DAG fantastici…
…e come utilizzarli

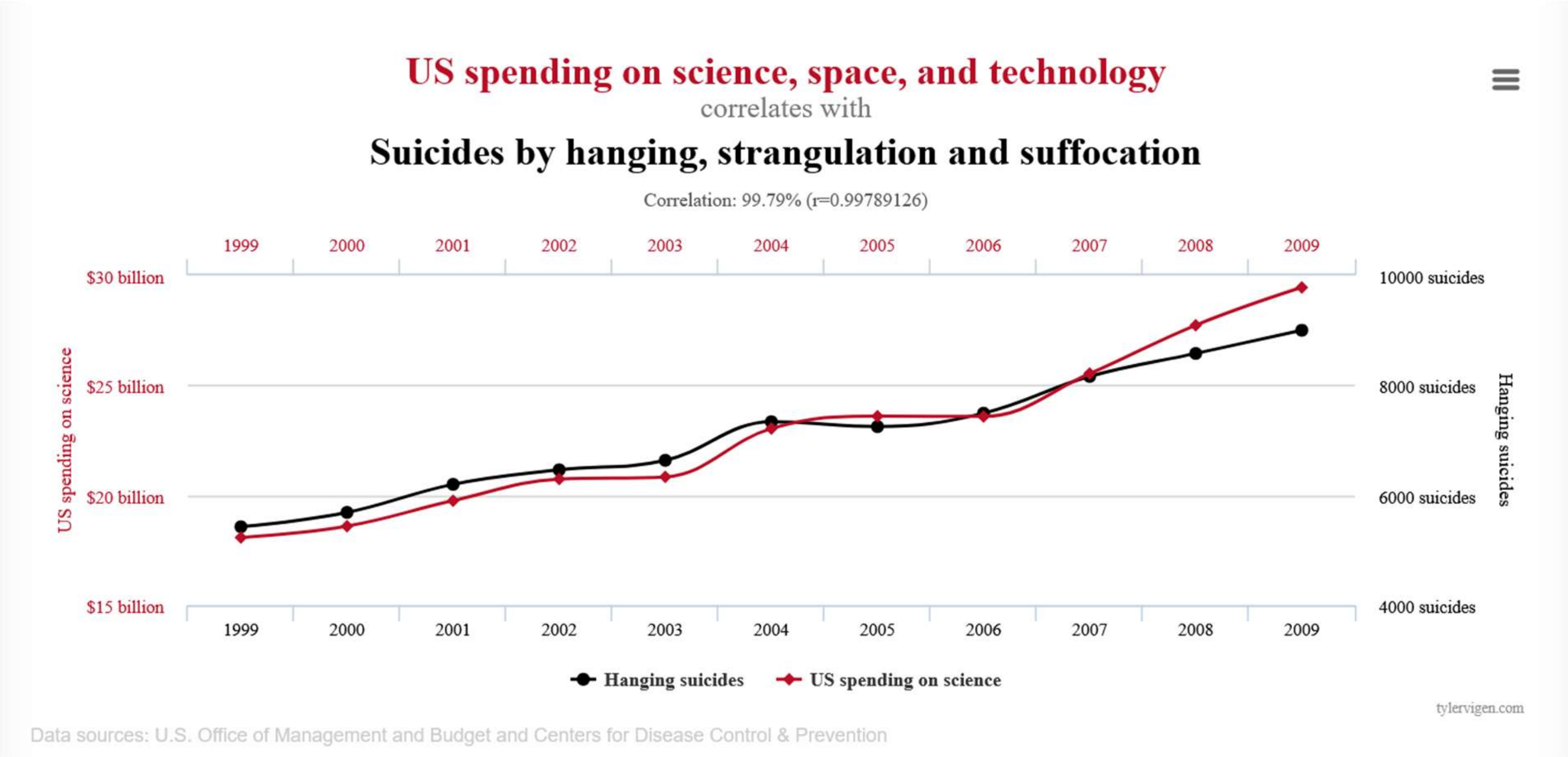
Metodi Quantitativi per le Biotecnologie

Corso di laurea magistrale in Biotecnologie per le Biorisorse e lo Sviluppo Ecosostentibile (A.A. 2023-2024)

