“I numeri, a torturarli a lungo, confessano qualsiasi cosa” (R. Coase).
Usi corretti o scorretti nell’uso degli strumenti statistici

Roberto Fini

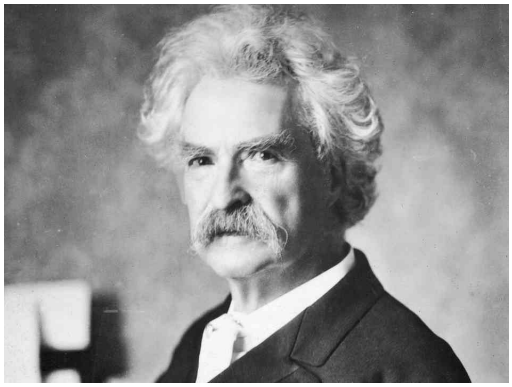


Figura1 – Mark Twain. “Ci sono tre tipi di bugie: le bugie, le bugie spudorate e le statistiche

Gli alligatori di New York e Nessie (e Nerone)

Di “bufale” negli ultimi anni si parla sempre più spesso. E con ragione: l’incredibile diffusione di strumenti che si servono della rete e dei social per diffondere notizie, non importa quanto vere o reali, rende effettivo il rischio che qualcuno prenda clamorosi abbagli e qualcun altro (se non le stesse persone…) li sfrutti a proprio uso e consumo.

Fino a qualche anno fa il termine “bufala” o, come oggi si tende a dire, la sua gemella “fake”, erano catalogate come “leggende metropolitane”. Ne ricordate qualcuna? Beh, giusto per rinfrescarvi la memoria: lo sapevate che le fogne di New York sono piene di alligatori? Nessuno li ha mai visti salvo una persona, che dichiarò di averne visto uno esplorandone i canali nel 1930. Da

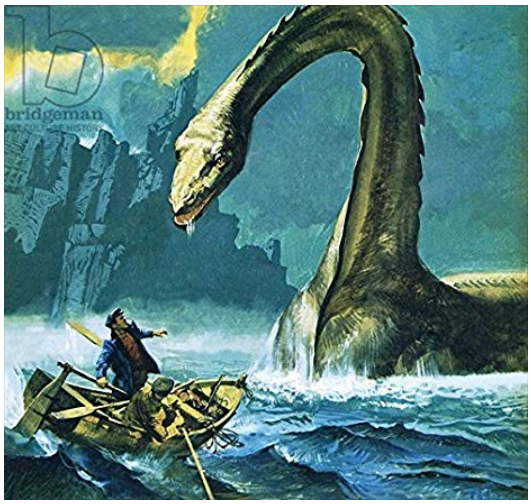


Figura 2 – Nessie: il mostro di Lock Ness

allora la notizia si è diffusa, è diventata "virale" come diremmo oggi, gli alligatori si sono moltiplicati fino a diventare una vera e propria specie endemica. Ma nessuno li vede davvero. Però state attenti potrebbero spuntare fuori nel vostro bagno[1].
A proposito di bagno: che non vi venga in mente di tuffarvi nelle acque del lago Lock Ness, in Scozia. Come tutti sanno, il lago è abitato da una mostruosa creatura che ogni tanto si fa viva in superfice. Il punto è che le numerose campagne di ricerca subacquea non ne hanno mai accertato l'esistenza. Come non esiste? E le foto scattate negli anni trenta? Ci dispiace darvi una così cocente delusione, ma gli stessi autori di quelle foto hanno confessato come fossero frutto di fotomontaggi. Però, se vi capita, fate un salto a Lock Ness e dintorni: di mostri neppure l'ombra, ma sarete in buona compagnia, visto che la leggenda di Nessie (così è stato battezzato *il* mostro) attira ogni anno un milione circa di turisti!
Qualcosa di meno esotico e più vicino a noi? D'accordo: tra gli imperatori romani che avete studiato in storia chi è stato il più "cattivo". Che domande? Nerone, certamente: si mise a suonare la cetra mentre Roma bruciava per un incendio da lui stesso fatto appiccare!
Beh, che Nerone non fosse esattamente una mammola in fiore e che non amasse particolarmente i cristiani (che furono artatamente accusati di aver provocato l'incendio) probabilmente è vero (o almeno verosimile), ma che si sia divertito a suonare mentre per sei giorni Roma bruciava, non è accertato. Vero: c'è Tacito che lo racconta e (per rimanere in ambito latino) "Tacito è uomo d'onore". E se non bastasse il più grande storico della romanità, la stessa cosa raccontano Svetonio, Plinio il Vecchio ed altri. Senonché non si può dare gran credito ad autori che, più o meno dichiaratamente, parteggiavano per l'aristocrazia senatoria ed erano contro Nerone che si appoggiava ai ceti popolari.
Avviciniamoci ai giorni nostri: all'alba dell'era di Internet, diciamo nel 1995, erano in pochi a dubitare del grande sviluppo che avrebbe avuto la rete. Si davano anche dei numeri: i più ipotizzavano una crescita di utilizzatori del 15% al mese. Non c'era convegno, comunicazione scientifica, piano di business che non desse per scontata una simile percentuale. Se una simile percentuale fosse stata anche solo vicina alla realtà e supponendo che nel gennaio del 1995 gli utenti fossero 1.000.000, il nuovo millennio si sarebbe aperto (marzo 2001) con il risultato che tutta la popolazione della Terra sarebbe stata connessa (cfr. figura 3).

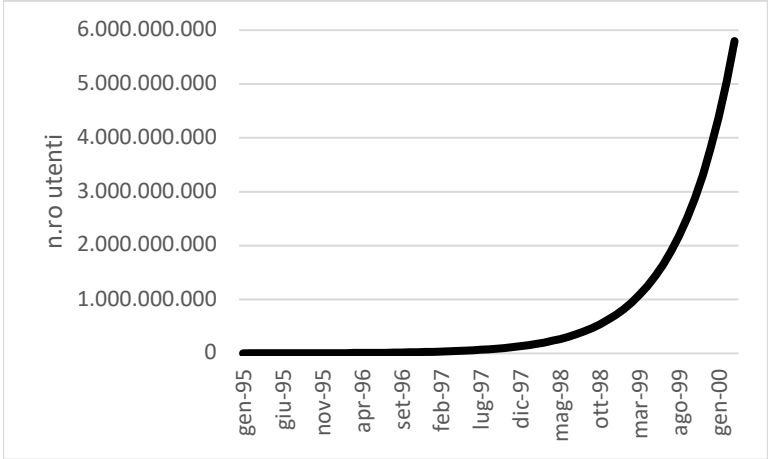


Figura 3 -

[1] Così il cantautore Samuele Bersani in una sua canzone (Coccodrilli, 2997): "In America lo sai che i coccodrilli/vengon fuori dalla doccia"

Ora, che Internet abbia conosciuto uno straordinario successo non c'è dubbio. Ma la realtà è ben lontana dalla fantasia di numeri sparati un po' a caso: secondo i più accreditati report, nel 2019 gli utenti Internet saranno diventati 4,3 miliardi. Una cifra incredibile, ma ben lontana da quanto si ipotizzava con una crescita del 15%, che avrebbe fatto raggiungere lo stesso traguardo nel gennaio del 2000. Sono passati quasi venti anni del nuovo millennio e a quello straordinario livello non si è ancora giunti.

Bene, penserete: l'entusiasmo può giocare brutti scherzi, ma è tutto sommato inoffensivo: in inglese si dice *wishfull thinking*, che liberamente tradotto potrebbe essere reso come "pensiero piegato ai desideri". Il problema che questa benevola interpretazione non fa i conti con la realtà. O peggio: può far prendere abbagli dalle conseguenze (perlopiù negative) notevoli.

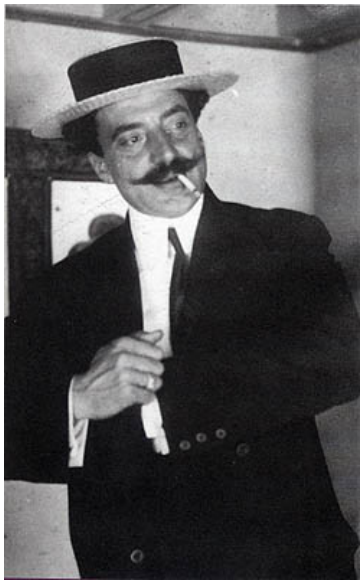


Figura 4 - Trilussa
"Sai ched'è la statistica? È na' cosa
che serve pe fà un conto in generale
de la gente che nasce, che sta male,
che more, che va in carcere e che spósa.
Ma pè me la statistica curiosa
è dove c'entra la percentuale,
pè via che, lì, la media è sempre eguale
puro co' la persona bisognosa.
Me spiego: da li conti che se fanno
seconno le statistiche d'adesso
risurta che te tocca un pollo all'anno:
e, se nun entra nelle spese tue,
t'entra ne la statistica lo stesso
perch'è c'è un antro che ne magna due.
Er compagno scompagno:
Io che conosco bene l'idee tue
so' certo che quer pollo che te magni,
se vengo giù, sarà diviso in due:
mezzo a te, mezzo a me... Semo compagni.
No, no - rispose er Gatto senza core -
io non divido gnente co' nessuno:
fo er socialista quanno sto a diggiuno,
ma quanno magno so' conservatore."

