

LA STATISTICA BAYESIANA IN MEDICINA. PARTE I: GLI STRUMENTI DI BASE

M. Nichelatti, C. Montomoli

Dipartimento di Scienze Sanitarie Applicate e Psicocomportamentali, Sezione di Statistica Medica ed Epidemiologia, Università degli Studi, Pavia

Bayesian statistics in medicine. Part I: the basic tools

Bayesian statistics, so called after the British scientist Thomas Bayes who first used it in the 18th century, is a widely applied analytical tool in many branches of the applied sciences but is seldom mentioned in the medical literature. This paper aims to present a simple overview of the Bayesian method by introducing its basic mathematical tools (in particular the conditional probability) and then applying them to evaluate the characteristics of diagnostic tests. (G Ital Nefrol 2008; 25: 342-6)

Conflict of interest: None

KEY WORDS:

Conditional probability, Bayesian statistics, Diagnostic tests

PAROLE CHIAVE:

Probabilità condizionale, Statistica bayesiana, Test diagnostici

✉ Indirizzo degli Autori:

Prof.ssa Cristina Montomoli
Dipartimento di Scienze Sanitarie Applicate e Psicocomportamentali
Università di Pavia
Via Agostino Bassi, 21
27100 Pavia
e-mail: cristina.montomoli@unipv.it

INTRODUZIONE

La statistica che più spesso viene utilizzata nelle pubblicazioni sulle riviste mediche è una statistica che viene definita *frequentista*. La definizione si basa su un concetto abbastanza semplice: la frequenza con cui si verifica un evento viene considerata equivalente alla probabilità che lo stesso evento si verifichi; per essere chiari, se è noto che in 3 partorienti su 100 può manifestarsi nefropatia gravidica, allora la probabilità che una partoriente presa a caso tra la popolazione soffra effettivamente di tale disordine è pari al 3%.

D'altra parte, quella frequentista non è la sola filosofia di approccio statistico; tra le filosofie alternative spicca quella detta bayesiana dal nome dello scienziato Thomas Bayes (1701-1761) che l'ha formulata per la prima volta in un articolo pubblicato postumo nel 1763 (Bayes T. *Essay towards solving a problem in the doctrine of chances*. Phil Trans Royal Soc 1763; 53: 370-418, ristampato in: *Biometrika* 1958; 45: 293-315). La filosofia bayesiana, pur essendo abbastanza intuitiva, può risultare inizialmente ostica, soprattutto perché si basa su presupposti come la probabilità condizionale, che è meno familiare di altre

definizioni di probabilità. L'uso della probabilità condizionale è indirizzato principalmente all'analisi dei risultati dei test diagnostici, consentendo di confrontare vari tipi di test, e di calcolare la probabilità di un evento in funzione dei livelli di esposizione. Inoltre, la probabilità condizionale è capace di contenere in sé l'informazione sulla sequenza temporale di due eventi: vedremo in seguito come l'informazione sulla probabilità di un evento B , dato il verificarsi di un evento A , precedente B , consentirà di calcolare, grazie ad una particolare funzione chiamata verosimiglianza, anche la probabilità "retrograda" che si sia verificato A , una volta noto il verificarsi di B .

LA PROBABILITÀ CONDIZIONALE

La statistica bayesiana - come quella frequentista - utilizza il linguaggio della probabilità, arricchendolo con nuove notazioni, che mostrano una utilità pratica immediata; lo strumento fondamentale per capire la statistica bayesiana è la definizione di probabilità condizionale. Sappiamo che la prevalenza della nefropatia diabetica in tutta la popolazione è di circa 16 casi ogni 1000

abitanti, mentre la prevalenza del diabete mellito è di 4 casi ogni 100 abitanti, quindi pari ad un caso ogni 25 abitanti. Se combiniamo le due informazioni possiamo ricavare la prevalenza della nefropatia diabetica nei soli pazienti diabetici, che è un dato che può rivelarsi molto più interessante ai fini epidemiologici.

