

Precisazioni su formula di Bayes e test diagnostico

1) Errata Corrigere (con ulteriori spiegazioni)

Attenzione! Le due slide relative all'esempio "Con che probabilità è infarto" contengono un errore: la formula di Bayes si applica quando le cause che possono spiegare l'evento osservato sono "esaustive", ossia eventi disgiunti che uniti ricostruiscono tutto l'insieme Ω – in termini più intuitivi, una e una sola delle due cause deve agire – e quindi la somma delle loro probabilità a priori deve essere pari a 1.

Ecco di seguito le slide corrette:

Correzione: le cause devono avere prob a priori con somma pari a 1

Importanza delle probabilità a priori

CON CHE PROBABILITA' E' INFARTO??

Un paziente si presenta dal medico per un dolore al braccio, temendo di avere un infarto in corso.

Il medico fa il seguente ragionamento:

- Se c'è un infarto, la probabilità di avere questo tipo di dolore è del 80%;
- D'altra parte, un'inflammation provocherebbe questo dolore nel 30% dei casi;

Il medico prosegue il ragionamento:

- Quest'uomo è giovane, magro, fa attività fisica ... la prob. di infarto in questi casi è bassa, 5%
- Invece, con lo sport che pratica, la prob. di infiammazione è 95%

E = dolore

C1 = infarto

C2 = infiammazione

$$p(E|C1) = 0.8$$

$$p(E|C2) = 0.3$$

VEROSIMIGLIANZE
delle ipotesi C1 e C2
dato E

$$p(C1) = 0.05$$

$$p(C2) = 0.95$$

Prob. a priori delle
ipotesi C1 e C2

Correzione: le cause devono avere prob a priori con somma pari a 1

La formula di Bayes

CON CHE PROBABILITA' E' INFARTO??

Per fare una diagnosi, il medico deve valutare tutti questi elementi, e valutare la probabilità che stia agendo la causa infarto avendo l'evidenza di un suo sintomo.

Quesito: $P(C1|E)$. Formula di Bayes:

$$\begin{aligned} &= \frac{p(E|C1) \cdot p(C1)}{p(E|C1) \cdot p(C1) + p(E|C2) \cdot p(C2)} \\ &= \frac{0.8 \cdot 0.05}{0.8 \cdot 0.05 + 0.3 \cdot 0.95} = 0.12 \end{aligned}$$

$p(C1|E)$: Prob. a posteriori
dell'ipotesi C1

E = dolore

C1 = infarto

C2 = infiammazione

$$p(E|C1) = 0.8$$

$$p(E|C2) = 0.3$$

VEROSIMIGLIANZE
delle ipotesi C1 e C2
dato E

$$p(C1) = 0.05$$

$$p(C2) = 0.95$$

Prob. a priori delle
ipotesi C1 e C2

Osserviamo però che il confronto delle probabilità a posteriori fra due cause mediante il rapporto, presentato nella slide successiva, vale anche se stiamo confrontando due cause non esaustive.

Infatti si può intendere che oltre a C1 e C2 possa agire una terza causa C3 (o molte altre cause), con $pr(C1)+pr(C2)+pr(C3)=1$; la modifica da fare riguarderebbe solo il denominatore della formula di Bayes, ma esso è uguale per $pr(C1|E)$ e $pr(C2|E)$ e si elide, in altri termini il rapporto dipende solo dai numeratori, che non cambiano.

Per cui è sempre valido che osservando l'evento E (i dati osservati su un campione, ad esempio), C1 è più probabile di C2 (come "urna" che genera quei dati) sia se la probabilità con cui li genera è più alta (maggiore verosimiglianza, $RR>1$), sia se a priori C1 è più probabile di C2.

$$\frac{p(C1|E)}{p(C2|E)} = \frac{p(E|C1)}{\underbrace{p(E|C2)}_{RR}} \cdot \frac{p(C1)}{p(C2)}$$

2) Esempio di stima del valore predittivo del test diagnostico sulla base dei dati osservati: il ruolo e la stima della Prevalenza della malattia

Consideriamo il seguente esercizio: un laboratorio per la diagnosi della celiachia ha sviluppato un nuovo metodo diagnostico basato su un'analisi del sangue. Per stimarne sensibilità e specificità, ricorre ad una sperimentazione controllata.

Seleziona 20 soggetti a cui la presenza di celiachia è stata confermata dalla biopsia ("malati"), e 60 soggetti non celiaci ("sani") che si sono rivolti al laboratorio per sintomi che potevano far sospettare una celiachia. Sottopone tutti i soggetti al nuovo test, i risultati sono riportati in tabella:

	Test +	Test -	Tot
Malati	17	3	20
Sani	5	55	60
tot	22	58	80

Stime:

sensibilità = $17 / 20 = 0.85$

specificità = $55 / 60 = 0.92$

Per il valore predittivo del test positivo, occorre conoscere la PREVALENZA nella popolazione, e applicare la FORMULA di BAYES.

Supponendo di sapere che la prevalenza sia pari all'8%, si ha:

$$p(M | T+) = \frac{sens \cdot prev}{sens \cdot prev + (1 - spec) \cdot (1 - prev)} = \frac{0.85 \cdot 0.08}{0.85 \cdot 0.08 + (1 - 0.92) \cdot (1 - 0.08)} = 0.47$$

Ossia un soggetto sottoposto a questo nuovo test e risultato positivo ha il 47% di probabilità di essere celiaco (ossia che la celiachia sia confermata dalla biopsia).