Per comprenderlo torniamo ai "favolosi" anni novanta, quando la scommessa sulle magnifiche sorti e progressive dell'era Internet era il mantra più ricorrente: se la crescita di Internet è di una tal portata, avranno pensato in molti, allora è lì che bisogna investire; è lì che stanno i soldi! Giusto? No! Clamorosamente sbagliato! In una manciata di anni sorsero migliaia di imprese, per lo più di software e di servizi alla rete. E, in altrettanto breve tempo, chiusero i battenti. Alla fine del millennio si parlò di una vera e propria *bolla delle dot.com*, che ebbe notevoli conseguenze sul piano economico, specie in USA e UK.

I big data e la scienza *d'er pollo*

Può apparire paradossale (e forse lo è), ma con la crescita della disponibilità di dati e di strumenti automatici in grado di processarli i margini di errore non sono diminuiti, come non è diminuita la fede verso i numeri: "lo dicono le statistiche…"; "una fonte autorevole sostiene che…", "sembra accertato che esista una relazione fra i due fenomeni in questione…".

In realtà, il difetto, quando c'è, sta nel manico: può riguardare il modo con cui è stato raccolto il campione, oppure nella sua numerosità; più raramente si tratta di errori di calcolo (visto oltretutto che ormai gran parte delle statistiche quantitative vengono trattate con strumenti automatici che minimizzano la probabilità di errore umano).

Ma a volte succede qualcosa che è talmente paradossale da suscitare una sin troppo facile ironia. In questi casi hanno buon gioco tutti i numerosissimi critici del metodo statistico. I quali trovano molte buone orecchie ad ascoltare: già perché se è numerosa la schiera di coloro che hanno una fiducia cieca nel "dato statistico" senza preoccuparsi di

verificarne la fonte, allo stesso modo è folta la schiera di coloro pronti a scagliare feroci strali contro la veridicità di questa o quella statistica e per (indebita) estensione alla statistica in generale.
A volte si tratta di strizzare l'occhio all'ignoranza di alcune persone: lo fa, simpaticamente ma non per questo in modo più corretto, il poeta vernacolare Trilussa. Si può forse dire che ha fatto più male la poesia di Trilussa sulla statistica del pollo di mille errori nel rilevamento dei dati o sul loro significato. Tutti sanno di che si tratta: è una critica in dialetto romanesco sulla significatività della media per cui se *risurta che te tocca un pollo all'anno:/ e, se nun entra nelle spese tue, /t'entra ne la statistica lo stesso /perch'è c'è un antro che ne magna due.*
Trilussa era impareggiabile nel mettere alla berlina certi comportamenti o certi personaggi: sapeva di trovare molti consensi nelle persone "di buon senso". Nello specifico, sapeva che tutti quelli che il pollo non riuscivano a mangiarlo, mentre vedevano altri che se ne ingozzavano, avrebbero gradito una critica alla media statistica. Fingeva di ignorare, il buon Trilussa, uomo di cultura e senatore del Regno d'Italia, che il metodo statistico aveva da tempo elaborato strumenti e misure che permettevano di integrare la media aritmetica per giungere a misurazioni meno grossolane di quelle della *statistica d'er pollo*.
A volte le statistiche sono falsificate ad arte, come nel caso delle notizie durante le guerre. Uno che di queste cose se ne intendeva, il premier inglese Churchill durante la seconda guerra mondiale, sosteneva che le statistiche dovevano essere piegate alle esigenze belliche: "il nemico ha subìto ingenti perdite", "i nostri reparti hanno compiuto un ripiegamento per meglio organizzare la controffensiva", ecc.

Figura 5 – Winston Churchill: "Le sole statistiche di cui possiamo fidarci sono quelle che noi abbiamo falsificato"

L'uso politico delle statistiche non è certo un'invenzione dello statista inglese: ognuno cerca di piegare i numeri a proprio uso e consumo. Qualcuno lo fa con decenza e persino con una dose di sfacciata eleganza; in altri casi in modo che se falsificare i dati fosse un reato chi lo fa meriterebbe l'ergastolo.
Un aspetto vale la pena di sottolineare: quando la statistica è "fatta male", con campioni poco significativi oppure con strumenti matematici inadeguati, o peggio ancora poggia su presupposti falsi, a rimetterci è l'intera scienza statistica.
Nelle prossime pagine prenderemo in considerazione alcuni importanti strumenti della statistica descrittiva cercando di indicarne il modo corretto di usarli, senza cadere in trappole cognitive e in tranelli logici in cui si rischia facilmente di incappare.

Gente fortunata…

Già, gente fortunata quella che si trova oggi a dover analizzare dati economico-sociali di tipo quantitativo! Fino a qualche decina di anni fa chi avesse voluto analizzare, per esempio, il livello dell'inflazione in Italia nel ventennio precedente, avrebbe dovuto armarsi di santa pazienza: avrebbe dovuto cercare una biblioteca che avesse le pubblicazioni contenenti il dato, ricopiarlo (a mano!) su un foglio e procedere poi ai calcoli che sarebbero stati utili. Anche semplici calcoli come l'inflazione media del ventennio che costituisce l'oggetto di studio, sarebbe stata un'operazione penosamente lenta. Se poi il nostro ricercatore avesse avuto la malaugurata idea di accompagnare il suo lavoro con dei grafici, avrebbe dovuto procurarsi un righello, una matita e un foglio di carta millimetrata!
Per fortuna le cose sono cambiate: oggi gran parte delle operazioni che abbiamo elencato prima sono automatizzate: una gran mole di dati è disponibile on line e facilmente scaricabile; software

come Excel, SPSS, Stata o R permettono di “processare” quei dati nel modo più opportuno. Calcoli, anche sofisticati e complessi, possono essere fatti con un semplice click e, sempre Excel (o simili), fornisce un’ampia scelta fra tipologie diverse di strumenti grafici.
Tutto più veloce, più comodo ed elegante, ma anche più efficace dal punto di vista dei risultati! Tutto questo grazie all’informatica e alla rete: le tecnologie informatiche permettono di avere hardware e software sempre più affidabili e *user-friendly*, mentre Internet consente di accedere “in remoto” ad una massa incredibile di dati prima irraggiungibili.
Tutto bene, tutto perfetto? Sì, certo! Fra l’altro è prevedibile che la mole di dati disponibili a partire da siti affidabili, sarà sempre maggiore e questo consentirà di fare valutazioni quantitative sempre più precise e quindi più affidabili. Il rovescio della medaglia sta nel modo con il quale questi dati vengono usati: non stiamo parlando di evidenti falsificazioni, né di statistiche anche vere ma patentemente “di parte”[2].
No, ci riferiamo alle parti di un’indagine statistica, dal modo con cui si raccoglie un campione fino al modo con cui si interpretano i dati ricavati, passando per aspetti apparentemente innocenti e “neutrali” come la loro rappresentazione grafica. Tutto può essere oggetto di errore come, al contrario, può avere una significatività considerevole.
Ed è ovviamente a quest’ultimo risultato che si deve tendere. Ma è tutt’altro che facile…

[2] Qualche mese fa il vostro autore notò il manifesto pubblicitario di una scuola privata che offriva la preparazione per i test di iscrizione alle facoltà di medicina e di fisioterapia. Il testo del manifesto recitava più o meno: frequentando i nostri corsi, uno studente su due passa il test”. Non abbiamo idea se il dato corrispondesse a verità, ma vogliamo supporre di sì. Ora immaginate che il marketer incaricato di preparare la campagna pubblicitaria di quella stessa scuola avesse proposto un testo che recitava: frequentando i nostri corsi, uno studente su due *non* passa il test”: dal punto di vista statistico le due espressioni si equivalgono, ma credete che il marketer in questione avrebbe continuato a lavorare per quella scuola privata se avesse proposto la seconda versione e non la prima?

Una brutta bestia: la correlazione

Cominciamo, non a caso, trattando di uno strumento statistico particolarmente importante, ma dove i tranelli logici sono spesso in agguato. È per di più uno strumento che trova un utilizzo particolarmente rilevante al giorno d'oggi e con ogni probabilità lo sarà ancora di più in futuro, grazie alla sempre crescente disponibilità di dati quantitativi. Facciamo riferimento alla possibilità di usare la grande massa di dati disponibili per stabilire dei collegamenti fra fenomeni. Operazione legittima ed altamente utile: se due fenomeni ci appaiono collegati nella loro variabilità (per esempio al crescere di valore nel tempo di uno cresce anche il valore dell'altro), allora saremo tentati di affermare che esiste una correlazione fra l'uno e l'altro dei due fenomeni. Osservate i grafici contenuti in figura 6.