Chiamando $Pr(D)$ la probabilità che un soggetto sia diabetico (quindi la prevalenza del diabete), e $Pr(N)$ la probabilità che un soggetto abbia la nefropatia diabetica (quindi la prevalenza della nefropatia), avremo $Pr(D) = 0.04$ e $Pr(N) = 0.016$, da cui vogliamo ricavare la prevalenza della nefropatia diabetica nei soli diabetici, cioè la probabilità che un soggetto abbia la nefropatia diabetica, a condizione che lo stesso soggetto sia anche diabetico: questa probabilità viene definita con la notazione $Pr(N|D)$ (l'espressione si legge "probabilità di N, dato D"), dove la barra verticale è il simbolo di probabilità condizionale. Per definizione, la probabilità condizionale è data dal rapporto:

$$Pr(N|D) = \frac{Pr(N \cap D)}{Pr(D)}$$

in cui il termine $Pr(N \cap D)$ identifica la probabilità di essere simultaneamente sia nefropatici, sia diabetici. Il calcolo della probabilità condizionale in questo caso è abbastanza semplice: assumiamo che la popolazione totale sia di 500000 persone; tra essi, date le prevalenze note delle due patologie, ci saranno 20000 soggetti diabetici e 8000 soggetti con nefropatia diabetica, per cui si potrà scrivere:

$$Pr(N|D) = \frac{\text{nefropatici che sono anche diabetici}}{\text{diabetici}} = \frac{8000}{20000} = 0.4 = 40\%$$

in altre parole, il 40% circa dei pazienti diabetici sarà nefropatico, il che equivale a dire che per ottenere la probabilità condizionale abbiamo calcolato la prevalenza della nefropatia solo nei soggetti già diabetici. Quindi, la probabilità condizionale contiene anche un riferimento alla sequenza temporale degli eventi studiati: un soggetto diventa *prima* diabetico, e *poi* verrà eventualmente colpito *anche* dalla nefropatia diabetica.

Per inciso, lo stesso risultato l'avremmo potuto ottenere utilizzando direttamente le due prevalenze, infatti, visto che:

$$\frac{\text{nefropatici}}{\text{diabetici}} = \frac{\text{nefropatici}}{\text{popolazione}} \times \frac{\text{popolazione}}{\text{diabetici}} = \frac{\text{nefropatici}/\text{popolazione}}{\text{diabetici}/\text{popolazione}}$$

allora:

$$Pr(N|D) = \frac{Pr(N)}{Pr(D)} = \frac{0.016}{0.04} = 0.4 = 40\%$$

La probabilità condizionale è quindi la probabilità calcolata non rispetto alla popolazione totale, ma rispetto alla popolazione che obbedisce alla condizione imposta (Fig. 1). I soggetti diabetici sono ovviamente meno numerosi della popolazione totale, quindi la probabilità di nefropatia diabetica è maggiore nei pazienti diabetici che in tutta la popolazione.

Per approfondire il concetto di probabilità condizionale, consideriamo una famiglia con due figli: la probabilità che siano entrambi maschi la possiamo esprimere con la notazione $Pr(M_2, M_1)$, dove il numero in pedice identifica il figlio. D'altra parte, avremmo potuto esprimere questa situazione anche come probabilità condizionale nella forma $Pr(M_2 | M_1)$, per specificare che si sta parlando della probabilità che il secondo figlio sia maschio, una volta accertato che sia un maschio anche il primo. Anche se le due notazioni sembrano descrivere l'identica cosa, il valore delle due probabilità è profondamente differente, e quindi: $Pr(M_1, M_2) \neq Pr(M_2 | M_1)$.

