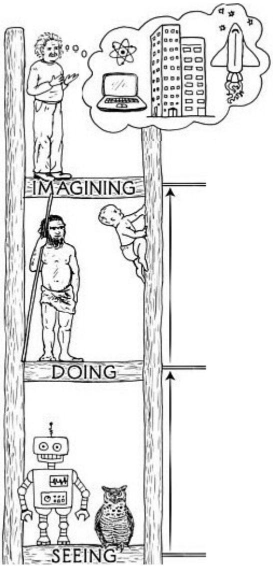
Counterfactual reasoning

Charles K. Assaad, Emilie Devijver, Eric Gaussier

emilie.devijver@univ-grenoble-alpes.fr

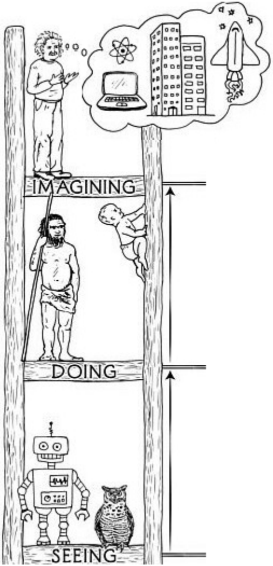
Table of contents



Counterfactuals

Interventions

Associations



Counterfactuals
I took an aspirin, and my headache is gone: would I have had a headache had I not taken that aspirin?

Interventions
It I take an aspirin now, will I wake up with a headache?
P(*headache*|do(*aspirin*))

Associations
I took an aspirin after dinner, will I wake up with a headache?

A first example

I took an aspirin, and my headache is gone: would I have had a headache had I not taken that aspirin?

- T: observed treatment (aspirin)
- Y: observed outcome (headache)
- i: used in subscript to denote a specific individual (me)
- $Y_i(1)$: potential outcome under treatment for individual i
- $Y_i(0)$: potential outcome under no treatment for individual i

$$do(T = 1) \rightarrow Y_i(1) = 1$$
$$do(T = 0) \rightarrow Y_i(0) =?$$

A first example

I took an aspirin, and my headache is gone: would I have had a headache had I not taken that aspirin?

- T: observed treatment (aspirin)
- Y: observed outcome (headache)
- i: used in subscript to denote a specific individual (me)
- $Y_i(1)$: potential outcome under treatment for individual i
- $Y_i(0)$: potential outcome under no treatment for individual i

$$\begin{array}{rl} \text{factual} & do(T = 1) \rightarrow Y_i(1) = 1 \\ \text{counterfactual} & do(T = 0) \rightarrow Y_i(0) = ? \end{array}$$

Definition

$$Y(t) \mid T = t', Y = y'$$

where t is the hypothetical condition, and $T = t', Y = y'$ is the observation.

Interest in an individual level

From an experimentalist perspective, there is a profound gap between population and individual levels of analysis: the do(x)-operator captures the behavior of a population under intervention, whereas $Y_x(u)$ describes the behavior of a specific individual under such interventions.

Definition

$$Y(t) \mid T = t', Y = y'$$

where t is the hypothetical condition, and $T = t', Y = y'$ is the observation.

Interest in an individual level

From an experimentalist perspective, there is a profound gap between population and individual levels of analysis: the do(x)-operator captures the behavior of a population under intervention, whereas $Y_x(u)$ describes the behavior of a specific individual under such interventions.

Fundamental law of counterfactuals

T, Y be two variables, not necessarily connected by a single equation, described in a structural model M.
Let M_t stand for the modified version of M, with the equation of T replaced by $T = t$.
Formal definition of $Y_t(u)$: $Y_t(u) = Y_{M_t}(u)$

Consistency rule: if $T = t$, then $Y_t = Y$.

Example with binary treatment

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

$$Y = TY_1 + (1 - T)Y_0$$

For example,

- Y: being happy or unhappy (1 or 0)
- T: get a dog or don't (1 or 0)
- U: unobserved variable describing the individual (1 if dog-person or 0 if anti-dog person)

then $Y_1 = U$ and $Y_0 = 1 - U$.