Università di Verona

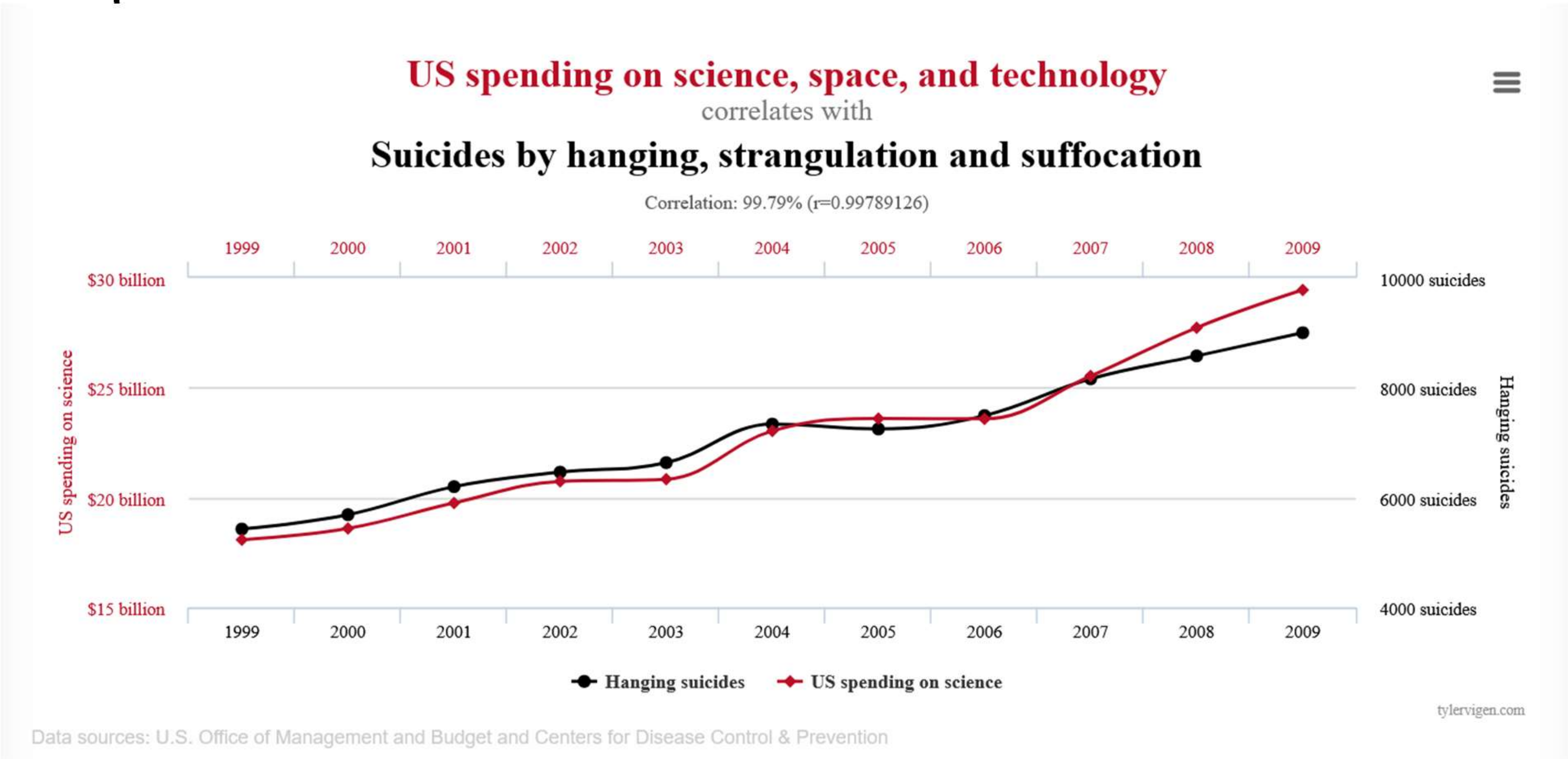
Matteo Migliorini (matteo.migliorini@univr.it)- Assegnista

