What paradox?

Causal Models and Statistical Confusion

Oisín Ryan

Department of Methodology and Statistics
Faculty of Social Sciences
Utrecht University

April 7, 2022

Drug	No drug

	Drug	No drug
Male		
Female		

	Drug	No drug
Male	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
Female	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)

	Drug	No drug
Male	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
Female	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

	Drug	No drug
Male	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
Female	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

Should we prescribe the drug?

	Drug	No drug
Low Blood Pressure	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
High Blood Pressure	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

	Drug	No drug
Low Blood Pressure	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
High Blood Pressure	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

Should we prescribe the drug?

Simpsons Paradox

Statistical phenomena where a relationship which is present when aggregating over the population may be reversed or absent when looking at sub-populations

Simpsons Paradox

Statistical phenomena where a relationship which is present when aggregating over the population may be reversed or absent when looking at sub-populations

Berksons Paradox

Two phenomena which are statistically *independent* in the general population are statistically *dependent* in a sub-population

Simpsons Paradox

Statistical phenomena where a relationship which is present when aggregating over the population may be reversed or absent when looking at sub-populations

Berksons Paradox

Two phenomena which are statistically *independent* in the general population are statistically *dependent* in a sub-population

Lord's Paradox

The relationship between a categorical exposure and a continuous outcome is reversed when we condition on a third variable

Simpsons Paradox

Statistical phenomena where a relationship which is present when aggregating over the population may be reversed or absent when looking at sub-populations

Berksons Paradox

Two phenomena which are statistically *independent* in the general population are statistically *dependent* in a sub-population

Lord's Paradox

The relationship between a categorical exposure and a continuous outcome is reversed when we condition on a third variable

Confusing, but **not a paradox**

Simpsons Paradox

Statistical phenomena where a relationship which is present when aggregating over the population may be reversed or absent when looking at sub-populations

Berksons Paradox

Two phenomena which are statistically *independent* in the general population are statistically *dependent* in a sub-population

Lord's Paradox

The relationship between a categorical exposure and a continuous outcome is reversed when we condition on a third variable

Confusing, but **not a paradox**

You're asking a question that statistics alone is not equipped to answer

Estimand	Estimator	Estimate

Estimand **Estimator** **Estimate**

Statistics in a nutshell

Estimand

Estimator

1 Prepare Chocolate Cake Batter

Preheat oven to 350 degrees, and prepare Yo's Ultimate Chocolate Cake batter. Prepare your pans with parchment. Pour 2 ½ lbs into each 7" round pan, 1 ½ lbs into your 6" round pan, and divide the remaining batter evenly between your 5" round pans.

2 Bake Cakes

Bake your 7" round cakes for 50 minutes, your 6" round cake for 40 minutes, and your 5" round cakes for 30 minutes, or until a toothpick comes out clean. Set aside to cool completely in their pans on a wire rack.

3 Prepare Fillings & Simple Syrup

Prepare your dark chocolate ganache, Italian meringue buttercream, and simple syrup. Set aside until you're ready to decorate.

4 Level Cakes

Remove your cooled cakes from their pans and level them with a ruler and serrated knife.

Estimate

Estimand

Estimator

1. **Prepare Chocolate Cake Batter**
   Preheat oven to 350 degrees, and prepare Yo's Ultimate Chocolate Cake batter. Prepare your pans with parchment. Pour 2 ½ lbs into each 7" round pan, 1 ½ lbs into your 6" round pan, and divide the remaining batter evenly between your 5" round pans.

2. **Bake Cakes**
   Bake your 7" round cakes for 50 minutes, your 6" round cake for 40 minutes, and your 5" round cakes for 30 minutes, or until a toothpick comes out clean. Set aside to cool completely in their pans on a wire rack.

3. **Prepare Fillings & Simple Syrup**
   Prepare your dark chocolate ganache, Italian meringue buttercream, and simple syrup. Set aside until you're ready to decorate.

4. **Level Cakes**
   Remove your cooled cakes from their pans and level them with a ruler and serrated knife.

