LA PROBABILITÀ:
CAPIRE LA REALTÀ E PRENDERE DECISIONI MIGLIORI[1]

LEONARDO GRILLI

Dipartimento di Statistica, Informatica, Applicazioni "G. Parenti", Università di Firenze

1. *Il ruolo del caso nella nostra vita: ignorarlo o comprenderlo?*

L'atteggiamento ambiguo degli esseri umani di fronte al caso è ben descritto da Jeffrey Rosenthal [5]: "Gli esseri umani sono sempre stati affascinati e, contemporaneamente, terrorizzati dal caso. La verità è che, quando entra in gioco il caso, possiamo fuggire, ma non possiamo nasconderci. Moltissimi aspetti della nostra vita sono determinati da eventi che non controlliamo completamente, e l'incertezza è ineliminabile. Abbiamo due opzioni: possiamo lasciare che l'incertezza prevalga su di noi o possiamo imparare a comprendere il caso. Se optiamo per la seconda soluzione, faremo scelte migliori e impareremo a sfruttare la casualità per i nostri scopi."

Lo strumento per la comprensione del caso è la probabilità. Il concetto di probabilità è tuttora oggetto di controversie sia tra i filosofi che tra i matematici e gli statistici. Una definizione molto generale, dovuta al matematico italiano Bruno De Finetti, afferma che la probabilità è il *grado di fiducia* che un *individuo razionale* attribuisce al verificarsi di un evento. Formalmente la probabilità di un evento E si indica con $P(E)$ ed è un numero compreso tra 0 (impossibile) e 1 (certo). In alternativa la probabilità può essere espressa in termini percentuali, variando quindi tra 0% (impossibile) e 100% (certo). Per assegnare un valore alla probabilità l'individuo razionale usa tutte le informazioni disponibili sulla struttura dell'esperimento e sulle sue precedenti realizzazioni. In alcune situazioni vi sono delle semplici regole per determinare il valore della probabilità: si tratta dell'approccio classico e di quello frequentista.

L'*approccio classico* si applica ad un esperimento aleatorio i cui risultati possibili sono in numero finito ed *equiprobabili* (= stessa probabilità). In questa situazione la probabilità di un evento è il rapporto tra il numero di casi favorevoli e il numero di casi possibili. Ad esempio, nel lancio di un dado la probabilità di ottenere un valore dispari è 0.5, dato dal rapporto tra i 3 casi favorevoli e i 6 casi possibili (assumendo che il dado sia bilanciato).

L'*approccio frequentista* si applica quando un esperimento aleatorio viene ripetuto con una serie di prove *indipendenti* e *in identiche condizioni*. In questa situazione la probabilità di un evento

[1] Lezione tenuta il 13.01.2014 presso il Liceo Scientifico "Copernico", Prato, e il 10.02.2014 presso il Liceo Classico Statale "Michelangiolo", Firenze.

è calcolata come rapporto tra il numero di prove in cui si è verificato l'evento e il numero totale di prove (formalmente la probabilità in senso frequentista è definita come il limite a cui tende il suddetto rapporto quando il numero di prove tende a infinito). Ad esempio, supponiamo di lanciare un dado 50 volte ottenendo un valore dispari in 23 casi: la stima della probabilità è 23/50 = 0.46. Se dopo 100 lanci i valori dispari osservati sono 49, la stima della probabilità viene aggiornata in 49/100 = 0.49. La stima frequentista della probabilità è dunque variabile, ma al crescere del numero di prove si stabilizza attorno ad un valore (0.5 nell'esempio del dado, se questo è bilanciato).

2. *Eventi indipendenti e la regola moltiplicativa*

Due eventi sono *indipendenti* quando il verificarsi dell'uno non modifica la probabilità del verificarsi dell'altro. Ad esempio, consideriamo i seguenti eventi: E_1 = «indossare calzini a righe», E_2 = «arrivare tardi all'appuntamento con la fidanzata», E_3 = «litigare con la fidanzata». Verosimilmente E_1 è indipendente sia da E_2 che da E_3, mentre E_2 non è indipendente da E_3!

