Come *non* usare la correlazione

Roberto Fini

Dove si narra di incendi & vigili del fuoco

Immaginate di leggere sul giornale una notizia di questo tipo: *"Furioso incendio distrugge una fabbrica di vernici. Impiegate sul posto numerose squadre di vigili del fuoco. Molte le vittime fra i lavoratori"*. Certo, una di quelle notizie che nessuno vorrebbe mai leggere, ma capita purtroppo. Ora immaginate di voler cercare di capire cosa sia successo, cosa ha provocato così tanti morti. Un corto circuito? Una disattenzione? Un attentato?
Beh, se vi accontentate di poco già il titolo del giornale vi può dare qualche indizio. Già perché una correlazione c'è. È lì, bella e pronta: molti pompieri accorsi sul luogo dell'incendio, molte vittime. Da persona coscienziosa quale siete non vi accontentate: avete bisogno di qualche ulteriore verifica. Andate in Google, digitate qualcosa come "incendi con vittime" e cosa scoprite? Scoprite che ogni vota che un incendio ha provocato dei morti, erano state impegnate molte squadre di vigili del fuoco.
Dunque, la relazione esiste ed è consistente! Quei morti pesano sulla coscienza dei pompieri! Si potrebbe persino ipotizzare che se fossero state impegnate meno squadre di vigili del fuoco i morti darebbero stati di meno. E se nessun pompiere fosse accorso, di morti non ce ne sarebbero stati!
Avrete certo compreso che si tratta di un ragionamento per assurdo, ma per comprenderne bene il meccanismo lasciateci ancora insistere su che cosa si basa l'evidente errore. È assolutamente vero, in effetti, che tra la presenza di molti pompieri e l'alto numero di vittime sembra esserci una relazione di causa-effetto. Allora dove sta l'errore? Che cosa abbiamo trascurato nel nostro ragionamento?
Come spesso accade quando si cerca di mettere in rapporto due fatti, ci si accontenta di verificare la relazione più evidente, che è anche in genere la più facile perché è lì sotto il nostro naso.
L'errore consiste nel fermarsi subito nell'indagine, trascurando un altro elemento che sarebbe stato invece determinante: le dimensioni dell'incendio. Vediamo di esplicitare i passaggi logici: ogni volta che si verifica un incendio di grandi proporzioni (c), accorrono molte squadre di vigili (a), ma in un incendio di grandi dimensioni è probabile che ci siano molte vittime (b)!
Cioè, non è vera l'ipotesi che

$$b = f(a)\,,$$

ma è vero che

$$a = f(c)$$

e che, contemporaneamente,

$$b = f(c)\ .$$

Così è tutto più chiaro. Ed è elementare nella sua semplicità, ma è sempre opportuno mettere in evidenza i trabocchetti logici in cui è facile cadere quando si cerca di spiegare eventi e fenomeni utilizzando relazioni fra di essi che, all'apparenza, sono convincenti.

Rondini & matrimoni

Un altro esempio? Si è osservato che sembra esistere una relazione diretta fra numero di stormi di rondini in cielo (x) e numero di matrimoni celebrati (y). Cioè: tanto maggiore è la quantità di rondini, tante più coppie convoleranno a nozze. In formula:

$$y = f(x).$$

Possibile? Beh, i dati sono inequivocabili: sembra che la relazione esista e che, come nel caso precedente, sia piuttosto consistente. Dunque, potremmo ipotizzare che se per qualche evento naturale le rondini smettessero di solcare i nostri cieli, nessuna coppia deciderebbe di sposarsi. Se la relazione fosse vera sarebbe un bel guaio!

Per fortuna (sia delle rondini che delle coppie di fidanzati) le cose non stanno esattamente come la relazione precedente ipotizza. In questo caso l'evento che si è omesso riguarda l'andamento stagionale (z) che collega i due fenomeni: si dà il caso, infatti, che le rondini compaiono più numerose nei nostri cieli in primavera ed in autunno, cioè si ha che

$$x = f(z).$$

Al tempo stesso possiamo facilmente verificare che primavera ed autunno sono le stagioni scelte da molte coppie per sposarsi. Cioè:

$$y = f(z).$$

Non c'è alcun legame diretto fra x ed y. Invece c'è un legame che lega, contemporaneamente x a z ed y a z. Solo in questo modo può evitare di fare brutte figure!

