

F. Counterfactuals

- So far in the course we have only discussed predictive probabilities ('effects of causes'):
 - "what is the probability that Nokia moves its headquarters away from Finland if we raise the taxes?"
 - "what will happen if we don't stop global warming?"
 - "what is the probability that the patient recovers if we decide to give / not to give this new drug?"
 - ...

These can all be written in the form $P(y \mid \text{do}(x), z)$

- What about counterfactuals?
 - "would I have caught the bus had I ran?"
 - "had Pitkämäki won gold in Helsinki 2005, had it not rained?"
 - "how would the second world war have ended, had Hitler managed to build nuclear weapons?"
- If we want some probabilities for the following statements ('causes of effects') being true, this also requires counterfactual reasoning...
 - "the rain was the reason that Pitkämäki did not get a medal"
 - "the medicine the doctor prescribed caused the death of the patient"
 - "unemployment would be much lower if the main opposition party had been in power"

These cannot be written in the form $P(y \mid \text{do}(x), z)$

- Example:

One summer 100 persons came to the hospital with a very serious disease. The doctor randomly prescribed a new drug to half (50) of the patients, the rest received a placebo. In both groups half (25) recovered, the rest died:

		$y = 1$	$y = 0$
		recovered	died
$x = 1$	medicine	0.25	0.25
$x = 0$	placebo	0.25	0.25

Now a relative of one of the dead blames the hospital, claiming that the patient might have recovered had she received the drug (rather than the placebo that she actually received). Is it possible to prove her claim right or wrong based on the data?

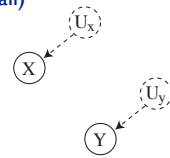
- **Model 1:** (the drug has no effect at all)

$$P(u_x = 1) = 0.5$$

$$P(u_y = 1) = 0.5$$

$$x := u_x$$

$$y := u_y$$



- **Model 2:** (the drug has an effect: those who without medicine would die actually recover, but on the other hand those who without medicine would recover now die)

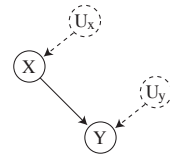
$$P(u_x = 1) = 0.5$$

$$P(u_y = 1) = 0.5$$

$$x := u_x$$

$$y := (1 - x)u_y + x(1 - u_y)$$

(Note: XOR-function, one input decides whether to flip the bit of the other!)



- Both models produce the same $P(x, y)$ (i.e. they cannot be distinguished by observation)
- Both models produce the same $P(y | \text{do}(x))$ (i.e. they cannot be distinguished by randomized experiments)
- But: the models give different answers to the relative's claim:
 - If model 1 is correct then the claim is false, the patient would have died even if given the drug
 - If model 2 is correct then the claim is correct, the patient would indeed have recovered if the drug had been given



- **Example 2 (Dawid):**

X : Decision to take aspirin or not (binary)

Y : Log-time until headache goes away (continuous)

We will model Y , conditional on setting X , as normal:

$$p(y | \text{do}(x)) = (2\pi)^{-1/2} \exp \left[-\frac{(y - \mu_x)^2}{2} \right]$$

Now, deciding whether or not to take aspirin is a standard decision-theoretic problem, which involves deciding which of the two distributions one 'prefers'. (The parameters μ_x can be estimated using randomized controlled trials.) This is an 'effects of causes' problem.

But what about answering counterfactuals ('causes of effects')? (on next slide)



Suppose I took aspirin ($\text{do}(X = 1)$) and observed my headache go away on log-scale as $Y = y$, how long would it have taken for my headache to go away had I not taken the aspirin ($\text{do}(X = 0)$)?

Using the simple model we have, one cannot answer this question.

Potential response approach (Rubin 1974; Rubin 1978):

Prior to taking our decision, there exists definite values Y_0 and Y_1 that define what will happen under either decision (i.e. either value of X). These values are of course unknown to us so they are random variables. For any given decision we only get to see the realization of one of them. Nevertheless, they both 'exist' and so there is a joint distribution $p(y_0, y_1)$ that governs their behavior.

Now, given that we decide to take aspirin and we observe the result, the answer to the counterfactual query of how long the headache would have lasted had we not taken aspirin is given by $p(y_0 | y_1)$



But how can we know $p(y_0, y_1)$?

One **cannot estimate it** from observed data, not even from randomized controlled trials, because one only gets to see one of the two values for any given data point! (I.e. we can estimate the marginals $p(y_0)$ and $p(y_1)$ but not the joint distribution.)

So somehow the correlation structure (dependency) between the 'potential response' variables must come from some 'background knowledge'... we discuss this in more detail a bit later.

- **Pearl:** To answer **counterfactuals** one needs **functional** causal models (equiv to potential response models). Causal Bayesian networks are not enough...

