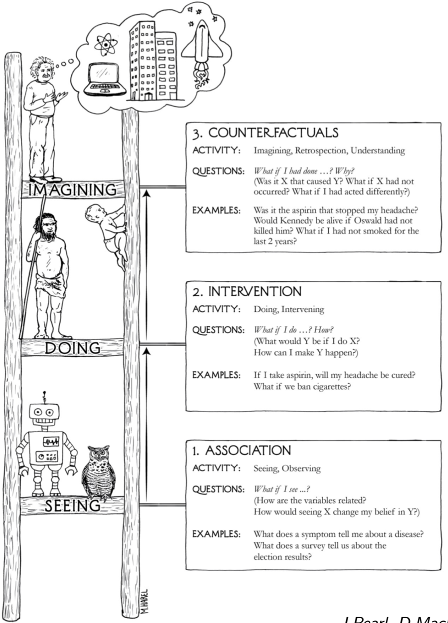
Some insights about observational biases and causal modeling

vittorio.perduca@u-paris.fr

École de l'Inserm Liliane Bettencourt, February 2022

Pearl's ladder of causation



J.Pearl, D.Mackenzie, The Book of Why.

Goals

- Introduce and formalize confounding and selection biases
- Introduce DAGs and *d*-separation
- Define causal models through DAGs and the do operator
- Discuss rules to identify causal effects from observational data

Contents

Example I: do vitamin-based supplements protect from the flu?

- The dataset `vitamines.csv` contains $n = 200$ (simulated) observations for variables
  - X: vitamin yes/no (1/0)
  - Y: flu yes/no (1/0)
  - Z: lifestyle healthy/unealthy (1/0).
- These data were simply **observed**, i.e. they do not come from an interventional experiment.
- Load the data in `R` and estimate the conditional probabilities

$$P(Y = 1|X = 1), \quad P(Y = 1|X = 0),$$

  the relative risk $\mathrm{RR} = \frac{P(Y=1|X=1)}{P(Y=1|X=0)}$, and test the independence between X and Y.

How to do it in R

- Download the dataset in your favorite folder, say the `Desktop`, then load it into `R` with

```
setwd("~/Desktop") # Unix systems
setwd("C:/Users/account_name/Desktop") # Windows
d <- read.table(file = "vitamines.csv",
                sep = ",",
                header = TRUE)
```

- To compute the proportion of people with $Y = 1$ among those with $X = 1$ and to test the independence between X and Y:

```
mean(d$grippe[d$vitamines == 1])
# alternatively:
prop.table(table(d$vitamines, d$grippe))
# rows: vitamins, columns: flu
chisq.test(table(d$vitamines, d$grippe))
```

A randomized trial

- From the preliminary analysis of these **observational** data, it looks like that taking vitamin supplements is associated to a lower flu risk.
- But can we conclude that supplements protect from the flu?
- This is a **causal** question!
- One way to answer is to carry out an interventional experiment.
- The dataset `vitamines_trial.csv` contains $n = 200$ observations from a (simulated) randomized trial in which participants were assigned to the arm $X = 0$ or $X = 1$.
- Re-run the previous analysis on this dataset and comment the results.

Confounding bias

- Results from the interventional data are clear: no effect of vitamins on flu risk whatsoever!
- The association seen in the observational data is an example of **spurious association** due to **confounding**.
- Can you see why? Reconsider the observational data and estimate the following conditional probabilities

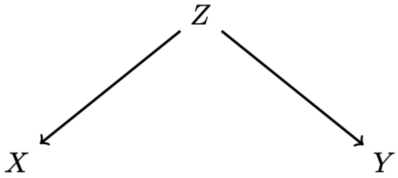
$$P(X = 1|Z = 0), \quad P(Y = 1|Z = 0)$$

and

$$P(X = 1|Z = 1), \quad P(Y = 1|Z = 1).$$

Underlying causal model

- It looks like people with an healthy lifestyle tend to take supplements **and**, because of more hygienic behaviors, are at lower flu risk.
- The actual model used to simulate the data is



Crucial question:

- **Assuming this causal diagram**, is it possible to estimate the effect of X on Y from the observational dataset?