Problema…



… quindi investire in scienza aumenta i casi di suicidio ???… O viceversa???…

… Risposta: nessuno dei due



Obiettivi della lezione

- Cogliere la differenza tra inferenza causale (causa-effetto) e statistica
- Cosa sono i DAG e perché sono utili per fare inferenza causale
- Riconoscere le principali strutture nei DAG e i relativi flussi associativi
- Distinguere tra effetti diretti e totali di una variabile; interpretare i coefficienti di regressione
- Applicare queste nozioni utilizzando i pacchetti R `dagitty` e `ggdag`

Inferenza causale vs. statistica

Inferenza Causale vs. statistica

- La statistica misura l'**associazione** tra due variabili. **NON** misura la **causalità** tra le due!

$$A \rightarrow B \qquad = \qquad B \leftarrow A \qquad = \qquad A \text{ ind. } B$$

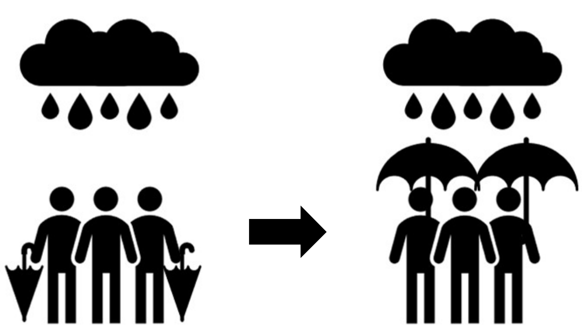
Inferenza Causale (causa-effetto)

- Def: “It is the reasoning to the conclusion that something is, or is likely to be, the cause of something else”

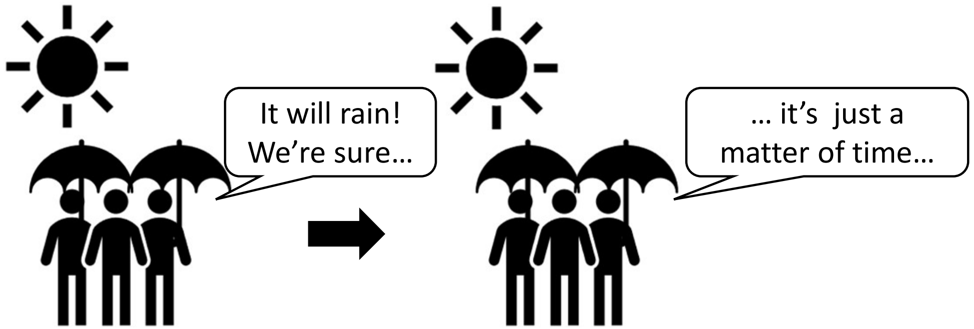
- Studia I rapporti di causa-effetto e le associazioni che scaturiscono da questi

- Rapporto di casusa-effetto: considerando due fenomeni A e B, A è causa di B se modificando A viene a modificarsi B, ma viceversa la modifica di B non influenza A.

Inferenza Causale (causa-effetto) - esempio



La pioggia fa aprire gli ombrelli alle persone…



… non viceversa…

Inferenza Causale vs. statistica

- La statistica misura l'**associazione** tra due variabili. **NON** misura la **causalità** tra le due!

$$A \rightarrow B \qquad = \qquad A \leftarrow B \qquad = \qquad A \text{ ind. } B$$

- L'inferenza causale presuppone che una variabile sia responsabile della variazione di un'altra variabile.

$$A := B \qquad \neq \qquad B := A$$

Inferenza Causale vs. statistica

- La statistica misura l'**associazione** tra due variabili. **NON** misura la **causalità** tra le due!

$A \rightarrow B \quad = \quad A \leftarrow B \quad = \quad A \text{ ind. } B$

A causa B — B causa A — A e B sono indipendenti (associazione spuria)

- L'inferenza causale presuppone che una variabile sia responsabile della variazione di un'altra variabile.

