UNIVERSITÀ DI PAVIA

Probabilistic Graphical Models and Causal Inference

Episode 2: Causal Graphical Models

Marco Piastra

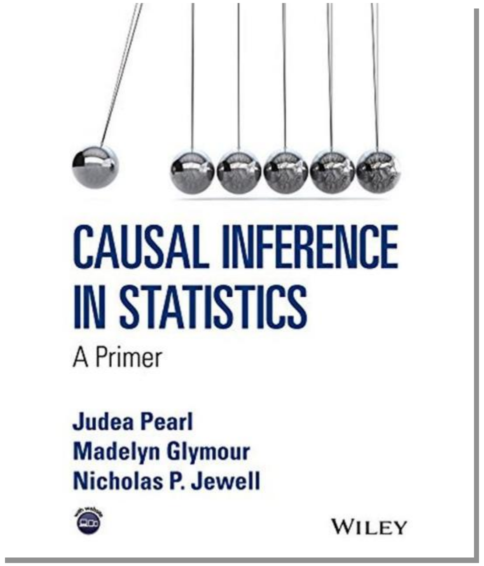
This presentation can be downloaded at:
https://vision.unipv.it/AI/AIRG.html

Very Good Readings

- **Causal Inference in Statistics**
  A Primer

  Judea Pearl, Madelyn Glymour and Nicholas P. Jewel

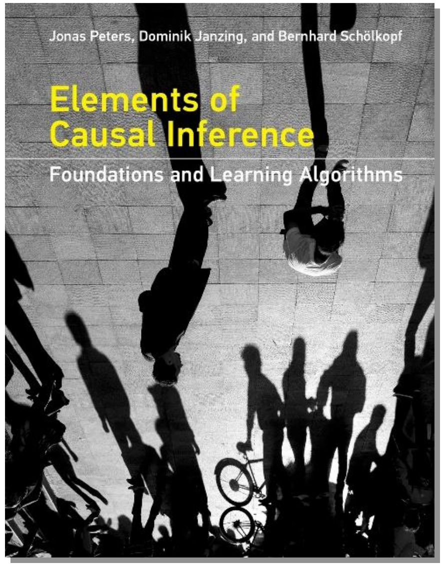
  Wiley, 2016



- **Elements of Causal Inference**
  Foundations and Learning Algorithms

  Jonas Peters, Dominik Janzing and Bernhard Schölkopf

  MIT Press, 2017



Expected Value of a Random Variable

Basic definition

$$\mathbb{E}_X[X] := \sum_{x \in \mathcal{X}} x\ P(X = x)$$

In a more concise notation:

$$\mathbb{E}[X] := \sum_{x} x\ P(x)$$

Continuous case

$$\mathbb{E}_X[X] := \int_{x \in \mathcal{X}} x\ p(x) dx$$

Probability density

Expectation is a linear operator

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$$

$$\mathbb{E}[cX] = c\mathbb{E}[X]$$

Conditional expectation

$$\mathbb{E}_X[X|Y = y] = \mathbb{E}[X|Y = y] := \sum_{x \in \mathcal{X}} x\ P(X = x|Y = y)$$

Simpson's Paradox

Causes and Effects: the Simpson's Paradox [1922]

- Does physical exercise prevent cholesterol?

Apparently not: correlation is *positive*

$$\rho(X,Y) := \frac{\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

Pearson correlation

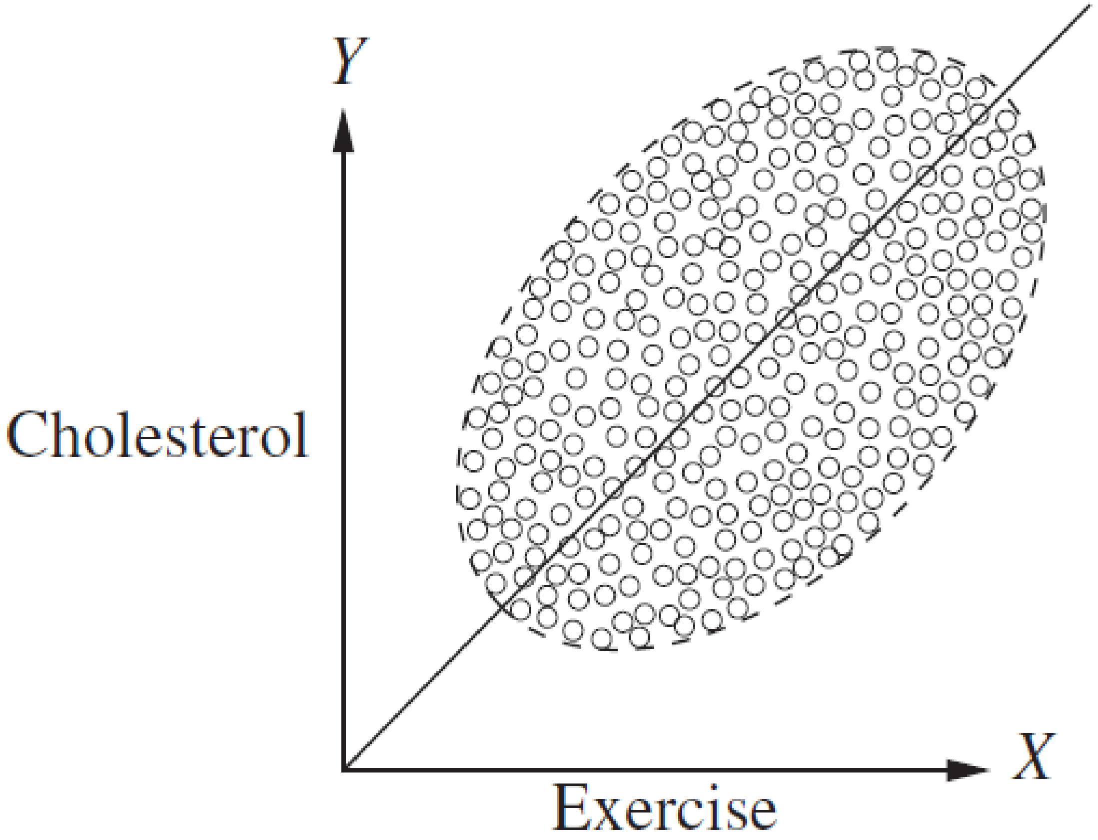
where:

$$\mu_X := \mathbb{E}_X[X]$$

$$\sigma_X := \sqrt{\mathrm{Var}(X)} = \sqrt{\mathbb{E}[(X-\mu_X)^2]}$$

standard deviation

In words:
Does <u>more</u> physical exercise *cause* <u>more</u> cholesterol?



[Image from Pearl, J. et al., "Causal Inference in Statistics: A Primer", Wiley, 2016]

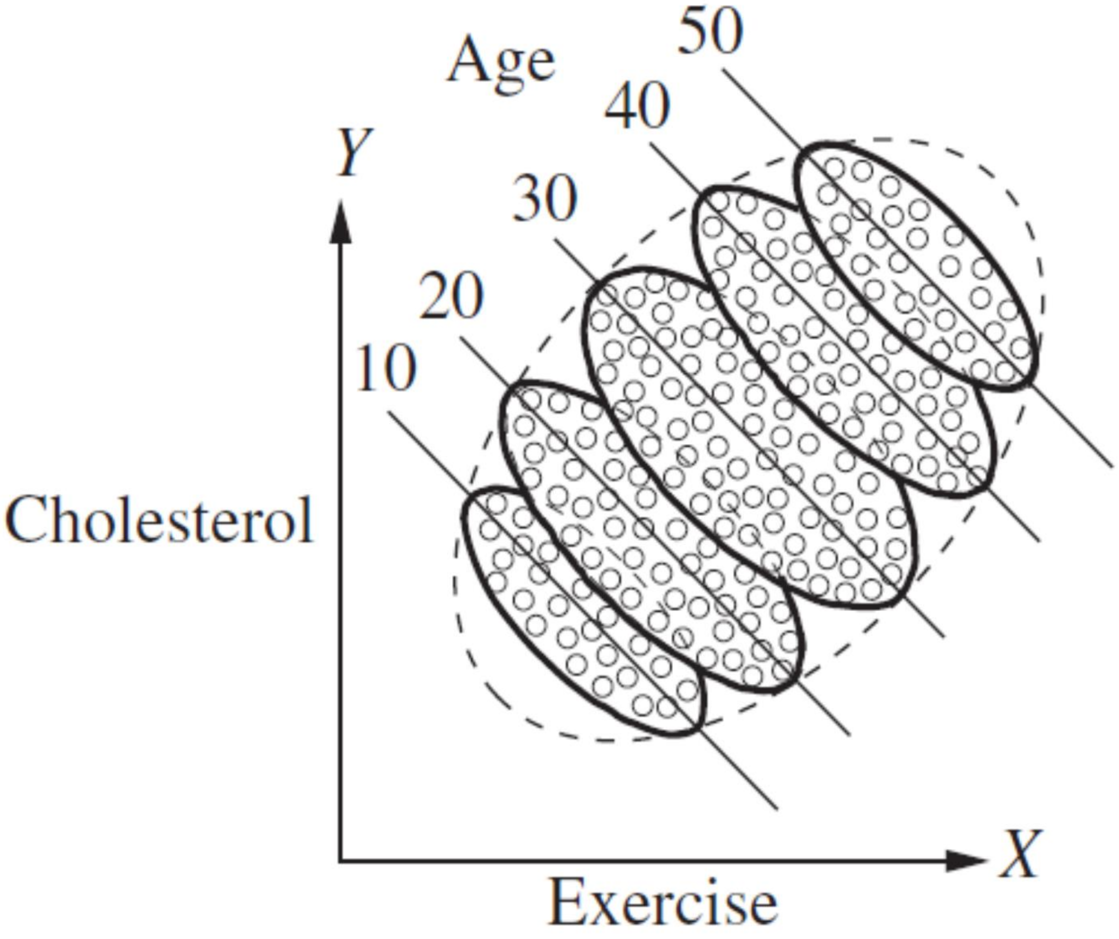
Causes and Effects: the Simpson's Paradox [1922]