Observations: $T = 0$ and $Y = 0$

$U = 1$ and $Y_u(1) = 1$

Example with binary treatment

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

$$Y = TY_1 + (1 - T)Y_0$$

For example,

- Y: being happy or unhappy (1 or 0)
- T: get a dog or don't (1 or 0)
- U: unobserved variable describing the individual (1 if dog-person or 0 if anti-dog person)

then $Y_1 = U$ and $Y_0 = 1 - U$.

Observations: $T = 0$ and $Y = 0$

$U = 1$ and $Y_u(1) = 1$

Example with binary treatment

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

$$Y = TY_1 + (1 - T)Y_0$$

For example,

- Y: being happy or unhappy (1 or 0)
- T: get a dog or don't (1 or 0)
- U: unobserved variable describing the individual (1 if dog-person or 0 if anti-dog person)

then $Y_1 = U$ and $Y_0 = 1 - U$.

Observations: $T = 0$ and $Y = 0$

$U = 1$ and $Y_u(1) = 1$

Example with binary treatment

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

$$Y = TY_1 + (1 - T)Y_0$$

For example,

- Y: being happy or unhappy (1 or 0)
- T: get a dog or don't (1 or 0)
- U: unobserved variable describing the individual (1 if dog-person or 0 if anti-dog person)

then $Y_1 = U$ and $Y_0 = 1 - U$.

Observations: $T = 0$ and $Y = 0$
$U = 1$ and $Y_u(1) = 1$

Example with binary treatment

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

$$Y = TY_1 + (1 - T)Y_0$$

For example,

- Y: being happy or unhappy (1 or 0)
- T: get a dog or don't (1 or 0)
- U: unobserved variable describing the individual (1 if dog-person or 0 if anti-dog person)

then $Y_1 = U$ and $Y_0 = 1 - U$.

Observations: $T = 0$ and $Y = 0$
$U = 1$ and $Y_u(1) = 1$

General steps for computing deterministic counterfactuals

1. **Abduction:** use the observations to determine the value of U
2. **Action:** modify the model M by removing the structural equations for the variables in T and replacing them with the appropriate functions $T = t$, to obtain the modified model M_t
3. **Prediction:** use the modified model M_t and the value of U to compute the value of $Y(t)$, the consequence of the counterfactual

Example with binary treatment, cont'd

What if we can't solve for U?

$$Y = \begin{cases} 1 & \text{if individual always happy} \\ 0 & \text{if individual never happy} \\ T & \text{if individual dog-needer} \\ 1 - T & \text{if individual dog-hater} \end{cases}$$

For example,

- Y: being happy or unhappy (1 or 0)
- T: get a dog or don't (1 or 0)
- U: unobserved variable describing the individual (1 if dog-person or 0 if anti-dog person)

Observations: $T = 1$ and $Y = 0$: $Y_u(1) = 0$. What is $Y_u(0)$?
We don't know if the individual is never happy or a dog-hater.

Example with binary treatment, cont'd

What if we can't solve for U?

$$Y = \begin{cases} 1 & \text{if individual always happy} \\ 0 & \text{if individual never happy} \\ T & \text{if individual dog-needer} \\ 1 - T & \text{if individual dog-hater} \end{cases}$$

For example,

- Y: being happy or unhappy (1 or 0)
- T: get a dog or don't (1 or 0)
- U: unobserved variable describing the individual (1 if dog-person or 0 if anti-dog person)

Observations: $T = 1$ and $Y = 0$: $Y_u(1) = 0$. What is $Y_u(0)$?
We don't know if the individual is never happy or a dog-hater.

Example with binary treatment, cont'd

What if we can't solve for U?

$$Y = \begin{cases} 1 & \text{if individual always happy} \\ 0 & \text{if individual never happy} \\ T & \text{if individual dog-needer} \\ 1 - T & \text{if individual dog-hater} \end{cases}$$

For example,

- Y: being happy or unhappy (1 or 0)
- T: get a dog or don't (1 or 0)
- U: unobserved variable describing the individual (1 if dog-person or 0 if anti-dog person)

Observations: $T = 1$ and $Y = 0$: $Y_u(1) = 0$. What is $Y_u(0)$?
We don't know if the individual is never happy or a dog-hater.

