

## Capitolo 7. Probabilità, verosimiglianze e teorema di Bayes.

### Probabilità, verosimiglianza e teorema di Bayes

Se A e B sono due eventi **mutualmente esclusivi**, la probabilità che si verifichi l'uno o l'altro è data dalla somma delle rispettive probabilità individuali:

$$p(A \text{ o } B) = p(A) + p(B)$$

Due eventi mutualmente esclusivi possono essere anche **esaustivi**. In tal caso:

$$p(A \text{ o } B) = p(A) + p(B) = 1$$

Esempio:

- I semi cuori e quadri delle carte sono mutualmente esclusivi ma non esaustivi (mancano picche e fiori)
- Maschio e femmina sono mutualmente esclusivi ed esaustivi

Se il fatto che si verifichi un evento non influenza la probabilità che dopo si verifichi l'altro, allora si dice che gli eventi sono **indipendenti**. La probabilità che si verifichino entrambi (simultaneamente o in successione, ma restando indipendenti) è quindi data dal prodotto delle rispettive probabilità individuali **a priori**:

$$p(A \text{ e } B) = p(A) \times p(B)$$

Se invece gli eventi non sono indipendenti, nel senso che il fatto che si sia verificato l'uno influenza la probabilità che si verifichi l'altro, allora si parla di **probabilità condizionata** o **congiunta** o **a posteriori**, espressa mediante la formula:

$$p(A \text{ e } B) = p(A) \times p(B/A)$$

$$p(B \text{ e } A) = p(B) \times p(A/B)$$

ove  $p(A/B)$  sta per la probabilità di A dato B. Altrettanto dicasi per  $p(B/A)$ .

Nota bene:  $p(B/A)$  è diverso da  $p(A/B)$ . Le due espressioni non sono complementari né esiste tra loro alcuna precisa relazione definibile a priori.

Ad esempio, è ovvio che la probabilità di estrarre dal mazzo di 40 carte una carta rossa che sia anche il re di quadri ( $K\spadesuit$ ) si riferisce ad eventi tra loro non indipendenti, in quanto lo stesso  $K\spadesuit$  è una carta rossa. Allora la probabilità di estrarre dal mazzo di 40 carte una carta rossa che sia anche il  $K\spadesuit$  è data da:

$$p(\text{carta rossa e } K\spadesuit) = p(\text{carta rossa}) \times p(K\spadesuit/\text{carta rossa}) = 20/40 \times 1/20 = 1/40$$

Il risultato è apparentemente sciocco, in quanto sappiamo già che nel mazzo composto da 40 carte c'è un solo  $K\spadesuit$ . Ma il metodo diventa utile in tutte quelle situazioni in cui non abbiamo conoscenze dirette dei fenomeni che ci interessano.

Esempi di eventi <i>indipendenti</i>	Esempi di eventi <i>dipendenti</i>
Regina - Carta di picche	Occhi chiari - Capelli chiari
Domenica - Giorno dispari	Carta di cuori - Carta rossa
Malato di raffreddore - Calvo	Pioggia - Giornata nuvolosa

E' un grande vantaggio potersi basare su probabilità a posteriori, anziché su probabilità a priori. Consideriamo l'ultimo esempio. Se si sa in anticipo che la giornata è di pieno sole, senza neppure una nuvola, è assai più facile prevedere che non piova. Come dire:

$$p(\text{Pioggia}/\text{Nessuna nuvola})=0.$$

Diversamente, non sapendo niente sulla presenza di nuvole in cielo, la probabilità che piova dipenderà unicamente dalla frequenza di giornate piovose determinata nell'arco dell'anno o della stagione (la probabilità di osservare un evento è pari alla frequenza con cui quell'evento si manifesta in una serie molto lunga, teoricamente infinita di prove). Tale probabilità a priori è ovviamente meno attendibile della probabilità a posteriori basata sulla presenza o meno di nuvole in cielo.

Da notare che nel caso del mazzo di carte, del tiro di dadi, ecc. è facile calcolare le probabilità indipendenti e congiunte in quanto conosciamo perfettamente la popolazione. In altre situazioni le probabilità sono da determinare empiricamente in base alle frequenze dei vari fenomeni in quanto non possono essere definite su base teorica.

A questo punto, entriamo nella sfera delle applicazioni biomediche, riprendendo quanto detto in precedenza. Poniamo al posto degli eventi A e B la condizione di essere

- affetto o non affetto da una certa malattia, rispettivamente M+ o M-
- positivo o negativo ad un certo test o sintomo clinico specifico per quella malattia, rispettivamente T+ o T-

Poiché malattia e test non sono indipendenti, la probabilità di essere malato dopo aver constatato la positività al test sarebbe:

$$p(M+ \text{ e } T+) = p(M+) \times p(T+/M+)$$

Tuttavia lo stesso test risulta positivo anche su una certa frazione di soggetti sani, per cui possiamo scrivere:

$$p(M- \text{ e } T+) = p(M-) \times p(T+/M-)$$

Ricordiamo:  $p(T+/M+)$  e  $p(T+/M-)$  non sono uguali né complementari. Appartengono a due diverse distribuzioni! E tuttavia la condizione di positività al test compare in entrambe le espressioni, per cui dobbiamo tener conto di tutte e due le espressioni. La domanda

### quale è la probabilità che un soggetto positivo al test sia malato ?

è più che legittima, anche se non possiamo utilizzare probabilità indicate da tabelle o funzioni di distribuzioni. Per rispondere dobbiamo ponderare le due formule che si riferiscono alla probabilità di essere malato o sano e insieme positivo al test (che si riferiscono a distribuzioni diverse).