La diversità dei risultati è dovuta al fatto che nel primo caso (due figli maschi, senza ulteriori vincoli) il primo evento (primo figlio maschio) non ha nessun effetto sul verificarsi del secondo evento (secondo figlio maschio), e viceversa: in altre parole, i due eventi sono tra loro indipendenti, per cui la probabilità che si verifichino simultaneamente è data dal prodotto delle singole probabilità, ovvero:

$$Pr(M_1, M_2) = Pr(M_1) Pr(M_2) = 0.5 \times 0.5 = 0.25 = 25\%$$

nel secondo caso, invece, si sta parlando della probabilità della nascita di un secondo figlio maschio sapendo che il primo (già nato) è un maschio, quindi il primo evento condiziona - per così dire - il secondo. La possibile sequenza del sesso di due figli è M_1, M_2 , M_1, F_2 , F_1, M_2 e F_1, F_2 : vi sono quindi quattro alternative, e nel caso precedente ne interessava una sola (M_1, M_2), per cui la probabilità dell'evento era, appunto $Pr(M_1, M_2) = 1/4 = 0.25 = 25\%$. Nel secondo caso, invece, impone che il primo figlio sia maschio rende possibili solo due alternative M_1, M_2 e M_1, F_2 , per cui la probabilità dell'evento di interesse ($M_2 | M_1$) diventa:

$$Pr(M_2 | M_1) = \frac{Pr(M_1 \cap M_2)}{Pr(M_1)} = \frac{\frac{1}{4}}{\frac{1}{2}} = 0.5 = 50\%$$

La probabilità condizionale è uno strumento molto utile quando ci si occupa di esposizione ad un determinato rischio: i valori della probabilità di tumore ai polmoni in un fumatore $Pr(K | F^+)$ e in un non fumatore $Pr(K | F^-)$ daranno informazioni più utili di quelle otte-

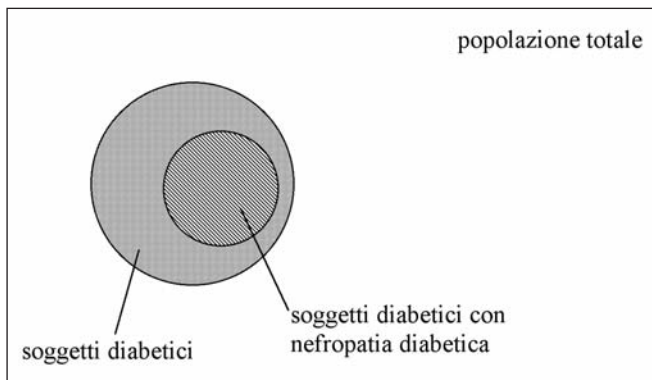


Fig. 1 - Diagramma che mostra come la popolazione dei soggetti affetti da nefropatia diabetica sia un sottoinsieme (una sottopopolazione) dei diabetici, mentre i soggetti diabetici sono a loro volta un sottoinsieme dell'intera popolazione.

nibili dalla conoscenza delle generica probabilità di tumore $Pr(K)$ in tutta la popolazione.

LA PROBABILITÀ CONDIZIONALE APPLICATA AI TEST DIAGNOSTICI

La probabilità condizionale è indispensabile nell'analisi dell'efficienza dei test diagnostici. È noto che ogni test diagnostico, oltre a riconoscere i veri sani (negativi al test) ed i veri malati (positivi al test) produce anche degli errori, ovvero dei falsi positivi (soggetti sani che il test positivo riconosce come malati) e dei falsi negativi (soggetti effettivamente malati che il test negativo riconosce erroneamente come sani). Possiamo allora definire la *sensibilità* $Pr(T^+ | M^+)$ di un test come la probabilità che un soggetto risulti positivo al test quando tale soggetto sia effettivamente malato, e la *specificità* $Pr(T^- | M^-)$ come la probabilità che un soggetto risulti negativo quando sia effettivamente sano, per cui avremo:

$$Pr(T^+ | M^+) = \frac{\text{veri positivi}}{\text{veri positivi} + \text{falsi negativi}} = \frac{VP}{VP + FN}$$

$$Pr(T^- | M^-) = \frac{\text{veri negativi}}{\text{veri negativi} + \text{falsi positivi}} = \frac{VN}{VN + FP}$$