Adjusting for confounding

- The idea is to assess the association between X and Y holding Z fixed (*ceteris paribus*)
- For $z \in \{0, 1\}$, estimate the conditional probabilities

$$P(Y = 1|Z = z, X = 1), \quad P(Y = 1|Z = z, X = 0)$$

and the relative risk

$$\mathrm{RR}_z = \frac{P(Y = 1|Z = 1, X = 1)}{P(Y = 1|Z = 1, X = 0)}$$

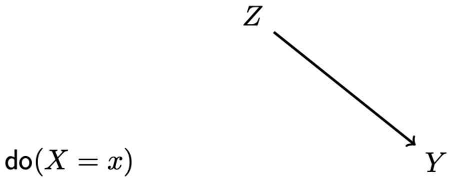
- Test the independence between X and Y while taking into account Z with the Cochran-Mantel-Haenszel test (function `mantelhaen.test(...)`). We say that Z is an **adjustment** variable.

Why does this work?

- In this model, holding $Z = z$ blocks the spurious (i.e., non-causal) path between X and Y.

Causal model under intervention

- This is exactly what happens in a trial, where it is the investigator who chooses (albeit randomly) the X value for each participant.
- The causal model under intervention $\text{do}(X = x)$ is



- Note the missing edge from Z to X: under this intervention, Z no longer influences X

Example II: comparing kidney stone removal modus operandi

- Summary statistics from observational data:

	$d < 2$cm		$d \geq 2$cm		all ds	
	success	failure	success	failure	success	failure
m_1	81	6	192	71	273	77
m_2	234	36	55	25	289	61

- Modus operandi vs Success, unconditional inference:

$$P_n(\text{success}|m_1) = \frac{273}{350} \approx 78\% < P_n(\text{success}|m_2) = \frac{289}{350} = 83\%$$

- It look likes chances of success are higher with m_2...

Simpson's paradox

- Modus operandi vs Success conditionally on Stone size:

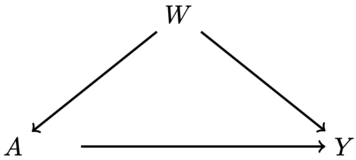
$$P_n(\text{success}|d < 2, m_1) = \frac{81}{87} \approx 93\% > P_n(\text{success}|d < 2, m_2) = \frac{234}{270} \approx 87\%$$
$$P_n(\text{success}|d \geq 2, m_1) = \frac{192}{263} \approx 73\% > P_n(\text{success}|d \geq 2, m_2) = \frac{55}{80} \approx 69\%$$

- For each stone size, chances of success are higher with m_1. Note the **association reversal**!
- This is an instance of the so called Simpson's paradox.
- Note that the total probability law implies

$$\begin{aligned} P_n(\text{success}|m_1) &= P_n(\text{success}, d < 2|m_1) + P_n(\text{success}, d \geq 2|m_1) \\ &= P_n(\text{success}|d < 2, m_1) \times P_n(d < 2|m_1) \\ &\quad + P_n(\text{success}|d \geq 2, m_1) \times P_n(d \geq 2|m_1) \\ &= \frac{81}{87} \times \frac{87}{350} + \frac{192}{263} \times \frac{263}{350} \qquad (1) \\ &= 78\%. \end{aligned}$$

Spurious association, again

- The previous data were generated according to the causal model



- According to this model:
  - Modus operandi A is influenced by Stone size W: *m_2 is preferred with small stones and m_1 with larger stones*.
  - Success Y is determined by A and W: *m_1 and small stones increase chances of success*.
- But then what happens when we compute $P_n(\text{success}|m_1)$? **Knowing** the modus operandi $A = m_1$ says something about
  - Y because of the directed causal path $A \rightarrow Y$
  - W, which in turns allow to predict Y.
- It is the latter non-causal path $A \leftarrow W \rightarrow Y$ that distorts the observed association!