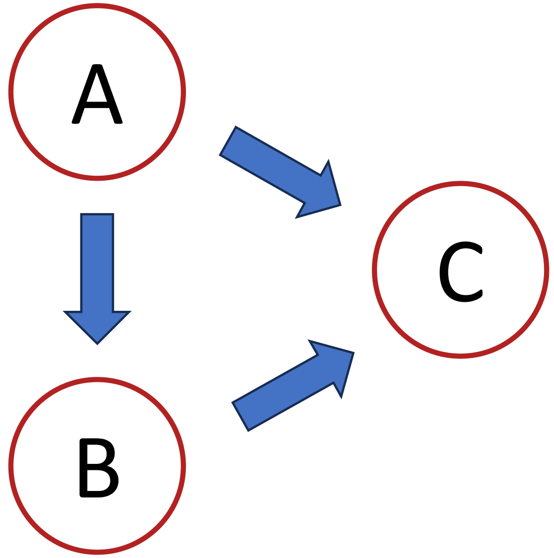
$A := B \quad \neq \quad B := A$

A è causato da una qualche funzione di B

Cosa sono i DAG?

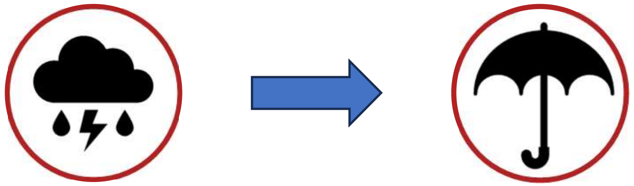
Grafi Aciclici Diretti – DAG

- Grafo: struttura orientata definita da vertici (cerchi) e archi (frecce)
- Diretto (o orientato): gli archi hanno una direzionalità
- Aciclico: selezionando un qualsiasi vertice, non è possibile ritornarvici seguendo l'orientazione degli archi



Grafi Aciclici Diretti – DAG

- Riprendendo l'esempio di prima in cui la pioggia causa l'apertura degli ombrelli, il DAG è il seguente:



- In genere il DAG si costruisce sulla base delle conoscenze scientifiche, non sui dati.

In che modo ci aiutano con l’inferenza causale?

Elemental confounders (EC)

- I DAG ci consentono di vedere come fluisce l'associazione tra variabili e di capire come bloccare le associazioni spurie.

- Prima però dobbiamo capire come l'associazione tra le variabili *fluisce* all'interno del DAG. Ci sono <u>quattro strutture base a tre nodi</u> (elemental confounders) che compongono un DAG:
  - Pipe X -> Z -> Y
  - Fork X <- Z -> Y
  - Collider X -> Z <- Y
  - Descendant Z -> X -> Y

NB: X è la variabile indipendente; Y è la nostra variabile dipendente; Z è un'altra variabile indipendente misurata che possiamo o meno inserire all'interno della regressione.

Elemental confounders (EC)

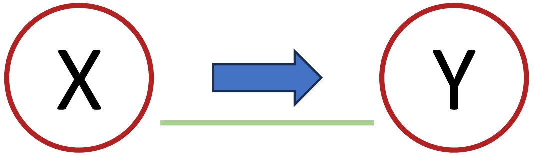
- Per comprendere il flusso associativo è utile ragionare in termini di regressione lineare multivariabile.

$$Y = \alpha + \beta_k \, x_k$$

- Spiegando Y in funzione di X e aggiungendo o meno una terza variabile di controllo Z scopriremo che i flussi associativi possono essere aperti o bloccati in base alla struttura che collega i tre nodi

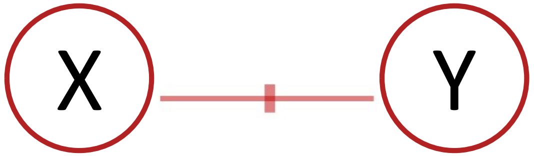
- Possiamo spiegare i flussi associativi considerando la significatività dei coefficienti di X e di Z

