# Indice

*Introduzione parte II* 2

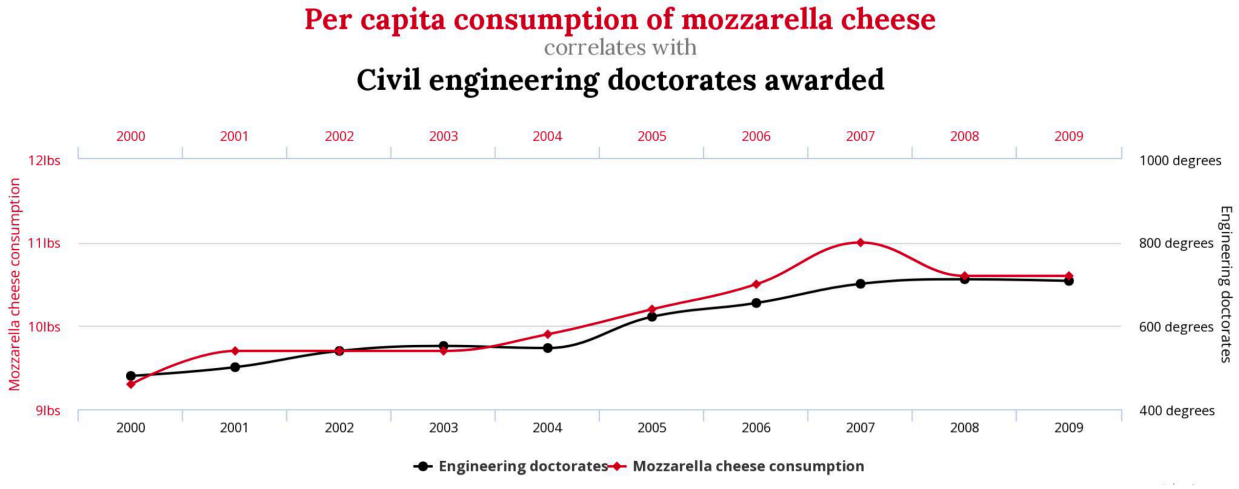
*DAGs* 4

*Modello del Controfattuale* 7

# Introduzione parte II

Questa parte di corso si concentra sul condurre inferenza su i dati disponibili. In particolare, la statistica classica ci può aiutare a trovare correlazione tra i dati, ci avverte che correlazione non è causalità[1], ma non ci dice quale sia la causa di un fenomeno e come si identifica. Le risposte a queste domande non risiedono nei meri dati.

[1] Per esempio, se i lavoratori $L$ di un'azienda sono proporzionali al fatturato $F$ tramite una costante di proporzionalità $k$ , allora $L = kF$, $L - kF = 0$ e $F = L/k$ sono tutte corrette, non abbiamo informazioni su quale sia il driver, se il fatturato o il numero di lavoratori.




Correlazione significa co-variabilità: variabili correlate tendono a variare assieme. La causalità implica una relazione tra due variabili di cui una è causa dell'altra. Si noti che la correlazione è condizione necessaria ma non sufficiente per la causalità.

## *Ceteris Paribus Condition*

Prendiamo come esempio l'analisi tra il debito sottoscritto dagli studenti e i loro risultati accademici, e supponiamo di trovare una correlazione negativa. Supporre allora che il debito abbia un'influenza negativa sulla performance accademica degli studenti può essere fuorviante: il fatto che lo studente venga da una famiglia meno abbiente può infatti essere ragionevolmente la causa di ambi i fenomeni. Questo vuol dire che il confronto non è *a parità di condizioni* i.e. *ceteris paribus*. Il problema principale dell'inferenza nasce dal fatto che non possiamo confrontare i risultati accademici degli stessi studenti se si indebitassero o meno. L'econometria cerca quindi di costruire situazioni simili ad hoc, ovvero usa i dati per trovarsi il più possibile nella situazione di "a pari condizioni", in mo-

do da evitare i problemi che possono generare stime scorrette dell'effetto causale. Questo lavoro sui dati viene descritto dal *do operator*.

## *Do Operator*

Immaginiamo di predirre $Y$ usando $X$ su certi dati osservazionali. In generale

$$P(Y|X) = \frac{P(X,Y)}{P(X)} \tag{1}$$

Questa non è però una relazione causale, ma solo condizionata. Occorre dunque introdurre il do operator, condizionando quindi anziché su $X$ su do$(X)$. In altre parole, dovremmo calcolare $P(Y|\text{do}(X))$, che è la distribuzione condizionata di $Y$ se fossimo in grado di intervenire nel processo di generazione dei dati e imporre $X = x$.

