DATI, STATISTICHE, NOTIZIE FALSE E DISTORTE

di Giovanni Alfredo Barbieri, Esperto di cultura statistica

I caratteri delle leggi statistiche

In occasione del discorso inaugurale della sua presidenza della Royal Statistical Society britannica, il 28 giugno 2017, il professor David Spiegelhalter (2017) ha rilevato che molte voci competono per la nostra attenzione e, in misura crescente, si avvalgono di dati statistici per dare un fondamento apparentemente oggettivo alle loro affermazioni. Le parti politiche vi giocano un ruolo importante, proponendo spesso soluzioni e ricette presentate come *evidence-based* e dunque inoppugnabili. Le tecnologie hanno moltiplicato le fonti: una cosa buona per il dibattito democratico, inficiata però dall'assenza di controlli di qualità. Il risultato è che invece che dall'evidenza scientifica, l'opinione pubblica è spesso guidata dalle emozioni.

All'estremo ci sono quelle che ormai tutti chiamiamo '*fake news*': notizie non solo false, ma *costruite* come false, con l'intento esplicito di trarci in inganno. Alcune contromisure – come il *fact-checking* – sono già in atto, anche se non sempre risultano efficaci, perché la 'notizia' ha sempre un'eco maggiore della smentita. Non è questo il rischio maggiore, dunque, secondo Spiegelhalter. Ben più pericolose sono «la manipolazione e l'esagerazione, attraverso un'interpretazione inappropriata, di 'fatti' che possono essere tecnicamente corretti ma sono distorti da quelle che potremmo chiamare 'pratiche discutibili di interpretazione e comunicazione'».

Ma perché troviamo tanto difficile discernere un'informazione veridica e attendibile da una falsa e di fonte dubbia, soprattutto quando si tratta di

dati statistici? Più in generale, perché risulta così arduo, anche per persone colte e istruite, trarre conclusioni corrette sulla base del calcolo delle probabilità?

Sono due questioni diverse, anche se collegate strettamente. Partiamo dalla seconda.

La statistica si è sviluppata relativamente tardi: i fondamenti della geometria sono stati sistematizzati da Euclide, oltre duemila anni fa; il calcolo delle probabilità si fa risalire a uno scambio di lettere tra Pascal e Fermat del 1654.

Come mai? Perché le leggi statistiche sono molto diverse dalle altre leggi scientifiche. Nella geometria euclidea, la somma degli angoli interni di un triangolo è sempre eguale all'angolo piatto. In fisica, un sifone funziona sempre allo stesso modo, quale che sia il materiale di cui è fatto il tubo – gomma piombo o ceramica – e quale che sia il fluido – acqua olio o vino. Invece, nel lancio di una moneta, la probabilità che venga testa è del 50%: ma che significa? Non ha senso, per un singolo lancio. E non mi garantisce che, su 100 lanci, in 50 venga testa. Deve entrare in gioco un diverso modo di ragionare. Lo stesso termine 'probabilità' rinvia a questa differenza: il riferimento non è alla dimostrazione di un teorema (prova matematica), o all'affermazione di un'autorità, o all'asserzione di un testimone (prova estrinseca), ma al contributo alla conferma di un'ipotesi in un contesto induttivo (Hacking 1975).

Questa caratteristica delle leggi statistiche non è una loro debolezza, ma un loro punto di forza. Non operiamo quasi mai in un contesto di certezze (secondo un noto detto, sono ineluttabili soltanto la morte e le tasse, e in Italia ci sono parecchi dubbi sulla certezza delle seconde). Disporre di strumenti che ci consentono di assumere decisioni informate in condizioni

di incertezza e di rischio merita dunque di essere considerata una grande conquista della scienza e dell'intelletto umano, che ci facilita la vita sia nelle grandi scelte sia in quelle di tutti i giorni.