Questo modo di procedere non valuta probabilità derivanti da un'unica distribuzione, ma **mescola** probabilità derivanti da diverse distribuzioni. Tuttavia la probabilità che si ricerca ha un significato reale e legittimo, in quanto deriva da un quesito o un'ipotesi valida, anche se la risposta non potrà essere ricavata dall'analisi di una distribuzione o popolazione omogenea. Per non equivocare oltre sul termine di probabilità, la statistica distingue i termini di:

- **Verosimiglianza** quando ci si riferisce a probabilità eterogenee attinte da diverse distribuzioni, valutabili attraverso metodi di ponderazione.
- **Probabilità** vera e propria quando ci si riferisce a probabilità omogenee tratte da un'unica distribuzione, valutabili in base alle frequenze rilevate in una serie infinita o estremamente lunga di prove.

In inglese le cose vanno meglio in quanto verosimile si dice *likely* che vuol dire anche probabile; e verosimiglianza si dice **likelihood**, che vuol dire anche probabilità. In italiano verosimile e verosimiglianza sono usati come sinonimi di credibile e credibilità, più che di probabile e probabilità.

Il concetto della verosimiglianza è generalmente accettato dalla statistica formale, anche se esistono posizioni più o meno favorevoli all'allargamento del suo uso a situazioni soggettive.

Detto in altre parole, il fatto di non poterci basare su una sola distribuzione dipende dalla impossibilità di sottoporre un gran numero di soggetti ad un test e poi vedere quanti di questi sono realmente malati. Occorrerebbe, per ogni test e per ogni malattia, uno screening su diverse migliaia di soggetti.

La soluzione al problema ci viene data dal teorema di Bayes. La verosimiglianza (=probabilità ottenuta soppesando elementi provenienti da distribuzioni differenti) viene ponderata ponendo al numeratore la probabilità di essere malato e poi positivo al test essendo malato, ed al denominatore la sommatoria di tutte le probabilità di malattia e non malattia assieme alla positività per il test per le rispettive condizioni:

$$p(M+/T+) = \frac{p(M+) \cdot p(T+/M+)}{[p(M+) \cdot p(T+/M+)] + [p(M-) \cdot p(T+/M-)]}$$

$$= \frac{\text{frequenza delle persone malate che sono al tempo stesso positive al test}}{\text{somma delle frequenze di tutte le persone malate e non malate che sono al tempo stesso positive al test}}$$

in cui

p(M+)	Frequenza dei malati o prevalenza della malattia (beninteso della malattia M).
p(M-)	Frequenza dei sani (non malati di quella malattia M).
p(T+/M+)	Frequenza di test positivi su malati = test veri positivi. E' detto anche <i>sensibilità</i> o <i>potenza</i> del test.
p(T+/M-)	Frequenza di test positivi su sani = test falsi positivi. E' l'errore di I tipo. Il suo complementare (veri negativi) è detto anche <i>specificità</i> del test.

Per capire meglio la situazione, concediamoci un esempio meno serio. Dalla cucina di una casa, in cui sono tenuti 8 cani e 2 gatti è sparita una bistecca. La bistecca era cruda, pronta da cucinare, posta al centro di un tavolo. Mi pongo la domanda se a rubare la bistecca sia stato un cane o un gatto. Sia i cani che i gatti erano liberi di circolare in casa. Non avendo altri elementi oggettivi a disposizione (tracce del furto, comportamenti strani, ecc.) ma avendo invece tempo a disposizione, posso ricorrere al teorema di Bayes. Per prima cosa devo stabilire, con una serie di prove, quale sia la frequenza con cui un cane riesce a rubare la bistecca dal centro del tavolo. Esprimerò questa come:

p(B/C): probabilità che, essendo l'animale un cane, prenda la bistecca.

Su cento prove, i cani rubano la bistecca 10 volte, quindi  $p(B/C) = 0.10$ .

Ripetendo le prove con dei gatti otterrò:

p(B/G): probabilità che, essendo l'animale un gatto, prenda la bistecca.

Su cento prove, i gatti rubano la bistecca 25 volte, quindi  $p(B/G) = 0.25$ .

Inoltre conosco la frequenza dei cani (numero di cani su tutti gli animali sospettabili del furto):

$P(C) = 8/10 = 0.8$

e la frequenza dei gatti (numero di gatti su tutti gli animali sospettabili del furto):

$p(G) = 2/10 = 0.2$ .

