# Overview and Outline

Sources: (a) Elements of Causal Inference, Peters et al., 2017. → (aka "text")
(b) Causality, J. Pearl, 2009. → (aka "causality book")

Note: Though we mostly follow (a), we should emphasize that many of the original papers (and theoretical framework) are by Pearl and colleagues. We follow (a), as it is easier to follow for someone with a prob./ML background.

## Correlation vs. Causation: Three motivating examples

### (1) Rain and Wetness:

Consider the following thought experiment:
Every day, I stand outdoors for an hour. I record the following pair of (Bernoulli) r.v.s:

$$(R, W) = (\text{Did it rain during the hour}, \text{Did I get wet during the hour})$$

$$R = \begin{cases} 0 & \text{no rain} \\ 1 & \text{rain} \end{cases} \qquad W = \begin{cases} 0 & \text{not wet} \\ 1 & \text{got wet} \end{cases}$$

For now, considering the population case (infinite samples), we can verify if R and W are correlated, simply by checking:

$$P(R=0 \mid W=0) \stackrel{?}{=} P(R=0 \mid W=1)$$

or

$$P(W=0 \mid R=0) \stackrel{?}{=} P(W=0 \mid R=1)$$

In either case, these are "testable" from infinite or "sufficient" amount of finite data (w.h.p.).

Now, moving to causality. I would like to test the following: I getting wet causes rain to occur, i.e., "wetness causes rain".

To test the truth of this statement from the previous data (aka observational data) seems impossible. However if we could conduct the following four experiments, we can test for the truth of the following two statements:

aka interventions

(A) "Wetness causes rain"

(B) "Rain causes wetness"

(I) For one year, each day I stand outside under an umbrella (whatever the weather conditions might be)

II Each day, I stand outside and get a pail of water poured on me.

III Each day, using a cloud seeder, rain is forced to occur.

IV Each day, using a giant fan, all clouds are blown out of the city.

(Of course, there are some implicit assumptions, that the actions that are taken does not alter the system beyond the specific variable — wetness in I, II and rain in III, IV — that is being altered.)

Given the above data, now compute and check:

computed with (interventional) data from experiment I

$$P^{\text{I}}(R=1 \mid W=0) \stackrel{?}{=} P^{\text{II}}(R=1 \mid W=1)$$

If these are equal, then Wetness DOES NOT influence Rain.

$$W \xrightarrow{?} R$$

Similarly using III, IV, we can test for "Rain causes wetness".

In this case, we will likely compute to see that:

$$P^{\text{III}}(W=1 \mid R=1) \neq P^{\text{IV}}(W=1 \mid R=0).$$

Thus, we can conclude that Rain has a causal effect on wetness.

$$R \longrightarrow W$$

Summary: Correlations can be tested with observational data. Causality is defined and tested through interventional data, i.e., data that is generated through intervening / actively modifying a variable (in this 2-variable example).

Note: We will later see that the dist. w.r.t the interventions I – IV are denoted by

$$P^{\text{I}}(R=1 \mid W=0) \equiv P^{do(W:=0)}(R=1 \mid W=0),$$

and similarly for other interventions

we are "do"ing $W \leftarrow 0$,
i.e., forcibly setting $W \leftarrow 0$.

Finally, the structural relationship between the variables after the above conclusion can be represented by a **Directed Graphical Model**:

$$R \longrightarrow W$$

(Also, in experiments I and II, we effectively deleted this edge and worked with $R \cdots \not\rightarrow W$)

where nodes represent the variables and directed edges indicate the direction of causation. A complete model for the discussion above would also specify the "noise" distributions, allowing one to fully specify the joint distribution on $(R, W)$, e.g.:

$$R := N_1 \qquad\qquad N_1, N_2 \text{ indep.}$$
$$W := \max\{R, N_2\} \qquad\qquad \text{Bernoulli}(p_i) \text{ noise}$$

This is called the **Structural Causal Model (SCM)**.

---

## 2 The Kidney Stone Dataset (Simpson's Paradox)

(Ref: Examples 6.37 and 6.16 in text, originally from Charig et. al. 1986)

Kidney stone recovery data from 700 patients

Successful Recovery Statistics

| | Overall Success | Patients with Small stones | Patients with Large stones |
|---|---|---|---|
| Treatment a: Open Surgery | 78% (273/350) | 93% (81/87) | 73% (192/263) |
| Treatment b: (small puncture surgery - percutaneous nephrolithotomy) | 83% (289/350) | 87% (234/270) | 69% (55/80) |

Apparent paradox: Overall, treatment b seems more effective. However, digging into the data, for each class (small / large kidney stone), treatment a is better.