Debunking Simpson's paradox (I)

- The gold standard to decide what is the best modus operandi would be conducting a trial where the investigator intervenes by imposing m_1 or m_2.
- Under such intervention, A is no longer influenced by W: the non-causal path from A to Y is thus blocked.
- Mathematically, we are interested in the law of Y after the intervention:

$$\mathbb{P}_n(\text{success}|\text{do}(m_1))$$

- Can we estimate this post-intervention law **without doing an actual trial**?

Debunking Simpson's paradox (II)

- The post-intervention model is

$$\begin{tikzcd} & W \arrow[dr] & \\ \mathrm{do}(A = m_1) \arrow[rr] & & Y \end{tikzcd}$$

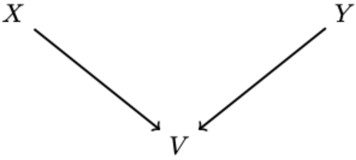
- We will see that from this model we obtain

$$\begin{aligned} \mathbb{P}_n(\text{success}|\text{do}(m_1)) &= P_n(\text{success}|d < 2, m_1) \times P_n(d < 2) + \\ &\quad + P_n(\text{success}|d \geq 2, m_1) \times P_n(d \geq 2) \\ &= \frac{81}{87} \times \frac{357}{700} + \frac{192}{263} \times \frac{343}{700} \approx 83\%. \end{aligned} \tag{2}$$

- Note that we have taken W into account: we will see that W is an adjustment variable.
- Compare equations (2) and (1): can you see why $P_n(\text{success}|m_1) < \mathbb{P}_n(\text{success}|\text{do}(m_1))$?
- Show that $\mathbb{P}_n(\text{success}|\text{do}(m_2)) \approx 78\%$.

Example III: does the flu protect against appendicitis?

- Consider variables
  - X: flu yes/no (1/0)
  - Y: appendicitis yes/no (1/0)
  - V: hospitalization yes/no (1/0)
- with the following causal model



  - $X \sim \mathcal{B}(0.3)$, i.e. a toss of a biased coin with $P(X=1)=0.3$
  - $Y \sim \mathcal{B}(0.05)$
  - $(V|X=x, Y=y) \sim \mathcal{B}(p_{x,y})$ with $p_{x,y}$ given by

	$X=0$	$X=1$
$Y=0$	0.05	0.2
$Y=1$	1	1

Generating the data

- Simulate $n = 1000$ data points (x, y, v) according to the model
- How to do it in R:

```
n <- 1000
x <- rbinom(n, size = 1, prob = .3)
# alternative:
# x=sample(0:1,size=n,replace=TRUE,prob=c(.7,.3))
# and similarly for y
v <- 1
v[which(x==0 & y==0)] <- rbinom(length(which(y==0 & x==0)),
                                        size = 1, prob = .05)
# and similarly for simulating v when x = 0 and y = 1...
```

Berkson's paradox

- Compute the estimates $P_n(Y = 1|X = 0)$ and $P_n(Y = 1|X = 1)$ and show these are essentially the same.
- Test the independence between X and Y.
- Show that $P_n(Y = 1|X = 0, V = 1) >> P_n(Y = 1|X = 1, V = 1)$.
- Test the independence between X and Y while taking into account V with the Cochran-Mantel-Haenszel test.
- It look likes that X and Y are independent and become dependent while conditioning on V:
  - in the general population the probability of appendicitis is the same irrespective of the flu
  - but **at the hospital**, the flu seems to protect against appendicitis!
- This is an instance of the Berkson's paradox or **selection bias**.

Debunking Berkson's paradox (I)

- Truth is that X and Y are independent because this is how we generated them.
- We will see that conditioning on V makes X and Y dependent because it *opens* the non-causal path $X \to V \leftarrow Y$.
- If the question is whether the flu protects against appendicitis, we should rather look at the consequences of the (unethical) interventions $\mathrm{do}(X = 0)$ and $\mathrm{do}(X = 1)$:

$$\mathbb{P}_n(Y = 1 | \mathrm{do}(X = 0)) \text{ and } \mathbb{P}_n(Y = 1 | \mathrm{do}(X = 1)). \quad (3)$$

- It is simple to simulate interventions $\mathrm{do}(X = 0)$: simply replace the code generating `x` with

```
x <- rep(0, n)
```

and generate `y` and `v` as before.