EC – base con due nodi



- X causa Y
- coefficiente di X significativo

$$Y \not\perp\!\!\!\perp X \qquad [\text{x and y are assocaited}]$$



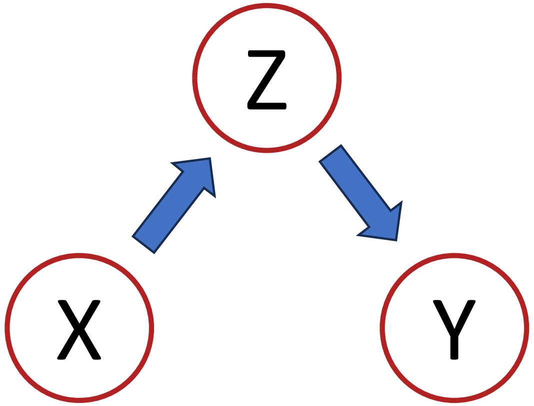
- Non ci sono rapporti di causa-effetto tra X e Y
- Coefficiente di X non significativo

$$Y \perp\!\!\!\perp X \qquad [\text{x and y are not associated}]^{*}$$

*a eccezione di associazioni spurie (improbabili, ma esistono)

EC - Pipe

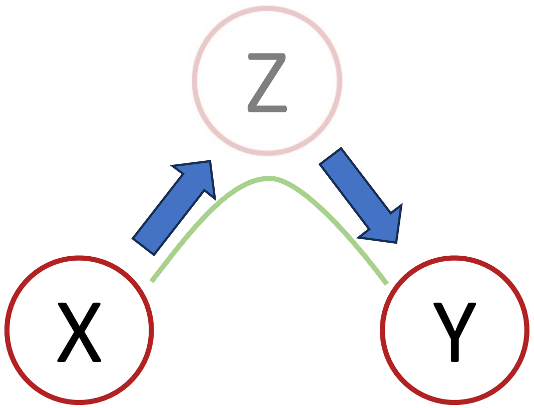
- PIPE: X è una causa di Y, mediata da Z



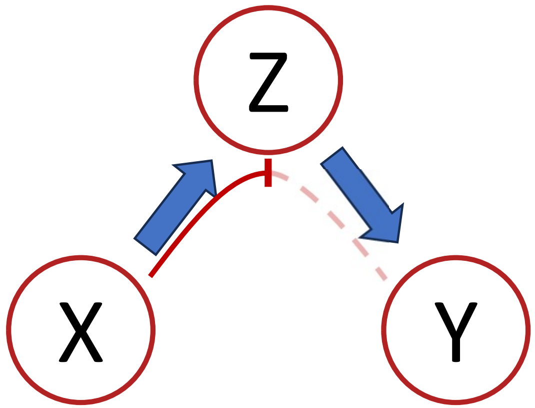
- Per capire come fluisce l'associazione, facciamo due regressioni lineari: nella prima Y è spiegato solo da X, nella seconda è spiegato sia da X che da Z:
    1. Y ~ X → Il coefficiente di X è significativo (associazione aperta)
    2. Y ~ X + Z → Il coefficiente di X non è significativo (associazione chiusa)

EC - Pipe

- Indipendenze condizionali della PIPE:



Y ~ X (non stratifico/condiziono per Z)
Flusso associativo aperto (linea verde)



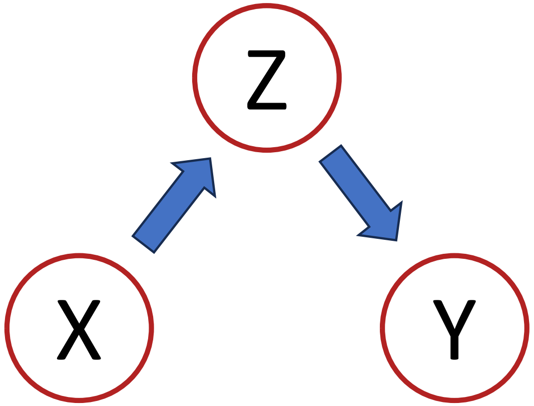
Y ~ X (stratifico/condiziono per Z)
Flusso associativo chiuso (linea rossa)

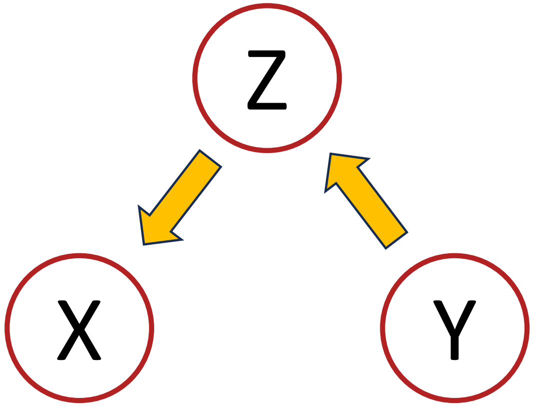
$$Y \not\!\perp\!\!\!\perp X \qquad \text{[x and y are assocaited]}$$

$$Y \perp\!\!\!\perp X|Z \qquad \text{[x and y are not associated, conditional on Z]}$$

EC - Pipe

- E se ho pipe con direzionalità opposte?