SE INVECE commettiamo l'ERRORE di non usare la prevalenza nota per quella popolazione, ma di stimarla dai dati, possiamo usare la formula che la calcola direttamente come [Attenzione!! Anche per essa c'era un piccolo errore sulle slide, vd. sotto]:

$$p(M | T+) = (\text{veri positivi}) / \text{positivi} = vp / (vp + fp) = 17 / 22 = 0.77$$

Perchè è un errore: perchè questo esperimento è stato ‘costruito’ dagli investigatori scegliendo loro il numero di “malati” e di “sani” da includere, pertanto la proporzione dei malati sul totale non rispecchia quella “vera”, esistente in natura (a meno di un caso fortuito, o se è stato fatto apposta, cosa che qui non abbiamo motivo di pensare; spesso negli studi si fissa un numero di ‘casi’ – ‘malati’ e per confronto si fissa un numero multiplo di ‘controlli’ – ‘non malati’, in questo caso in rapporto 1:3; altre volte il rapporto è 1:1 o 1:2)

Qui la prevalenza stimata (erroneamente) dalla tabella, pari (per costruzione!) a $20/80=25\%$, è superiore a quella vera (8%); ne consegue che anche la probabilità a posteriori di essere celiaci sapendo di avere test positivo è molto maggiore del valore predittivo vero (77% anzichè 47%).

Vediamo **un altro esempio in cui si può assumere che la prevalenza stimata dal campione sia quella della popolazione:**

Un grande studio prospettico di coorte (termini che si conosceranno con il corso di epidemiologia!) ha osservato per 2 anni una piccola comunità di 100 individui esposti ad una sostanza tossica che può causare in breve tempo uno specifico tipo di tumore del sangue. Alla fine dello studio, si hanno 22 soggetti che presentano quel tipo di tumore. Ogni soggetto era stato sottoposto all’inizio dello studio a un test di screening, che li classificava in ‘positivi’ o ‘negativi’, dove il risultato ‘positivo’ indicava un rischio particolarmente accresciuto.

I dati sono:

	Test +	Test -	Tot
Malati	13	9	22
Sani	25	53	78
tot	38	62	100

Stime:

sensitività = $13 / 22 = 0.59$

specificità = $53 / 78 = 0.68$

prevalenza stimata: $22/100 = 22\%$

Vista la natura dello studio, il dato sulla prevalenza alla fine dei 2 anni rispecchia quello ‘in natura’, per questa piccola popolazione. Dunque possiamo usare equivalentemente la FORMULA di BAYES, o quella molto più semplice: verifichiamo che danno lo stesso risultato:

$$p(M | T+) = \frac{\text{sens} \cdot \text{prev}}{\text{sens} \cdot \text{prev} + (1 - \text{spec}) \cdot (1 - \text{prev})} = \frac{0.59 \cdot 0.22}{0.59 \cdot 0.22 + (1 - 0.68) \cdot (1 - 0.22)} = 0.34$$

e anche = (veri positivi) / positivi = $13 / 38 = 0.34$

Questo studio indica che in una popolazione esposta alla sostanza tossica (e simile a quella studiata) il rischio di avere quel tumore del sangue è del 22%, e che il test di screening (che ha sensibilità e specificità medie, rispettivamente uguali a 59% e 68%) permette di migliorare moderatamente l’individuazione dei soggetti più a rischio: infatti il valore predittivo del test positivo è pari a 34%, ossia un soggetto che risulti positivo ha il 34% di probabilità di avere il tumore in oggetto. Senza fare il test, la probabilità sarebbe come per tutti del 22%.

Calcoliamo anche la probabilità a posteriori di tumore se il test risulta negativo (usiamo la formula generale, che può essere applicata a qualsiasi esempio):

$$p(M | T-) = \frac{(1 - \text{sens}) \cdot \text{prev}}{(1 - \text{sens}) \cdot \text{prev} + \text{spec} \cdot (1 - \text{prev})} = \frac{(1 - 0.59) \cdot 0.22}{(1 - 0.59) \cdot 0.22 + 0.68 \cdot (1 - 0.22)} = 0.14$$

Un soggetto di quella popolazione che abbia test negativo ha una probabilità di tumore pari a 14%.

Vediamo la slide con segnalata la correzione:

Stima di Sensitività e Specificità

Si vuole stimare la sensibilità e la specificità di una nuova tecnica diagnostica per immagini, alternativa ad una con risultato certo, ma meno invasiva / costosa. Si prende quindi un campione di n soggetti che, sottoposti alla "vecchia" tecnica, vengono classificati in "malati" e "non malati"; li si sottopongono poi alla "nuova" diagnostica, ottenendo i seguenti risultati:

	Test +	Test -	
Malati	vp	fn	m
Non Malati	fp	vn	n-m
	vp+fp	fn+vn	n

n soggetti di cui m malati, gli altri $(n-m)$ non malati

Dei malati, vp hanno Test+ e fn hanno Test- (vp sono i "veri positivi", fn sono i "falsi negativi")

etc

$$\text{sensitività} \approx \frac{vp}{m} = \frac{vp}{vp + fn}$$

$$\text{specificità} \approx \frac{vn}{n-m} = \frac{vn}{vn + fp}$$

Attenzione! Il valore predittivo si può calcolare solo se conosciamo la **prevalenza** della malattia. **Solo se possiamo pensare di stimarla dal campione, ($=m/n$)** allora si ha:

$$p(M | T+) \approx \frac{vp}{vp + fp} \quad \text{red arrow}$$

(stesso risultato con la formula di Bayes)