In questo modo, per ogni test si potrà ottenere una Tabella a due vie che riassume la situazione nel modo seguente

	Positivi al test	Negativi al test	Totale
Malati	VP	FN	VP + FN
Sani	FP	VN	VN + FP
Totale	VP + FP	VN + FN	VP + FP + VN + FN

in cui la somma $VP + FP + VN + FN$ rappresenta l'intera la popolazione, mentre la prevalenza della malattia per la cui diagnosi viene effettuato il test è ovviamente data dal rapporto:

$$\text{prevalenza} = \frac{\text{malati}}{\text{popolazione}} = \frac{VP + FN}{VP + FN + VN + FP}$$

Per un test clinico è importante conoscere anche il *valore predittivo positivo* $Pr(M^+ | T^+)$, ovvero la probabilità che un soggetto sia effettivamente malato essendo risultato positivo al test (cioè la percentuale dei positivi malati rispetto a tutti i positivi al test), ed il *valore predittivo negativo* $Pr(M^- | T^-)$, ovvero la probabilità che un soggetto sia sano essendo risultato negativo al test (cioè la percentuale dei negativi sani rispetto a tutti i negativi al test), quindi:

$$Pr(M^+ | T^+) = \frac{\text{veri positivi}}{\text{veri positivi} + \text{falsi positivi}} = \frac{VP}{VP + FP}$$

$$Pr(M^- | T^-) = \frac{\text{veri negativi}}{\text{veri negativi} + \text{falsi negativi}} = \frac{VN}{VN + FN}$$

Un'ulteriore misura di validità che viene usata per i test diagnostici è l'*accuratezza*, data dal rapporto:

$$\frac{\text{diagnosi esatte}}{\text{diagnosi totali}} = \frac{VP + VN}{VP + FN + VN + FP}$$

che in pratica fornisce il numero di volte che il test ha dato un risultato esatto sul totale dei test effettuati.

I valori delle varie probabilità condizionali sono variamente affidabili: in generale, sensibilità e specificità sono i fattori più utilizzati per valutare un test, ma sono altrettanto importanti anche i valori predittivi positivo e negativo; mentre i primi due dicono quale sia la probabilità per un malato e per un soggetto sano di risultare rispettivamente positivo o negativo al test (potremmo chiamarla *probabilità forward*, in quanto l'essere malato o sano precede l'esecuzione del test), gli altri due dicono quale sia la probabilità che un soggetto sia malato o sano data la rispettiva positività o negatività al test (in questo caso potremmo parlare di una *probabilità backward*, dato che il fatto di essere realmente malati o sani viene valutato dopo la positività o la negatività al test).

Per fare un esempio numerico, ipotizziamo la situazione descritta nella Tabella seguente:

	Positivi al test	Negativi al test	Totale
Malati	230	20	250
Sani	750	5000	5750
Totale	980	5020	6000

otteniamo i valori:

$$\text{sensibilità} = \frac{VP}{VP + FN} = \frac{230}{250} = 92.0\% ;$$

$$\text{specificità} = \frac{VN}{VN + FP} = \frac{5000}{5750} = 86.9\% ;$$

$$\text{prevalenza} = \frac{VP + FN}{VP + FN + VN + FP} = \frac{250}{6000} = 4.2\%$$

$$\text{valore predittivo positivo} = \frac{VP}{VP + FP} = \frac{230}{980} = 23.5\%$$

$$\text{valore predittivo negativo} = \frac{VN}{VN + FN} = \frac{5000}{5020} = 99.6\%$$

$$\text{accuratezza} = \frac{VP + VN}{VP + FN + VN + FP} = \frac{5230}{6000} = 87.2\%$$

TEST DI VERIFICA

1) La probabilità di due eventi indipendenti:

- È pari alla somma delle singole probabilità
- È pari al prodotto delle singole probabilità
- È pari al rapporto delle singole probabilità
- È pari al valore della probabilità dell'evento a probabilità minore.