Debunking Berkson's paradox (II)

- Simulate $n_0 = 500$ observations under the intervention $\mathrm{do}(X = 0)$ and $n_1 = 500$ observations under the intervention $\mathrm{do}(X = 1)$.
- Calculate quantities in equation (3) and conclude.
- We will introduce formal arguments to show that the considered causal diagram implies:

$$\mathbb{P}(Y = 1|\mathrm{do}(X = x)) = P(Y = 1) \text{ for } x \in \{0, 1\}$$

Facultative exercise

- Rather than comparing $P_n(Y = 1|X = 0, V = 1)$ and $P_n(Y = 1|X = 1, V = 1)$ it is interesting to look at

$$\mathbb{P}_n(Y = 1|\text{do}(X = 0), V = 1)$$
$$\mathbb{P}_n(Y = 1|\text{do}(X = 1), V = 1)$$

- Compute these quantities from the previous slide's simulations and interpret the results.
- We will show that we do not need to do real-life interventions to compute the post-intervention laws: the post-intervention laws can be computed from observational data as follows

$$\mathbb{P}(Y = 1|\text{do}(X = x), V = 1) = \frac{P(Y = 1)P(V = 1|Y = 1, X = x)}{P(V = 1|X = x)}$$

for $x \in \{0, 1\}$.

First summary

- Confusion and selection biases might arise when analyzing observational data.
- Association (correlation) is not causation.
- We are often interested in questions of causal nature.
- Gold standard to provide answers to causal questions is carrying out interventions (e.g. randomized trials).
- Answering causal questions using observational data seem to require *adjusting* on selected variables.
- Is there always such a set of adjustment variables? And how to find it?

Directed Acyclic Graphs (DAGs)

- We model variables and the relations between them with diagrams called DAGs.
- A DAG is a graph with arrows and no cycles (i.e., starting from a vertex, it is not possible to go back to it following the direction of the arrows).

DAGs + joint probability distributions

- Let $\mathcal{G}$ be a (DAG). Each vertex represents a random variable.
- For each variable X we consider its parents $pa(X)$ in $\mathcal{G}$ and the conditional probability distribution $P(X|\text{pa}(X))$
- We suppose that each variable is independent from all other non-descendant variables given its parent (Markov property):

$$X \perp (\text{nondesc}(X) \setminus \text{pa}(X)) \,|\text{pa}(X)$$

- It follows that the joint probability distribution is

$$P(X_1, \ldots X_n) = \prod_{i=1}^{n} P(X_i|\text{pa}(X_i)) \tag{4}$$

- The pair $(\mathcal{G}, P)$ is called a Bayesian network.

Why are Bayesian networks so useful

- Bayesian networks are very convenient for modeling, because their topology encodes all possible independence relations between subsets of variables.
- The correspondence between the topology of $\mathcal{G}$ and the independence relations characterizing P is given by the rules of d-separation.
- We start by looking at this correspondence for three special DAGs:
  - $i \rightarrow w \rightarrow j$ and $i \leftarrow w \rightarrow j$
  - $i \rightarrow w \leftarrow j$

Open and blocked paths

We need the following definitions characterizing the paths in $\mathcal{G}$:

1. We say that the paths $i \to w \to j$ and $i \leftarrow w \to j$ are **opened** and that they are *blocked* once conditioning on w.

- Using Markov property it is easy to show that if the paths $i \to w \to j$ and $i \leftarrow w \to j$ correspond to the whole DAG:
  - variables i and j are not independent.
  - variables i and j are independent conditional on variable w. Intuitively, conditioning on w *blocks* the information flow.