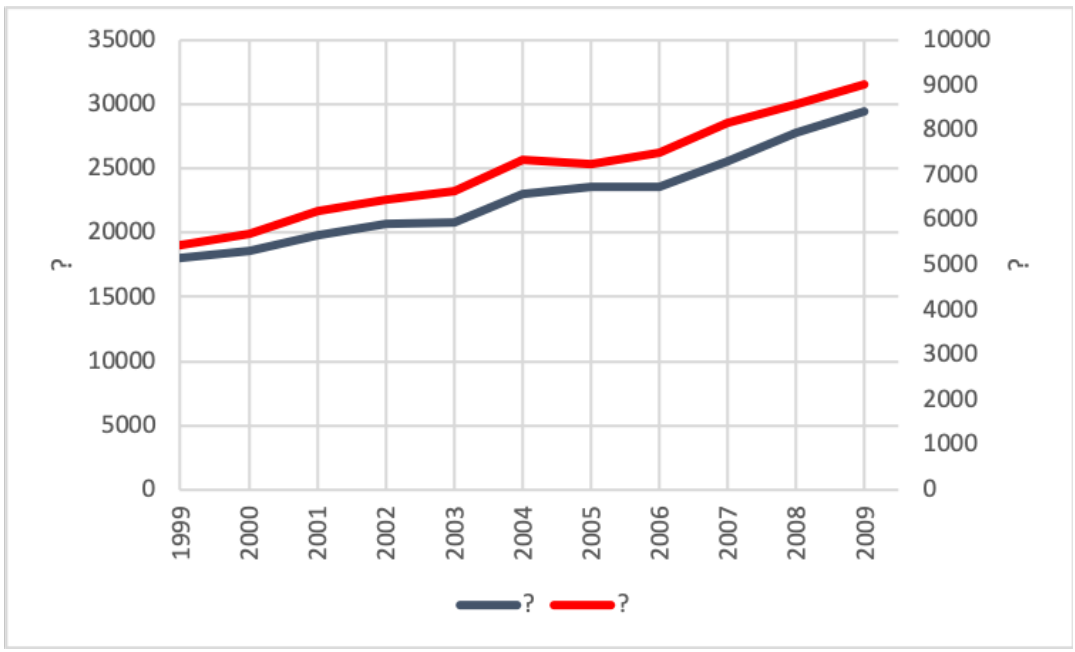


Fig. 6

Come vedete non abbiamo indicato né le denominazioni degli assi e neppure il significato della legenda; il tutto è stato sostituito da ineleganti (e anche un po' sadici) punti interrogativi. Cosa ne dite? I fenomeni rappresentati dai due grafici vi sembrano collegati? Certo che sì, direte senza tema di smentita!

Avreste ragione, se non fosse per una minuzia, un trascurabile particolare che è in grado però di mettere in discussione le più granitiche certezze. Di che si tratta? Si tratta del fatto che i due fenomeni che appaiono a prima vista così strettamente collegati sono la spesa pubblica annua in USA nei campi delle scienze e della tecnologia (asse Y di sinistra) e … i morti per suicidio da impiccagione, strangolamento e soffocamento (asse Y di destra)! La figura 7 svela il segreto che avevamo gelosamente custodito.

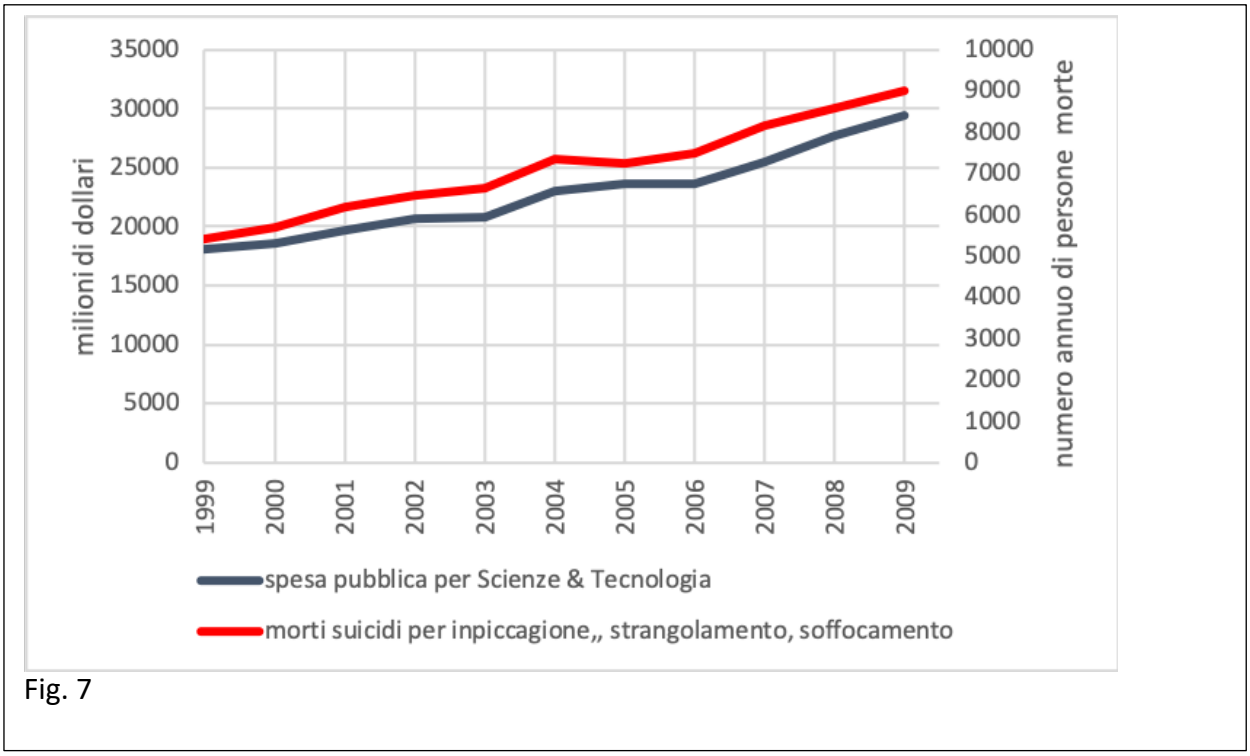


Fig. 7

Ora, a meno che voi non riusciate a trovare una qualche sensata ragione che dimostri che le due variabili rappresentate sono davvero collegate, allora dovremo ammettere che si tratta di una coincidenza. Una coincidenza può sempre esserci, starete certamente pensando. Se poi c'è chi vuole mettere in difficoltà l'ingenuo lettore e presenta correlazioni scelte ad arte, allora è chiaro che gioca sporco!
Per dimostrarvi che in realtà chi scrive non vuole carpire la vostra buona fede né trattarvi da ingenui, vi presentiamo un altro caso (figura 8).
Questa volta non trovate criptici punti interrogativi: sugli assi e sulla legenda tutto è chiaro, solare, evidente: i due grafici rappresentano l'andamento di due variabili: il consumo di formaggio in USA (asse Y di destra) e … i morti perché si sono attorcigliati nelle lenzuola! Vi assicuriamo che negli USA vengono contate anche questo tipo di morti. I grafici dimostrano che i due fenomeni sono andati entrambi crescendo nel corso del tempo: si può ragionevolmente ipotizzare che non si tratti di una pura e semplice coincidenza? Tutto è possibile ovviamente: si può pensare che il consumo di formaggio provochi un aumento del colesterolo e che a sua volta questo provochi un aumento nel numero degli individui sovrappeso; i quali probabilmente faranno qualche fatica a muoversi agilmente. Qualcuno potrebbe persino non essere più in grado di liberarsi dalla stretta mortale di maligne lenzuola.