Tasso di mortalità & riti matrimoniali

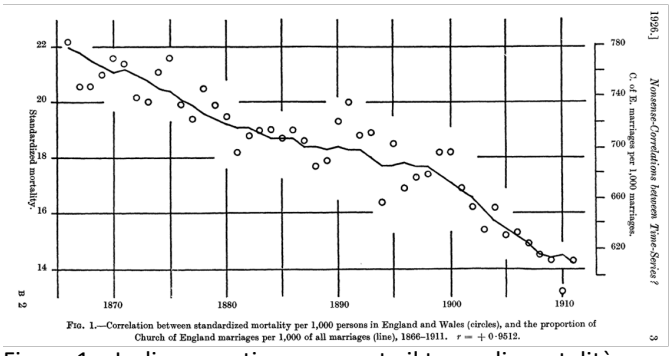

Figura 1 – La linea continua presenta il tasso di mortalità, mentre i "pallini" raffigurano la percentuale di matrimoni celebrati con rito anglicano.

Ora osservate il grafico riprodotto in figura 1. Venne pubblicato a corredo di un articolo scritto dallo statistico inglese George Udny Yule (1871-1951) nel 1926. Lo studioso si occupava già da molto tempo di correlazioni fra fenomeni e in questo articolo si propone di illustrare i pericoli legati alla creazione di legami impropri fra eventi che in realtà non hanno nulla a che vedere uno con l'altro. Torniamo alla figura: si tratta di una serie storica nella quale vengono presentati i tassi di mortalità (asse Y di sinistra) e la percentuale di matrimoni con rito anglicano (asse Y di destra). È evidente: una correlazione c'è! In realtà fra i due fenomeni non esiste nessun rapporto; e come potrebbe? Si tratta di eventi del tutto indipendenti uno dall'altro: dunque, il fatto che i loro andamenti siano simili è una semplice coincidenza.

A questo tipo di coincidenze viene dato il nome di *correlazione spuria*.

La "maledizione di Ramsey"

Nella grande maggioranza dei casi le correlazioni spurie sono frutto di pure e semplici coincidenze statistiche e per di più sono facilmente spiegabili (pensate al caso della "correlazione" fra numerosità degli stormi di rondine e matrimoni celebrati). In altri casi c'è poco da spiegare: si

tratta di fenomeni non correlati che per puro caso hanno andamenti analoghi. In altri casi si tratta di costruzioni a tavolino, a volte fatte per gioco e in altri casi per danneggiare qualcuno.
Ricordate la commedia *La patente* di Pirandello? Il personaggio principale, Rosario Chiàrchiano, ha fama di essere uno "iettatore" e per questo viene evitato da tutti. Prende allora una decisione: chiede ad un giudice di certificare questa sua fama con un attestato, in modo tale da poterla usare in modo, diciamo così, professionale.
Si tratta di un caso paradossale? Certo, ma osservate la tabella 1: si riferisce alla vicenda di Aaron Ramsey, un bravo centrocampista che attualmente gioca nella Juventus, ma che ha militato per molti anni nella squadra inglese dell'Arsenal. Aaron segna spesso e i suoi gol sono stati di frequente decisivi per la sua squadra. Ma, almeno degli ambienti dei VIP, più di qualcuno dovrebbe sperare in un infortunio del giocatore che lo tenga fuori dal campo per un periodo di tempo illimitato.

Figura 2 – Ramsey esulta dopo aver segnato un gol. Lui non ne è consapevole, ma un VIP sta per morire…

Perché tanta paura dei risultati calcistici di Aaron? Perché con una quasi perfetta coincidenza ad ogni rete segnata dal nostro (colonne 1 e 2 della tabella) qualche personaggio famoso muore (colonne 3 e 4 della tabella)! Non abbiamo evidenze statistiche che lo provino , ma sarebbe comprensibile se fra i VIP di mezzo mondo non ci fossero molti tifosi dell'Arsenal…

Tabella 1 – "La maledizione di Ramsey": rapporto fra gol segnati da Aaron Ramsey nell'Arsenal e morte di un personaggio famoso