Example with binary treatment, cont'd

What if we can't solve for U?

$$Y = \begin{cases} 1 & \text{if individual always happy} \\ 0 & \text{if individual never happy} \\ T & \text{if individual dog-needer} \\ 1 - T & \text{if individual dog-hater} \end{cases}$$

For example,

- Y: being happy or unhappy (1 or 0)
- T: get a dog or don't (1 or 0)
- U: unobserved variable describing the individual (1 if dog-person or 0 if anti-dog person)

Observations: $T = 1$ and $Y = 0$: $Y_u(1) = 0$. What is $Y_u(0)$? We don't know if the individual is never happy or a dog-hater.

Example with binary treatment, cont'd

We add a probability distribution over U:

$$P(U \text{ always happy }) = 0.3$$
$$P(U \text{ never happy }) = 0.2$$
$$P(U \text{ dog-needer}) = 0.4$$
$$P(U \text{ dog-hater}) = 0.1$$
$$P(U \text{ never happy} | T = 1, Y = 0) = 0.2/(0.2 + 0.1) = 2/3$$
$$P(U \text{ dog-hater} | T = 1, Y = 0) = 0.1/(0.2 + 0.1) = 1/3$$
$$P(Y_u(0)) = 1/3$$

Example with binary treatment, cont'd

We add a probability distribution over U:

$$
\begin{aligned}
P(U \text{ always happy }) &= 0.3\\
P(U \text{ never happy }) &= 0.2\\
P(U \text{ dog-needer}) &= 0.4\\
P(U \text{ dog-hater}) &= 0.1\\
P(U \text{ never happy} | T = 1, Y = 0) &= 0.2/(0.2 + 0.1) = 2/3\\
P(U \text{ dog-hater} | T = 1, Y = 0) &= 0.1/(0.2 + 0.1) = 1/3\\
P(Y_u(0)) &= 1/3
\end{aligned}
$$

Example with binary treatment, cont'd

We add a probability distribution over U:

$$
\begin{aligned}
P(U \text{ always happy }) &= 0.3 \\
P(U \text{ never happy }) &= 0.2 \\
P(U \text{ dog-needer}) &= 0.4 \\
P(U \text{ dog-hater}) &= 0.1 \\
P(U \text{ never happy} | T = 1, Y = 0) &= 0.2/(0.2 + 0.1) = 2/3 \\
P(U \text{ dog-hater} | T = 1, Y = 0) &= 0.1/(0.2 + 0.1) = 1/3 \\
P(Y_u(0)) &= 1/3
\end{aligned}
$$

General steps for computing probabilisttic counterfactuals

1. **Abduction:** use the observations to update the distribution of U
2. **Action:** modify the model M by removing the structural equations for the variables in T and replacing them with the appropriate functions $T = t$, to obtain the modified model M_t
3. **Prediction:** use the modified model M_t and the updated distribution of U to compute the value of $Y(t)$, the consequence of the counterfactual

Another example: fully specified linear model M

$$
\begin{aligned}
X &= U_X && \text{Encouragement} \\
H &= 0.5X + U_H && \text{Homework} \\
Y &= 0.7X + 0.4H + U_Y && \text{Exam score} \\
\sigma_{U_i U_j} &= 0 \text{ for all } i, j \in \{X, H, Y\}
\end{aligned}
$$

$$
\begin{tikzcd}
X \arrow[r] \arrow[rr, bend left, leftrightarrow] & H \arrow[r] & Y
\end{tikzcd}
$$

Observation: a student named Joe, $X = 0.5$, $H = 1$, $Y = 1.5$

Another example: fully specified linear model M

$$X = U_X \qquad \text{Encouragement}$$
$$H = 0.5X + U_H \qquad \text{Homework}$$
$$Y = 0.7X + 0.4H + U_Y \qquad \text{Exam score}$$
$$\sigma_{U_i U_j} = 0 \text{ for all } i, j \in \{X, H, Y\}$$