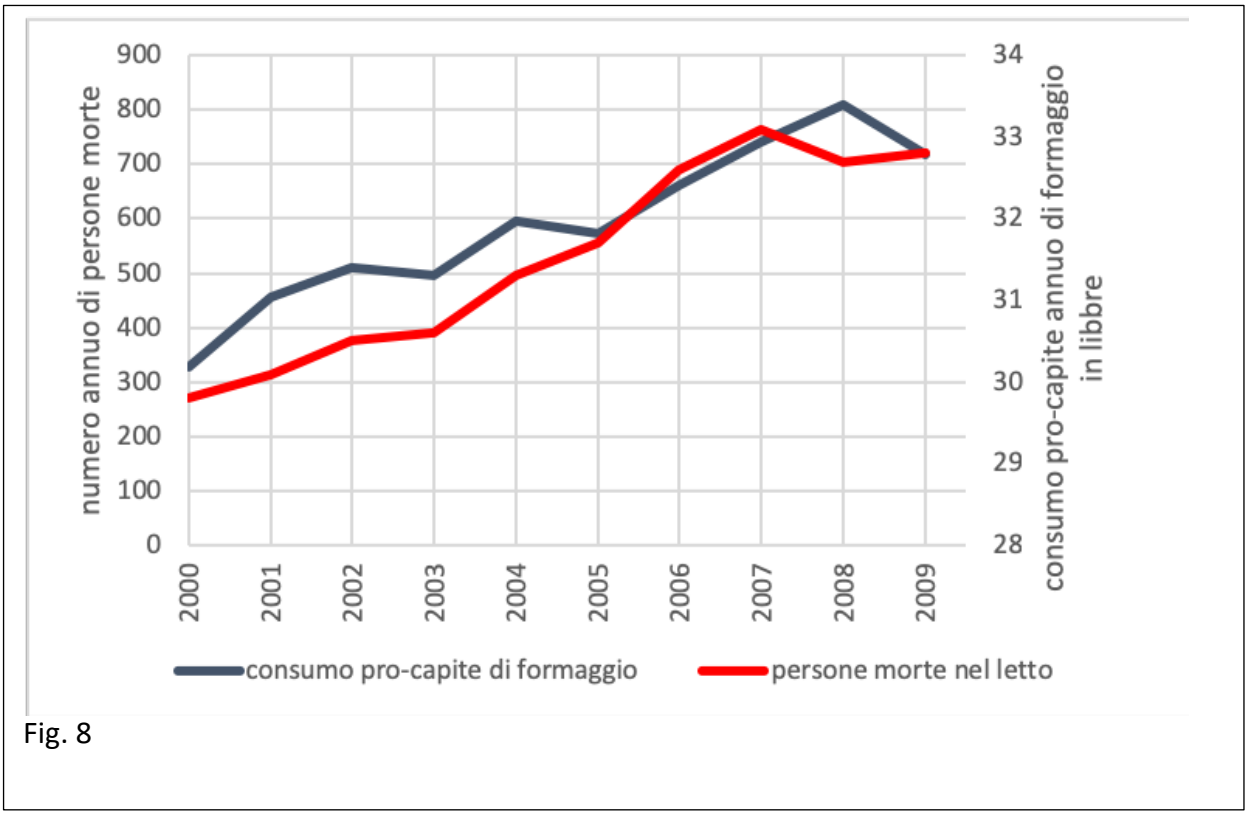


Fig. 8

Può essere, ma ve la sentite di mettere in così stretta relazione i due fenomeni? No, vero? Saggia decisione! Anche in questo caso una semplice coincidenza? Non proprio: così come nel caso di suicidi e spesa pubblica in scienze e tecnologie, anche nel caso di consumo di formaggio e morti da attorcigliamento di lenzuola, si tratta di qualcosa di ulteriore rispetto alla pura coincidenza.
Immaginate che vi sia un omino che, una volta raccolti i dati relativi ad un qualunque fenomeno (per esempio il valore delle spese in scienza e tecnologia) si metta alla ricerca di tutto ciò che appare collegato a questa variabile.
Impossibile, dite? Se pensate questo, ci dispiace ma sbagliate di grosso. O meglio se non credete alla storiella di un omino perverso avete assolutamente ragione. Ma esistono software che lo sostituiscono egregiamente! In altri termini: cercate nel web un fenomeno quantitativo, ne scaricate i dati e poi lanciate il programma che spazza la rete fino a trovare fenomeni che quantitativamente mostrano una relazione con il fenomeno di riferimento. Poi create il grafico e, voilà, il gioco è fatto! Avete la vostra bella correlazione con tanto di rappresentazione grafica!
Se invece usate un po' di buonsenso quando si tratta di mettere in relazione dei fenomeni, magari farete un po' più di fatica, ma vi risparmierete qualche brutta figura. Purtroppo, è difficile che troviate omini di buona volontà che vi spieghino quando una correlazione è ragionevole, ma il sale della ricerca sta proprio in questo: nel cercare ipotesi ragionevoli e nel dimostrare che un fenomeno si lega ad un altro in modo consistente (e convincente).

Nubi e correlazione

Supponete di aver trovato due variabili quantitative che hanno una correlazione piuttosto evidente e vi sembra che tale correlazione abbia buone basi logiche che le permettono di stare in piedi. Instiamo su questo aspetto delle relazioni fra variabili perché se, a volte, rappresentano una trappola logica da evitare come le false correlazioni di cui abbiamo fornito un paio di esempi, in altri casi rappresentano un bel passo avanti nella conoscenza della natura di un fenomeno.
In particolare, spesso gli economisti sono interessati alle relazioni fra due (o più) variabili. Ad esempio, un economista potrebbe chiedersi se il livello di istruzione formale (titolo di studio)

raggiunto da una persona è associato a livelli di salario più alti; oppure se le differenze negli investimenti di ogni Paese sono associate a tassi di crescita del PIL. Domande simili a queste sono il pane quotidiano per gli economisti ed è per questo che occorre spendere qualche altro ragionamento in proposito.
Presentiamo come esempio un caso di correlazione ben più interessante rispetto a quelli già visti: i geografi e gli economisti dello sviluppo avanzano l'ipotesi che si possa stabilire una relazione diretta fra tassi di deforestazione di un territorio e densità abitativa del territorio stesso. Per verificare questa ipotesi supponete di avere a disposizione il dato sulla deforestazione (in ettari) e quello sulla densità demografica (n.ro di abitanti per 1.000 ettari) di una settantina di Paesi tropicali.
Per verificarlo graficamente utilizziamo un grafico "a dispersione" (grafico XY nel linguaggio Excel), che consente di costruire una "nube" di punti in cui ogni puntino corrisponde ad un Paese. Nel grafico, l'asse X considera la densità abitativa per ettaro e l'asse Y la percentuale di area forestale deforestata negli ultimi dieci anni; l'incrocio dei due valori, indicata da un pallino, corrisponde ad un Paese. Inoltre, abbiamo aggiunto una finezza da nerd: abbiamo calcolato i valori medi delle due variabili, che compaiono nel grafico come un pallino più grande degli altri e, sulla base del dato medio, abbiamo individuato quattro aree in cui è diviso il piano cartesiano. (figura 9)[3].

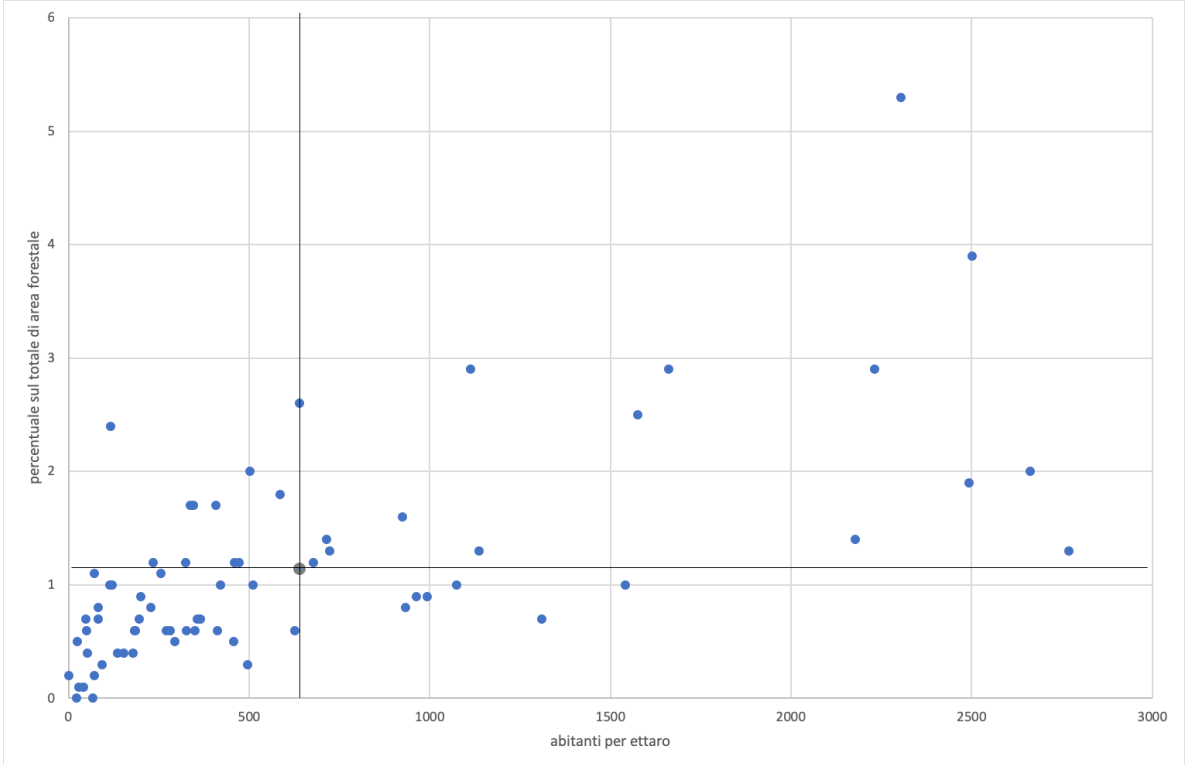


Figura 9

Osservate con attenzione il grafico: potete individuare un addensamento maggiore di punti? Certo! Un gran numero di punti/Paese (26 per l'esattezza) si trovano nel quadrante in basso a