Estimate

Credit to Peter Tennant @PWGTennant

	Drug	No drug
Male	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
Female	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

	Drug	No drug
Male	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
Female	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

Conditional Probabilities:

$$P(R = r | D = d, S = s)$$

	Drug	No drug
Male	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
Female	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

Marginal Probabilities:

$$P(R = r|D = d)$$

Estimand	**Estimator**	**Estimate**

Estimand	Estimator	Estimate
$P(R=1 \mid D=1, S=0)$	# Recovered takers Male / # Drug takers Male	$.93$

Statistics in a nutshell

Estimand	Estimator	Estimate
$P(R = 1 \mid D = 1, S = 0)$	# Recovered takers Male / # Drug takers Male	$.93$
$P(R = 1 \mid D = 1)$	# Recovered drug takers / # Drug takers	$.78$

What's the paradox?

Two different sets of **estimands** yield two different sets of **estimates**

- No paradox there!

What's the paradox?

Two different sets of **estimands** yield two different sets of **estimates**

- No paradox there!

We are not interested in either of these estimands *for their own sake*

What's the paradox?

Two different sets of **estimands** yield two different sets of **estimates**

- No paradox there!

We are not interested in either of these estimands *for their own sake*

We are interested in a **causal effect**

- Does taking the drug cause recovery?
- **Causal Estimand**
- But we have no way of expressing this in the language of statistics

What's the paradox?

Two different sets of **estimands** yield two different sets of **estimates**

- No paradox there!

We are not interested in either of these estimands *for their own sake*

We are interested in a **causal effect**

- Does taking the drug cause recovery?
- **Causal Estimand**
- But we have no way of expressing this in the language of statistics

Statistical estimand $\leftarrow$? $\rightarrow$ Causal Estimand

Causal Graphs

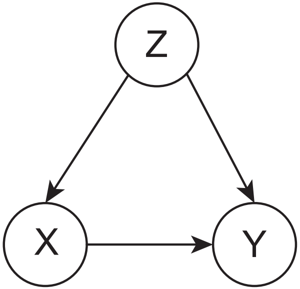
Causal Graphs

A causal graph is a diagram representing (our beliefs about) which variables share causal relations with each other

Causal Graphs

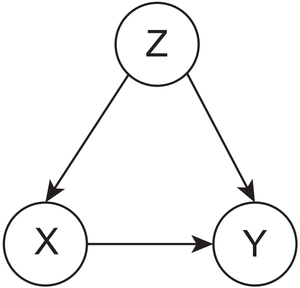
A causal graph is a diagram representing (our beliefs about) which variables share causal relations with each other

- The arrow $X \to Y$ represents our belief that X is a direct cause of Y
- We omit an arrow if expert knowledge tells us that one variable does not directly cause another. The *absence* of an arrow is a strong statement



A causal graph is a diagram representing (our beliefs about) which variables share causal relations with each other

- The arrow X $\rightarrow$ Y represents our belief that X is a direct cause of Y
- We omit an arrow if expert knowledge tells us that one variable does not directly cause another. The *absence* of an arrow is a strong statement



Directed Acyclic Graph (DAG) or Bayesian Network

Why Causal Models?

This machinery is useful for three important and closely related reasons:

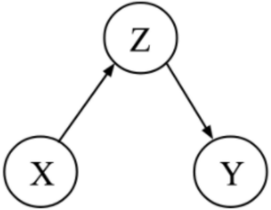
1. Causal models map causal dependencies onto statistical dependencies
   - *Regardless* of distributions and functional forms
2. Causal models allow us to define **causal effects** in the language of interventions and probabilities
3. Causal models tell us which when and how statistical estimands can act as causal estimands

Why SCMs?

This machinery is useful for three important and closely related reasons:

1. **Causal models map causal dependencies onto statistical dependencies**
    - *Regardless* of distributions and functional forms
2. Causal models allow us to define **causal effects** in the language of interventions and probabilities
3. Causal models tell us which when and how statistical estimands can act as causal estimands

Chain

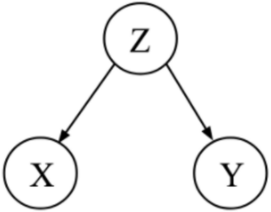


X: Smoking
Z: Tar
Y: Cancer

$X \not\perp\!\!\!\perp Y$

$X \perp\!\!\!\perp Y \mid Z$

Fork

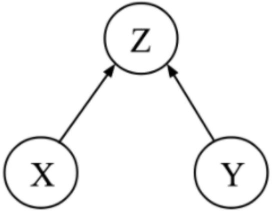


X: Storks
Z: Environment
Y: Babies

$X \not\perp\!\!\!\perp Y$

$X \perp\!\!\!\perp Y \mid Z$

Collider



X: Attractiveness
Z: Being Single
Y: Intelligence

$X \perp\!\!\!\perp Y$

$X \not\perp\!\!\!\perp Y \mid Z$

Why SCMs?