- Does physical exercise prevent cholesterol?

    Maybe yes if we consider another variable...

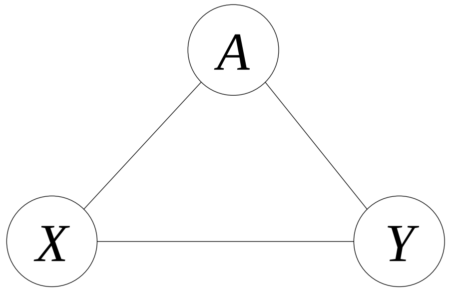
    Correlation in Age subgroups is *negative*

    In words:
    Does <u>more</u> physical exercise *cause* <u>less</u> cholesterol?

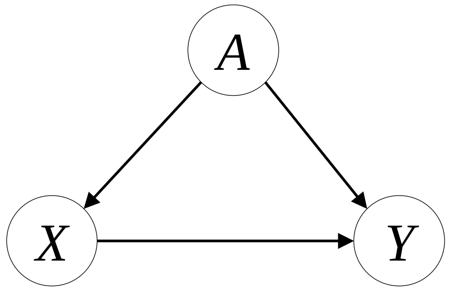


[Image from Pearl, J. et al., "Causal Inference in Statistics: A Primer", Wiley, 2016]

- Does physical exercise prevent cholesterol?

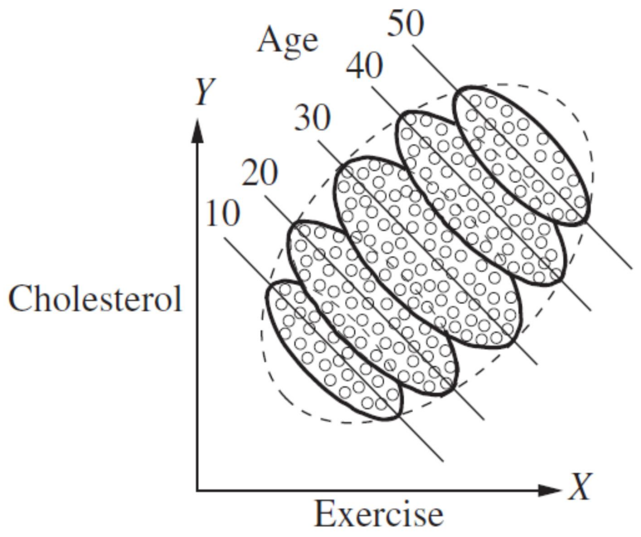


Undirected structure (a clique): no independence assumptions.
All DAGs built form it will be equivalent (just different factorizations)



Does this DAG make more sense from a causal viewpoint?

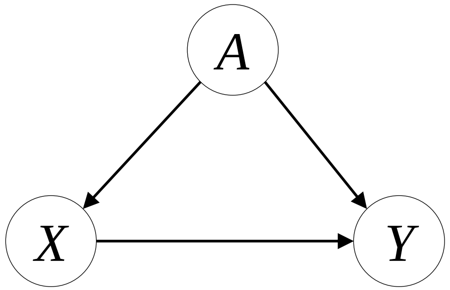
And what is the meaning of this, after all?

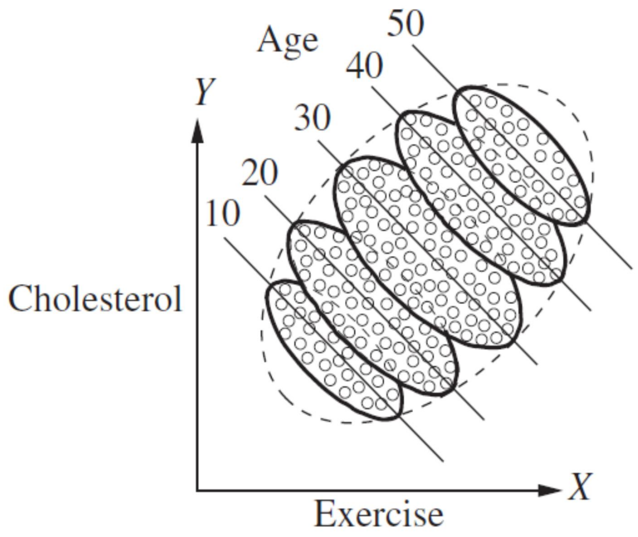


[Image from Pearl, J. et al., "Causal Inference in Statistics: A Primer", Wiley, 2016]

Causes and Effects: *say it with graphs*

- What is a cause?





A variable X is said to be a *cause* of a variable Y
if Y can change in response to changes in X

In a *Causal Graphical Model* (CGM), each parent is a *direct cause* of all its children

The notation is the same as with Probabilistic Graphical Model (PGM), the intended meaning is different

[Image from Pearl, J. et al., "Causal Inference in Statistics: A Primer", Wiley, 2016]

Reichenbach's Common Cause Principle

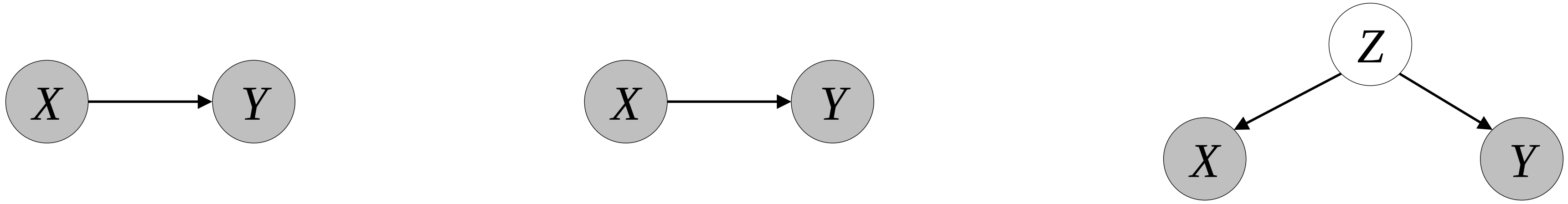
Given two <u>observed</u> variables X and Y
that are *dependent* on each other

$$\langle X \not\perp Y \rangle$$

There exists a third variable Z which *causally influences* both of them
and makes them *conditionally independent*

$$\langle X \perp Y \mid Z \rangle$$

As a special case, Z may coincide with either X or Y



Caution: such principle is not undisputed...