Observation: a student named Joe, $X = 0.5$, $H = 1$, $Y = 1.5$

Another example: fully specified linear model M

$$
\begin{aligned}
X &= U_X && \text{Encouragement}\\
H &= 0.5X + U_H && \text{Homework}\\
Y &= 0.7X + 0.4H + U_Y && \text{Exam score}\\
\sigma_{U_i U_j} &= 0 \text{ for all } i, j \in \{X, H, Y\}
\end{aligned}
$$

Observation: a student named Joe, $X = 0.5$, $H = 1$, $Y = 1.5$
What would Joe's score have been had he doubled his study time?
$U_X = 0.5$, $U_H = 0.75$, $U_Y = 0.75$
$Y_{H=2}(U_X = 0.5, U_H = 0.75, U_Y = 0.75) = 1.90$

Another example: fully specified linear model M

$$
\begin{aligned}
X &= U_X & \text{Encouragement}\\
H &= 0.5X + U_H & \text{Homework}\\
Y &= 0.7X + 0.4H + U_Y & \text{Exam score}\\
\sigma_{U_i U_j} &= 0 \text{ for all } i, j \in \{X, H, Y\}
\end{aligned}
$$

Observation: a student named Joe, $X = 0.5$, $H = 1$, $Y = 1.5$
What would Joe's score have been had he doubled his study time?
$U_X = 0.5$, $U_H = 0.75$, $U_Y = 0.75$
$Y_{H=2}(U_X = 0.5, U_H = 0.75, U_Y = 0.75) = 1.90$

Another example: fully specified linear model M

$$
\begin{aligned}
X &= U_X && \text{Encouragement} \\
H &= 0.5X + U_H && \text{Homework} \\
Y &= 0.7X + 0.4H + U_Y && \text{Exam score} \\
\sigma_{U_i U_j} &= 0 \text{ for all } i, j \in \{X, H, Y\}
\end{aligned}
$$

Observation: a student named Joe, $X = 0.5$, $H = 1$, $Y = 1.5$
What would Joe's score have been had he doubled his study time?
$U_X = 0.5, U_H = 0.75, U_Y = 0.75$
$Y_{H=2}(U_X = 0.5, U_H = 0.75, U_Y = 0.75) = 1.90$

Another example: fully specified linear model M

$$
\begin{aligned}
X &= U_X && \text{Encouragement}\\
H &= 0.5X + U_H && \text{Homework}\\
Y &= 0.7X + 0.4H + U_Y && \text{Exam score}\\
\sigma_{U_i U_j} &= 0 \text{ for all } i, j \in \{X, H, Y\}
\end{aligned}
$$

Observation: a student named Joe, $X = 0.5$, $H = 1$, $Y = 1.5$
What would Joe's study time have been had he doubled his score?

Another example: fully specified linear model M

$$
\begin{aligned}
X &= U_X && \text{Encouragement} \\
H &= 0.5X + U_H && \text{Homework} \\
Y &= 0.7X + 0.4H + U_Y && \text{Exam score} \\
\sigma_{U_i U_j} &= 0 \text{ for all } i, j \in \{X, H, Y\}
\end{aligned}
$$

Observation: a student named Joe, $X = 0.5$, $H = 1$, $Y = 1.5$
What would Joe's study time have been had he doubled his score?
$U_X = 0.5$, $U_H = 0.75$, $U_Y = 0.75$

Another example: fully specified linear model M

$$
\begin{aligned}
X &= U_X && \text{Encouragement} \\
H &= 0.5X + U_H && \text{Homework} \\
Y &= 0.7X + 0.4H + U_Y && \text{Exam score} \\
\sigma_{U_i U_j} &= 0 \text{ for all } i, j \in \{X, H, Y\}
\end{aligned}
$$

Observation: a student named Joe, $X = 0.5$, $H = 1$, $Y = 1.5$
What would Joe's study time have been had he doubled his score?
$U_X = 0.5$, $U_H = 0.75$, $U_Y = 0.75$
$H_{Y=2}(U_X = 0.5, U_H = 0.75, U_Y = 0.75) = 1$