# DAGs

Nel 2000 J. Pearl ha introdotto un chiaro e innovativo approccio grafico al tema della causalità: *Directed Acyclic Graphs*. L'obiettivo di un DAG è quello di disegnare un sistema causale per rappresentare esplicitamente tutte le cause dell'outcome d'interesse. Il modello di causalità è semplificato in quanto:

- Assume un effetto omogeneo su tutte le osservazioni
- Utilizza solo outcome osservabili e non potenziali
- Si concentra sull'effetto medio incondizionato del trattamento
- Non specifica il tipo di relazione tra le variabili

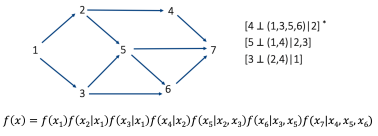
Un po' di terminologia riguardante i grafici aciclici orientati:

- Orientato: tutte le relazioni puntano da una causa a un effetto
- Aciclico: partendo da un qualunque vertice non possiamo tornare allo stesso
- Percorso: una qualsiasi sequenza di vertici orientati in una qualsiasi direzione
- Percorso diretto (o causale): un percorso in cui tutti i nodi puntano in una sola direzione

Vengono inoltre detti:

- Genitore: una causa diretta di una variabile
- Figlio: l'effetto diretto di una certa variabile
- Antenato: una causa diretta o indiretta di una certa variabile
- Discendente: un effetto diretto o indiretto di una certa variabile

### *DAG - Modello di Markov*




Un DAG è un modello in cui le variabili dipendono solo dai propri genitori e sono indipendenti dalle altre variabili. La funzione di probabilità di massa è

$$f(x) = \Pi_{i=1}^{N} f(x_i|x_{iPA}) \tag{2}$$

Se la relazione tra A e B è mediata da C, C è detto mediator

Se A e B concausano C, C è un collider
Se A e B sono causati da C, C è un cofounder
Attenzione perché il cofounder è esattamente ciò che crea problemi nell'analisi: infatti è facile confondere la correlazione tra A e B per causalità, mentre la causa della correlazione è C. Un percorso causale può essere immediato $A \to B$ o mediato, come per esempio $A \to B \to C$. L'effetto totale di A su B è la combinazione di tutti i percorsi immediati e mediati da A a B.

L'utilizzo dei DAG permette di offrire una rappresentazione delle relazioni causali, chiarire le domande di ricerca ed evidenziare i concetti rilevanti, rendere esplicite le assunzioni dei nostri modelli, indentificare appropriatamente le variabili da inserire nell'analisi e ottenere risultati più affidabili, riducendo possibili bias. Per costruire un DAG:

- Articolare la domanda di ricerca identificando la causa e l'effetto a cui si è interessati («qual è l'effetto di A su B?»)
- Identificare altre variabili rilevanti per la relazione, come collider e mediator
- Identificare variabili cofounder
- Identificare eventuali variabili non misurate o non misurabili
- Identificare possibili processi di selezione nello studio

Negli anni sono stati creati strumenti a supporto della costruzione di un DAG come `http://www.dagitty.net/`

### Paradosso di Simpson

Situazione statistica nella quale un trend o una relazione che è osservata tra diversi sottogruppi sparisce quando i gruppi sono combinati. In altre parole, dividendo i dati in gruppi, le conclusioni sono diverse rispetto a quelle derivanti da un'analisi aggregata. Una back-door path (BDP) è un percorso che è diretto verso D da una parte e termina verso Y dall'altra, cioè si ha una connessione tra D ed Y che non segue il percorso delle frecce. Lasciare un BDP aperto introduce un bias e non permette di indentificare l'effetto causale per via della presenza del cofounder. Per identificare la causalità, bisogna quindi chiudere tutti i BDP aperti.

In questo esempio avremmo potuto controllare sia per C che per A, identificando correttamente la causalità tra D ed Y, ma allora perché non controllare per tutte le variabili sempre, senza perdere tempo capendo le relazioni tra le varaibili? Perché controllare variabili senza criterio può invalidare l'analisi qualora:

- La variabile controllata sia un mediator, cioè un discendente di D in un percorso diretto verso Y
- La variabile controllata sia un collider in una back-door path da D a Y

Per fare un esempio si supponga che il merito creditizio (A) e la validità di un progetto (B) influenzino la ricezione di un prestito (P = 1 se il prestito è concesso, 0 altrimenti). Ipotizzando che A e B siano indipendenti

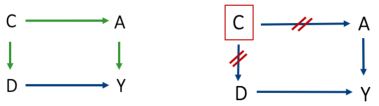



Figura 1: Sulla sinistra, un BDP aperto. Sulla destra, controllando C, eliminiamo l'impatto di C su D ed A. La relazione fra D e Y è ora determinata solo da un percorso diretto.