Usiamo questi strumenti quotidianamente, e spesso inconsapevolmente, applicando quello che Kahneman (2011) chiama 'pensiero veloce'. E di solito non commettiamo errori. A volte, però, ci troviamo in contesti più difficili, magari concepiti ad arte per confonderci (pensate al gioco del lotto e ai suoi derivati, creati per finanziare l'erario e non per far vincere i cittadini). In questi casi l'intuito statistico ci trae in inganno.

Dati statistici e percezioni

Questa risposta, ancorché parziale, alla seconda domanda, apporta elementi utili per la prima: quando il 'pensiero veloce' ci porta fuori strada, le nostre percezioni risultano errate.

Lo scorso anno Einaudi ha pubblicato la traduzione italiana di un libro di Bobby Duffy, *I rischi della percezione*. Duffy – ci informa l'editore – «è Professor of Public Policy e direttore del Policy Institute presso il King's College di Londra. In passato è stato direttore generale dell'Ipsos MORI Social Research Institute e direttore globale dell'Ipsos Social Research Institute.» Ipsos MORI è un'importante società britannica di ricerche di mercato, parte del gruppo francese Ipsos, che opera anche in Italia (il suo volto più noto è Nando Pagnoncelli).

Dal 2012 Ipsos MORI svolge un'indagine, che ormai coinvolge 50.000 intervistati in 13 paesi, sulla *percezione* della situazione economica e sociale, come documentata dalle statistiche. Alla domanda «Su 100 persone in età lavorativa quante, secondo lei, sono disoccupate e cercano lavoro?», gli italiani intervistati hanno risposto (in media): 49. All'epoca

della rilevazione, invece, erano 12 – quattro volte di meno. Siamo cioè risultati, per questa domanda e per l'insieme delle domande della rilevazione Ipsos-MORI, il Paese con le percezioni più sbagliate.

Perché? Facile rispondere dando la colpa al basso grado di istruzione della popolazione, all'analfabetismo di ritorno, alla superficialità e alla ricerca delle notizie a effetto dei mezzi d'informazione. Probabilmente c'è del vero in ciascuna di queste spiegazioni. Ma vorrei provare ad aggiungere un'altra ragione. E con questo cercare di rispondere alla prima delle domande che ci siamo posti all'inizio: perché troviamo tanto difficile discernere un'informazione veridica e attendibile da una falsa e di fonte dubbia, soprattutto quando si tratta di dati statistici?

La statistica applica ai fenomeni collettivi un approccio scientifico, in cui sono irrinunciabili la *quantificazione* (cioè la loro traduzione in termini di quantità, introducendo precisi elementi di valutazione) e la *misurazione* (cioè la determinazione della loro misura, confrontandoli con uno standard predefinito).

Per fare questo, la statistica usa una strategia di classificazione: ogni entità deve essere assegnata a un contenitore, in modo che il contenuto di questo possa essere quantificato e misurato. Affinché queste operazioni possano svolgersi correttamente, senza errori di calcolo, è necessario che le classi siano esaustive (ogni entità, nessuna esclusa, deve essere assegnata a una classe) e mutuamente esclusive (se un'entità appartiene a una classe, non può appartenere a nessun'altra classe). Pensate alle caratteristiche della popolazione rispetto allo stato occupazionale: ogni individuo deve essere assegnato a una delle tre classi definite (occupato, disoccupato o inattivo, senza possibili vie di mezzo) e se è assegnato a una classe non può essere allo stesso tempo assegnato a un'altra (o sei occupato, oppure sei disoccupato o inattivo). Inoltre, le classi devono

essere organizzate gerarchicamente (ogni classe di livello superiore contiene un numero intero e finito di classi del livello inferiore, e ogni classe di livello inferiore appartiene a una e una sola classe di livello superiore). Pensate alla classificazione amministrativa del territorio: ogni comune appartiene a una e a una sola provincia, che è composta da un numero intero e finito di comuni; a sua volta, ogni provincia appartiene a una e a una sola regione, che è composta da un numero intero e finito di province; e dunque è vero anche che ogni comune appartiene a una e a una sola regione, che è composta da un numero intero e finito di comuni…

Questa strategia di classificazione ha molti vantaggi. In primo luogo, permette di trattare quantitativamente entità che, per quanto simili, differiscono in caratteristiche considerate irrilevanti ai fini di un particolare trattamento statistico. È una strategia di digitalizzazione in senso lato, che permette di minimizzare gli errori e di migliorare il rapporto tra segnale e rumore: è una strategia che la scienza e la tecnologia utilizzano pervasivamente e che noi stessi applichiamo ogni volta che parliamo, scriviamo, leggiamo, facciamo musica… È anch'essa una componente del 'pensiero veloce'.