Another example: fully specified linear model M

$$
\begin{aligned}
X &= U_X && \text{Encouragement}\\
H &= 0.5X + U_H && \text{Homework}\\
Y &= 0.7X + 0.4H + U_Y && \text{Exam score}\\
\sigma_{U_iU_j} &= 0 \text{ for all } i, j \in \{X, H, Y\}
\end{aligned}
$$

Observation: a student named Joe, $X = 0.5$, $H = 1$, $Y = 1.5$
What would Joe's study time have been had he doubled his score?
$U_X = 0.5, U_H = 0.75, U_Y = 0.75$
$H_{Y=2}(U_X = 0.5, U_H = 0.75, U_Y = 0.75) = 1$

Counterfactual conditions are on the future, not on the past!

Another example: fully specified linear model M, cont'd

Again, in that case, some questions can't be explicitly determined.

- Suppose Joe had a scored $Y = y$ in the exam. What is the probability that Joe's score would be $Y = y'$ had he had five more hours of encouragement training?
- What would his expected score be in such hypothetical world?

We do not have information on X, H: we cannot therefore determine uniquely the value u that pertains to Joe.

Another example: fully specified linear model M, cont'd

Again, in that case, some questions can't be explicitly determined.

- Suppose Joe had a scored $Y = y$ in the exam. What is the probability that Joe's score would be $Y = y'$ had he had five more hours of encouragement training?
- What would his expected score be in such hypothetical world?

We do not have information on X, H: we cannot therefore determine uniquely the value u that pertains to Joe.

Counterfactuals in linear models

Theorem
Let τ be the slope of the total effect of X on Y,

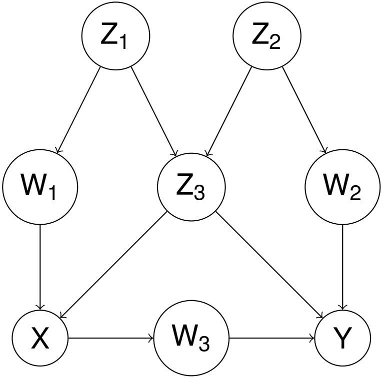
$$\tau = E(Y|do(x+1)) - E(Y|do(x))$$

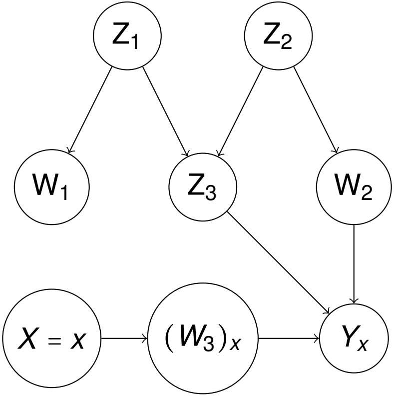
then, for any evidence $Z = e$, we have

$$E(Y_{X=x}|\, Z = e) = E(Y|Z = e) + \tau(x - E(X|Z = e))$$

Proof on board

Graphical representations of counterfactuals





Backdoor criterion

Theorem 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:

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

It helps when estimating the probabilities of counterfactuals from observational studies.

$$\begin{aligned} P(Y_x = y) &= \sum_z P(Y_x = y|Z = z)P(Z = z) \\ &= \sum_z P(Y_x = y|Z = z, X = x)P(Z = z) \\ &= \sum_z P(Y = y|Z = z, X = x)P(Z = z). \end{aligned}$$

Backdoor criterion

Theorem 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:

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

It helps when estimating the probabilities of counterfactuals from observational studies.

$$\begin{aligned}
P(Y_x = y) &= \sum_z P(Y_x = y|Z = z)P(Z = z) \\
&= \sum_z P(Y_x = y|Z = z, X = x)P(Z = z) \\
&= \sum_z P(Y = y|Z = z, X = x)P(Z = z).
\end{aligned}$$

Difference between post intervention and pre intervention

Example with college, skill and salary

References

- *Causal inference in statistics: a primer*, Pearl, Glymour and Jewell, 2016
- *Causality*, Pearl, 2000
- *Causation, Prediction, and Search*, Spirtes, Glymour, Scheines, 1993