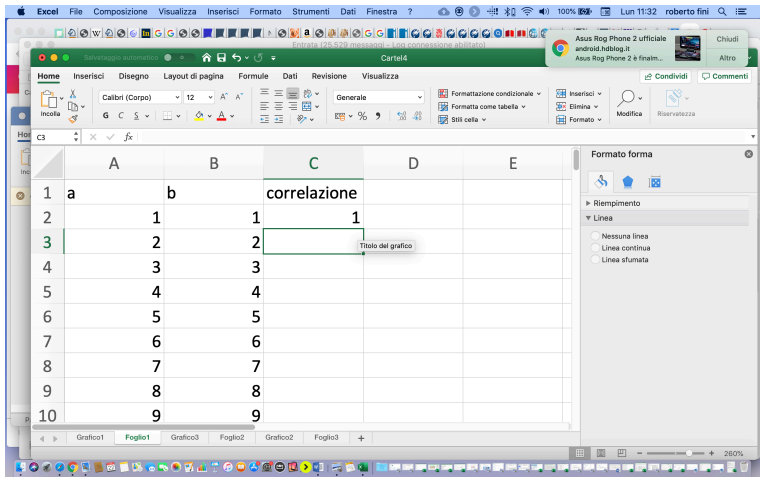
[3] Costruendo un'apposita "macro" secondo le sintassi di Excel avremmo potuto anche indicare il nome di ogni Paese (che risulterebbe a fianco di ciascun puntino) ma le etichette avrebbero ingombrato inutilmente il grafico: quel che ci interessa è verificare se fra le due variabili considerate è possibile stabilire una correlazione e non stabilire a quale Paese corrisponde ciascun punto.

sinistra del grafico; un altro buon numero (15) si trova nel quadrante in alto a destra. Queste due aree possono considerarsi quelle in cui l'ipotesi di partenza è verificata: ad una bassa/alta densità abitativa corrisponde una bassa/alta percentuale di deforestazione. Le altre due aree del grafico contengono dati "anomali" (*outlier*): in alto a sinistra si situano Paese con bassa densità abitativa e alta percentuale di deforestazione; mentre il quadrante in basso a destra "contiene" i Paesi ad alta densità abitativa e bassa deforestazione.
Gli economisti sono di solito interessati a rintracciare nei dati comportamenti generali o, quanto meno, tendenze. Da questo punto di vista il grafico della figura 4 fornisce un'indicazione preziosa: la nube di punti è addensata fra le aree in basso a sinistra/in alto a destra. Ben più della metà dei Paesi (41) si trova in questi due quadranti: potremmo quindi dire che, in linea di massima, la relazione fra le due variabili è verificata. Tuttavia, dobbiamo pur sempre tenere conto che una consistente minoranza di Paesi presentano dagli *outlier* e in molti casi studiare le ragioni per cui un Paese presenta valori anomali può essere persino più interessante del dato di tendenza, specie ovviamente quando gli *outlier* presentano dati fortemente anomali, cioè collocano un Paese alla periferia dell'area considerata.

Esplorare (bene) la correlazione

Per comprendere appieno il significato di correlazione e l'uso dei grafici XY (o a nube di punti o a dispersione), vi proponiamo alcuni esempi puramente teorici:

a. Correlazione positiva massima;
b. Correlazione negativa massima;
c. Correlazione nulla.



	A	B	C	D	E
1	a	b	correlazione		
2	1	1	1		
3	2	2			
4	3	3			
5	4	4			
6	5	5			
7	6	6			
8	7	7			
9	8	8			
10	9	9			

Figura 10 -

Correlazione massima positiva

Supponete di avere due serie di numeri (a e b) tra 1 e 50, tali per cui ad ogni valore della colonna a corrisponde un identico valore della colonna b (cfr. il particolare in figura 10). Per ora ignorate il valore che compare nella colonna C sia in questo che nei fogli successivi. Come vedete al valore 1 presente nella colonna di a corrisponde il valore 1 nella colonna di b; al valore 2 nella colonna di a corrisponde un identico valore nella colonna di b, ecc.
Se a questo punto lanciamo il grafico XY relativo alle due serie di dati, esso avrà un aspetto come quello presentato nella figura 11.
Come si vede facilmente dal grafico di figura 11, al crescere dei valori di a, crescono con lo stesso incremento i valori di b.
I punti che rappresentano le coordinate di a e di b sono allineati lungo un'immaginaria linea retta crescente da sinistra verso destra.
Come è evidente, la correlazione fra la serie di valori di a e la serie di valori di b è massima.

Correlazione massima negativa

Situazione completamente opposta è quella che si trova visualizzata nel foglio Excel della figura 12. Qui la correlazione esiste al pari del caso precedente, ma è di tipo inverso: i valori massimi di a corrispondono i valori minimi di b, cosicché dove a=50, b=1 (anche in questo caso ignorate, per ora il valore che compare nella cella della correlazione). Così come avevamo fatto per il caso della correlazione diretta (figure 10 e 11), è possibile tracciare il grafico XY della correlazione inversa qui ipotizzata (fig. 13).

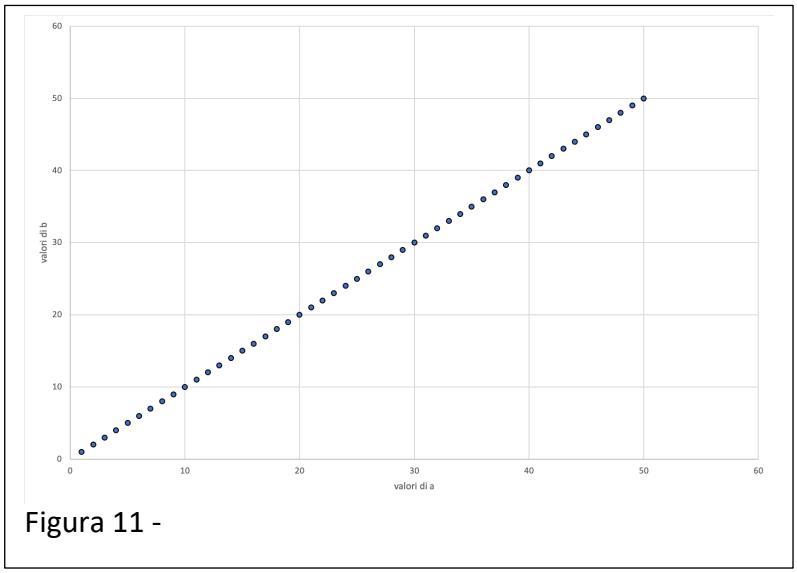
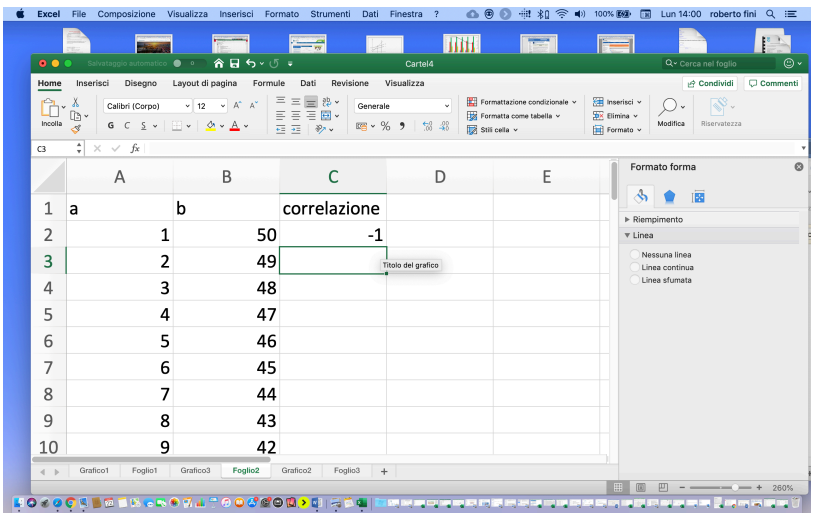


Figura 11 -



	A	B	C	D	E
1	a	b	correlazione		
2	1	50	-1		
3	2	49			
4	3	48			
5	4	47			
6	5	46			
7	6	45			
8	7	44			
9	8	43			
10	9	42			

Figura 12 -

Anche in questo caso i punti sono allineati lungo un'ipotetica retta ma, al contrario rispetto al caso precedente, tale retta ha un orientamento decrescente da sinistra verso destra.

