# *Il paradosso di Simpson e i dati sulle vaccinazioni anticovid*

Quentin Berger (*Sorbonne Université*),
Francesco Caravenna (*Università di Milano - Bicocca*).

*Questo articolo è apparso in francese sul sito The Conversation (3 novembre 2021). Una traduzione in italiano è stata pubblicata sul sito di Internazionale (9 dicembre 2021).*

La statistica può essere la sorgente di risultati del tutto controintuitivi, benché dimostrati rigorosamente, detti *paradossi*. Questo termine indica infatti risultati che non sono falsi o incompatibili con altri risultati noti, ma che risultano contrari alla nostra intuizione.

### Il paradosso di Simpson

Uno dei paradossi statistici più sorprendenti è il paradosso di Simpson. Esso afferma che, analizzando una popolazione composta da diversi gruppi, è possibile che all'interno di ogni gruppo si osservi uno stesso fenomeno, mentre nella popolazione totale si osserva *il fenomeno opposto*. Questo paradosso è all'origine di molti errori di interpretazione, anche da parte di matematici esperti.

Presentiamo qui un esempio notevole, che emerge dai dati sulle vaccinazioni anti Covid-19 in Inghilterra. Esaminando i rapporti sui decessi in persone positive alla variante Delta del Covid-19, nel periodo tra giugno e settembre 2021, possiamo estrarre le informazioni seguenti (presentiamo i dati, le referenze complete e i calcoli in un'appendice):

1. nella popolazione sotto i 50 anni, la percentuale di decessi è circa 1,8 volte *più* elevata tra i non vaccinati rispetto ai vaccinati;

2. nella popolazione sopra i 50 anni, la percentuale di decessi è circa 6,3 volte *più* elevata tra i non vaccinati rispetto ai vaccinati;

3. invece, nella popolazione totale, la percentuale di decessi è circa 1,3 volte *meno* elevata tra i non vaccinati rispetto ai vaccinati.

Si impongono due considerazioni. Innanzitutto, la terza affermazione sembra contraddire le prime due: come è possibile che la vaccinazione riduca la percentuale di decessi sia nelle persone sotto i 50 anni sia in quelle sopra i 50 anni, ma la aumenti tra le persone di ogni età?

In secondo luogo, e in modo più inquietante, a seconda che ci basiamo sui dati relativi alle persone sotto i 50 anni e sopra i 50 anni separatamente, o che consideriamo le persone di ogni età, giungiamo a conclusioni opposte sull'efficacia del vaccino. In effetti, guardando i punti 1 e 2, il vaccino sembra efficace nel ridurre la mortalità sia nella popolazione sotto i 50 anni che in quella sopra i 50 anni. Se invece guardiamo il punto 3, riferito alla popolazione nel suo insieme, sembra legittimo concludere che il vaccino sia inefficace, per non dire pericoloso... Qual è la conclusione corretta?

### Spiegazione del paradosso

I dati precisi sono presentati in un'appendice, ma è utile dare una spiegazione generale del modo in cui questo paradosso può manifestarsi.

L'osservazione cruciale è che nel periodo in esame *la percentuale di persone vaccinate è molto più elevata sopra i 50 anni* (circa il 95% secondo i dati del NHS) che sotto i 50 anni (circa il 50%). Di conseguenza, tra le persone non vaccinate la grande maggioranza ha meno di 50 anni e ha un tasso di mortalità basso in ragione dell'età, mentre *tra le persone vaccinate la maggioranza ha più di 50 anni e ha dunque un tasso di mortalità più elevato* (anche se fortemente ridotto dal vaccino). Questo spiega perché, guardando alla popolazione nel suo complesso, la percentuale di decessi tra le persone non vaccinate può risultare inferiore a quella tra le persone vaccinate.

Un modo intuitivo di visualizzare questo paradosso è descritto nell'immagine seguente, in cui usiamo dati fittizi per rendere il fenomeno più chiaro:

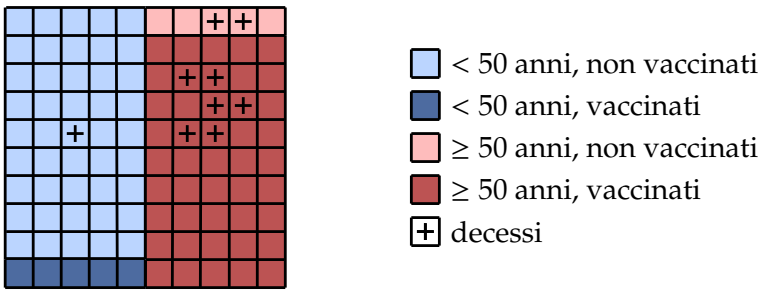
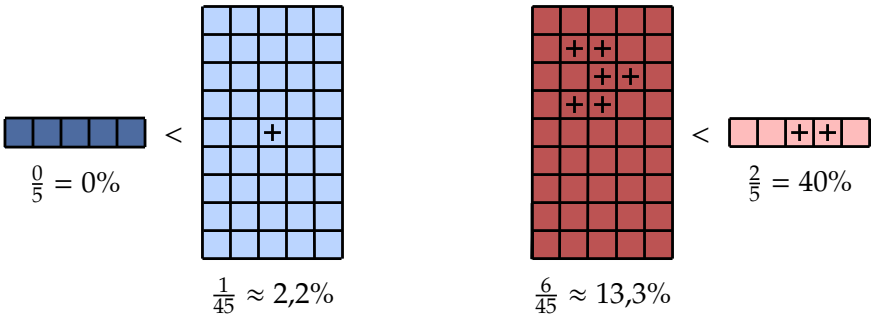