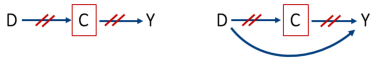



Figura 2: Se controllassimo per C, perderemmo l'intero (caso 1) o parte (caso 2) impatto di D su Y

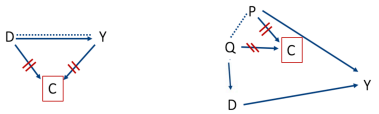
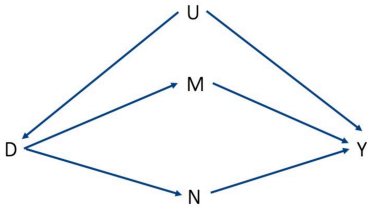



Figura 3: Controllare un collider può portare a creare un'associazione fra le variabili che lo concausano, anche quando queste sono indipendenti tra loro. Questo può portare all'apertura di un percorso controllando per un collider.

e controllando per P osserveremmo una falsa correlazione tra A e B. In particolare, se selezioniamo solo le imprese con P = 1, se esse hanno un progetto poco convincente dovranno avere un merito creditizio molto alto, creando una relazione di correlazione inversa tra le due variabili incorrelate.  Criterio del back-door Condizioni sufficienti per l'identificazione dei nessi causali nei DAG:

- Identificare tutti i back-door path da D a Y
- Identificare i mediator
- Bloccare solo i BDP aperti, controllandone le variabili cofounder
- Stare attento a non controllare per mediator e collider

Criterio del front-door Potrebbe esistere una variabile U non osservabile (e quindi non controllabile), allora il BDP $D \leftarrow U \rightarrow Y$ non può essere chiuso.
Nell'esempio di destra, posso semplicemente stimare l'effetto di D su M ed N, e poi di queste su Y. Combinando l'effetto delle due avremo l'effetto di D su Y. Per poter applicare il criterio del front-door, le variabili M ed N devono essere




- Esaustive: la combinazione di M ed N catturano tutto l'effetto di D su Y
- Isolate: tutti i back-door path da M a N sono bloccati un volta controllata D ( cioè non esistono cofounder per M e N)

# Modello del Controfattuale

Dati:
$Y_{0i}$ = Stato di salute della persona $i$ senza copertura assicurativa
$Y_{1i}$ = Stato di salute della persona $i$ con copertura assicurativa
L'effetto causale della copertura assicurativa è $Y_{1i} - Y_{0i}$. Se potessi osservare la stessa persona potrei così stimare l'impattto della copertura assicurativa. Questo è impossibile nella realtà, quindi ci limitiamo a confrontare persone diverse con e senza polizza. Supponiamo quindi di osservare due persone, $i$ che sottoscrive la polizza (trattata) e $j$ no (controllo). Nella realtà conosciamo solo $Y_{1i}$ e $Y_{0j}$, e non le altre. Un modo per confrontare la validità della polizza può essere $Y_{1i} - Y_{0j}$, che può essere riscritto come: $Y_{1i} - Y_{0j} = Y_{1i} - Y_{0i} + Y_{0i} - Y_{0j}$, ovvero l'effetto causale del trattamento su $i$, più la differenza tra $i$ e $j$ se entrambi scegliessero di non sottoscrivere la copertura assicurativa. Questo secondo termine descrive la fragilità di $i$ rispetto a $j$, ed è ciò che, nell'obiettivo d'individuare la relazione causale tra trattamento e outcome, viene definito come *selection bias* Il confronto univariato quindi non permette di valutare in modo corretto la relazoine causale tra trattamento e outcome a causa del selection bias, ovvero della differenza tra trattate e controllo prima del trattamento.