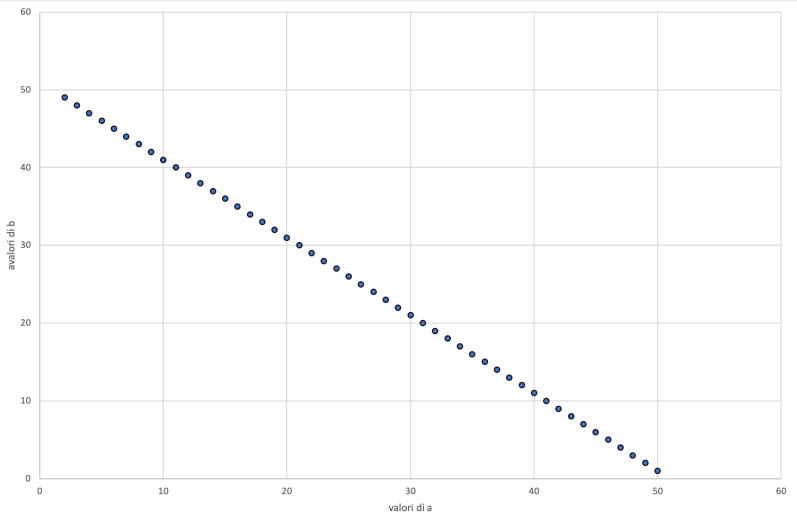
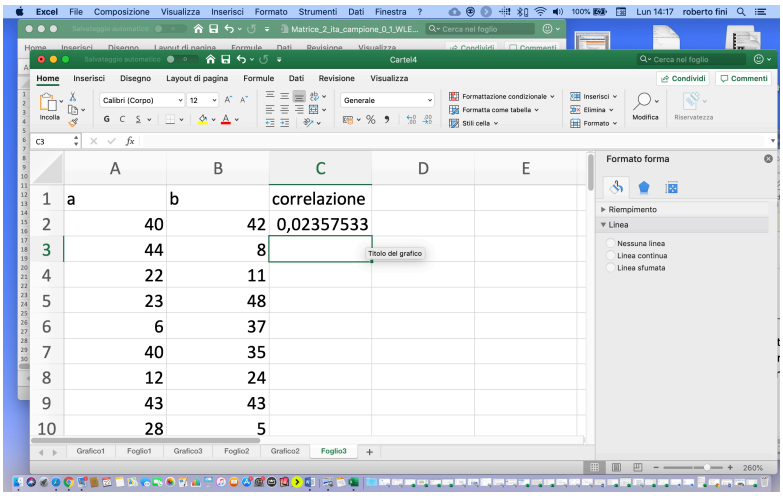


Figura 13

Come avevamo accennato, i due casi presentati sopra si riferiscono a due situazioni astratte: è ben difficile trovare delle correlazioni così strette fra fenomeni di tipo economico-sociale. Potremmo persino spingerci a dire che molti fenomeni di questo tipo rientrano fra quelli che non hanno correlazioni oppure che, se le hanno sono piuttosto deboli.

Correlazione nulla



	A	B	C	D	E
1	a	b	correlazione		
2	40	42	0,02357533		
3	44	8			
4	22	11			
5	23	48			
6	6	37			
7	40	35			
8	12	24			
9	43	43			
10	28	5			

Figura 14 -

La situazione di due fenomeni che non hanno una correlazione è quella evidenziata nella schermata parzialmente rappresentata in figura 14. Come potete vedere, fra i valori di a e di b non c'è alcuna evidente relazione[4]: non è possibile legare ai valori delle due colonne nessun tipo di regola che li colleghi.
Anche in questo caso, comunque, è possibile generare un grafico XY che mette in evidenza come la nube di punti appaia dispersa: i punti si trovano nell'area delimitata dai due massimi (a=50 e b=50) ma a parte questo si trovano in posizioni del piano cartesiano per le quali non è ragionevolmente possibile stabilire una regola che in qualche modo permetta di individuare una relazione (figura 15).

Misurare la correlazione

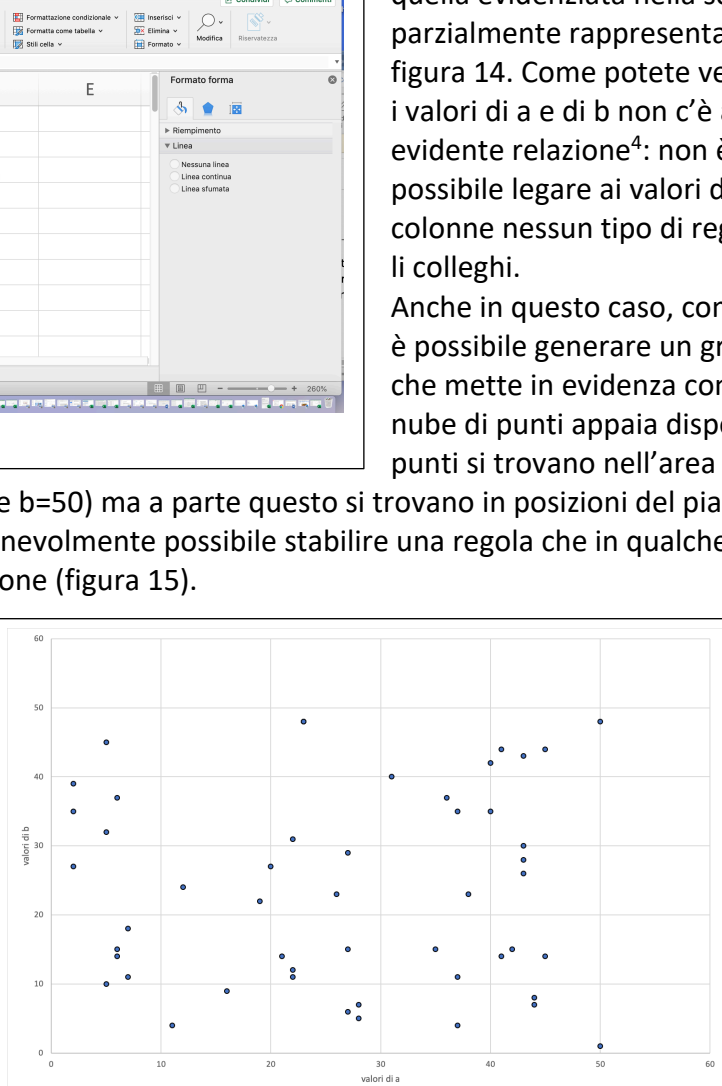


Figura 15

Tutto chiaro vero? Se è così proviamo a complicarvi leggermente la vita. È arrivato il momento di prendere in esami i valori della correlazione che trovate nella colonna 3 dei precedenti fogli di lavoro.
I valori che trovate sono generati grazie ad una utilissima funzione di Excel[5]: in pratica, il software mette a confronto le coppie d valori presenti nelle colonne prescelte (nel nostro caso le prime due colonne di ciascun foglio) e "controlla" presenza di un'eventuale correlazione. Come potete facilmente verificare riprendendo le figure 10 e 12, la

[4] In effetti per generare i valori che compaiono sul foglio è stata utilizzata la funzione Excel "casuale.tra", che permette di generare due set di numeri casuali privi di alcuna relazione fra di loro.
[5] Si tratta della funzione "correlazione"

correlazione diretta “perfetta” vale 1 mentre quella inversa (sempre “perfetta”) vale -1. In formula, date due variabili X e Y il valore della correlazione ρ[6]:

$$-1<\rho XY<1 .$$

Se
$\rho XY=1$ si ha correlazione diretta perfetta;
$\rho XY=-1$ si ha correlazione inversa perfetta;
$\rho XY=0$ le due variabili sono incollerate.
Come abbiamo osservato è ben difficile che la correlazione (diretta o inversa) sia perfetta, assuma cioè valori pari, rispettivamente, a 1 o -1. È più probabile che i valori di correlazione siano uguali ad un valore posto tra i due estremi. In riferimento alla correlazione diretta[7], possiamo stabilire una regola generale per cui se:
$0<\rho XY<0,3$ si ha correlazione debole;
$0,3<\rho XY<0,7$ si ha correlazione moderata;
$\rho XY>0,7$ si ha correlazione forte.
In pratica, se due fenomeni assumono un valore di correlazione inferiore o uguale a 0,3 possiamo considerarli non correlati (soprattutto, ovviamente se il valore di ρ è vicino allo 0). Solo con valori di ρ vicini o superiori a 0,7 possiamo ipotizzare che i due fenomeni abbiano una correlazione consistente. Ne discende che in caso di valori intermedi situati nell’intervallo

$$-0,7<\rho XY<0,7$$

anche quando esiste una certa correlazione (cioè i valori sono vicini a 0,7 o -0,7) è bene operare ulteriori verifiche prima di spingersi ad affermare che la correlazione individuata ha una qualche consistenza dal punto di vista socio-economico.

La correlazione e il buon senso

Siamo consapevoli del fatto che l’ultimo paragrafo possa aver spaventato quelli fra i nostri lettori che non hanno dimestichezza con numeri e formule. Tuttavia, una lettura attenta di quanto abbiamo dedicato alla correlazione dovrebbe essere sufficiente a far comprendere sia il concetto che l’utilità come strumento logico prima che statistico.
Non siete convinti? Beh, qualche ragione l’avete se è così. Siamo i primi a riconoscere che la correlazione va usata *cum gran salis* e che non è sufficiente disporre di un comodo strumento software (come la funzione *correlazione* in Excel) per risolvere ogni problema. A costo di essere considerati dei masochisti, vi diamo un suggerimento a vantaggio di coloro che hanno scarsa fiducia nei suoi strumenti e considerano la statistica come la scienza *d’er pollo*. Salvo poi farvi ricredere (o almeno lo speriamo…).
Dunque, tornate per un momento con la mente alla relazione fra spesa pubblica per scienze e tecnologia in USA e il numero di suicidi per impiccagione, strangolamento e soffocamento sempre in USA (argomento leggero e piacevole già affrontato in figura 7). Sapete quanto vale la correlazione fra le due variabili? All’incirca 0,99! Una correlazione apparentemente fortissima e che non dovrebbe lasciare dubbi sul fatto che i due fenomeni indagati viaggino insieme.
Bene, dirà qualcuno che non vedeva l’ora di trovare un argomento da usare contro la correlazione fra fenomeni: se il risultato matematico è 0,99 e siamo di fronte a due fenomeni che in tutta evidenza non hanno una connessione logica, allora questo famoso ρ non vale granché! Non è che abbiate tutti i torti, *apparentemente*: già perché come ogni tanto viene ricordato, anche un orologio fermo segna due volte al giorno l’ora giusta. Il fatto, cioè che la correlazione appaia forte