Causation, Dependence, Correlation

Correlation and Dependence

$$\rho(X,Y) := \frac{\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

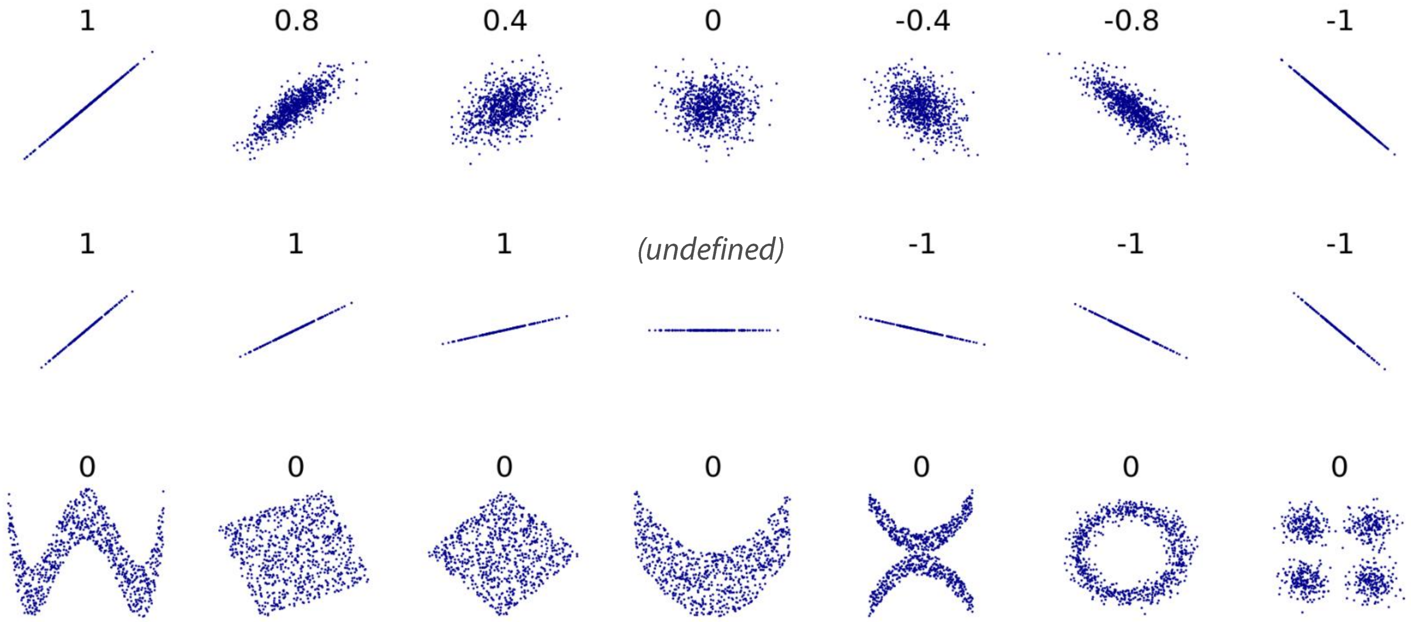
Pearson correlation

where:

$$\mu_X := \mathbb{E}[X]$$

$$\sigma_X := \sqrt{\mathrm{Var}(X)} = \sqrt{\mathbb{E}[(X-\mu_X)^2]}$$

Standard deviation



[Image from https://en.wikipedia.org/wiki/Pearson_correlation_coefficient]

Correlation and Dependence

$$\rho(X,Y) := \frac{\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

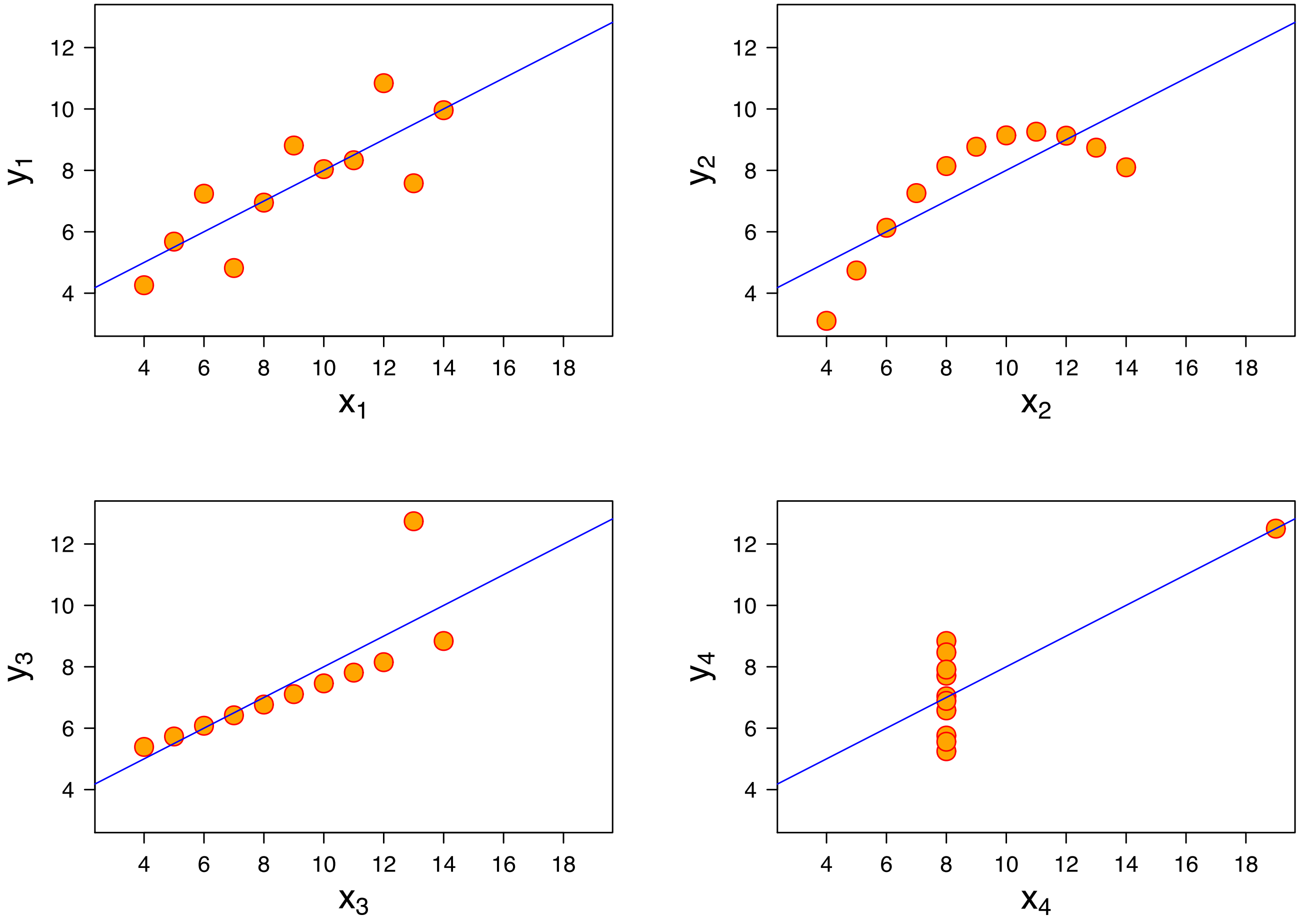
Pearson correlation

where:

$$\mu_X := \mathbb{E}[X]$$

$$\sigma_X := \sqrt{\mathrm{Var}(X)} = \sqrt{\mathbb{E}[(X-\mu_X)^2]}$$

Standard deviation



Anscombe Quartet: eleven points in 2D space - same mean, variance and correlation

[Image from https://en.wikipedia.org/wiki/Anscombe%27s_quartet]

Correlation and Dependence

Zero correlation does NOT imply independence

Does <u>independence</u> imply <u>zero correlation</u>?

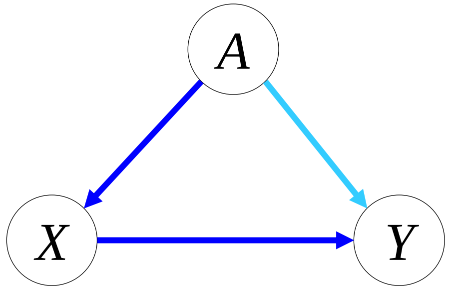
$$\rho(X,Y) := \frac{\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

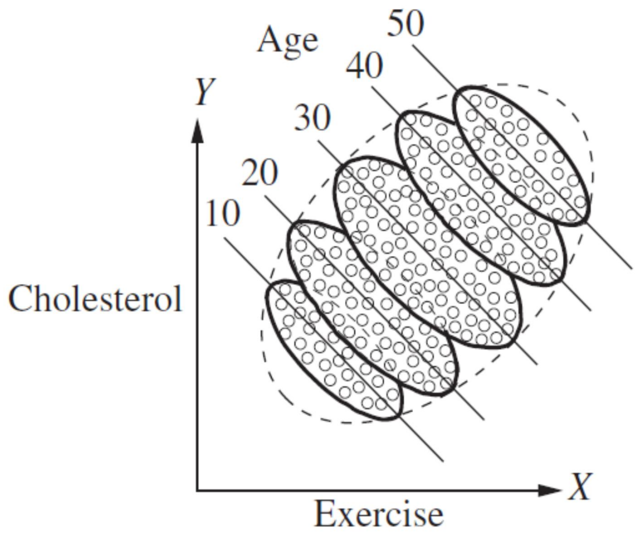
$$\begin{aligned}
\mathrm{Cov}(X,Y) &:= \mathbb{E}[(X-\mu_X)(Y-\mu_Y)] \\
&= \mathbb{E}[XY - X\mu_Y - Y\mu_X + \mu_X\mu_Y] \\
&= \mathbb{E}[XY] - \mu_Y\mathbb{E}[X] - Y\mu_X\mathbb{E}[Y] + \mu_X\mu_Y \\
&= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y] \\
&= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \\
&= \sum_{x,y} xy\ P(x,y) - \sum_x x\ P(x) \sum_y y\ P(y) \\
&= \sum_{x,y} xy\ P(x,y) - \sum_{x,y} xy\ P(x)P(y)
\end{aligned}$$