...**but:** yes they are enough, as long as we have hidden variables (which we often need anyway) which are **persistent!** For instance, model 2 can be written as a causal Bayesian network as follows:

$$P(u_x = 1) = 0.5$$

$$P(u_y = 1) = 0.5$$

$$P(x = 1 \mid u_x = 1) = 1$$

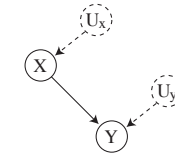
$$P(x = 1 \mid u_x = 0) = 0$$

$$P(y = 1 \mid x = 0, u_y = 0) = 0$$

$$P(y = 1 \mid x = 1, u_y = 0) = 1$$

$$P(y = 1 \mid x = 0, u_y = 1) = 1$$

$$P(y = 1 \mid x = 1, u_y = 1) = 0$$



You might say:

“But the previous example is, in practice, a functional model, since all the probabilities are either 0 or 1.”

Yes, but imagine changing these to 0.001 and 0.999. In this situation, we could obviously say that the patient would very probably have survived had she been given the medicine.

So: Counterfactuals do not require functional models, but rather can equally well be answered using causal Bayesian networks which **include hidden variables which can be persistent.**

I.e.: Functional causal models and causal Bayesian networks are still equivalent in this way, even with regards to counterfactuals!

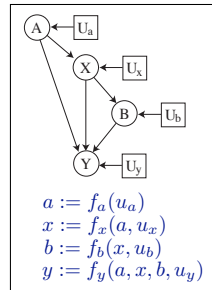
Causal Bayesian network vs functional causal model:

- A causal Bayesian network can always be represented as a **functional model**, such that the ‘unexplained’ undeterminacy is explained using the ‘disturbance’ variables. (Note: many different representations, which are equivalent in predictive terms but not necessarily equivalent for counterfactuals!)
- A functional model can always be represented as a **causal Bayesian network**, such that the disturbance variables are represented using the conditional distributions or as hidden variables. (Note: for counterfactual reasoning one must be careful about which variables are persistent and which are not!)

(From here on, for simplicity, we will follow Pearl and describe how to calculate counterfactuals using functional models)

Counterfactuals, in a functional model:

- Observe $X = x$ and $Y = y$
- What is the probability, that Y would have attained the value y' if X had been x' ? (here y and y' can be equal but $x' \neq x$)
- Variables A and B can be either observed or hidden, but **the full model (graph, functions, and $P(\mathbf{U})$) is assumed to be known**
- Interpreting the question: We assume a **minimal** change of mechanism, i.e. we set X into state x' without changing anything else, i.e. $\text{do}(X = x')$
- Interpreting the question: We assume that the disturbance variables $\mathbf{U} \setminus U_x = \{U_a, U_b, U_y\}$ are **persistent**, i.e. do not change

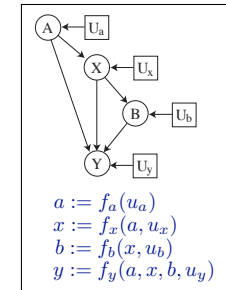


- With these specifications, we have a well defined probability:

$$P(Y_{x'} = y' \mid X = x, Y = y)$$

$Y_x(u)$ = 'the value of Y , when the disturbance variables attain the values u and X is set to equal x '.

- Example:
 X : rain
 Y : medal/no medal for Pitkämäki
 A : season
 B : slippery

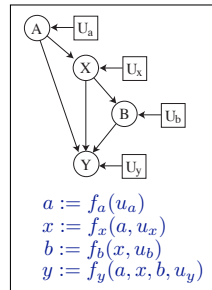


- But, how to calculate $P(Y_{x'} = y' \mid X = x, Y = y)$?

(Pearl theorem 7.1.7)

Three steps:

- ('abduction'): Calculate the probability distribution over all disturbances, given the evidence e , i.e. $P(\mathbf{U} \mid e)$
- ('action'): Change the model by the intervention $\text{do}(X = x')$, i.e. remove all arrows into X and set its value to x'
- ('prediction'): Using the updated model, and the probability distribution $P(\mathbf{U} \mid e)$, calculate $P(Y = y')$

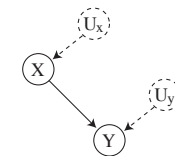


Example

- The model is given:

u_x	0	1	
$P(u_x)$	0.3	0.7	
u_y	0	1	2
$P(u_y)$	0.2	0.5	0.3

	u_y		
$y :=$	0	1	2
$x := u_x$	0	1	0
x	1	1	0



- What is the probability that y had attained the value 1 if x had been 1, given that we observed $x = 0$ and $y = 0$?

$$i. P(u_y \mid e) = \frac{u_y}{0 \quad 1 \quad 2} \Big| \frac{0}{0.5/0.8} \quad \frac{2}{0.3/0.8}$$

ii. $\text{do}(x = 1)$, i.e. cut all arrows into x

$$iii. P(Y_{X:=1} = 1 \mid X = 0, Y = 0) = 0.5/0.8 = 0.625$$



Example 2

[on the board]