(1) Squadre in campo	(2) Data della partita	(3) Personaggio	(4) Data della morte
Liechtenstein vs Galles	14/10/2009	Andres Montes[1]	16/10/2009
Galles vs Scozia	14/11/2009	Antonio de Nigris[2]	16/11/2009
Arsenal vs M. United	01/05/2011	Osama Bin Laden[3]	02/05/2011
Tottenham vs Arsenal	02/10/2011	Steve Jobs[4]	05/10/2011
O. Marsiglia vs Arsenal	19/10/2011	Muammar Gheddafi[5]	20/10/2011
Sunderland vs Arsenal	11/02/2012	Whitney Houston[6]	11/02/2012
Regno Unito vs Corea	04/08/2012	Chavela Vargas[7]	05/08/2012
Scozia vs Galles	22/03/2013	Bebo Valdes[8]	22/03/2013
Arsenal vs Wigan	14/05/2013	Jorge Rafael Videla[9]	17/05/2013
O. Marsiglia vs Arsenal	18/09/2013	Ken Norton[10]	18/09/2013
Cardif vs Arsenal	30/11/2013	Paul Walker[11]	30/11/2013
Hull city vs Arsenal	20/04/2014	"Huracan" Carter[12]	20/04/2014
Norwich vs Arsenal	11/05/2014	H.G. Giger[13]	12/05/2014
Arsenal vs Manchester City	10/08/2014	Robin Williams[14]	11/08/2014

[1] Giornalista sportivo spagnolo
[2] Calciatore messicano
[3] Terrorista internazionale
[4] CEO di Apple
[5] Dittatore libico
[6] Attrice e cantante USA
[7] Cantante messicana
[8] Musicista cubano
[9] Presidente dell'Argentina
[10] Pugile e attore USA
[11] Attore USA
[12] Pugile USA
[13] Pittore surrealista svizzero
[14] Attore USA

Probabilmente avete già scoperto il trucchetto: si prendono le date dei gol di Ramsey e si verifica chi è morto nel giorno stessa o al massimo qualche giorno dopo. Certamente (o quasi) si troverà qualche personaggio famoso che ha tirato le cuoia in quegli stessi giorni! Vi pare impossibile? Se è cosi vi proponiamo un giochetto facile facile (tabella 2)[15]: cercate in rete le date dei gol di Cristiano Ronaldo e poi, in Google, scrivete le date di ogni gol. Perbacco! anche Ronaldo sembra essere un infallibile iettatore: ogni volta che segna qualcuno di famoso muore…
Per di più il Nostro segna di frequente: una vera strage!

Tabella 2 – "La maledizione di Ronaldo": rapporto fra gol segnati da Cristiano Ronaldo e morte di un personaggio famoso

(1) Squadre in campo	(2) Data della partita	(3) Personaggio	(4) Data della morte
Portogallo vs Kazakistan	20/08/2003	Ian MacDonald[16]	20/08/2003
Portogallo vs Albania	11/10/2003	Paolo Bozzi[17]	11/10/2003
Portogallo vs Inghilterra	18/02/2004	Prosperino Gallipoli[18]	18/02/2004
…	…	…	…
Portogallo vs Svizzera	05/06/2019	Takko Ishimori[19]	05/06/2019
Portogallo vs Paesi Bassi	09/06/2019	Iain Banks[20]	09/06/2019

Prendere lucciole per lanterne

Va bene: torniamo seri! Proviamo a vedere se esiste un rapporto fra disoccupazione e possesso di un titolo di studio uguale o superiore a ISCED3[21]: il quesito cui ci proponiamo di rispondere e se esiste un legame inverso fra due variabili come avere un titolo di studio almeno ISCED3 e tasso di disoccupazione. Si tratta di una valutazione importante nell'ambito delle politiche educative: se il legame dovesse esistere, allora incoraggiare i giovani a prendere almeno il diploma di scuola superiore[22], potrebbe tradursi in un elemento di lotta alla disoccupazione.
Dunque, altro che rondini svolazzanti in cielo e matrimoni! Qui si parla di ben altro! Si parla del destino dei nostri giovani e dell'intera società! Quindi mettiamoci a tavolino e facciamo le persone serie[23]. Supponiamo una situazione nella quale prendiamo in considerazione due coorti di età, quella dei giovani (19-35 anni) e quella degli anziani (36-65 anni). In ciascuna delle due coorti demografiche vi saranno un certo numero di lavoratori (occupati + disoccupati) che hanno conseguito ISCED3 o superiori e altri che non hanno raggiunto questo obiettivo (tabella 3 relativa ad un ipotetico anno t_1):

Tabella 3 – Numero di lavoratori per coorti demografiche e titolo di studio conseguito

Lavoratori	ISCED 1 – 2	ISCED 3 – 5	Totale
Giovani 19 – 35 anni	20	80	100
Anziani 36 – 65 anni	120	30	150