[6] Per indicare la correlazione convenzionalmente si usa la lettera dell’alfabeto greco ρ , che si legge “rho”
[7] Ma, con le opportune modifiche di segno, le stesse regole sono applicabili alla correlazione inversa

anche di fronte a fenomeni che non sono logicamente collegati, non esclude che in altri casi possa considerarsi uno strumento di indagine statistica efficace e persino potente.
Niente può sostituire il buon senso quando si cercano collegamenti fra fenomeni o fra comportamenti: si può trovare qualche ragione di buon senso che spieghi l'apparente legame fra spesa pubblica e suicidi? Probabilmente no! Se il valore di ρ è alto deve essere una coincidenza, oppure il campione su cui si è operato non era significativo o correttamente selezionato.
Per spiegarci bene su questo aspetto prendiamo l'altro esempio palesemente assurdo che abbiamo fatto nei paragrafi precedenti. Ricordate? Si tratta della correlazione fra consumo di formaggio e morti per essersi attorcigliati nelle proprie lenzuola (!). Se vi prendeste la briga di calcolare il valore di ρ, anche in questo caso scoprireste che assume un valore molto alto, intorno a 0,99. È possibile? Potremmo ipotizzare che il formaggio aumenta il colesterolo e che questo provochi un eccesso di grassi nel sangue e che a sua volta questo determini movimenti più lenti ed impacciati.
Un'ipotesi di questo genere, però, non regge alla prova del buon senso. Dunque, come spiegare una correlazione così forte? Un modo è quello concludere che si tratti di una pura e semplice coincidenza. Un'altra spiegazione, più plausibile, è che il campione sul quale si è lavorato non era stato scelto con adeguata scientificità. Tanto per rimanere nell'assurdo: supponete che il campione sia stato costituto da un solo individuo e che fosse un buon mangiatore di formaggio. E supponete anche che finisca i suoi giorni ingloriosamente soffocato dalle proprie lenzuola. Domandina facile facile: a quanto sarà uguale il risultato in termini di probabilità che un consumatore di formaggio muoia a causa delle sue lenzuola? 100%! Elementare Watson!
Che un campione statistico sia formato da una sola persona è palesemente assurdo e se succedesse dovremmo diffidare non della statistica, ma di chi propone di utilizzarne il risultato in un contesto scientifico. Più probabilmente cerca di abbindolarvi con il trucchetto di presentare dati "incontrovertibili". Ma di statistiche "fatte male", sciatte, superficiali, è pieno il mondo!
E, dunque, attenzione! Ma, per usare un modo di dire di dubbio gusto ma efficace: non buttate via il bambino con l'acqua sporca!

"Post hoc ergo propter hoc"

Bene! Credete di avere capito tutto della correlazione? Sapete come funziona, il significato di quel maledettissimo ρ ; siete persino consapevoli che ci sono numerosi trabocchetti ed insidie logiche quando si tratta di usare la correlazione come strumento di ricerca. Perfetto! Ora, pensate, possiamo passare ad altro (magari di più simpatico…). No, ci dispiace per voi ma non è così: ci manca un passo importante, uno di quei passi che possono mettere in crisi tutto quello che credete di aver capito sull'argomento.
Di che si tratta? Si tratta del *post hoc ergo propter hoc*[8] (per gli amici *post hoc*). Ecco, starete pensando ci mancava il *latinorum* di manzoniana memoria. Potreste anche avere ragione, ma lasciateci spiegare di che cosa si tratta e poi decidete se vale la pena di approfondire l'argomento. Vi accorgerete fra l'altro che si tratta di un errore logico piuttosto frequente che ha che fare più che con la statistica in senso stretto con il modo con cui ragioniamo quando siamo davanti a due (o più) fenomeni.
Dunque, vediamo di partire in modo soft: Qualcuno si ricorda il film "Operazione sottoveste"? è un film USA del 1959 ed è ambientato a bordo di un sommergibile americano operante nei mari del Sud durante la seconda guerra mondiale. Ma non è un film di guerra: è una divertente commedia con Cary Grant e Tony Curtis. Come in tutte le commedie che si rispettano finisce bene: gli ufficiali sposano le belle ausiliarie e i buoni alla fine vincono la guerra.

[8] Letteralmente: "dopo di ciò e di conseguenza necessariamente a causa di ciò".

Ma non è di questo che vogliamo parlare: nel film il sottomarino viene seriamente danneggiato durante un bombardamento giapponese. L'equipaggio lo rimette insieme alla meglio in modo da riprendere il mare al più presto. Quando tutto sembra pronto il comandante (Cary Grant) ordina di avviare i motori: dalla sala-macchine provano e riprovano, ma non succede niente. I diesel non vogliono partire. Mentre equipaggio e comandante pensano di rinunciare, sul molo presso cui è ormeggiato il sottomarino si materializza uno stregone (o che tale sembra) ingaggiato dall'ufficiale maneggione (Tony Curtis) con due aiutanti.
Lo stregone improvvisa una danza propiziatoria, recita una nenia indiavolata in una strana lingua, getta sul sottomarino qualcosa che assomiglia a cenere. Alla fine, con meraviglia di tutti (anche dello stregone), i motori si avviano e il battello può prendere il largo per andare a far guai (ma ne farà ben pochi in realtà). Del resto, restando nell'ambito delle religioni primitive, se uno stregone fa la danza della pioggia e poco dopo piove, la gente della sua tribù crederà (con qualche ragione apparente) che sia stato lui con la sua danza che ha fatto piovere.
Ora, nessuno potrebbe credere che davvero lo stregone abbia fatto partire i motori o abbia fatto piovere: siamo gente di mondo e a queste cose non crediamo! Ma in un errore simile a quello di chi crede che lo stregone abbia fatto piovere tendiamo a cadere facilmente. Si tratta, per l'appunto, dell'errore del post hoc: se qualcosa succede dopo che qualcosa è già successo, pensiamo che ciò che è avvenuto dopo sia stato causato dall'avvenimento precedente. Lo stregone danza e poi piove? Lo stregone ha fatto piovere!
Dicevamo: sarà anche possibile tra qualche popolo superstizioso, ma noi no ci caschiamo! Niente stregoni alle nostre latitudini! Vero, ma sentite questa storiella: il dott. Sorcerer è consulente del ministro dell'Economia. In una riunione di alto livello propone al suo capo una ricetta semplice ed immediata per risollevare le finanze disastrate del Paese: "Riduciamo le aliquote fiscali, così facciamo aumentare i redditi netti dei contribuenti; in questo modo aumenteranno i consumi e avremo un risparmio sulla spesa pubblica assistenziale.". Ragionevole, non trovate? Potrebbe persino funzionare. Ma come purtroppo spesso accade, il diavolo si nasconde nei dettagli.
Il dettaglio che il dott. Sorcerer consiste in un classico caso di errore post hoc: quello di supporre che la riduzione delle imposte abbia determinato un aumento dei consumi ed una riduzione conseguente della spesa pubblica assistenziale. Un'altra ipotesi, altrettanto legittima rispetto a quella del buon Sorcerer, suggerita dal maligno capo dell'opposizione Mr. Beelzebub è la seguente: "Si potrebbe supporre che il sistema economico del nostro Paese presentasse già elementi di crescita e il reddito pro-capite si stava innalzando; la riduzione delle aliquote fiscali ha avuto come effetto una riduzione delle entrate tributarie che, al contrario, sarebbero potute aumentare!".
Ha ragione Sorcerer o Beelzebub? Non si può dirlo a priori, ma certamente Sorcerer compie un classico errore di post hoc: poiché la riduzione delle aliquote fiscali è arrivata prima dell'aumento dei consumi, allora questo effetto è stato causato da quella riduzione. Fosse così facile la vita avremmo risolto ogni problema!
Ah, a proposito: sorcerer è il vocabolo inglese che sta per stregone (quanto al maligno Beelzebub, beh non è necessario nessun suggerimento per svelarne l'identità, vero?).
La questione legata all'errore del post hoc è molto importante: tutti noi tendiamo a cadere nella trappola di un simile errore e dunque vale la pena di insistere sull'insidia che ci fa credere ciò che non è.
Dunque, altro esempio tratto da vicende reali: molto tempo fa accadde che qualcuno decise di impegnarsi a fondo per cercare di capire se i fumatori di sigarette avessero rendimenti scolastici più bassi dei non fumatori. Se si fossero trovate delle evidenze statistiche in tal senso, sarebbe stata giustificata una campagna volta a scoraggiare il consumo di sigarette quanto meno fra i giovani studenti. Nobile impegno!