This is the well-known Simpson's paradox, which shows that splitting data into categories can lead to a reversal of trend for every category, in comparison to the overall trend.

A causal perspective (see Pearl's book) argues that there is no paradox. More details on the history of Simpson's paradox and a causality perspective in paper below:

Edited version forthcoming, The American Statistician, 2014.

TECHNICAL REPORT
R-414
December 2013

# Understanding Simpson's Paradox

**Judea Pearl**
Computer Science Department
University of California, Los Angeles
Los Angeles, CA, 90095-1596
judea@cs.ucla.edu
(310) 825-3243 Tel / (310) 794-5057 Fax

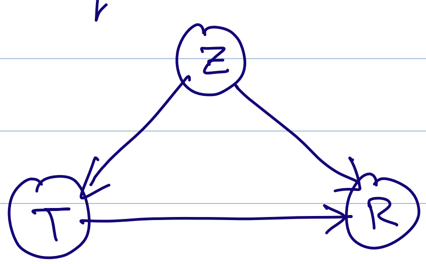
Simpson's paradox is often presented as a compelling demonstration of why we need statistics education in our schools. It is a reminder of how easy it is to fall into a web of paradoxical conclusions when relying solely on intuition, unaided by rigorous statistical methods.[1] In recent years, ironically, the paradox assumed an added dimension when educators began using it to demonstrate the limits of statistical methods, and why causal, rather than statistical considerations are necessary to avoid those paradoxical conclusions (Arah, 2008; Pearl, 2009, pp. 173–182; Wasserman, 2004).

In this note, my comments are divided into two parts. First, I will give a brief summary of the history of Simpson's paradox and how it has been treated in the statistical literature in the past century. Next I will ask what is required to declare the paradox "resolved," and argue that modern understanding of causal inference has met those requirements.

The intuition for the effect with kidney stones: Larger stones are more difficult to treat (and thus, lower success rate) with either treatment. The doctors thus prescribe Treatment a (which is perhaps more complicated/expensive) more frequently for larger stones compared to patients with smaller stones.

A causal, quantitative model of above:

$\mathbb{C}$ (causal model)




$Z =$ size of stone $\in \begin{cases} 0 & \text{small} \\ 1 & \text{large} \end{cases}$

$R =$ recovery / success of treatment $\in \begin{cases} 0 & \text{did not succeed} \\ 1 & \text{succeed} \end{cases}$

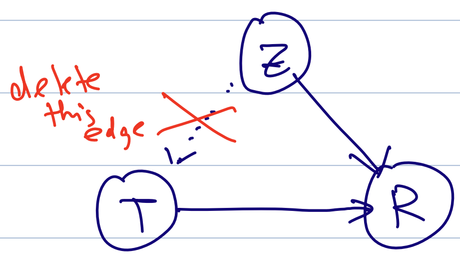
$T =$ Treatment $\in \begin{cases} 0 & \text{Treatment a} \\ 1 & \text{Treatment b} \end{cases}$

In the model above, the size of stone (Z) affects both treatment choice (T) as well as recovery (R).

In the dataset, an (incorrect) reading suggests that if we did not know the size of the kidney stone, treatment b is preferred. We will argue that this interpretation is incorrect.

What we need to answer is the following:

Suppose that we could intervene, and remove the effect of the kidney stone size on the choice of treatment, i.e., work with:




$C'$ (updated causal model)

Then, with this new model, we would compare

new model

$$P^{C'}(R=1 \mid T=0)$$ (treatment a)

vs.

$$P^{\mathbb{C}'}(R=1 \mid T=1) \qquad \text{(treatment } b\text{)}.$$

In the language of causality (Pearlian), this is written as (using notation from text; Pearl's notation is a bit different)

$$P^{\mathbb{C};\, do(T:=0)}(R=1)$$

$$P^{\mathbb{C};\, do(T:=1)}(R=1).$$

This notation emphasizes that these quantities are prob associated with the ORIG. causal model $\mathbb{C}$, but with an intervention do(...)

One of the main goals of the causal calculus is to compute the above "interventional" probabilities but using only observational data

We will later see that computations with the new model can sometimes be reduced to computations with the original model, BUT with a different "Total Probability Theorem (TPT)"

Here, it turns out:

$$P^{\mathbb{C}'}(R=1 \mid T=0) = P^{\mathbb{C};\, do(T:=0)}(R=1)$$

$$= P^{C}(R=1 \mid T=0, Z=0)\, P^{C}(Z=0)$$

note that this is NOT the usual TPT; conditioning $Z$ on $T$ is missing here.