[15] La tabella non è completa: i punti di sospensione rappresentano un invito a completarla aggiungendo righe con le altre indicazioni relative ai gol di Cristiano
[16] Critico musicale inglese
[17] Psicologo italiano
[18] Missionario italiano
[19] Doppiatore giapponese
[20] Scrittore scozzese
[21] Nella codifica internazionale dei titoli di studio ISCED1 corrisponde alla nostra licenza elementare, ISCED2 alla licenza di scuola media inferiore, ISCED3 al diploma di scuola superiore, ISCED4 al titolo di studio ottenuto frequentando corsi superiori non universitari (ITS), ISCED5 alla laurea, ISCED6 al dottorato di ricerca
[22] O addirittura rendere obbligatoria la frequenza scolastica fino all'età legale corrispondente a quella di ottenimento del titolo, come è orientamento di moltissimi Paesi.
[23] Con poche varianti l'esempio che segue è tratto da Wikipedia alla voce *Paradosso di Simpson*. Il paradosso di Simpson (dal cognome dello statistico che se ne occupò a lungo) indica una situazione in cui una relazione fra due eventi appare modificata, e a volte perfino invertita, dai dati in possesso a causa di altri fenomeni che non sono stati presi in considerazione nell'analisi (variabili nascoste).

Totale giovani + anziani	140	110	250

Ora prendiamo in considerazione i tassi di disoccupazione in ciascuna delle due coorti di età in relazione al titolo di studio raggiunto. In quasi tutti i Paesi si verifica una situazione nella quale il numero dei diplomati e laureati nella generazione degli anziani è molto inferiore rispetto ai pari titolo nella generazione dei giovani. Inoltre, per motivi legati alle caratteristiche dei mercati del lavoro, il tasso di disoccupazione riscontrabile fra i giovani è più elevato rispetto a quello che si riscontra fra gli anziani. La tabella 4 presenta questa situazione, come risultato dell'analisi di serie storiche pluriennali (t_n):

Tabella 4 – Tasso di disoccupazione per coorti demografiche e titolo di studio conseguito		
Tasso di disoccupazione	ISCED 1 – 2	ISCED 3 – 5
Giovani 19 – 35 anni	30%	15%
Anziani 36 – 65 anni	5%	3,33%

Per quanto i dati delle tabelle 3 e 4 siano del tutto ipotetici, essi sono sostanzialmente vicini alla realtà statistica di molti Paesi industrializzati:

a. Il numero (assoluto) di lavoratori giovani è inferiore al numero (assoluto) di lavoratori anziani (tabella 3);
b. Il tasso di disoccupazione fra i lavoratori giovani è superiore al tasso di disoccupazione presente fra i lavoratori anziani (tabella 4);
c. Inoltre, per entrambe le coorti demografiche, il tasso di disoccupazione di coloro che hanno conseguito almeno ISCED3 è all'incirca la metà rispetto a quello di coloro che non lo hanno raggiunto (cfr. ancora tabella 4).

Avendo a disposizione questi dati, è possibile calcolare il numero di disoccupati in un anno determinato (t_1):

Tabella 5 – Numero di disoccupati previsti nell'anno t_1 in base alla serie storica t_n			
Disoccupati	ISCED 1 – 2	ISCED 3 – 5	Totale
Giovani 19 – 35 anni	6	12	18
Anziani 36 – 65 anni	6	1	7
Totale	12	13	25

A questo punto il tasso di disoccupazione nell'anno t_1 è facilmente utilizzando i dati della tabella 3 e della tabella 5. Dunque, ci aspettiamo che nell'anno considerato la percentuale di disoccupati sia quella riscontrabile nella tabella 6:

Tabella 6 – Tassi di disoccupazione nell'anno t_1 in base al titolo conseguito	
Percentuale di disoccupati	
ISCED 1 – 2	12/140 = 8,6%
ISCED 3 – 5	13/110 = 11,8%

I dati della tabella 6 probabilmente procureranno qualche sconcerto: il tasso di disoccupazione fra i titolari di ISCED3 e più, invece di essere la metà circa rispetto a quello di coloro che hanno titoli di studio inferiori, è in realtà più alto di circa un terzo. Dunque: studiare non sembra essere un buon affare, almeno dal punto di vista occupazionale.

Ma le cose non sono come sembra: non che i dati della serie storica della tabella 4 relativa a t_n siano sbagliati, né lo sono quelli della tabella 6 riguardanti l'anno t_1. Il fatto è che occorre un piccolo supplemento di indagine. Il paradosso (solo apparente) è dovuto al fatto che il tasso di disoccupazione fra i giovani, cioè la coorte di età che si caratterizza per avere una più alta percentuale di diplomati/laureati, è nettamente superiore rispetto a quello degli anziani che, al contrario, presentano una più bassa percentuale di ISCED3-5 (cfr. tabella 4 relativa alla serie storica t_n).