Ma queste operazioni di codifica hanno comunque un costo, il costo dell'astrazione: si astrae dalla complessità implicita nel grande numero di variabili e attori del processo modellizzato, ma questo significa che c'è un livello di dettaglio in cui le ipotesi assunte sono false. Un grande statistico, George Box, ha riassunto la questione così: «*All models are wrong, but some are useful*» (Box 1979).

Fin qui tutto bene, dunque? Non proprio. Il problema è che questa strategia si riflette nel linguaggio che la statistica parla, all'interno della sua comunità scientifica, ma anche comunicando a tutti noi i risultati della sua attività, cioè le informazioni statistiche.

Mi spiego con un esempio: quando l'Istat comunica i dati periodici sul mercato del lavoro, e dà il numero dei disoccupati, fa di necessità riferimento alle definizioni e ai criteri di classificazione che ha adottato per stimarlo, perché ha dovuto collocare ogni singolo individuo del campione (e della popolazione) in uno dei contenitori della classificazione: occupato, disoccupato o inattivo. Non possono esserci casi dubbi. Ma i casi dubbi e le sfumature esistono, e su quelli occorre decidere – anche se sulla base di regole certe, definite a livello internazionale (per consentire confronti tra paesi diversi) e costanti nel tempo (per consentire confronti tra periodi diversi). *De-cidere*: cioè tagliare via, come ci ricorda l'origine della parola. Pagare un prezzo, come dicevamo poco fa: sacrificare la grande diversità dei casi individuali per 'costringerli' dentro uno schema di classificazione.

Chi, ad esempio, è senza lavoro e desidera lavorare, ma non ha svolto azioni concrete di ricerca di lavoro, è classificato come 'inattivo' e non come 'disoccupato'. Chi ha lavorato anche una sola ora alla settimana nel periodo di riferimento è classificato come 'occupato' e non come 'disoccupato'. Sono situazioni al margine, quantitativamente poco importanti, e non spostano in misura sensibile i dati statistici su occupazione e disoccupazione. Ma influenzano fortemente la nostra percezione, strutturalmente più sensibile ai dettagli che al quadro d'insieme: è la nostra mente che funziona così. Anche questo è uno dei modi in cui opera il 'pensiero veloce'.

Avvicinare il linguaggio della statistica a quello naturale

Ma non si tratta soltanto questo. Il modo in cui di solito organizziamo i concetti nel nostro pensiero non è quello delle classificazioni esaustive, mutuamente esclusive e organizzate gerarchicamente. Nel 'linguaggio

naturale', quello che applichiamo nella vita quotidiana, organizziamo i concetti diversamente, per somiglianze, analogie, metafore. Mi avvarrò di un celebre esempio proposto da Wittgenstein (1953). Pensiamo ai processi che chiamiamo 'giochi': «giochi da scacchiera, giochi di carte, giochi di palla, gare sportive, e via discorrendo». Per applicare un criterio di classificazione come quello cui abbiamo fatto riferimento finora ci dovrebbe essere qualche elemento che li accomuna tutti: ma non c'è. Quello che c'è sono quelle che Wittgenstein chiama 'somiglianze di famiglia': vediamo «somiglianze emergere e sparire, […] una rete complicata di somiglianze che si sovrappongono e si incrociano a vicenda, […] somiglianze in grande e in piccolo». Estendiamo i concetti – conclude Wittgenstein – «come, nel tessere un filo, intrecciamo fibra con fibra. E la robustezza del filo non è data dal fatto che una fibra corre per tutta la sua lunghezza, ma dal sovrapporsi di molte fibre una all'altra».