Causal explanation, causal utterances and their interpretation

- "X is a cause of Y", if there exist two values x and x' of X and a value u of U such that $Y_x(u) \neq Y_{x'}(u)$
- "X is a cause of Y in the context $Z = z$ " if there exist two values x and x' of X and a value u of U such that $Y_{xz}(u) \neq Y_{x'z}(u)$
- "X is a direct cause of Y" if there exist two values x and x' of X and a value u of U such that $Y_{xr}(u) \neq Y_{x'r}(u)$, where r is some realization of $(V \setminus \{X, Y\})$
- "X is an indirect cause of Y" if X is a cause of Y and X is not a direct cause of Y
- "Event $X = x$ always causes $Y = y$ " if
 1. $Y_x(u) = y$ for all u ; and
 2. there exists a value u of U such that $Y_{x'}(u) \neq y$ for some $x' \neq x$

- "Event $X = x$ may have caused $Y = y$ " if
 1. $X = x$ and $Y = y$ are true; and
 2. there exists a value u of U such that $X(u) = x$, $Y(u) = y$ and $Y_{x'}(u) \neq y$ for some $x' \neq x$
- "The unobserved event $X = x$ is a likely cause of $Y = y$ " if
 1. $Y = y$ is true; and
 2. $P(Y_x = y, Y_{x'} \neq y \mid Y = y)$ is high for all $x' \neq x$
- "Event $Y = y$ occurred despite $X = x$ " if
 1. $X = x$ and $Y = y$ are true; and
 2. $P(Y_x = y)$ is low

and we might also mention the concepts...

- 'sufficient cause'
- 'necessary cause'

When do we need counterfactuals?

- Pearl:
 - **Law:** did the defendant's actions cause the death of the victim, or would he have died anyway? (e.g.: the responsibility of the doctor for the death of a patient)
 - **The importance of context in decision-making:** Decisions are seldom made in an information vacuum, but rather one always has to take into account the prevailing circumstances

- Criticism directed towards counterfactuals

“If we cannot distinguish models, which give different answers to counterfactuals, even with randomized experiments, how can we ever be sure that the answers we get are correct?”

Example. XOR-networks from before:



Example: heads or tails? We bet 5 euros, pick heads. Turns out tails, and we lose our money. What would have happened if we had picked tails, would we have won?

Intuitively the answer is yes...

But: can we be sure that the variable ‘the result of the coin toss is tails’ is persistent, but the variable ‘the result of the coin toss is the same as our guess’ is not? Could be difficult to argue that we know which one is persistent and which is not.

“Even if we knew the full model, how can we be sure that all the disturbance variables are persistent?”

Example: Often the disturbance variables are decisions made by other people, decisions which they take after we make ours. If indeed we have ‘free will’ to choose at the time of the decision, don’t they also?

...but the strongest argument (in my view) against counterfactuals is:

A counterfactual question is never the right question!

Example:

A player takes decisions on behalf of his team. The rules of the game are that the player must guess the result of the throw of a die. He can either guess ‘it’s a one’ or ‘it’s anything other than a one’. If the player guesses correctly the team wins a euro. It is known that the die is fair.

The player choose ‘it’s anything other than a one’. The result is a one. Can the team now blame the player for the failure. Was his decision bad?

Answer (in my mind): Of course not. It was an excellent decision.

Example 2:

A doctor performs exactly as dictated by all rules and the latest medical knowledge, and is very careful in all procedures. But the patient dies nonetheless. Turns out (by later research and understanding) that the decisions the doctor took 'caused' the death, in the sense that had he taken different decisions the patient would have lived. Is now the doctor guilty, should he pay fines or spend time in jail?

Answer (in my mind): Of course not.

Example 3:

The defendant tried to kill the victim, but failed. Should he be punished as hard as if he had succeeded.

Answer (in my mind): Of course he should.

In all the above examples the common idea is that the merit of the decisions should be based on the available knowledge at the time of the decision and on the expected result of the decision!

What about the second argument: context in decision-making?

- Of course the context should be taken into account when making decisions, but it does not require counterfactuals!

Example:

“Would I have caught the bus had I ran?” The question is not in any way interesting except in terms of helping me make decisions in the future, in a similar context.

So the real question should be: “When I find myself in a similar situation, with what probability will I catch the bus if I decide to run?” This is a regular predictive do-probability.

But why is finding the 'causes' of events then so important?

- Example: What was the cause of the plain crash?

The answer seems (to me) that we do not usually know the full model, and the point of the investigation is to find out more about causal relationships (in general) so that we can make the appropriate interventions so that such accidents don't happen again.

If we really new the whole causal model precisely, then trying to determine the 'cause' of some particular event is not actually important... (?)

Required reading

- Part 4 of chapter 1 of 'Causality'...
<http://bayes.cs.ucla.edu/BOOK-2K/ch1-4.pdf>
- Email exchange between Pearl and myself (Feb 2006)
<http://www.cs.helsinki.fi/patrik.hoyer/teaching/causality/emails/>
- Sections 4 and 5 of Glenn Shafer's paper:
<http://www.glennshafer.com/assets/downloads/other12.pdf>

Recommended reading

- Chapters 2 and 3 of Dawid's text:
<http://www.ucl.ac.uk/Stats/research/Resrpts/psfiles/rr279.pdf>