L'indipendenza è la condizione alla base della *regola moltiplicativa* $P(A \text{ e } B) = P(A)\times P(B)$. Infatti, la probabilità che due eventi possibili A e B si verifichino entrambi è uguale al prodotto delle singole probabilità se e solo se i due eventi sono indipendenti. Un errore comune consiste nell'applicare la regola moltiplicativa quando la condizione di indipendenza è palesemente violata.

I calcoli probabilistici errati di solito non hanno gravi conseguenze (se non un brutto voto all'esame di Statistica), ma a volte gli esiti sono drammatici, come nel caso giudiziario di Sally Clark. Nel 1999 una corte Britannica condannò Sally Clark per l'omicidio dei due suoi bambini morti improvvisamente all'età di 11 e 8 settimane per cause ignote. Non furono trovate prove dell'omicidio, né una valida motivazione. L'accusa sosteneva il soffocamento volontario, la difesa parlava di cause naturali, la cosiddetta sindrome della "morte in culla" (SIDS). La prova principale per la condanna di Sally Clark fu una stima probabilistica contenuta nella perizia di un pediatra, secondo cui la probabilità che si verifichino due casi di "morte in culla" è di circa 1 su 73 milioni. Questo valore è stato ottenuto in due passi: prima il pediatra ha esaminato i dati sulla sindrome della "morte in culla", osservando che colpisce un bambino ogni 8500, per cui la probabilità di un caso è stimata pari a 1/8500; poi il pediatra ha calcolato la probabilità congiunta di due casi di "morte in culla" tramite la regola moltiplicativa, ottenendo l'impressionante stima di 1 su 73 milioni.

Come evidenziato dalla Royal Statistical Society, il calcolo del pediatra era completamente errato per vari motivi, fra cui il fatto che non vi erano le condizioni per applicare la regola moltiplicativa. Infatti, una seconda "morte in culla" non è indipendente dalla prima perché vi sono cause genetiche: in una famiglia in cui si è già verificato un caso, la probabilità che si verifichi un secondo caso è stimata in circa 1/100 (si tratta è una *probabilità condizionata*). Pertanto, la probabilità congiunta di due casi di "morte in culla" non è data da 1/8500 ×1/8500, ma da 1/8500 ×1/100, il cui risultato è 1/850000. Un primo appello nel 2000 ha confermato la sentenza, ma un secondo appello nel 2003 ha assolto Sally Clark, rilasciata dopo tre anni di carcere. Questa esperienza ha segnato irrimediabilmente la donna, deceduta nel 2007. Il sito web `www.sallyclark.org.uk` riporta i dettegli di questa triste vicenda, nella speranza che il dibattito pubblico contribuisca ad evitare il ripetersi di errori così gravi.

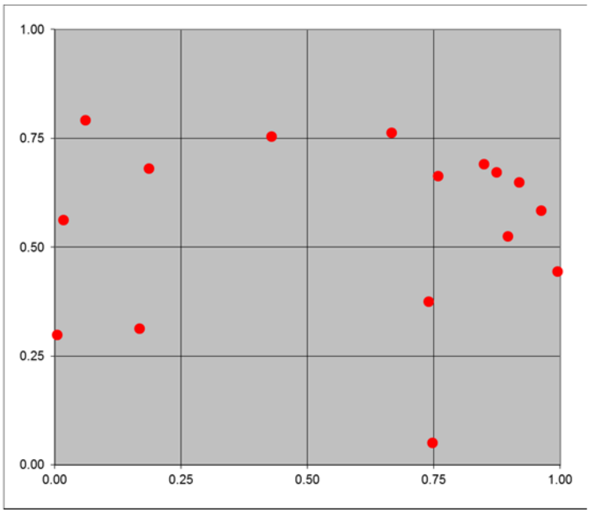
3. *Coincidenze e raggruppamenti casuali*