Ripetendo lo schema fatto sopra per l'esempio della malattia:

p(C)	Frequenza dei cani
p(G)	Frequenza dei gatti
p(B/C)	Frequenza di cani che rubano la bistecca (cani ladri positivi)
p(B/G)	Frequenza di gatti che rubano la bistecca (gatti ladri positivi)

A questo punto potrò applicare la formula di Bayes per conoscere la probabilità che la bistecca sia stata presa da un gatto:

$$p(G/B) = \frac{p(G) \times p(B/G)}{[p(G) \times p(B/G)] + [p(C) \times p(B/C)]} = \frac{0.2 \times 0.25}{[0.2 \times 0.25] + [0.8 \times 0.10]} = \frac{0.05}{0.05 + 0.08} = 0.385$$

e la probabilità che la bistecca sia stata presa da un cane:

$$p(C/B) = \frac{p(C) \times p(B/C)}{[p(C) \times p(B/C)] + [p(G) \times p(B/G)]} = \frac{0.8 \times 0.10}{[0.8 \times 0.10] + [0.2 \times 0.25]} = \frac{0.08}{0.08 + 0.05} = 0.615$$

Noterete che i denominatori sono uguali e che i due risultati sono ovviamente complementari:  $0.385 + 0.615 = 1$ . Inoltre noterete che nel denominatore abbiamo riunito tutte le popolazioni sospettabili, anche se si tratta di popolazioni diverse.

Ma la cosa più importante è che sui cani, che singolarmente sono meno ladri dei gatti, cadono i maggiori sospetti. Questo è dovuto al maggior numero di cani rispetto ai gatti. La formula di Bayes mette insieme tutti i pezzi del problema per darci la migliore risposta con i dati a disposizione.

Come promemoria, ricordiamo che le probabilità condizionate riguardano la stessa distribuzione quando hanno lo stesso denominatore. Pertanto:

- $p(T+/M+)$  e  $p(T+/M-)$   
veri positivi e falsi positivi, **non appartengono** alla stessa distribuzione,  
  
mentre invece
- $p(T+/M+)$  e  $p(T-/M+)$   
veri positivi e falsi negativi, **appartengono** alla stessa distribuzione, così come
- $p(T+/M-)$  e  $p(T-/M-)$   
falsi positivi e veri negativi, **appartengono** alla stessa distribuzione.

		Risultato del test		
		Positivo T+	Negativo T-	
Condizione	Sano M-	falso-positivo $T+/M-$	vero-negativo $T-/M-$ <i>specificità</i>	<b>stessa distribuzione</b>  $\Leftrightarrow$ complementari probabilità totale = 1
	Malato M+	vero-positivo $T+/M+$ <i>sensibilità</i>	falso-negativo $T-/M+$	<b>stessa distribuzione</b>  $\Leftrightarrow$ complementari probabilità totale = 1
		$\Downarrow$ non complementari: elementi di <b>diverse distribuzioni</b> verosimiglianza totale = 1	$\Downarrow$ non complementari: elementi di <b>diverse distribuzioni</b> verosimiglianza totale = 1	

Nel caso in cui l'espressione del sintomo fosse diversa per  $n$  diversi gradi o tipi della malattia, non più solo per le sole due condizioni di malato e sano, potremmo estendere la formula a tutte le diverse condizioni:

$$p(M_i / T+) = \frac{p(M_i) \cdot p(T+ / M_i)}{\sum_{i=0}^n [p(M_i) \cdot p(T+ / M_i)]}$$

in cui

$p(M_i)$	Frequenza della malattia di grado/tipo $i^{mo}$ . $M_0$ è la condizione di malattia di grado zero = salute.
$p(T+/M_i)$	Frequenza di test positivi su soggetto con malattia di grado/tipo $i^{mo}$ . Ovviamente $p(T+/M_0)$ è la frequenza di test positivi su soggetti sani.

Supponiamo ad esempio di individuare tre gradi di una certa malattia, oltre alla condizione di salute. Dovremo in tal caso conoscere i valori di:

$p(M_0)$  frequenza dei sani (per quella malattia  $M$ , ma potrebbero avere altri mali)

$p(M_1)$  frequenza di soggetti con il grado 1 della malattia  $M$

$p(M_2)$  frequenza di soggetti con il grado 2 della malattia  $M$

$p(M_3)$  frequenza di soggetti con il grado 3 della malattia  $M$

$p(T+/M_0)$  frequenza del test positivo per i sani

$p(T+/M_1)$  frequenza del test positivo per i soggetti con il grado 1 della malattia  $M$

$p(T+/M_2)$  frequenza del test positivo per i soggetti con il grado 2 della malattia  $M$

$p(T+/M_3)$  frequenza del test positivo per i soggetti con il grado 3 della malattia  $M$

Applicheremo quindi il test 4 volte per conoscere le varie probabilità che un soggetto positivo al test sia sano  $M_0$ , o malato  $M_1$ , o malato  $M_2$ , o malato  $M_3$ .