2. We say that the path $i \to w \leftarrow j$ is **blocked**. Conditioning on the **collider** w *opens* the path.

- It is easy to show that if the path $i \to w \leftarrow j$ corresponds to a whole DAG:
  - variables i and j are independent
  - variables i and j are not independent conditional on w.

d-separation

By definition, we say that a set of nodes W in $\mathcal{G}$ **blocks** a path p if

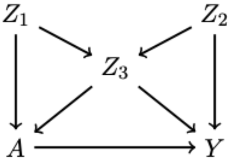
1. p contains at least one sequence $i \to w \to j$ or $i \leftarrow w \to j$, with $w \in W$;

OR

2. p contains at least one collider w (i.e., a sequence $i \to w \leftarrow j$) that is outside W and has no descendant in W.

The set W is said to d-separate A and Y in the graph $\mathcal{G}$.

Illustrating d-separation



- Consider the path $p_1 : A \leftarrow Z_3 \rightarrow Y$:
  p_1 is blocked by $\{Z_3\}$.
- $p_2 : A \leftarrow Z_1 \rightarrow Z_3 \rightarrow Y$:
  p_2 is blocked by $\{Z_1\}$, $\{Z_2\}$ and $\{Z_1, Z_3\}$.
- $p_3 : A \leftarrow Z_3 \leftarrow Z_2 \rightarrow Y$:
  p_3 is blocked by $\{Z_3\}$, $\{Z_2\}$ and $\{Z_3, Z_2\}$.
- $p_4 : A \leftarrow Z_1 \rightarrow Z_3 \leftarrow Z_2 \rightarrow Y$:
  p_4 is not blocked by $S = \{Z_3\}$ since no emitting-arrow node of p_4 is in S, and p_4 contains a collider Z_3 which is in S. However, p_4 is blocked by $\{Z_1\}$ or $\{Z_1, Z_3\}$ or even $\emptyset$.

Probabilistic implications of d-separation

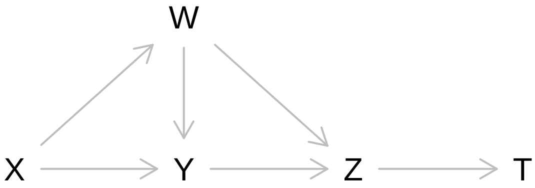
d-separation allows the identification of **all** the conditional independence relationships implied by the structure of the DAG:

A and Y are independent in P conditionally on W, and we write $(A \perp Y | W)_P$,

$$\Leftrightarrow$$

W d-separates A and P in $\mathcal{G}$.

Exercise: *d*-separation



1. Name all of the parents of Z; name all the ancestors of Z.
2. Name all the children of W; name all the descendants of W.
3. List all simple paths between X and T (i.e., no node should appear more than once).
4. List all the directed paths between X and T.
5. Does $\{Z\}$ *d*-separate X and T? And $\{W\}$? And $\{W, Y\}$?
6. List all the open paths between X and T (i.e., the paths that are not blocked by $\emptyset$).
7. List all the paths between X and T blocked by $\{Y\}$.
8. List all *minimal* conditional independencies between pairs of non-adjacent variables implied by the DAG. We say that the conditional independence statement "A independent of B given a set of variables W" is minimal if A and B are no longer independent given a subset of W. You can use the tool `dagitty.net`

do operator

The do operator, implements mathematically the notion of intervention in a Bayesian network.

- Consider a DAG $\mathcal{G}$ and a joint probability distribution P over its vertices $A, X_1, \dots, X_n$.
- Let a be a fixed value. By definition, the distribution of $X_1, \dots, X_n$ following the intervention $\text{do}(A = a)$, is obtained by

1. Removing all arrows pointing towards A in $\mathcal{G}$
2. Setting $A = a$ in all the conditional probability distributions appearing in the right hand side of factorization (4)

- In particular the **post-intervention distribution** is

$$\mathbb{P}(X_1, \dots, X_n | do(A = a)) = \prod_i P(X_i | \text{pa}(X_i))|_{A=a} \quad (5)$$