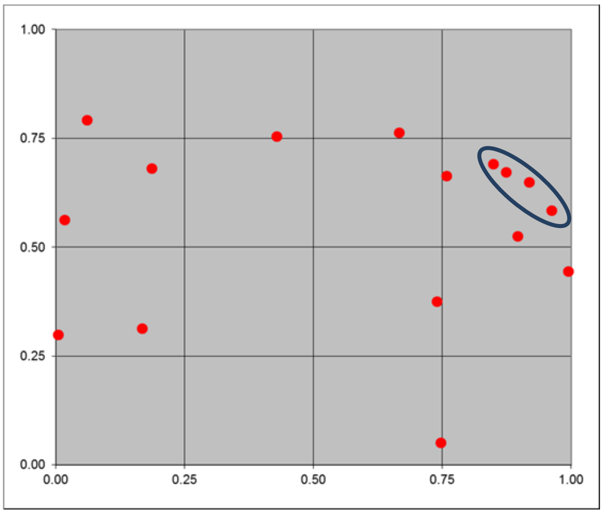
Alcuni eventi appaiono sorprendenti perché intuitivamente riteniamo che abbiano una probabilità molto piccola. Tuttavia, se eseguiamo correttamente il calcolo, in molti casi scopriamo che la probabilità non è così piccola [2]. Ad esempio, nell'estrazione del gioco del Lotto del 24 dicembre 2013 il numero 57 è uscito, come primo estratto, sulle ruote di Bari e Milano – che coincidenza! Tuttavia, per capire se si tratta davvero di un evento raro dobbiamo calcolare la sua probabilità. A questo punto ci rendiamo conto che la probabilità dipende dalla formulazione della domanda: spesso l'ambiguità sta proprio nel fatto che nella nostra mente la domanda non è espressa in modo preciso. Consideriamo tre formulazioni alternative:

- Formulazione n. 1 – Probabilità che esca il numero 57 come primo estratto sia a Bari che a Milano: $1/90 \times 1/90 = 1/8100 = 0.00012$;
- Formulazione n. 2 – Probabilità che esca lo stesso numero (non necessariamente il 57, può essere anche il 12 o il 63) come primo estratto sia a Bari che a Milano: $1/90 = 0.01111$;
- Formulazione n. 3 – Probabilità che, considerando tutte le 11 ruote, vi siano almeno due numeri primi estratti identici: P(almeno due numeri identici) = $1–P$(tutti numeri diversi) = $1–90/90 \times 89/90 \times 88/90 \times ... \times 80/90 = 1–0.5292 = 0.4708$.

L'ultimo calcolo mostra che non dovremmo sorprenderci quando osserviamo due numeri primi estratti identici perché questo accade quasi una volta ogni due estrazioni!

Un caso particolarmente interessante di coincidenze ingannevoli è quello dei *raggruppamenti casuali*. Il tema è introdotto da Carlo Rovelli [4] con l'esempio dei chicchi di riso lanciati sul pavimento: anche se il riso viene lanciato a caso, i chicchi non si distribuiscono uniformemente perché si osserva che alcune mattonelle risultano vuote, mentre su altre finiscono molti chicchi. L'esperimento del riso si può effettuare in modo virtuale tramite un foglio di calcolo elettronico: ad esempio, possiamo disegnare un quadrato di lato unitario (pavimento) suddiviso in 16 quadratini (mattonelle) e generare in modo casuale le coordinate di 16 punti (chicchi di riso). Per generare le coordinate si può usare la distribuzione uniforme tra 0 e 1 della funzione CASUALE() sia per la x che per la y. Siccome la superficie è divisa in 16 quadratini, ci aspettiamo in media 1 punto per quadratino, ma non è quello che accade. La figura a sinistra mostra il risultato di una estrazione casuale delle coordinate di 16 punti.