Covariance

So, the answer is *<u>yes</u>: the last term must be zero if the two variables are independent*

Causation and Dependence





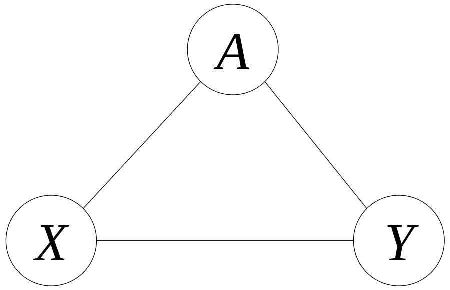
In a *Causal Graphical Model* (CGM), each parent is a *direct cause* of all its children

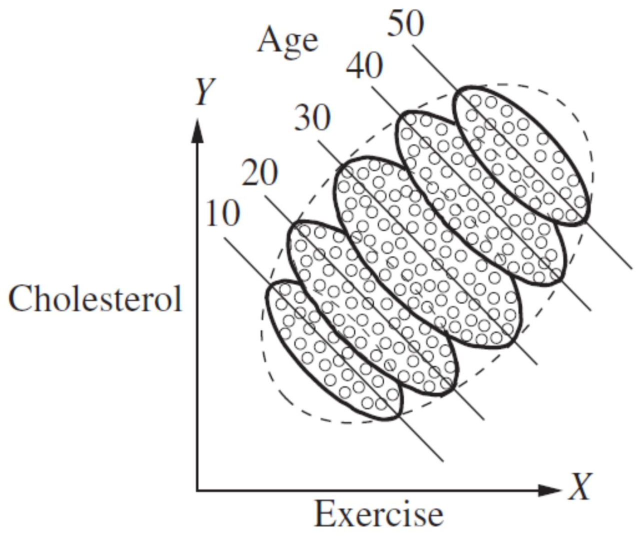
A chain of causes *along the arrows* is a causal path

- All variables along a path are dependent on each other
- The direction of arrows matters

Detecting the independence of random variables (absence of connections) from observations may be difficult, but even in that case ...

[Image from Pearl, J. et al., "Causal Inference in Statistics: A Primer", Wiley, 2016]





From Episode 1 of this course

- **Markov Equivalence Class**

  Two graphical models share <u>the same</u> independence assumptions when:

  1) they share the same *undirected* structure (i.e., *skeleton*)

  2) they share the same *joins* (a.k.a. *colliders*)

  (*) This holds true when some independence is expressed (i.e., if some links are missing).
  Any DAG built out of a clique will be equivalent, regardless of joins
  (i.e., no independence assumptions represented anyway)

[Image from Pearl, J. et al., "Causal Inference in Statistics: A Primer", Wiley, 2016]

Causation and Correlation

- Between any two random variables:

  Causation, either direct or through a path, implies dependence

  Non-zero correlation implies dependence but not vice-versa

  Both correlation and independence are *commutative* relations

$$\rho(X,Y) := \frac{\mathbb{E}[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y}$$

Pearson correlation

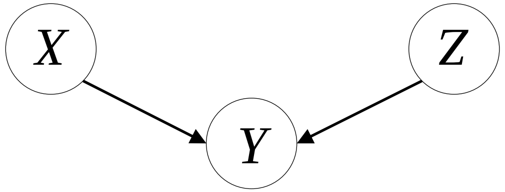
$$P(X,Y) = P(X)P(Y)$$

Independence

None of them will reveal the direction of arrows

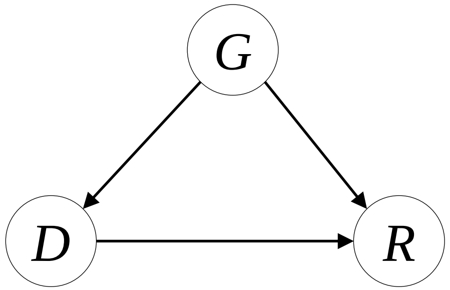
Except for independence in a non-clique collider

'join' or 'collider'



Say it with graphs

- What is a cause? (*Another example*)



G is *biological gender* (= Male/Female)
D is *drug administration* (= Yes(1)/No(0))
R is *recovery* from illness (= Yes(1)/No(0))

Experimental data

- In both groups, recovery rates are *higher* if drug is administered...
- ... while in the entire population, recovery rates are *lower*

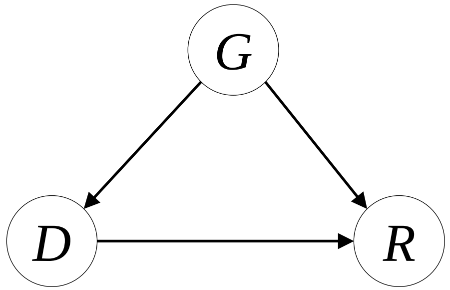
Females	$R = 0$	$R = 1$		Recovery Rate
$D = 0$	25	55	80	69%
$D = 1$	71	192	263	73%
	96	247	343	

Males	$R = 0$	$R = 1$		Recovery Rate
$D = 0$	36	234	270	87%
$D = 1$	6	81	87	93%
	42	315	357	

	$R = 0$	$R = 1$		Recovery Rate
$D = 0$	61	289	350	83%
$D = 1$	77	273	350	78%
	138	562	700	

[Data from Pearl, J. et al., "Causal Inference in Statistics: A Primer", Wiley, 2016]

- What is a cause? (*Another example*)



G is *biological gender* (= Male/Female)
D is *drug administration* (= Yes(1)/No(0))
R is *recovery* from illness (= Yes(1)/No(0))

Experimental data

- Note however that gender also influenced drug prescription...
- ... in fact, in this example, doctors were more likely to prescribe drug to males than to females

Females	$R = 0$	$R = 1$		Recovery Rate
$D = 0$	25	55	80	69%
$D = 1$	71	192	263	73%
	96	247	343	

Males	$R = 0$	$R = 1$		Recovery Rate
$D = 0$	36	234	270	87%
$D = 1$	6	81	87	93%
	42	315	357	

	$R = 0$	$R = 1$		Recovery Rate
$D = 0$	61	289	350	83%
$D = 1$	77	273	350	78%
	138	562	700	

[Data from Pearl, J. et al., "Causal Inference in Statistics: A Primer", Wiley, 2016]

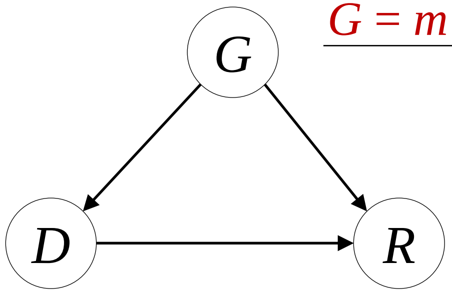
Causes and Effects: *say it with graphs*

- What is a cause? (*Another example*)

  Maximum Likelihood Estimation (CPTs)

$P(G)$

$G = f$	0.49
$G = m$	0.51



$P(D|G)$

	$G = f$	$G = m$
$D = 0$	0.23	0.76
$D = 1$	0.77	0.24

$P(R|G, D)$

	$G = f$, $D = 0$	$G = f$, $D = 1$	$G = m$, $D = 0$	$G = m$, $D = 1$
$R = 0$	0.31	0.27	0.13	0.07
$R = 1$	0.69	0.73	0.87	0.93

Females	$R = 0$	$R = 1$		Recovery Rate
$D = 0$	25	55	80	69%
$D = 1$	71	192	263	73%
	96	247	343	

Males	$R = 0$	$R = 1$		Recovery Rate
$D = 0$	36	234	270	87%
$D = 1$	6	81	87	93%
	42	315	357	

	$R = 0$	$R = 1$		Recovery Rate
$D = 0$	61	289	350	83%
$D = 1$	77	273	350	78%
	138	562	700	

[Data from Pearl, J. et al., "Causal Inference in Statistics: A Primer", Wiley, 2016]

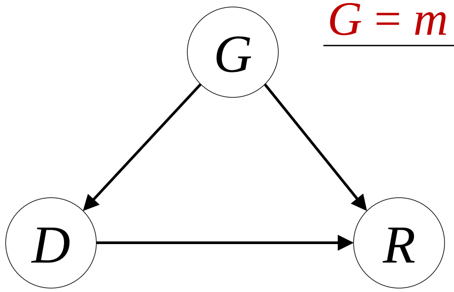
Causes and Effects: *say it with graphs*

- What is a cause? (*Another example*)

  Maximum Likelihood Estimation (CPTs)

$P(G)$

$G = f$	0.49
$G = m$	0.51



$P(D|G)$

	$G = f$	$G = m$
$D = 0$	0.23	0.76
$D = 1$	0.77	0.24

$P(R|G, D)$

	$G = f$, $D = 0$	$G = f$, $D = 1$	$G = m$, $D = 0$	$G = m$, $D = 1$
$R = 0$	0.31	0.27	0.13	0.07
$R = 1$	0.69	0.73	0.87	0.93