Causal DAGs and causal effects

- A causal DAG is simply a Bayesian network equipped with the do operator.
- Assuming a causal DAG, we can assess the consequences of an intervention without the need to actually implement it in the real life: **if all variables are observed** we can estimate the post-intervention distribution from equation (5) using observational variables only.
- This allows to estimate the post-intervention distribution of the outcome of interest using the total probability law. For instance with four discrete variables $A, X_1, X_2, X_3 = Y$:

$$\mathbb{P}(X_3|\text{do}(A=a)) = \sum_{x_1,x_2} \mathbb{P}(X_3, X_1 = x_1, X_2 = x_2|\text{do}(A=a))$$
$$\sum_{x_1,x_2} \prod_i P(X_i|\text{pa}(X_i))|_{A=a,X_2=x_2,X_3=x_3}$$

- In turns this allow to estimate causal effects such as the average treatment effect

$$\mathbb{P}(Y|\text{do}(A=1)) - \mathbb{P}(Y|\text{do}(A=0)) \tag{6}$$

Identification problem

- But is it still possible to estimate

$$\mathbb{P}(Y = y|\text{do}(A = a)) = \mathbb{P}(y|\text{do}(a)) \tag{7}$$

  without implementing the intervention in the real life if some of the variables are not observed?
- In other words: can we express (7) as a function of the distribution P of a subset of the observational variables?
- This is called the identification problem.
- A sufficient condition to identify (7) is the existence of **adjustment variables** that have been observed.

Common folklore about adjustment

- What are adjustment variables?
- Common folklore about adjustment, such as
  - adjusting for more variables is better
  - one should adjust for all variables related to both A and Y
  - adjusting for pre-treatment variables is always safe
  - adjusting for descendants of A is always bad
  - mutual adjustment works
  - ... are generally false!
- We need a formal definition.

Adjustment sets

- By definition, W is an adjustment set w.r.t. (A, Y) in a causal DAG if

$$\mathbb{P}(y|do(a)) = \begin{cases} P(y|a) & \text{if } W = \emptyset \\ \sum_w P(y|a,w)P(w) = E\{P(y|a,W)\} & \text{otherwise} \end{cases}$$

- In this definition, we supposed that the variables in W are discrete, if the variables in W are continuous, simply replace sums with integrals.
- Note that the right hand sides depend only on the distribution P of observational variables.
- Hence, if all the variables in W have been measured, we can identify the target $\mathbb{P}(y|do(a))$.
- But how do we find such adjustment sets W?

Adjustment variables and the linear model

Important fact:

Suppose that Y and A are continuous variables and the true underlying data-generating mechanism is linear, i.e. each variable is generated as a linear combination of its parent plus a random noise.

$\Rightarrow$

if W is an adjustment set w.r.t. (A, Y), the average total effect of A on Y defined in (6) is the coefficient of A in the linear regression `Y ~ A + W`.

Back-door criterion

Many criteria exist to find adjustment sets, here we introduce the most popular one.

- We say that W satisfies the back-door criterion w.r.t.(A, Y) if

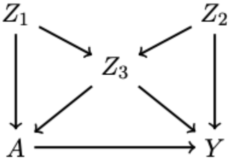
1. W does not contain any descendant of A
2. W blocks all **back-door paths** between A and Y, that is all paths terminating with an arrow pointing to A.

- It can be proven that if W satisfies the back-door criterion, then W is an adjustment set w.r.t. (A, Y):

$$\mathbb{P}(Y = y|\text{do}(A = a)) = \sum_{w} P(y|A = a, W = w)P(W = w)$$

- Algorithms exist to find subsets of variables satisfying the back-door criterion, e.g. in the `R` package `dagitty`.