This machinery is useful for three important and closely related reasons:

1. Causal models map causal dependencies onto statistical dependencies
    - *Regardless* of distributions and functional forms
2. **Causal models allow us to define causal effects in the language of interventions and probabilities**
3. Causal models tell us which when and how statistical estimands can act as causal estimands

Causal Effects in SCMs

The **do-operator** $do(X = x)$ represents a "surgical intervention" to set the value of the variable X to a constant value x

- $do(D = 1)$ - the act of intervening such that everyone takes an aspirin

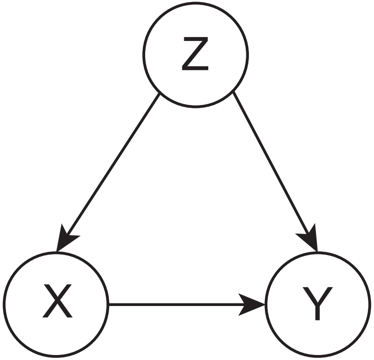
The **do-operator** $do(X = x)$ represents a "surgical intervention" to set the value of the variable X to a constant value x

- $do(D = 1)$ - the act of intervening such that everyone takes an aspirin

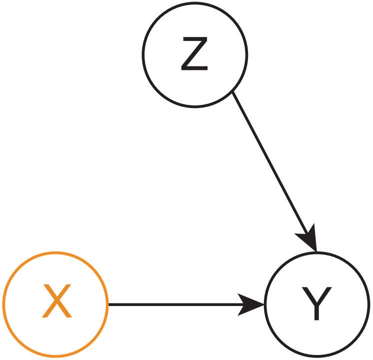
In the graph, a $do-$ operation on X cuts-off all incoming ties

Two versions of the causal system

Observing



Intervening



Causal Effects in SCMs

We can use the do-operator to define our **causal estimand**

Causal Effect of Drug-Taking on Recovery:

$$CE = P[R \mid do(D = 1)] - P[R \mid do(D = 0)]$$

We can use the do-operator to define our **causal estimand**

Causal Effect of Drug-Taking on Recovery:

$$CE = P[R \mid do(D = 1)] - P[R \mid do(D = 0)]$$

Inference problem: “Seeing” is not always the same as ”doing”

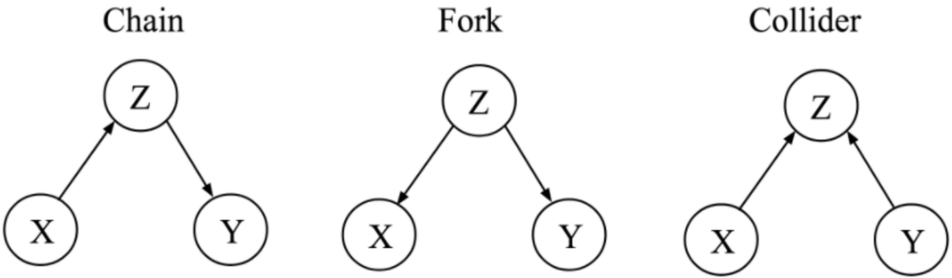
Observing $\neq$ Intervening:

$P[Y \mid X = x]$ is not **generally** the same as $P[Y \mid do(X = x)]$

Why SCMs?

This machinery is useful for three important and closely related reasons:

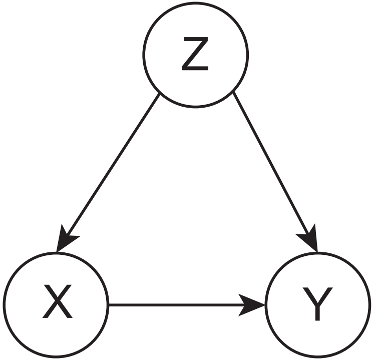
1. Causal models map causal dependencies onto statistical dependencies
   - *Regardless* of distributions and functional forms
2. Causal models allow us to define causal effects in the language of interventions and probabilities
3. **Causal models tell us which when and how statistical estimands can act as causal estimands**