Using Graphical Model as a predictor

Case 1: Gender is observed

$$P(R=1|G=0, D=0) = 0.69$$
$$P(R=1|G=0, D=1) = 0.73$$
$$P(R=1|G=1, D=0) = 0.87$$
$$P(R=1|G=1, D=1) = 0.93$$

Prescribe drug, regardless

Case 2: Gender is <u>not</u> observed

$$P(R|D) = \frac{\sum_G P(R|G,D)P(D|G)P(G)}{\sum_{G,R} P(R|G,D)P(D|G)P(G)}$$

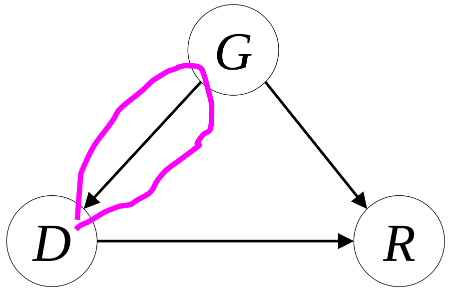
$$P(R=1|D=0) = 0.83$$
$$P(R=1|D=1) = 0.78$$

Do not prescribe drug, regardless

(ridiculous!)

Causes and Effects: *say it with graphs*

- What is a cause? (*Another example*)



G is *biological gender* (= Male/Female)
D is *drug administration* (= Yes(1)/No(0))
R is *recovery* from illness (= Yes(1)/No(0))

How can we solve the problem?

- The problem is due to the discrepancy in drug administration across genders
- An obvious solution would be *to repeat* the experiment with equal administration rates
- *In other words, we would sever this link*

Females	$R=0$	$R=1$		Recovery Rate
$D=0$	25	55	80	69%
$D=1$	71	192	263	73%
	96	247	343	

Males	$R=0$	$R=1$		Recovery Rate
$D=0$	36	234	270	87%
$D=1$	6	81	87	93%
	42	315	357	

	$R=0$	$R=1$		Recovery Rate
$D=0$	61	289	350	83%
$D=1$	77	273	350	78%
	138	562	700	

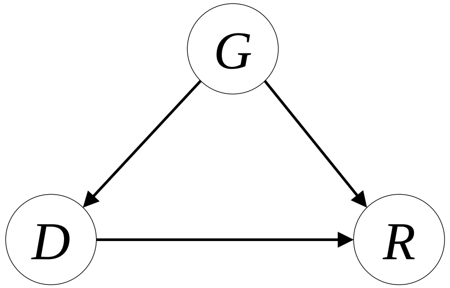
[Data from Pearl, J. et al., "Causal Inference in Statistics: A Primer", Wiley, 2016]

A Working Example
(see GeNIe 'berkeley' attachment)

Causation and observations

Causation and observations

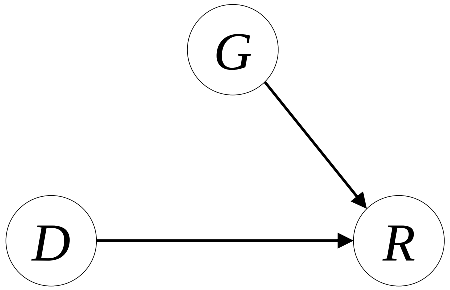
- **Confounders**



In this example, the problem is that G represents a 'common cause' of both D and R
It is a *confounder*, if we are interested in the causal link from D to R

In a controlled experiment, we could administer drug *at random*, regardless of G

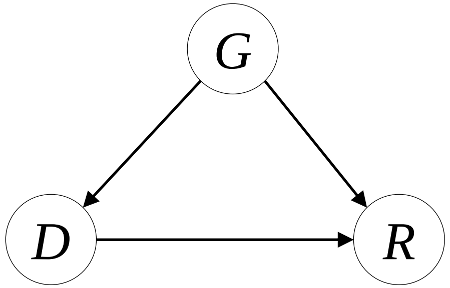
In this case we would have:



$$< D \perp G > \implies P(D \mid G) = P(G)$$

Can we always neutralize confounders in this way?

- **Counterfactuals, potential outcomes**



In many circumstances, data are acquired in an *uncontrolled* ways: they are mere *observations*

We might still circumvent the problem if we knew would have happened if actions were *different (i.e., counterfactuals* or *potential outcomes)*

It may be seen as a problem of missing data in the dataset:

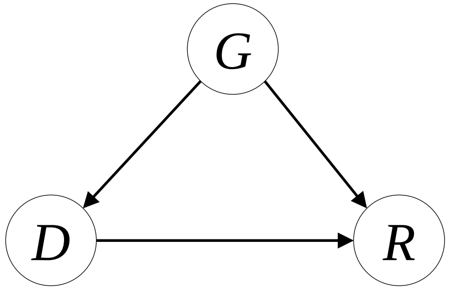
Subject	G	D	$R(D{=}0)$	$R(D{=}1)$
1	0	1	?	1
2	1	1	?	0
3	1	0	1	?
4	0	1	?	1
5	0	0	0	?
...	...	...	...	...
N	1	0	1	?

factual outcomes

counterfactual outcomes

Causation and observations

- **Counterfactuals, potential outcomes**



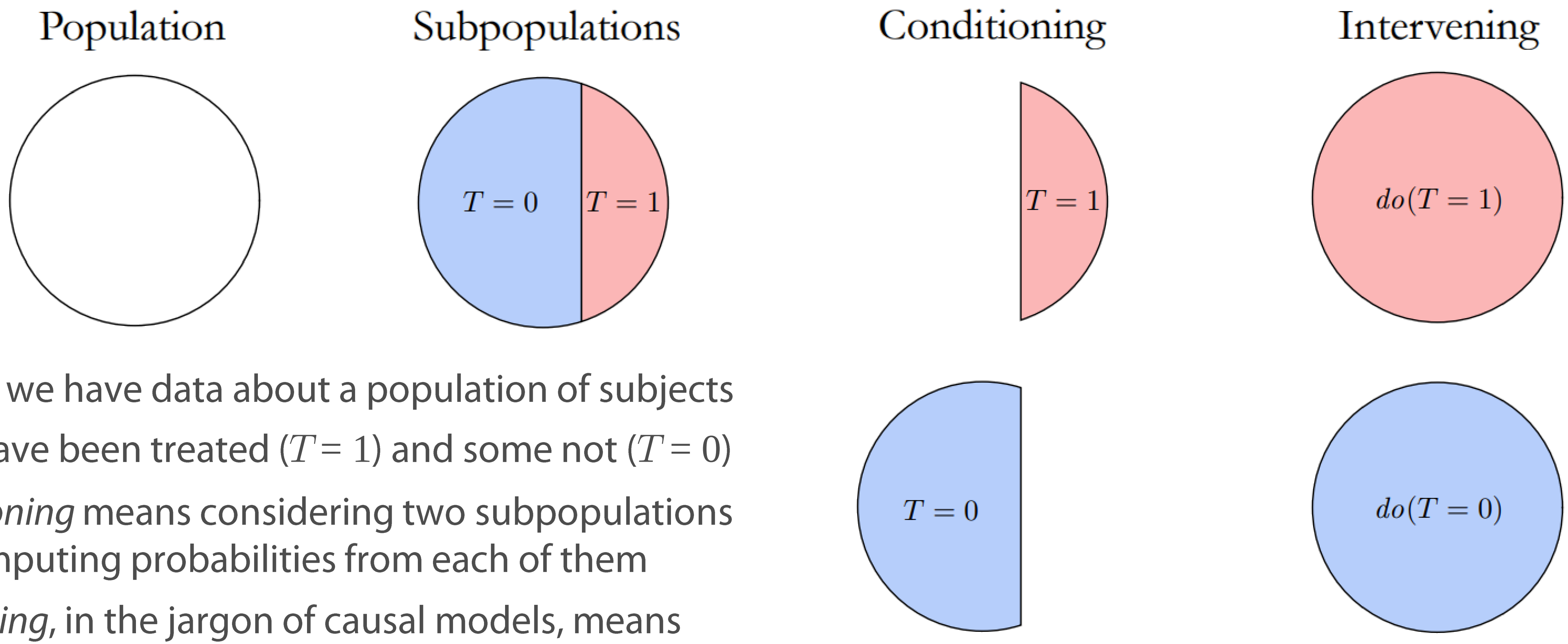
In many circumstances, data are acquired in an uncontrolled ways: they are mere *observations*

Can we work around all of this,
even with data from uncontrolled (i.e., observational) experiments?