Illustrating the back-door criterion



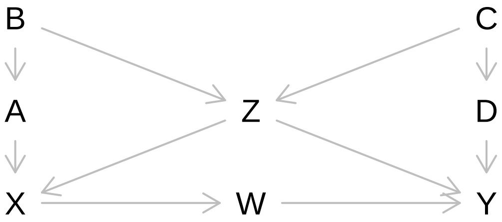
- The sets $\{Z_1, Z_2, Z_3\}$, $\{Z_1, Z_3\}$ and $\{Z_2, Z_3\}$ all satisfy the back-door criterion.
  $\Rightarrow$ Observing $\{Z_3, Z_1\}$ or $\{Z_3, Z_2\}$ is sufficient for the estimation of the causal effect of A on Y.
- The set $\{Z_1, Z_2\}$ does not satisfy the back-door criterion:
  - it does not block the path $A \leftarrow Z_3 \rightarrow Y$
- The set $\{Z_3\}$ does not satisfy the back-door criterion:
  - it does block paths $A \leftarrow Z_3 \rightarrow Y$, $X \leftarrow Z_1 \rightarrow Z_3 \rightarrow Y$ and $X \leftarrow Z_3 \leftarrow Z_2 \rightarrow Y$ ...
  - ... but it does not block the path $A \leftarrow Z_1 \rightarrow Z_3 \leftarrow Z_2 \rightarrow Y$

Intuition behind the back-door criterion

- Back-door paths induce spurious dependence between A and Y, while direct paths carry causal associations.
- Blocking back-door paths ensures that association measured after adjustment is truly causal
- In particular, the back-door criterion ensures that
  - all spurious paths from A and Y are blocked
  - all directed paths from A to Y are left untouched
  - no new spurious path is created
- One reason why we do not adjust for descendants of A is that this could block directed path from A to Y, thus invalidating point 2.

Exercise: back-door criterion

Consider the following DAG



1. List all minimal sets of variables that satisfy the back-door criterion to determine the causal effect of X on Y.
2. Suppose that we cannot measure Z. Can we still identify the causal effect of X on Y?
3. Choose an adjustment set. Write an equation giving the post-intervention density of Y in terms of conditional densities according to the back-door criterion.
4. List all sets of variables that satisfy the back-door criterion to determine the causal effect of W on Y.

Exercise: adjustment variables and the linear model

Consider the data generating mechanism $X \leftarrow Z \rightarrow Y$ with

$$\begin{cases} Z & \sim & \mathcal{U}_{\{0,1,2,3\}} \\ X & = & Z + \mathcal{N}(0, 0.2) \\ Y & = & Z + \mathcal{N}(0, 0.2) \end{cases}$$

where $\mathcal{U}_{\{0,1,2,3\}}$ means that Z is sampled by throwing an unbiased die with four faces.

1. Does X has a causal effect on Y?
2. Load the dataset `confusion_linear.csv` containing 200 data points (x, y, z) sampled from this model. We forget that we know the model that has generated the data and we analyse the available dataset.
3. Make a scatter plot of x and y.
4. Estimate the coefficient of X in the linear regression of Y as a function of X and test its significance. What do you observe?
5. Is Z an adjustment variable w.r.t. (X, Y)?
6. In the previous scatter plot, color the points according to the z values.
7. Estimate the coefficient of X in the linear regression of Y as a function of X and Z and test its significance. Comment the result.

Second summary and conclusions

- If we assuming a causal DAG and observe adjustment variables then we can identify the post-intervention distribution of interest without the need to carry out a real-life intervention.
- Warning: the inference quality will crucially depend on how well the assumed causal DAG match the data-generating mechanism.
- How do we learn the causal DAG?
  - From expert knowledge
  - Active field of research about data-driven methods
- Other methods exist to identify the causal targets of interest: matching, propensity scores, instrumental variables,. . .
- Two other approaches exist to define causal models:
  - structural equation modeling
  - counterfactual variables

References

- Pearl et al. Causal inference in Statistics. 2016, Wiley.
- Shalizi. Advanced data analysis from an elementary point of view. 2021. Chapters 19-21. Available at https://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/
- Matthews. Storks deliver babies (p= 0.008). Teaching Statistics, 2000.
- Hernán et al. A structural approach to selection bias. Epidemiology, 2004.
- Duan et al. Reflection on modern methods: understanding bias and data analytic strategies through DAG-based data simulations. International journal of epidemiology, 2021.

Aknowldegments

VP thanks Antoine Chambaz (Université de Paris) and Vivian Viallon (IARC) for sharing pedagogical material, including some of the previous slides, with him.