A fronte di un numero atteso di 1 punto per quadratino, osserviamo che la distribuzione dei punti è molto variabile: vi sono molti quadratini con 0 punti, mentre un quadratino ha addirittura 6 punti, con un rapporto *osservati su attesi* di 6 a 1. In questo esperimento abbiamo generato i punti casualmente e quindi sappiamo che quel rapporto così elevato è solo frutto del caso, ma studiando dati reali potremmo concludere che il quadratino con rapporto 6 a 1 ha una concentrazione anomala di punti. Nei dati reali i quadratini sono aree geografiche e i punti rappresentano eventi registrati in certo intervallo di tempo, ad es. casi di tumore, incidenti stradali, omicidi … allora il quadratino con rapporto 6 a 1 potrebbe generare un inutile allarme. È compito degli statistici valutare, di volta in volta, se un elevato rapporto *osservati su attesi* è spiegabile da un raggruppamento casuale oppure se è così grande da indicare una situazione anomala e far scattare l’allarme.

Un raggruppamento casuale genera un rapporto osservati su attesi ancora più impressionante se l’area di riferimento viene definita a posteriori, disegnandola attorno al gruppo di punti. Si tratta del cosiddetto *effetto cecchino*, a ricordo della storiella di quel tipo bizzarro che per mostrare la sua abilità di tiratore sparava un colpo a caso e poi disegnava il bersaglio intorno al foro. Ad esempio, la figura a destra mostra un ellisse che racchiude 4 punti ed ha un’area pari a circa 1/5 di quadratino, per cui il rapporto *osservati su attesi* è 4/(1/5) = 20, dunque 20 a 1 – davvero impressionante!

Leonard Mlodinow [3] riporta l’opinione di alcuni esperti a proposito delle conseguenze dei raggruppamenti casuali sulle politiche di sanità pubblica. Raymond Richard Neutra (Dipartimento della sanità della California) afferma che prendendo in esame un tipico registro dei tumori (una banca dati sui tassi di incidenza locali di dozzine di tumori diversi) per i 5000 distretti della California, ci attendiamo di trovarne 2750 con un’incidenza elevata di qualche forma di tumore, statisticamente significativa ma casuale. L’opinione pubblica è molto allarmata da queste situazioni di incidenza elevata e non ritiene verosimile che siano frutto del caso, per cui i dipartimenti di sanità sono indotti ad avviare costose indagini sulle possibili cause ambientali (es. inquinamento atmosferico, elettromagnetico etc.) che di solito non portano ad alcun risultato convincente. Alan Bender (Dipartimento della sanità del Minnesota) è convinto che queste indagini siano “un assoluto, totale e completo spreco dei dollari dei contribuenti”.

4. *Probabilità a priori e a posteriori: come valutare il risultato un test diagnostico?*

La probabilità che ognuno di noi assegna ad un evento dipende dalle informazioni a disposizione. Raccogliere informazioni attendibili consente di rendere più precisa la stima della probabilità: ad esempio, di fronte a certi sintomi il medico stima (di solito in modo inconscio) la probabilità che il paziente abbia una certa patologia e poi prescrive un test diagnostico (esame del sangue, radiografia etc.) per “raffinare” la stima iniziale. Si consideri che i test diagnostici sono fallibili, cioè non forniscono risultati sicuri, per cui anche dopo l’esito del test il medico assegna una probabilità che in genere è diversa sia da 0 che da 1. La probabilità prima del test diagnostico viene detta *a priori*, mentre quella successiva viene detta *a posteriori*. Come si passa dall’una all’altra?

Consideriamo una ipotetica malattia chiamata Stupidite che colpisce il 10% della popolazione (in termini medici questo 10% è detto *prevalenza*). Prendiamo una persona a caso dalla popolazione. Se non conosciamo l’esito del test diagnostico, la probabilità che abbia la malattia è detta a *probabilità a priori* e coincide con la prevalenza, in questo caso 10%.

Supponiamo di fare la diagnosi usando il test di laboratorio StupiTest. Come tutti i test diagnostici, lo StupiTest è utile ma non è perfetto: infatti, il test viene positivo (cioè, indica presenza di malattia) nel 70% dei soggetti malati (*sensibilità del test*) e nel 20% dei soggetti sani (*percentuale di falsi positivi*).