Causal Graphical Models at Work
(do-calculus)

Causation and Conditionals

- **Conditioning and Intervening**



Assume we have data about a population of subjects

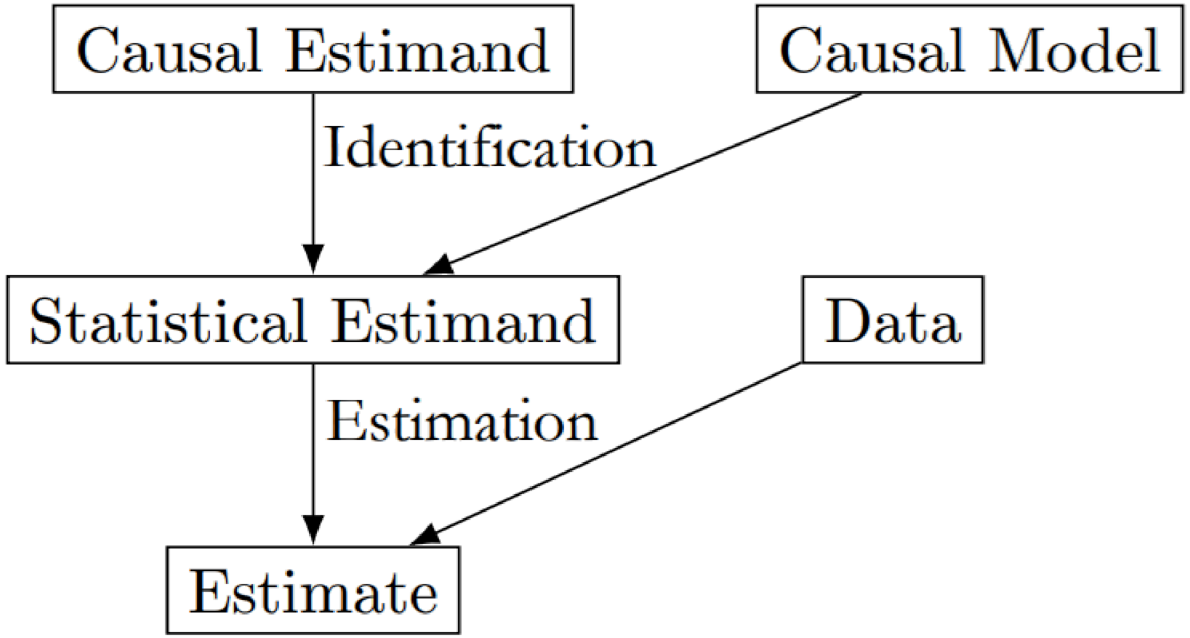
Some have been treated ($T = 1$) and some not ($T = 0$)

Conditioning means considering two subpopulations and computing probabilities from each of them

Intervening, in the jargon of causal models, means assuming that every subject in the population has been treated or not (*potential outcomes*)

[Image from https://www.bradyneal.com/causal-inference-course]

- **Causal Model and Estimation**



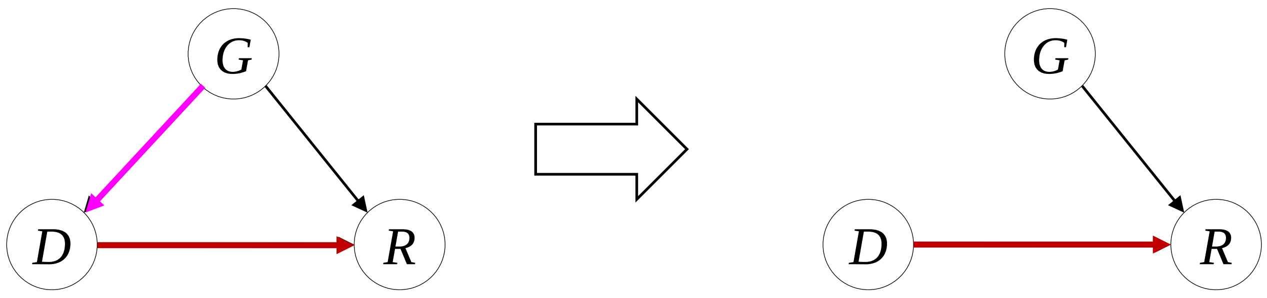
Basic principles:

1. Having selected what kind of causal effect we want to estimate
2. We start from a *Causal Graphical Model* (CGM)
3. To translate the estimate into a statistical estimand, (*Identification*)
4. We use then *observational* data to compute the estimate: a *probability* or an *expected value*

[Image from https://www.bradyneal.com/causal-inference-course]

The Magic of Controlled Experiments

- **When association is causation**



In this *Causal Graphical Model*:

1. The causal effect we are interested in is that of D over R
2. The link between G and D is *problematic*: we know that $P(D|G=0) \neq P(D|G=1)$
3. In a *controlled experiment*, D is administered at random , therefore

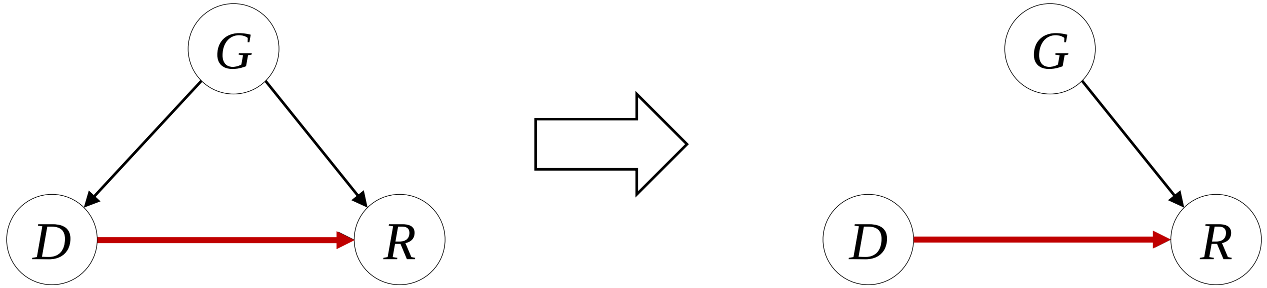
$$< D \perp G > \implies P(D|G=0) = P(D|G=1) = P(D)$$

4. In other words, the corresponding CGM 'loses' the problematic link and the estimate becomes

$$P(R|D) := \sum_G P(G)P(R|G,D)$$

The Magic of Controlled Experiments

- **When association is causation**

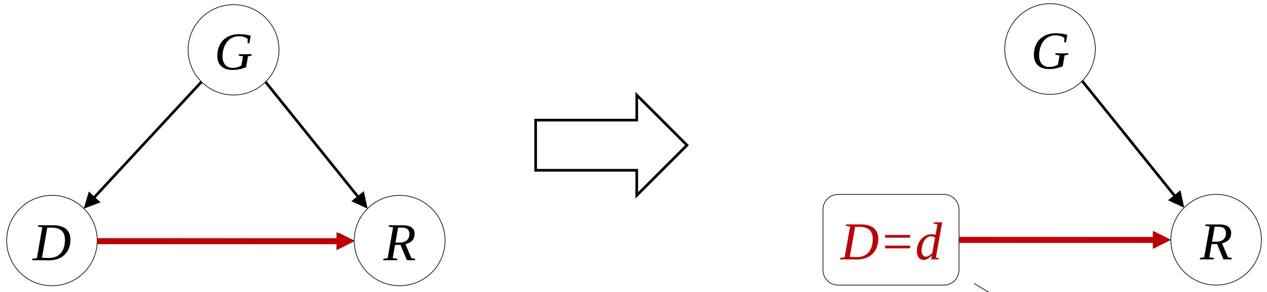


With *controlled experiments* (i.e., the 'gold standard' for testing) the principle is more general:

- by *randomizing* the administration of treatment
- we make the *effects* independent of any *confounders* (be them observed or not)

do-calculus

- **From Conditional (pre-intervention) to Intervention Probability**



A 'deterministic' node (i.e., not 'random' anymore)

In this *Causal Graphical Model* (for an uncontrolled experiment):

1. Conditional probability:

$$P(R|D=d) = \frac{\sum_G P(G)P(R|G, D=d)P(D=d|G)}{\sum_G P(G)P(D=d|G)}$$

2. Intervention (do-calculus, *this is new*)

$$P(R|do(D=d)) := \sum_G P(G)P(R|G, D=d)$$

These two expression would be identical if

$$P(D=d|G) = 1$$

which cannot be true in general

3. *This is equivalent to* $P(R|D=d)$ *in a modified CGM in which we 'enforce intervention'*

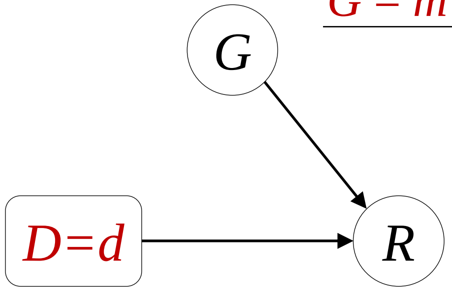
do-calculus

- **From Conditional (pre-intervention) to Intervention Probability**

  (same observational probabilities, from MLE)