In altri termini, non è corretto trascurare l'importanza di due relazioni fondamentali, quelle:

a. fra disoccupazione ed età;
b. fra età e titolo di studio.

L'errore sta proprio in questo! Ed è tanto più rilevante in quanto farebbe prendere decisioni di politica economica ed educativa sbagliate: dopotutto, se il titolo di studio più alto non garantisce contro la disoccupazione, che motivo abbiamo di incentivare i giovani a raggiungerlo? In questo caso l'errore è quasi immediatamente visibile, ma nelle statistiche che riportano serie storiche reali può succedere di non accorgersi delle relazioni implicite esistenti fra le variabili prese in considerazione, limitandosi ad analizzare i dati aggregati senza incrociarli con le variabili essenziali. La contraddizione non verrebbe percepita e si potrebbe arrivare a conclusioni opposte a quanto accade nella realtà, con conseguenze in termini di scelte da fare potenzialmente molto gravi.
L'insidia maggiore contenuta nel risultato a cui si giunge analizzando i dati come fatto nell'esempio appena riportato è costituita dal fatto che i dati prodotti non sono, in sé, sbagliati: devono però essere letti in modo diverso rispetto a come può fare un lettore superficiale. In termini diversi affermare che "tra le persone con diploma ISCED3 o superiore si registrano più disoccupati rispetto a quanto si verifica tra le persone con titoli ISCED1/2" non è sbagliato, mentre sarebbe sbagliato utilizzare una logica di causa-effetto che porti ad affermare che "avere un diploma è causa di maggiore disoccupazione".
Poiché molto spesso la tentazione di usare logiche di causa-effetto è troppo forte per potervi rinunciare facilmente, disponendo di tutti i dati che servono si potrebbe affermare senza incorrere in alcun errore, né statistico né concettuale (cfr. ancora i dati della tabella 6):

a. i giovani sono circa sei volte più esposti alla disoccupazione rispetto agli anziani…
b. … ma, sia fra i giovani che fra gli anziani, avere un diploma almeno ISCED3 riduce il rischio disoccupazionale all'incirca alla metà.

Le maledette correlazioni spurie: quando mangiare gelati fa male

Esistono diverse ragioni per procedere con molta attenzione sul sentiero della correlazione. In primo luogo, perché è tentatore come il canto delle sirene per Ulisse: trovare che esiste un rapporto fra il fenomeno x e il fenomeno y può apparire, in quanto tale, una grande scoperta; ma già questo dovrebbe indurre a qualche prima cautela: "non è che quella che mi appare come una correlazione è invece solo frutto di una coincidenza".
D'accordo: ammettiamo senza difficoltà che essere vittime di una coincidenza può apparirvi come pura e semplice sfortuna, ma la sfortuna esiste. Ed è bene tenerne conto quando si ha a che fare con dati statistici. Se poi alla sfortuna aggiungiamo un'interpretazione del rapporto causa-effetto tale per cui vogliamo credere che

$$y = f(x)\,,$$

allora vuol dire che vogliamo dare una robusta mano alla sfortuna.
Un ultimo esempio a questo proposito? D'accordo! Avreste mai pensato che il consumo di gelati è collegato al numero di incendi boschivi? No, vero? Eppure, sembra essere così: quanto più gelati si consumano in un determinato periodo, tanto maggiore è il numero di incendi di boschi. Dunque, vietare il consumo di gelati potrebbe essere una buona mossa per ridurre la devastazione dei nostri boschi.
Per fortuna degli amanti del gelato, la questione si pone in modo diverso: il massimo consumo di gelato si verifica durante la stagione calda, quando la siccità è spesso esca di incendi. I due fenomeni, consumo di gelato (g) e incendi boschivi (b), non sono direttamente collegati, o almeno non nel senso che

$$b = f(g)\,,$$

e neppure vale la relazione contraria:

$$g = f(b)\,.$$

È vero: i due fenomeni sembrano collegati l'uno rispetto all'altro, ma in realtà si tratta di due fenomeni del tutto indipendenti, entrambi semmai legati in modo determinante ad una terza variabile come le temperature medie stagionali.
Restiamo ancora sul consumo di gelati: perché è collegato con le morti per affogamento in piscina? Facile no?

Un monumento che ricorda i pericoli del consumo di gelato…