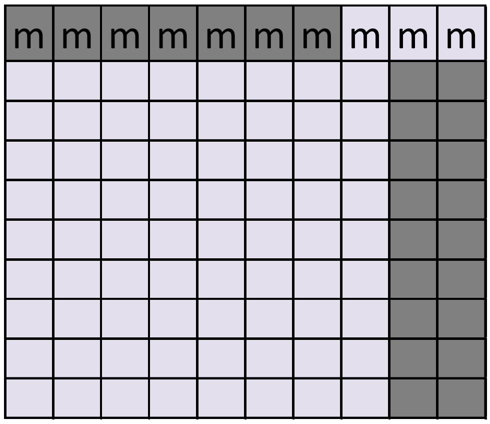
Sappiamo che Tizio ha effettuato il test, ottenendo un esito positivo: qual è la probabilità che Tizio sia davvero affetto da Stupidite? Dobbiamo calcolare la probabilità a posteriori. Prendiamo una persona a caso dalla popolazione. Se sappiamo che ha effettuato il test ed è risultato positivo, la probabilità che abbia la malattia è detta a probabilità a posteriori. Quanto vale la probabilità a posteriori? In questa situazione ipotetica, siccome il test è risultato positivo, la probabilità a posteriori deve essere maggiore di quella a priori (10%). È forse uguale alla sensibilità del test (70%)? La risposta si ottiene con la famosa formula di Bayes oppure con il metodo delle frequenze, cioè convertendo le probabilità in frequenze di una ipotetica popolazione – vediamo come si applica questo metodo. Ricapitoliamo:

- ✓ i soggetti malati sono il 10%;
- ✓ tra i malati, il test viene positivo nel 70% dei casi;
- ✓ tra i sani, il test viene positivo nel 20% dei casi.

Cosa ci aspettiamo sottoponendo al test 100 soggetti?

- ✓ 10 sono malati e 90 sono sani;
- ✓ tra i 10 malati, il test viene positivo a 7 soggetti;
- ✓ tra i 90 sani, il test viene positivo a 18 soggetti.

La situazione è illustrata dalla seguente matrice le cui 100 celle rappresentano i soggetti; le 10 celle con la lettera ‘m’ rappresentano i soggetti malati, mentre le 25 celle scure rappresentano i soggetti per i quali il test viene positivo (si noti che, per rappresentare le caratteristiche del test, le celle scure sono 7 tra le 10 con la lettera ‘m’ e 18 tra le altre 90).



Con questa impostazione la soluzione è semplice: infatti, il test viene positivo a 25 soggetti, di cui 7 malati e 18 sani, quindi se prendiamo un soggetto a caso tra i 25 che hanno effettuato il test con risultato positivo, la probabilità che sia malato è 7/25 = 28% – questa è la probabilità a posteriori, un valore intermedio tra la probabilità a priori (10%) e la sensibilità del test (70%).

Lo psicologo cognitivo G. Gigerenzer [1] ritiene che il metodo appena illustrato, basato sulle frequenze di una ipotetica popolazione, sia più facile da apprendere rispetto alla formula di Bayes e che quindi debba essere impiegato come strumento didattico per colmare la diffusa incapacità di calcolare correttamente i rischi, che riguarda perfino i professionisti.

Bibliografia

[1] Gigerernzer, G., Quando i numeri ingannano. Imparare a vivere con l'incertezza. R. Cortina, 2003.

[2] Hand, D., Il caso non esiste: Perché le cose più incredibili accadono tutti i giorni. Rizzoli, 2014.

[3] Mlodinow, L., La passeggiata dell'ubriaco. Le leggi scientifiche del caso. Rizzoli, 2009.

[4] Rovelli, C., Sì, no, anzi: probabilmente. Il Sole 24 ore del 20 gennaio 2013. Sito web: `http://foglianuova.wordpress.com/2013/01/20/carlo-rovelli-si-no-anzi-probabilmente`.

[5] Rosenthal, J.S., Le regole del caso: istruzioni per l'uso. Longanesi, 2006.