$P(G)$

$G = f$	0.49
$G = m$	0.51



$P(R|G, D)$

	$G = f$, $D = 0$	$G = f$, $D = 1$	$G = m$, $D = 0$	$G = m$, $D = 1$
$R = 0$	0.31	0.27	0.13	0.07
$R = 1$	0.69	0.73	0.87	0.93

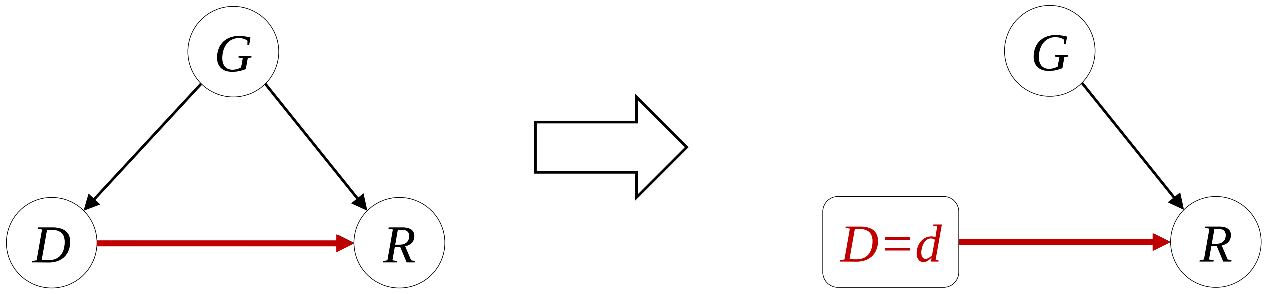
Using do-calculus

$$P(R=1|do(D=0)) = \sum_G P(G)P(R=1|G,D=0)$$
$$= 0.49 \cdot 0.69 + 0.51 \cdot 0.87 = 0.78$$

$$P(R=1|do(D=1)) = \sum_G P(G)P(R=1|G,D=1)$$
$$= 0.49 \cdot 0.73 + 0.51 \cdot 0.93 = 0.83$$

Prescribe drug, regardless

- **Compare two expressions**



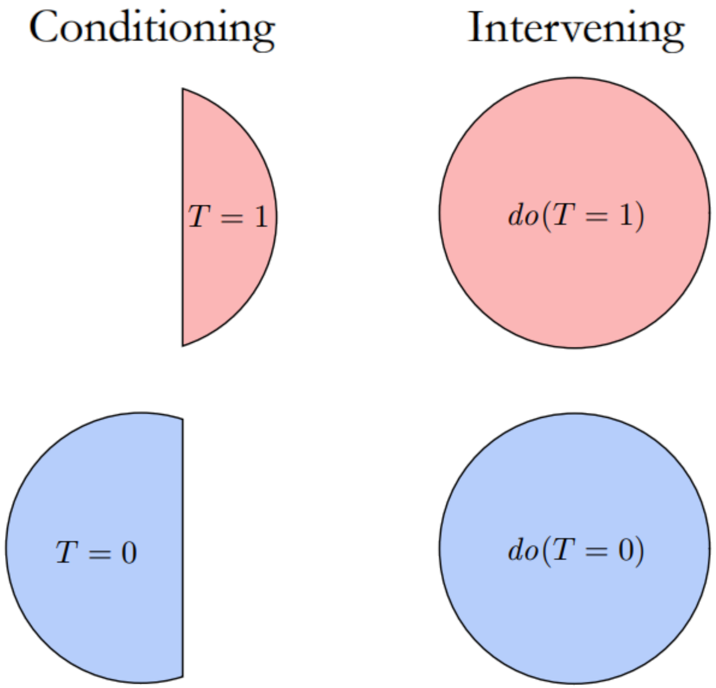
1. Conditioning:

$$P(R|D=d) = \frac{\sum_G P(G)P(R|G,D=d)P(D=d|G)}{\sum_G P(G)P(D=d|G)}$$

2. Intervening:

$$P(R|do(D=d)) := \sum_G P(G)P(R|G,D=d)$$

no normalization =
no conditional subspace



do-calculus:

Is it that simple?

(not so fast...)

do-Calculus

- **In general, in a Causal Graphical Model**

1. Joint Probability Distribution

$$P(X_1, X_2, \ldots, X_n) = \prod_i P(X_i \mid parents(X_i))$$

where $\{X_1, X_2, \ldots, X_n\}$ is the set of random variables in the model

2. Intervention (do-calculus):

$$P(\{X_i\}_{i \neq k} | do(X_k = x_k)) = \prod_{i \neq k} P(X_i \mid parents(X_i))\Big|_{X_k = x_k}$$

In general, do-calculus allows translating a ***causal estimand*** into a ***statistical estimand***, hence a *probability*

Under which conditions such translation is effective and justified?

Identification

- **Adjustment Set Criterion** [Shipster et al. 2010]

  In a Causal Graphical Model, the *causal effect* of T over Y is *identifiable*
  iff it exists an *adjustment set* $\boldsymbol{W}$ of variables such that:

  - no variable in $\boldsymbol{W}$ is on, or is a descendant of any variables on, a causal path (excluding the descendants of T alone)
  - the variables in $\boldsymbol{W}$ block *(in the sense of graphical models)* all the non-causal paths between T and Y

  This criterion is necessary and sufficient for effect identifiability

  Then:

$$P(Y|do(T=t)) = \sum_{\boldsymbol{W}} P(Y|T=t, \boldsymbol{W})P(\boldsymbol{W})$$

  In words, the causal effect can be estimated statistically, from data

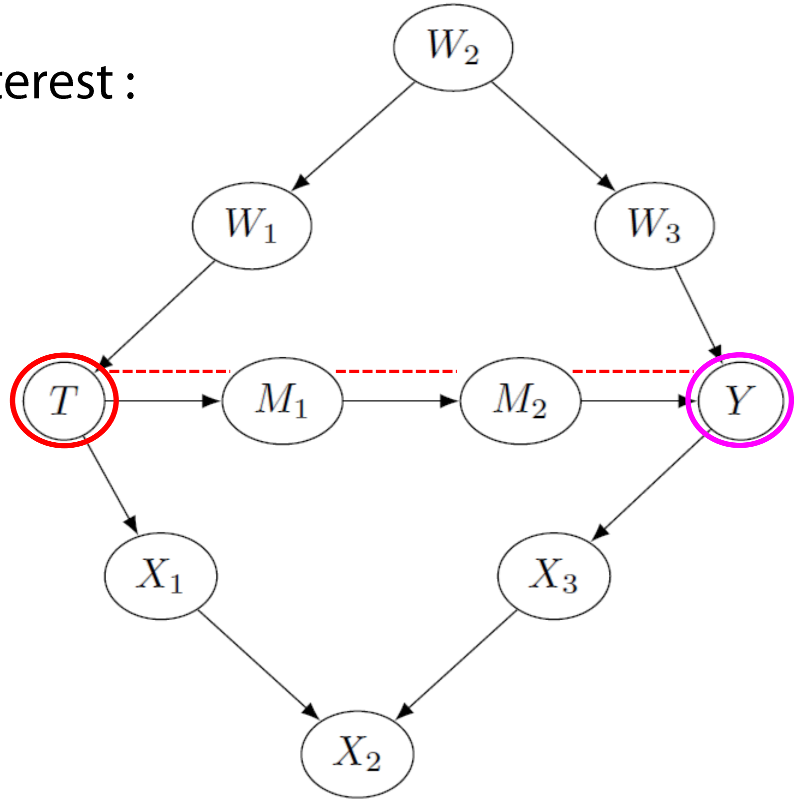
(*) *An earlier (and weaker) version of this is called 'back-door criterion'* [Pearl, 1993]

Identification

- **Identifiable Causal Effect**

In this example, assuming that T over Y is the *causal effect* of interest :

1. The one in red is the *causal path*
   (there could be more, in general)
2. None of M_1 or M_2 should be in the adjustment set W