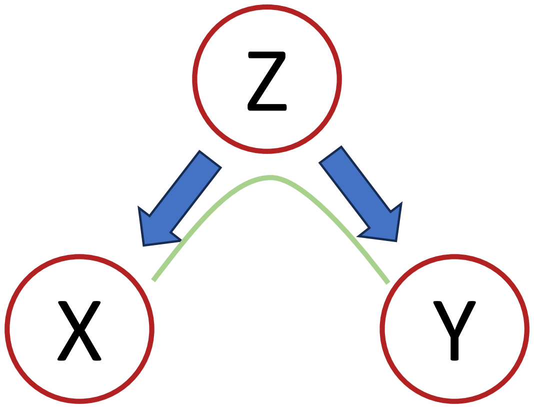




- ATTENZIONE: statisticamente parlando queste due strutture sono identiche!
- Le indipendenze condizionali sono le medesime: vi è associazione tra le variabili X e Y a meno che non stratifichi per la variabile Z

EC - Fork

- FORK: Z è una causa comune sia di X che di Y

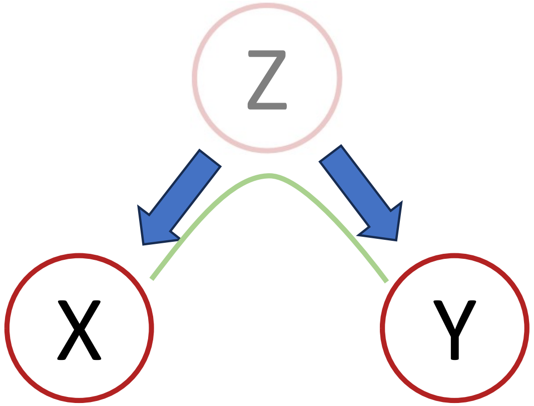


$Y \not\perp\!\!\!\perp X \qquad \text{[x and y are assocaited]}$

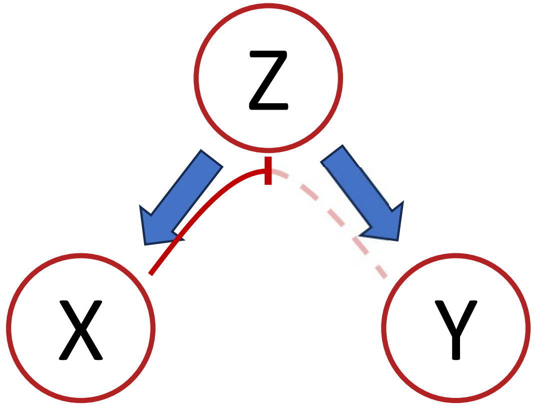
$Y \perp\!\!\!\perp X|Z \quad \text{[x and y are not associated, conditional on Z]}$

EC - Fork

- Indipendenze condizionali della fork:



Y ~ X (non stratifico/condiziono per Z)
Flusso associativo aperto (linea verde)



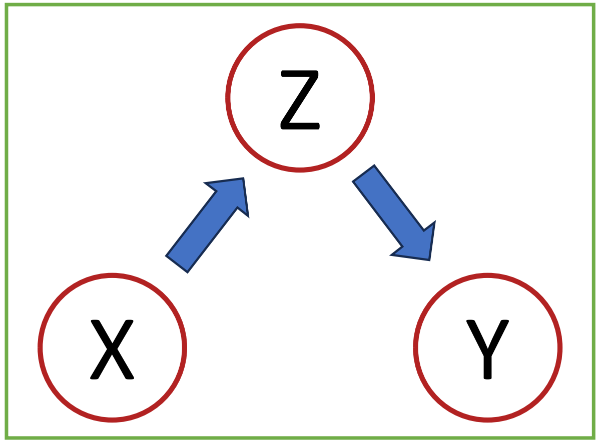
Y ~ X + Z (stratifico/condiziono per Z)
Flusso associativo chiuso (linea rossa)