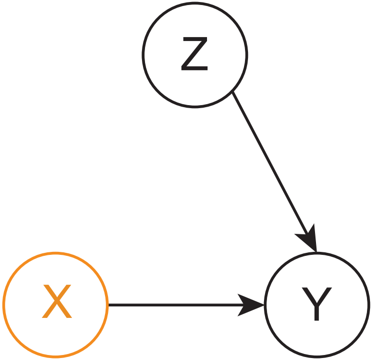
Chain	Fork	Collider
X: Smoking	X: Storks	X: Attractiveness
Z: Tar	Z: Environment	Z: Being Single
Y: Cancer	Y: Babies	Y: Intelligence
$X \not\perp\!\!\!\perp Y$	$X \not\perp\!\!\!\perp Y$	$X \perp\!\!\!\perp Y$
$X \perp\!\!\!\perp Y \mid Z$	$X \perp\!\!\!\perp Y \mid Z$	$X \not\perp\!\!\!\perp Y \mid Z$

Two versions of the causal system

Observing



Intervening



Statistics in a nutshell

Statistical Estimand

Estimator

1. **Prepare Chocolate Cake Batter**

   Preheat oven to 350 degrees, and prepare Yo's Ultimate Chocolate Cake batter. Prepare your pans with parchment. Pour 2 ½ lbs into each 7" round pan, 1 ½ lbs into your 6" round pan, and divide the remaining batter evenly between your 5" round pans.

2. **Bake Cakes**

   Bake your 7" round cakes for 50 minutes, your 6" round cake for 40 minutes, and your 5" round cakes for 30 minutes, or until a toothpick comes out clean. Set aside to cool completely in their pans on a wire rack.

3. **Prepare Fillings & Simple Syrup**

   Prepare your dark chocolate ganache, Italian meringue buttercream, and simple syrup. Set aside until you're ready to decorate.

4. **Level Cakes**

   Remove your cooled cakes from their pans and level them with a ruler and serrated knife.

Estimate

Causal Estimand	**Causal Model**	**Statistical Estimand**	**Estimator**	**Estimate**

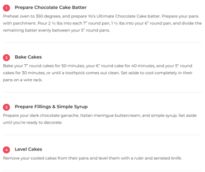
Causal Inference in a nutshell

Causal Estimand

Causal Model

Statistical Estimand

Estimator



1 **Prepare Chocolate Cake Batter**

Preheat oven to 350 degrees, and prepare Yo's Ultimate Chocolate Cake batter. Prepare your pans with parchment. Pour 2 ½ lbs into each 7" round pan, 1 ½ lbs into your 6" round pan, and divide the remaining batter evenly between your 5" round pans.

2 **Bake Cakes**

Bake your 7" round cakes for 50 minutes, your 6" round cake for 40 minutes, and your 5" round cakes for 30 minutes, or until a toothpick comes out clean. Set aside to cool completely in their pans on a wire rack.

3 **Prepare Fillings & Simple Syrup**

Prepare your dark chocolate ganache, Italian meringue buttercream, and simple syrup. Set aside until you're ready to decorate.

4 **Level Cakes**

Remove your cooled cakes from their pans and level them with a ruler and serrated knife.

Estimate

	Drug	No drug
Male	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
Female	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

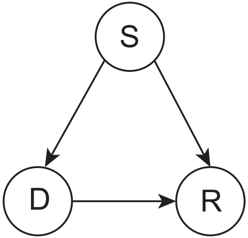
Statistical Estimand	Estimator	Estimate
$P(R=1 \mid D=1, S=1)$	# Recovered takers Male / # Drug takers Male	$.93$
$P(R=1 \mid D=1)$	# Recovered drug takers / # Drug takers	$.78$

Causal Inference in a nutshell

Causal Estimand

$$P[R \mid do(D=1)] - P[R \mid do(D=0)]$$

Causal Model



Statistical Estimand

$$P(R|D,S)$$

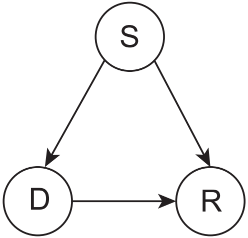
$$P(R|D)$$

Causal Inference in a nutshell

Causal Estimand

$$P[R \mid do(D=1)] - P[R \mid do(D=0)]$$

Causal Model



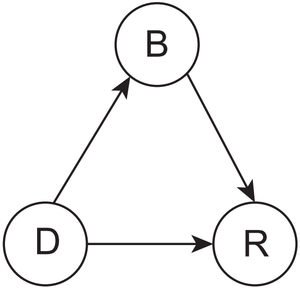
Statistical Estimand

$$P(R|D,S)$$

$$\cancel{P(R|D)}$$

Post-Treatment Blood Pressure:

- Statistical information is exactly the same
- The drug works in part by decreasing blood pressure
- We should **not** condition on blood pressure



	Drug	No drug
Low Blood Pressure	81 out of 87 recovered (93%)	234 out of 270 recovered (87%)
High Blood Pressure	192 out of 263 recovered (73%)	55 out of 80 recovered (69%)
Combined data	273 out of 350 recovered (78%)	289 out of 350 recovered (83%)

Example from Pearl, Glymour & Jewell, 2016

Simpsons Paradox

Statistical phenomena where a relationship which is present when aggregating over the population may be reversed or absent when looking at sub-populations

Absolutely not a paradox.