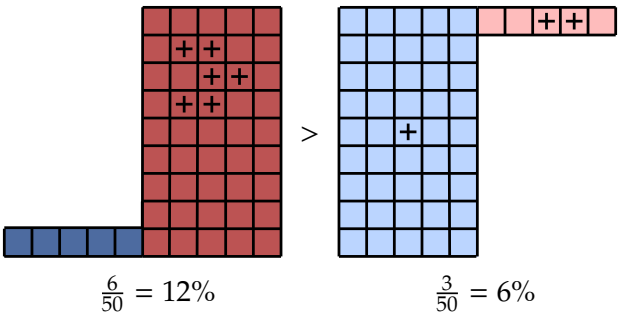

Illustrazione grafica del paradosso di Simpson (con dati fittizi): ogni persona è rappresentata da un quadrato, il colore indica la classe di età mentre la tonalità chiara o scura indica lo stato vaccinale; ogni croce indica un decesso.

Se consideriamo le persone sotto i 50 anni e quelle sopra i 50 anni come gruppi separati, si vede chiaramente che in entrambi i gruppi la percentuale di decessi è *inferiore tra i vaccinati*:




Sotto i 50 anni (in blu) la percentuale di decessi è più bassa tra i vaccinati (0%) che tra i non vaccinati (2,2%). Anche sopra i 50 anni (in rosso), la percentuale di decessi è più bassa tra i vaccinati (13,3%) che tra i non vaccinati (40%).

Se invece consideriamo persone di ogni età, la situazione si ribalta e la percentuale di decessi diventa *superiore tra i vaccinati*, come mostra la figura seguente:




Nella popolazione totale, la percentuale di decessi è più alta tra i vaccinati (blu scuro e rosso scuro, 12%) che tra i non vaccinati (blu chiaro e rosso chiaro, 6%).

Ciò è dovuto al fatto che la maggior parte dei vaccinati ha più di 50 anni.

### Che cosa possiamo concludere?

Da questo paradosso possiamo trarre un messaggio importante: bisogna fare molta attenzione quando si analizzano dati statistici che si riferiscono a gruppi con caratteristiche diverse. Nel nostro esempio, il paradosso di Simpson è dovuto al fatto che il tasso di vaccinazione varia molto con l'età, pertanto *è importante valutare l'efficacia del vaccino all'interno di un gruppo di persone con età il più possibile omogenee*.

Raggruppare gruppi con età molto diverse produce un fenomeno noto come "distorsione da selezione" (*selection bias*): l'insieme delle persone vaccinate è composto in gran parte da anziani, dunque più fragili, mentre l'insieme delle persone non vaccinate è composto per lo più da giovani, che sono meno fragili. Per questa ragione, confrontare vaccinati e non vaccinati tra le persone di ogni età corrisponde di fatto a *confrontare una popolazione mediamente anziana (i vaccinati) con una popolazione mediamente giovane (i non vaccinati)*. Affermare che la percentuale di decessi nell'intera popolazione è più elevata tra i vaccinati che tra i non vaccinati è dunque fuorviante, perché il confronto è falsato dalla grande variabilità del tasso di vaccinazione tra le diverse fasce d'età.

### Sulla difficoltà di interpretare le statistiche

I problemi legati alle distorsioni da selezione sono ben conosciuti in statistica e fanno parte degli errori di interpretazione più comuni.

Un esempio classico è legato allo statistico Abraham Wald: durante la seconda guerra mondiale, dopo avere analizzato gli aerei rientrati dai combattimenti, suggerì di rafforzare quelle parti dei velivoli che erano state *meno* colpite dai proiettili! Il ragionamento era che quelle potevano essere le parti più critiche, perché quando venivano colpite gli aerei avevano meno probabilità di ritornare dal combattimento. Wald aveva capito l'importanza di correggere la distorsione nota come "pregiudizio di sopravvivenza" (*survivorship bias*), che consiste nel fare analisi statistiche basandosi solo sui dati di chi sopravvive.

Le distorsioni da selezione, che siano consapevoli o meno, sono parte integrante della procedura di raccolta dei dati statistici, come è chiaro dall'esempio appena discusso. È importante rendersi conto di quali distorsioni sono presenti, per poterle correggere. Nel nostro esempio originale, confrontare la percentuale di decessi tra non vaccinati e vaccinati *nell'intera popolazione* comporta una distorsione dovuta all'età, come abbiamo visto. Un modo per correggerla è limitare il confronto a fasce di età il più possibile ristrette, all'interno delle quali il tasso di vaccinazione sia stabile.

Per concludere, i paradossi sono un modo efficace di ricordarci quali sono le insidie da evitare: con il loro aspetto sorprendente, ci colpiscono e ci aiutano ad affinare la nostra intuizione, o quantomeno a dubitarne. Ci ricordano che nessuno è infallibile e che spesso non è facile né immediato analizzare problemi anche semplici. Per questo i paradossi ci spingono ad approfondire le nostre riflessioni, con umiltà.

*Per gli amanti dei paradossi, ecco una lista dei più classici nell'ambito della teoria delle probabilità: il paradosso dei compleanni, il paradosso di Bertrand, il problema di Monty Hall, il dilemma del prigioniero, il paradosso dei figli, . . .*