Bene, risultò che era così: le ricerche statistiche sembravano dimostrare che una buona strada per avere buoni voti scolastici partisse dallo smettere di fumare. Oh, lo studio a supporto di una simile conclusione era serio: il campione era sufficientemente grande e ben stratificato. E la correlazione fra i due fenomeni, insuccesso scolastico/dipendenza da fumo, era sufficientemente alta. Dunque. perché dubitare della correttezza delle conclusioni cui perveniva lo studio?
Ancora una volta il problema sta nel cadere nella trappola del post ho: se B segue A (post hoc) allora vuol dire che (ergo) A è la causa di B (propter hoc). Nel caso del fumo e dei risultati scolastici, si parte da un'ipotesi indimostrata: poiché l'abitudine a fumare coincide con voti negativi, ne discende che il fumo è la causa del rendimento negativo.
Non discutiamo l'esistenza della correlazione, ma la direzione di tale correlazione: in assenza di indicazioni ulteriori non potrebbe essere altrettanto plausibile anche il contrario e cioè che siano i risultati negativi ad indurre le persone a fumare? Comprendiamo che una simile affermazione sembri poco sensata, ma eccovi una possibile ipotesi di lavoro: forse i voti bassi conseguiti dagli studenti potrebbe indurre a consolarsi con il tabacco. Chissà! Non c'è ragione di optare, a priori, per una o l'altra ipotesi, ma se si bada alla sostanza, questa deduzione ha la stessa probabilità dell'altra ed è in egual misura sostenuta dai dati.
Non dovete sposare l'una o l'altra linea di ragionamento ed è persino possibile che nessuno dei due fenomeni sia uno la causa dell'altro: tutti e due potrebbero essere le conseguenze di un terzo fattore. Potrebbe essere plausibile che un soggetto, più incline alla vita di gruppo, che prende meno sul serio i libri di scuola sia anche più probabilmente un fumatore? In questo caso la sequenza logica è: a) maggiori estroversione e tendenza alla vita di gruppo, b) minore attitudine allo studio scolastico, c) rendimenti scolastici negativi.

A mo' di epilogo: la (tristissima) storia del dott. Wakefield e della credulità popolare

Dunque, proviamo a riassumere i problemi legati alla correlazione e, in particolare, al problema del post hoc. È una questione importante perché si rischia di cadere nei tranelli di cui è disseminato il percorso logico che tenta di spiegare *se* e *in che modo* i fenomeni che stiamo esaminando siano collegati. E la verifica deve essere severa, a partire dal fatto che dobbiamo farlo con una mente sgombra da pre-giudizi.
Errori possibili? Uno è determinato dal fatto che la correlazione sia prodotta dal caso: supponiamo di prendere in esame due serie di dati e in base a questo esame si scopra che il fenomeno A è fortemente collegato al fenomeno B. Potremmo dirci soddisfatti? No: è buona norma tralasciare le serie numeriche esaminate in prima battuta e prendere in considerazione altre due serie, ovviamente riguardanti lo stesso fenomeno. Se il secondo esame conferma la correlazione scoperta nel primo caso, allora potremmo essere ragionevolmente certi del legame.
Inoltre, attenzione agli "opportunismi del ricercatore": costruire campioni statistici può essere costoso (se non in termini direttamente monetari, quanto meno in termini di tempo da impiegare), per cui si potrebbe essere tentati di fare riferimento a campioni piccoli: il problema è che con campioni piccoli si può dimostrare una correlazione fra qualsiasi coppia di fenomeni o di eventi.
Un genere diffuso di co-variazione tra fenomeni è quello in cui la correlazione è reale, ma non si può essere ragionevolmente certi su quale sia il fenomeno-causa e quale sia il fenomeno-effetto. A complicare le cose può verificarsi il caso in cui causa ed effetto scoperti in un primo esame possono invertire i loro ruoli, così: giusto per complicarci la vita. Supponiamo per esempio che si sia scoperta una forte correlazione fra reddito personale (Y) e possesso di pacchetti azionari (K) e che i dati a disposizione indichino il rapporto causale al tempo t:

$$K_t = f(Y_{t-1})$$

È ragionevole: più si guadagna e più si hanno margini finanziari per investire in azioni. Qui il nesso causale è ben chiaro. Ma più azioni si comprano e maggiore potrebbe essere il guadagno; dunque, al tempo t+1 potrebbe essere che

$$Y_{t+1} = f(K_t)$$

Dunque, non sarebbe corretto affermare né che:

$$K = f(Y)$$

e neppure che

$$Y = f(K)\ .$$

L'unica conclusione cui possiamo giungere è che i due fenomeni sono collegati:

$$Y \leftrightarrow K\ .$$

Forse il caso più insidioso è quello, piuttosto comune, in cui nessuna delle variabili considerate ha effetto sull'altra, tuttavia sembra esserci una reale correlazione. Su questo terreno è stato fatto molto lavoro sporco e molte delle correlazioni "scoperte" non hanno resistito ad indagini più approfondite. La questione del rapporto tra voti scolastici bassi (N) e forte dipendenza dal tabacco (T) al fumo rientra in questa categoria: affermare che c'è un legame e che questo legame è del tipo

$$N = f(T)$$

potrà fare effetto in una campagna contro il tabagismo, ma è una pura supposizione dal punto di vista statistico. Le informazioni "mediche" sono spesso costruite in modi simili e sono particolarmente pericolose perché quando si tratta della salute in genere le persone sono disposte a credere a qualunque panzana venga loro proposta.

Ricordate la storia dei vaccini & dell'autismo di qualche anno fa? No? Allora vale la pena di raccontare cosa successe: accadde che nel 1998 la prestigiosa rivista scientifica Lancet pubblicasse un articolo a firma del dott. Andrew Wakefield e di una dozzina di co-autori in cui si ipotizzava la relazione fra il vaccino MMR[9] e alcune gravi sindromi infantili. La pubblicazione dell'articolo venne accompagnata da una conferenza stampa di Wakefield e co-autori, evento quanto meno inusuale nella ricerca medica, ma dall'effetto dirompente. I ricercatori affermarono che, su un campione di 12 (dodici!) bambini ricoverati presso il reparto di gastroenterologia del Royal Free Hospital di Hampstead, a nord di Londra, erano stati riscontrati dei disturbi legati allo sviluppo che potevano essere stati causati dalla precedente somministrazione del vaccino MMR.

Comprensibilmente, apriti cielo! Lo studio, la conferenza stampa (e successivamente altri articoli, molti dei quali a firma dello stesso Wakefield) sostenevano che nell'arco di un paio di settimane dalla somministrazione del vaccino trivalente nel campione di piccoli ricoverati era insorta una sindrome che veniva descritta come una combinazione di serie patologie intestinali e quello che lo stesso Wakefield definiva "autismo regressivo", una versione tutta nuova del ben conosciuto disturbo mentale, a causa del quale i bambini (a cui in precedenza non era stata riscontrata alcuna forma di autismo) accusavano una progressiva perdita di capacità linguistico-motorie.

D'accordo, il campione era molto esiguo. Ma mettetevi nei panni di genitori che devono sottoporre il proprio figlio alla vaccinazione MMR: "E se toccasse proprio a mio figlio? Meglio rischiare l'insorgenza di una malattia come il morbillo, gli orecchioni o la rosolia che farlo ammalare di autismo!".

La campagna di Wakefield suscitò una vastissima eco presso l'opinione pubblica inglese e ben presto la diffidenza contro i vaccini attraversò l'Atlantico approdando in USA e poco più tardi nel resto d'Europa. Una parte del mondo dell'*entertainement* scese in campo a fianco della crociata di Wakefield, supportata anche da toccanti vicende personali come quella della *showgirl* Jenny

[9] Acronimo da Measles, Mumps e Rubella, cioè morbillo, orecchioni e rosolia

McCarthy[10]: i tassi di vaccinazione MMR in Gran Bretagna scesero sensibilmente a causa del panico scatenato dalla campagna di stampa anti-vaccinale.
Andrew Wakefield diventò ben presto una specie di eroe popolare, un Robin Hood che lottava contro le Big Pharma e cercava in tutti i modi la verità contro le speculazioni sulla salute. Il tutto, si badi, fondato su dati di quella iniziale ricerca e poco altro (se si esclude una gigantesca e ben orchestrata campagna di stampa). La sua immagine di giustiziere delle ragioni dei poveri è rimasta tale fino ad un'inchiesta del giornalista britannico Brian Deer nella quale si metteva in evidenza che Wakefield aveva qualche scheletro di troppo nel suo armadio: l'inchiesta di Deer rivelava una serie di conflitti di interesse che avevano permesso al "ricercatore" di guadagnare notevoli somme attraverso la campagna anti-vaccinale.
Alla fine, l'inganno venne scoperto e Wakefield venne espulso dall'ordine dei medici. Lancet pubblicò una smentita riguardo alle tesi contro i vaccini. Ma la diffidenza contro i vaccini resta tuttora una delle cause del rifiuto di una parte dei genitori a far vaccinare i propri figli e l'abbassamento, in qualche caso sotto il livello di guardia dell'effetto-gregge.
Il tutto a causa di una correlazione semplicemente inesistente…

[10] Al figlio di Jenny McCarthy, Evan, nato nel 1992 venne diagnosticato all'età di tre anni una grave disturbo autistico