Estendendo le nostre considerazioni a un campione di trattate e controllo composti da più di un'osservazione:

$$\mathbb{E}(Y|D=1) - \mathbb{E}(Y|D=0) = \text{Effetto causale}(k) + \text{selection bias} \quad (3)$$

Se l'unica ragione di selection bias fossero differenze che possono essere misurate, allora non è un problema utilizzare il modello controfattuale: infatti basta prendere campioni caratterizzati da trattate simili. Il trattamento può essere rappresentato con una variabile dicotomica: $D_i = 1$ se $i$ è trattato, $D_i = 0$ altrimenti. Quindi il gruppo dei $D_i = 0$ è definito gruppo di controllo (C), mentre il gruppo dei $D_i = 1$ è il gruppo di trattamento (T).

Definiamo l'*outcome potenziale*:
$Y_{1i} =$ outcome potenziale se l'unità $i$ è trattata ($D_i = 1$)
$Y_{0i} =$ outcome potenziale se l'unità $i$ non è trattata ($D_i = 0$
$Y_i = Y_{0i} + (Y_{1i} - Y_{0i})D_i$

L'impatto causale è $Y_{1i} - Y_{0i}$, ma non possiamo osservare entrambi per uno stesso campione, si necessita quindi di un controfattuale (ovvero un'approssimazione di $Y_0$) per valutare l'efficacia del trattamento: *problema fondamentale dell'inferenza causale*.
Per ottenere la stima analizziamo molteplici unità per ricavare l'effetto causale aggregato, se l'effetto del trattamento è omogeneo, corrisponde

all'effetto individuale (ovvero si può assumere k costante). Perché questo approccio sia valido occorre che non ci siano effetti di spillover tra gruppo di controllo e di trattamento. In caso di spillover, l'outcome del gruppo di controllo è influenzato da quello del gruppo trattamento, comportando un bias nel risultato. SUTVA (Stable Unit treatment value assumption): l'outcome potenziale di qualsiasi unità $i$ non varia con il trattamento assegnato ad altri, e per ogni unità non ci sono differenti versioni dello stesso trattamento che porta a differenti outcome potenziali. Un esempio di spillover può essere un gruppo di dipendenti che segue un corso di formazione e poi condivide le informazioni con altri dipendenti che non lo hanno seguito.

*Eterogeneità del trattamento*

ATE (Average Treatment Effect):

$$\mathbb{E} = \mathbb{E}(Y_1 - Y_0) = \mathbb{E}(Y_1) - \mathbb{E}(Y_0) \tag{4}$$

ATT (Average Treatment Effect on Treated):

$$\mathbb{E}(\delta|D = 1) = \mathbb{E}(Y_1 - Y_0|D = 1) - \mathbb{E}(Y_0|D = 1) \tag{5}$$

ATU/ATNT(Average Treatment Effect on Untreated):

$$\mathbb{E}(\delta|D = 0) = \mathbb{E}(Y_1 - Y_0|D = 0) = \mathbb{E}(Y_1|D = 0) - \mathbb{E}(Y_0|D = 0) \tag{6}$$

L'ATE è una combinazione lineare di ATT e ATU. Indicando con $\mu$ la percentuale di popolazione trattata: ATE $= \mu$ATT $+ (1 - \mu)$ATU, che può essere scritto anche come

$$\mathbb{E}(\delta) = Pr(D = 1)\mathbb{E}(\delta|D = 1) + Pr(D = 0)\mathbb{E}(\delta|D = 0) \tag{7}$$

Possiamo stimare ATT e ATU utilizzando uno stimatore ingenuo (*naive estimator*)

$$NE = \mathbb{E}(Y_1|D = 1) - \mathbb{E}(Y_0|D = 0) \tag{8}$$

$$NE = \mathbb{E}(Y_1|D = 1) - \mathbb{E}(Y_0|D = 1) + \mathbb{E}(Y_0|D = 1) - \mathbb{E}(Y_0|D = 0) \tag{9}$$

Dove i primi due termini sono l'ATT e gli ultimi due il Bias, che indica quanto sbaglieremmo se approssimassimo ciò che non osserviamo $\mathbb{E}(Y_0|D = 1)$ con ciò che osserviamo $\mathbb{E}(Y_0|D = 0)$

$$NE = \mathbb{E}(Y_1|D = 0) - \mathbb{E}(Y_0|D = 0) + \mathbb{E}(Y_1|D = 1) - \mathbb{E}(Y_1|D = 0) = ATU + BIAS \tag{10}$$

Conseguentemente, anche l'ATE stimato usando l'NE non sarà privo di bias, in quanto

$$NE = \pi(ATT + BIAS_{NEATT}) + (1 - \pi)(ATU + BIAS_{NEATU}) \tag{11}$$