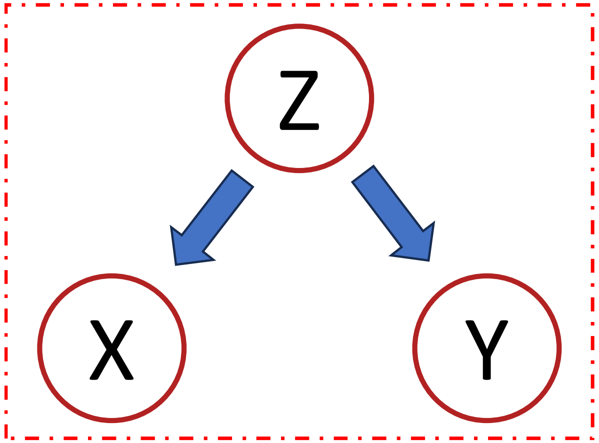
$Y \not\perp\!\!\!\perp X$ [x and y are assocaited]

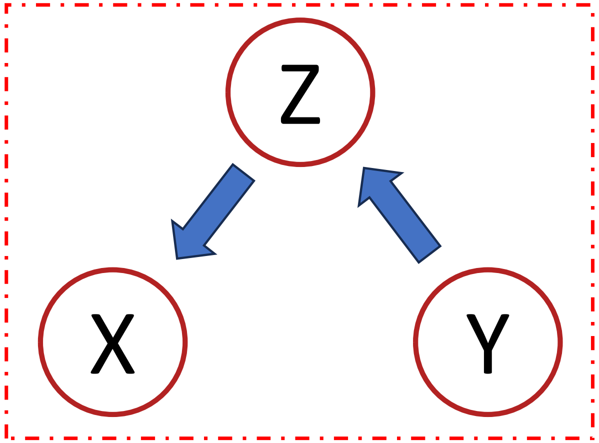
$Y \perp\!\!\!\perp X|Z$ [x and y are not associated, conditional on Z]

EC – Forks & Pipes

- FORK e PIPE hanno le stesse indipendenze condizionali !!!
- Senza conoscenze pregresse sono pertanto indistinguibili !!!







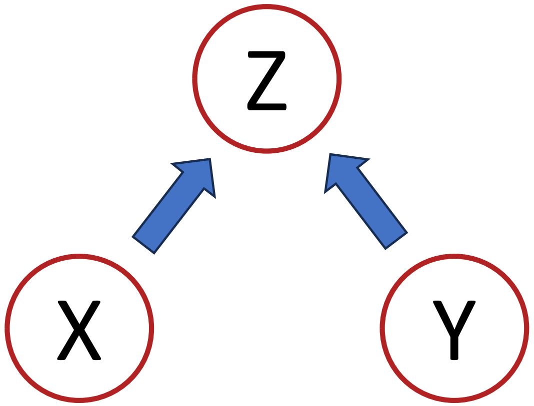
Unico caso in cui Y è una conseguenza (mediata da Z) di X

$$Y \not\perp\!\!\!\perp X \qquad \text{[x and y are assocaited]}$$

$$Y \perp\!\!\!\perp X|Z \qquad \text{[x and y are not associated, conditional on Z]}$$