Alla radice delle percezioni errate c'è dunque anche, e soprattutto, la distanza tra due linguaggi: quello tecnico, rigoroso e altamente formalizzato della statistica (e più in generale della scienza) e quello 'naturale', usato nella vita quotidiana e profondamente radicato nel nostro modo di pensare. Spesso in questa distanza si intrufolano e fanno leva le 'pratiche discutibili di interpretazione e comunicazione' di cui parlava Spiegelhalter.

Che cosa si può fare per colmarla e, dunque, per togliere spazio a chi la usa per proporre interpretazioni inappropriate delle informazioni statistiche? Non penso che ci sia una soluzione facile, ma propongo tre passi che – a mio parere – vanno nella direzione giusta:

- Il primo è che gli istituti di statistica si facciano carico dell'onere della traduzione dal linguaggio tecnico (indispensabile nel processo di produzione statistica dei dati, ma non in quello della

comunicazione al pubblico delle relative informazioni) al linguaggio naturale (conservando il necessario rigore). Chiedo agli istituti di statistica di dedicare alle parole della comunicazione dei dati la stessa attenzione che dedicano alla correttezza dei numeri.

- Il secondo è che la fiducia nell'autorevolezza e nella veridicità dei dati della statistica ufficiale deve essere conquistata quotidianamente: gli istituti di statistica si devono mostrare affidabili, competenti e indipendenti. Per questo occorre che si rendano 'vulnerabili', fornendo agli altri i mezzi per verificarlo. Mettendo in vista la cucina, come ormai fanno molti ristoranti.
- Il terzo è che il punto di forza degli istituti di statistica è il loro capitale umano: un indizio di questo potenziale è il divario tra i cittadini britannici che si fidano dell'accuratezza delle cifre ufficiali (78%) e quelli che si fidano dell'ONS, il loro istituto di statistica (90%) (NCSR 2017). Gli esperti in grado di mettere insieme dati e metadati, dando informazioni sul significato e l'utilità delle statistiche ufficiali, possono innescare un circolo virtuoso, in cui produttori e utenti co-evolvono nella loro capacità di utilizzare i dati.

***

Riferimenti bibliografici

Box, George E. P. (1979). "Some problems of statistics and everyday life". *Journal of the American Statistical Association*. 74 (365): 1–4. doi:10.2307/2286713. JSTOR 2286713.

Duffy, Bobby (2018). *The Perils of Perception: Why We're Wrong About Nearly Everything*. London: Atlantic Books. ISBN 9781786494573. [*I rischi della percezione: perché ci sbagliamo su quasi tutto*. Trad. F. Pè. Torino: Einaudi. 2019]

Hacking, Ian (1975). *The Emergence of Probability: A Philosophical Study of Early Ideas about Probability, Induction and Statistical Inference*. Cambridge UK): Cambridge University Press. ISBN 9780521866552.

Kahneman, Daniel (2011). *Thinking, Fast and Slow*. New York (NY): Farrar, Straus and Giroux. ISBN 9780374533557. [*Pensieri lenti e veloci*. Trad. L. Serra. Milano: Mondadori. 2012].

National Centre for Social Research (2017). Public confidence in official statistics. London: NCSR. Retrieved from http://natcen.ac.uk/our-research/research/public-confidence-in-officialstatistics.

Spiegelhalter, David (2017). "Trust in numbers [The address of the President delivered to The Royal Statistical Society on Wednesday, June 28th, 2017]". *Journal of the Royal Statistical Society*: Series A (Statistics in Society), October 2017: 948-965. https://doi.org/10.1111/rssa.12302

Wittgenstein, Ludwig (1953-1958). *Philosophische Untersuchungen. Philosophical investigations* (ed. Anscombe, G.E.M. e Rhees, R.). Oxford: Blackwell. ISBN 9781405159289. [*Ricerche filosofiche*. Trad. R. Piovesan e M. Trinchero. Torino: Einaudi, 1967].