- Confusion comes from a lack of clarity regarding our **causal estimand** and **causal model**

Statistical information *alone* cannot provide the answer

- Different DAGs can produce the exact same statistical dependencies in observational data

Causal models provide immediate conceptual clarity

- Miguel Hernan: Draw your assumptions before your conclusions!

Conclusions

Inappropriate reliance on (advanced) statistical modeling with no clear link to causal estimands or models

- Paradoxes and confusion result. Machine learning is no solution

Causal modeling can be powerful in reshaping how we approach statistical modeling

- Judea Pearl, Don Rubin, Jamie Robins, Miguel Hernan, Angrist & Imbens
- Example: Controlling for as many variables as possible is **an obviously terrible idea** when estimating causal effects

Researchers make causal inferences based on observational data **all the time**

- Better to be explicit and open about this so we can move forward

Thanks!
(o.ryan@uu.nl | oisinryan.org)

Shameless plug

My own research focuses on using these ideas to improve psychological and social science research

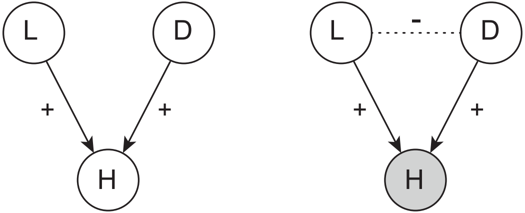
- Causal discovery (e.g. Ryan, Bringmann, Schuurman, in press)
- Causal estimands (e.g. Haslbeck*, Ryan*, Dablander* 2021)
- Constructing theories (Haslbeck*, Ryan*, Robinaugh*, Waldorp, Borsboom, 2021)
- Applications of causal inference (forthcoming)

Berksons Paradox

Two phenomena which are statistically *independent* in the general population are statistically *dependent* in a sub-population

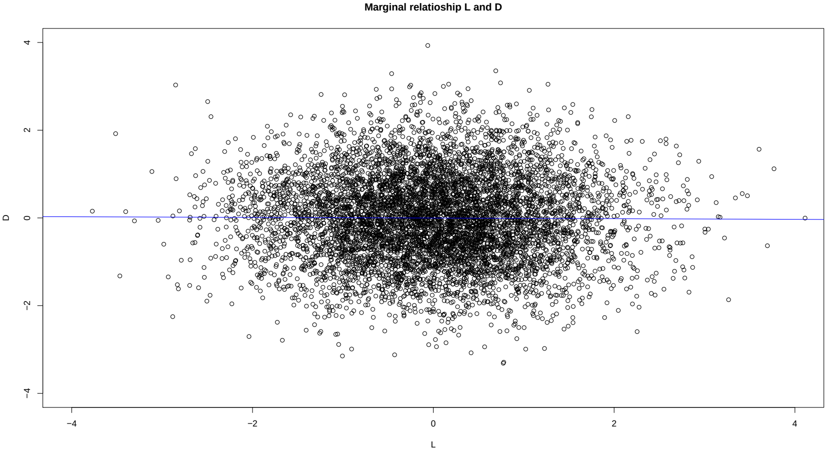
Classic example: We are interested in the relationship between *Lung Cancer* (L) and *Diabetes* (D)

- General population, these two variables are independent.
- In a sample of *hospital patients*, there is a negative dependency - patients who don't have diabetes are *more likely* to have lung cancer.



- Lung cancer L and diabetes D cause hospitalization H
- By taking participants from a hospital we *condition* on hospitalization $(H = 1)$
- If you are hospitalised, and you *don't* have diabetes, probably you do have lung cancer (Otherwise - why would you be in hospital?).
- $P(D|L = 1, H = 1) \neq P(D|do(L) = 1)$
- We have conditioned on a *collider*

Collider Bias



Collider Bias