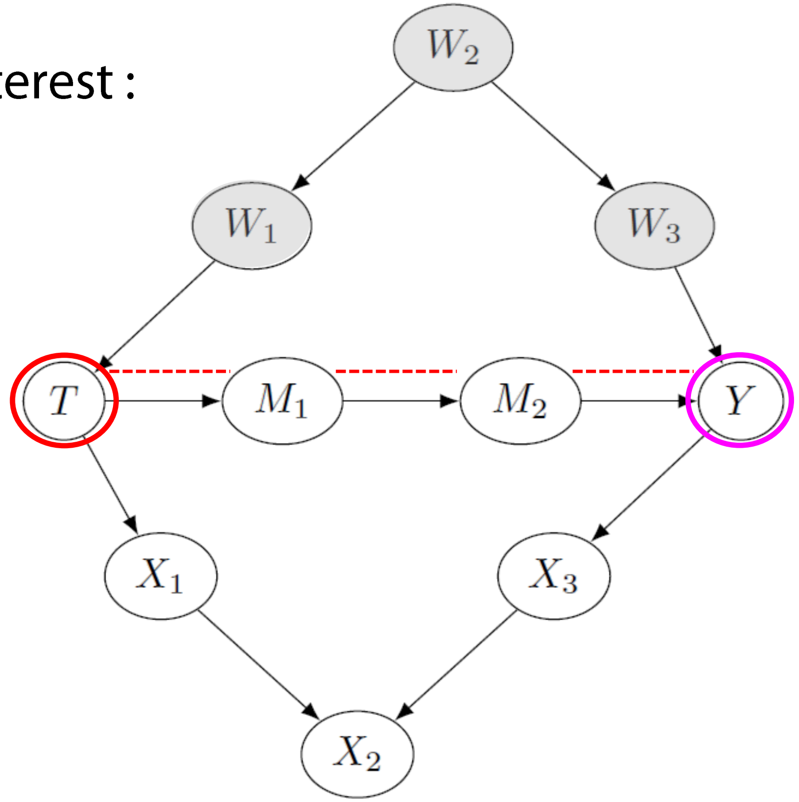


Identification

- **Identifiable Causal Effect**

  In this example, assuming that T over Y is the *causal effect* of interest :

  1. The one in red is the *causal path*
     (there could be more, in general)
  2. None of M_1 or M_2 should be in the adjustment set W
  3. Any non-empty subset of these three nodes
     is a valid *adjustment set* W

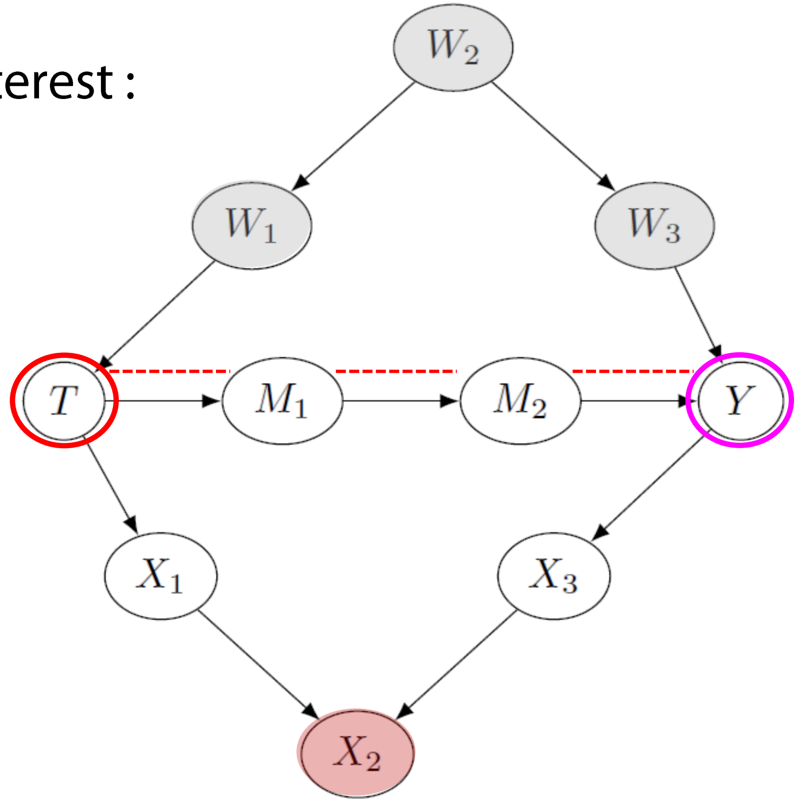


Identification

- **Identifiable Causal Effect**

In this example, assuming that T over Y is the *causal effect* of interest :

1. The one in red is the *causal path* *(there could be more, in general)*
2. None of M_1 or M_2 should be in the adjustment set W
3. Any non-empty subset of these three nodes is a valid *adjustment set* W
4. Adding node X_2 makes it invalid

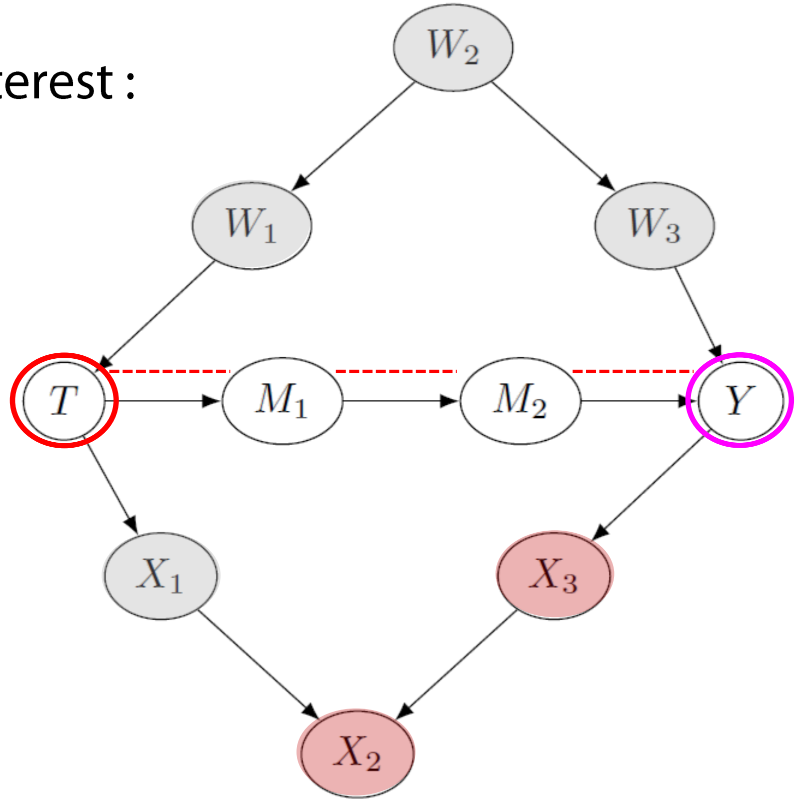


Identification

- **Identifiable Causal Effect**

  In this example, assuming that T over Y is the *causal effect* of interest :

  1. The one in red is the *causal path*
     (there could be more, in general)
  2. None of M_1 or M_2 should be in the adjustment set W
  3. Any non-empty subset of these three nodes
     is a valid *adjustment set* W
  4. Adding node X_2 makes it invalid
  5. Unlike with PGM, adding either X_1 or X_3
     does <u>not</u> make W valid again,
     since X_2 is a descendant of Y



- **Adjustment Set Criterion with *observed* and *unobserved* variables**

More in general, in practical cases,
there can be both observed and unobserved (possibly hidden) variables

An *adjustment set* can be composed of both:

$$\boldsymbol{W} = \boldsymbol{W}_{obs} \cup \boldsymbol{W}_{hid}$$

Then, if $\boldsymbol{W}$ satisfies the Adjustment Set Criterion:

$$P(Y|do(T=t), \boldsymbol{W}_{obs}) = \sum_{\boldsymbol{W}_{hid}} P(Y|T=t, \boldsymbol{W}_{hid}, \boldsymbol{W}_{obs}) P(\boldsymbol{W}_{hid})$$

When there are no *observed* variables in the adjustment set:

$$P(Y|do(T=t)) = \sum_{\boldsymbol{W}} P(Y|T=t, \boldsymbol{W}) P(\boldsymbol{W})$$

Likewise, when there are no *unobserved* variables in the adjustment set:

$$P(Y|do(T=t), \boldsymbol{W}) = P(Y|T=t, \boldsymbol{W})$$

Estimating Effects

Expected effects of different interventions can be estimated via do-calculus

In general, the *expected effect* on Y of treatment T will be

$$\mathbb{E}[Y|T=t, \boldsymbol{W}_{obs}] := \sum_{y \in \mathcal{Y}} y \; P(Y|do(T=t), \boldsymbol{W}_{obs})$$

where $\boldsymbol{W} = \boldsymbol{W}_{obs} \cup \boldsymbol{W}_{hid}$ is a valid *adjustment set*

Differences in effects can be measured by comparing expected effects.

As a special case, when $T \in \{0, 1\}$

- The *Conditional Average Treatment Effect* (CATE) is defined as:

$$\tau(\boldsymbol{W}_{obs}) := \mathbb{E}[Y|T=1, \boldsymbol{W}_{obs}] - \mathbb{E}[Y|T=0, \boldsymbol{W}_{obs}]$$

- The *Average Treatment Effect* (ATE) is defined as:

$$\mathbb{E}[\tau(\boldsymbol{W})] := \mathbb{E}[Y|T=1] - \mathbb{E}[Y|T=0]$$

Another Working Example
(see GeNIe 'berkeley_modified' attachment)