	0.3

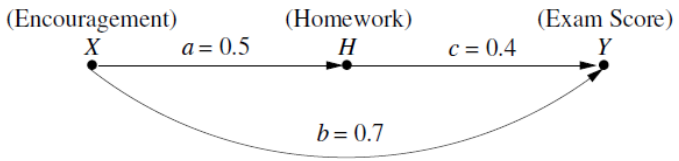


Figure 4.1: A model depicting the effect of Encouragement (X) on student's score

We use the model to fill the data from the U variables.

First item: Y_0 = 0.4 times 1 + 0.75= 1.05

Counterfactual in Experimental Settings

From this synthetic population, one can estimate the probability of every counterfactual query on variables X, Y,Z, assuming, of course, that we are in possession of all entries of the table.

Clearly the table is not available to us in either observational or experimental studies. This was deduced from the fully specified model from which we could infer the defining characteristics $\{U_X,U_H,U_Y\}$ of each participant, given the observations $\{X,H, Y\}$.

Without a parametric model, the observed behavior $\{X,H, Y\}$ tells very little of the potential outcome Y_1 or Y_0.

We know only the consistency rule: that $Y1$ must be equal to Y in case X = 1, and Y_0 must be equal to Y in case X = 0.

Yet we can say much at the population level estimating their probabilities or expectation. We can use
The adjustment formula of (4.16), where we were able to compute $E(Y\_1 - Y\_0)$ using the graph alone as we will see next.

Using Experimental Data

Assume that we have no information whatsoever about the underlying model. All we have are measurements on Y taken in an experimental study in which X is randomized over two levels, $X = 0$ and $X = 1$.

Table 4.4 Potential and observed outcomes in a randomized clinical trial with X randomized over $X = 0$ and $X = 1$

Participant	Predicted potential outcomes Y_0	Predicted potential outcomes Y_1	Observed outcomes Y_0	Observed outcomes Y_1
1	1.05	1.95	1.05	■
2	0.44	1.34	■	1.34
3	0.56	1.46	■	1.46
4	0.50	1.40	■	1.40
5	1.22	2.12	1.22	■
6	0.66	1.56	0.66	■
7	0.92	1.82	■	1.82
8	0.44	1.34	0.44	■
9	0.46	1.36	■	1.36
10	0.62	1.52	0.62	■
	True average treatment effect: 0.90		Study average treatment effect: 0.68	

Randomized: participants 1, 5, 6, 8 and 10 assigned to $X = 0$, and the rest to $X = 1$. The first two columns give the true potential outcomes (taken from Table 4.3) while the last two columns describe the information available to the experimenter.

The difference between the observed means in the treatment and control groups will converge to the difference of the population averages, $E(Y\_1 - Y\_0) = 0.9$ due to randomization.

Under randomization, the adjustment formula (4.16) is applicable with Z = {empty}, yielding $E[Y\_x] = E[Y \mid X = x]$.
So, Table 4.4 helps us understand what is actually computed when we take sample averages in experimental settings and how those averages are related to the underlying counterfactuals, Y_1 and Y_0.

ATE (Average Treatment Effect)

- No information on the underlying model, we can run experiments
  - What does random X do?
  - Removes arrows into X
  - Estimates Y_0 and Y_1
  - $E[Y_x] = \sum_z E[Y|z,x] \cdot P(z)$
    - $Z = \emptyset$
  - $E[Y_x] = E[Y|x]$
- Estimate $E[Y_1 - Y_0]$
  - Average observations
  - $= \sum Y_1/n - \sum Y_0/n$
  - $= 0.68$
  - Should be 0.9, why isn't it?

	Predicted potential outcomes		Observed outcomes	
Participant	Y_0	Y_1	Y_0	Y_1
1	1.05	1.95	1.05	■
2	0.44	1.34	■	1.34
3	0.56	1.46	■	1.46
4	0.50	1.40	■	1.40
5	1.22	2.12	1.22	■
6	0.66	1.56	0.66	■
7	0.92	1.82	■	1.82
8	0.44	1.34	0.44	■
9	0.46	1.36	■	1.36
10	0.62	1.52	0.62	■
	True average treatment effect: 0.90		Study average treatment effect: 0.68	

Outline

- Overview of last class:
  - Counterfactuals
  - Defining and computing counterfactuals.
  - The tree steps of computing counterfactuals (the deterministic case)
- Nondeterministic counterfactuals.
  - The 3-steps
  - Do operators are limited and Expressing do by counterfactuals
  - The graphical representation of counterfactuals
  - Counterfactuals in Experimental Settings
  - Practical use of counterfactuals

Practical Uses of Counterfactuals

- Recruitment program
- Additive Interventions
- Personal decision making
- Sex discrimination in hiring
- Mediation and path disabling

Recruitment Program Job Training Helps?

Example 4.4.1 A government is funding a job training program aimed at getting jobless people back into the workforce. A pilot randomized experiment shows that the program is effective; a higher percentage of people were hired among those who finished the program than among those who did not go through the program. As a result, the program is approved, and a recruitment effort is launched to encourage enrollment among the unemployed, by offering the job training program to any unemployed person who elects to enroll.

Enrollment is successful, and the hiring rate among the program's graduates turns out even higher than in the randomized pilot study. Success!!!

Critics say: Those who self-enroll, may be more intelligent, more resourceful, and more socially connected than the eligible who did not enroll and are more likely to have found a job regardless of the training.

The critics claim that what we need to estimate is the differential benefit of the program on those enrolled: the extent to which hiring rate has increased among the enrolled, compared to what it would have been had they not been trained.

ETT=

$X = 1$ represent training and $Y = 1$ represent hiring, the quantity that needs to be evaluated is the effect of training on the trained (ETT, better known as "effect of treatment on the treated,"

$ETT = E[Y_1 - Y_0 | X = 1]$ (4.20)

Here the difference $Y_1 - Y_0$ represents the causal effect of training (X) on hiring (Y) for a randomly chosen individual, and the condition $X = 1$ limits the choice to those actually choosing the training program on their own initiative. As in our freeway example of Section 4.1, we are

Personal Decision Making

Example 4.4.3 Ms. Jones, a cancer patient, is facing a tough decision between two possible treatments: (i) lumpectomy alone, or (ii) lumpectomy plus irradiation. In consultation with her oncologist, she decides on (ii). Ten years later, Ms. Jones is alive, and the tumor has not recurred. She speculates: Do I owe my life to irradiation?
Mrs. Smith, on the other hand, had a lumpectomy alone, and her tumor recurred after a year. And she is regretting: I should have gone through irradiation.
Can these speculations ever be substantiated from statistical data? Moreover, what good would it do to confirm Ms. Jones's triumph or Mrs. Smith's regret?

Sex Discrimination in Hiring

Example 4.4.4 Mary files a law suit against the New York-based XYZ International, alleging discriminatory hiring practices. According to her, she has applied for a job with XYZ International, and she has all the credentials for the job, yet she was not hired, allegedly because she mentioned, during the course of her interview, that she is gay. Moreover, she claims, the hiring record of XYZ International shows consistent preferences for straight employees. Does she have a case? Can hiring records prove whether XYZ International was discriminating when declining her job application?
At the time of writing, U.S. law doesn’t specifically prohibit employment discrimination on

Mediation and Path-disabling

Example 4.4.5 A policy maker wishes to assess the extent to which gender disparity in hiring can be reduced by making hiring decisions gender-blind, rather than eliminating gender inequality in education or job training. The former concerns the “direct effect” of gender on hiring, whereas the latter concerns the “indirect effect,” or the effect mediated via job qualification.