EC - Collider

- COLLIDER: X e Y sono entrambi cause di Z

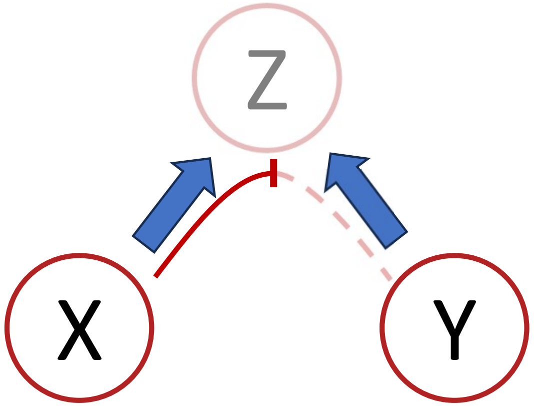


$$Y \perp\!\!\!\perp X \qquad \text{[x and y are not associated]}$$

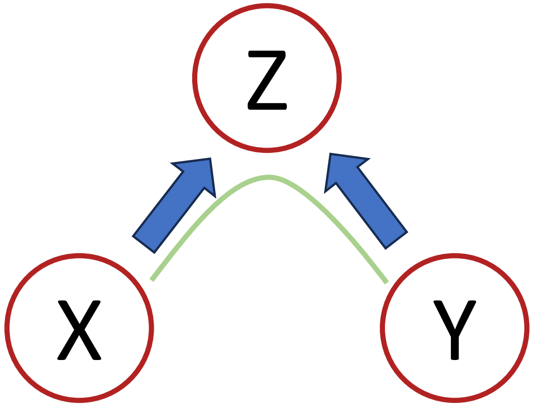
$$Y \not\perp\!\!\!\perp X|Z \qquad \text{[x and y are associated, conditional on Z]}$$

EC - Collider

- COLLIDER: X e Y sono entrambi cause di Z



Y ~ X (non stratifico/condiziono per Z)
Flusso associativo chiuso



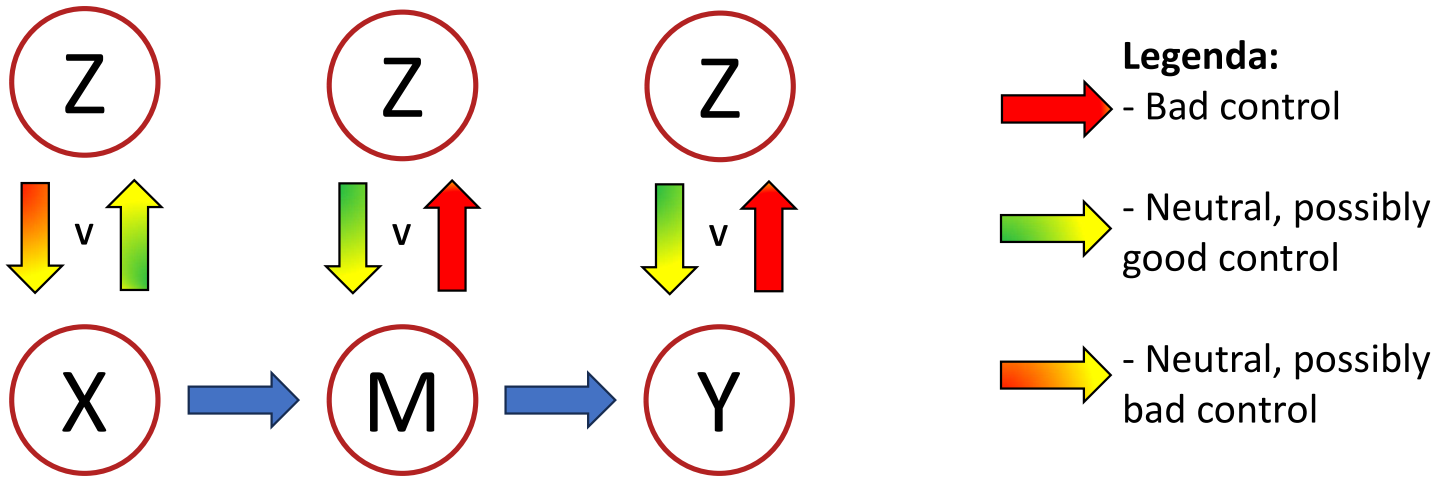
Y ~ X+Z (stratifico/condiziono per Z)
Flusso associativo aperto

$$Y \perp\!\!\!\perp X \qquad \text{[x and y are not associated]}$$

$$Y \not\perp\!\!\!\perp X|Z \qquad \text{[x and y are associated, conditional on Z]}$$

EC: Descendant

- DESCENDANT: quando una variabile è causa o conseguenza diretta di una variabile di interesse.
- In genere il loro comportamento è ambiguo; potrebbero aumentare la precisione, aumentare il bias o essere neutrali.
- Possono essere usati come proxy di una variabile non nota.



Ora proviamo su R