$$+ P^{C}(R=1 \mid T=0, Z=1)\, P^{C}(Z=1)$$

The "Usual" Total Prob. Theorem

$$\boxed{\begin{aligned} P^{C'}(R=1 \mid T=0) = {} & P^{C'}(R=1 \mid T=0, Z=0) \cdot P^{C'}(Z=0 \mid T=0) \\ & + P^{C'}(R=1 \mid T=0, Z=1)\, P^{C'}(Z=1 \mid T=0) \end{aligned}}$$

using original data without any new experiments

$$= 0.832$$

$$P^{C'}(R=1 \mid T=1) = P^{C;\,do(T:=1)}(R=1)$$

$$= P^{C}(R=1 \mid T=1, Z=0)\, P^{C}(Z=0)$$

$$+ P^{C}(R=1 \mid T=1, Z=1)\, P^{C}(Z=1)$$

$$= 0.782$$

∴ Treatment a is better than Treatment b even if we did not know the size of the stone. There is no paradox.

This difference: $P^{C;\,do(T:=0)}(R=1) - P^{C;\,do(T:=1)}(R=1)$ is called the **AVERAGE CAUSAL EFFECT (ACE)** for binary treatment choices.

3a. $X$ : treatment

$Y$ : effect

$Z$ : confounder $\rightarrow$ all other things that could potentially affect both $X, Y$.

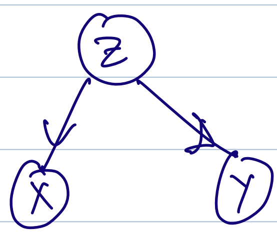
Goal: We want to determine if $X$ (the treatments) causes any change in $Y$ (the effect).

Supposing that we collect arbitrarily large dataset of $(x_i, y_i, z_i)$, sufficiently large to learn the joint dist. $p(x, y, z)$.

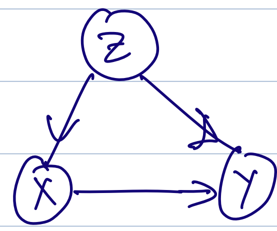
The question here becomes :

| | |
|---|---|
| $H_0 : X \perp\!\!\!\perp Y \mid Z$<br>independent    conditioned on | $\mathcal{H}_1 : X \not\!\perp\!\!\!\perp Y \mid Z$ |
| i.e., if we control for $Z$ (everything else in the world), then the treatment has no effect. | i.e., despite controlling for everything else, the treatment has an effect. |

Thus, testing for conditional independence (CI) is central to learning causal models; here distinguishing between




$H_0$ : null hypothesis (treatment has no effect)



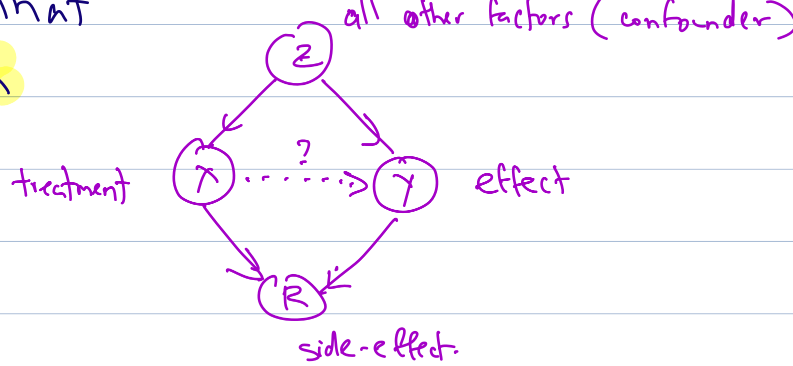

$H_1$ : alternative (treatment has effect)

Furthermore, we will see that we do not need any additional experiments / interventions in this case* to learn the presence / absence of the edge (with assumptions)

---

(3b.)

Suppose instead that the ground-truth model was:




In this case, suppose that the dotted edge was truly not present, (meaning the treatment did not truly influence the effect). Then, using the data $(x, y, r, z)$ and controlling on $(R, Z)$ will lead to a wrong conclusion. Namely, it will look like X has an effect on Y. This is because conditioning on (R and Z) conditionally correlates (X, Y) despite there being no causal relation (we will see Berkson's paradox later).

in linear regression setting, these are the independent variables.

Thus the covariates need to be carefully chosen, such that spurious conditional dependencies do not creep in.

---

Summary: Causal inference involves:

1. Reasoning about dependencies in a family of related distributions (the observation dist along with the intervention dists).

2. Controlling for confounding variables when reasoning about cause and effect.

Both these tasks require us to determine conditional independence relationships among the observed variables.

The roadmap from here on:

2. The mathematical plumbing needed for causal reasoning
   — Indep., Conditional Indep (CI), Directed Graphical Models.
3. Interventions. (Reasoning about given models).
4. Learning Causal Models
5. Instrument variables
6. CI Testing