2) La probabilità condizionale:

- È la probabilità che un evento condizioni il verificarsi di un evento successivo
- È la probabilità che un certo evento sia condizionato da un evento precedente
- È la probabilità del verificarsi di un dato evento, condizionato dal verificarsi di un altro evento
- È la probabilità con cui due eventi indipendenti si condizionano reciprocamente.

3) La sensibilità di un test diagnostico è:

- La probabilità che un positivo al test sia malato
- La probabilità che un malato sia positivo al test
- La probabilità che un negativo al test sia malato
- La probabilità che un positivo al test sia sano.

La risposta corretta alle domande sarà disponibile sul sito internet www.sin-italy.org/gjn e in questo numero del giornale cartaceo dopo il Notiziario SIN

QUALE TEST CONVIENE UTILIZZARE?

La sensibilità di un test dice qual è la probabilità che un malato risulti positivo al test, quindi con un test molto sensibile si avrà una piccola frazione di falsi

negativi; la specificità fornisce invece la probabilità che un soggetto sano risulti negativo al test, e quindi un test con specificità elevata avrà una frazione piccola di falsi positivi: in generale, il test diagnostico migliore sarà quello a maggiore sensibilità e specificità. Il problema sorge quando, per la stessa malattia, esistono due differenti test diagnostici, uno con elevata sensibilità e bassa specificità, e l'altro con elevata specificità e bassa sensibilità. La scelta del test ottimale, in questo caso, non dipende solamente dai valori numerici che assumono la sensibilità e la specificità, ma da una serie di valutazioni che devono prendere in considerazione anche il tipo di malattia e la sua prevalenza. Ad esempio, per quanto possa essere accurato, cioè, per quante diagnosi siano azzeccate, un test può essere del tutto inutile quando la prevalenza della malattia è molto bassa: quando questo si verifica, anche il valore predittivo positivo (che dipende dalla prevalenza) sarà molto basso, indipendentemente da sensibilità e specificità.

I test diagnostici con differenti caratteristiche di sensibilità e specificità potrebbero essere anche molto differenti qualitativamente; potrebbero infatti derivare dall'attività di un enzima, oppure da una ispezione ultrasonografica, oppure da un'analisi al microscopio ottico: in tal caso si parla di *test condizionalmente indipendenti*, ovvero di test in cui la diagnosi ottenuta con uno di essi non influenza la diagnosi che si otterrebbe con uno qualsiasi degli altri, perché ciascuno di essi si basa su una specifica caratteristica morfologica, funzionale o biochimica della malattia.

Un problema particolare sorge se si deve diagnosticare una malattia grave, come un tumore, disponendo di due (o più) test con caratteristiche diverse: potrebbe essere preferibile un test ad elevata sensibilità, in modo da avere pochi falsi negativi, ma per certi aspetti (psicologici, ad esempio), si potrebbe scegliere invece un test con elevata specificità, in modo da non fornire troppi falsi positivi. In situazioni del genere, la scelta dovrebbe cadere sulla combinazione di due o più test in sequenza: prescindendo dall'eventuale invasività e *compliance* dei test in oggetto (si dovrebbe sempre partire dal meno invasivo e "scomodo") si potrebbe - ad esempio - iniziare con quello a minore sensibilità: se un soggetto risultasse negativo, il test potrebbe essere ripetuto una seconda volta; se il soggetto risultasse positivo la prima o la seconda volta al test meno sensibile, dovrebbe essere sottoposto al test più sensibile. La migliore combinazione dipende comunque dalla patologia da diagnosticare, dal buon senso del medico, ed anche le caratteristiche individuali del paziente.

TEST DI VERIFICA

4) In un test a bassa specificità:

- a. La percentuale di falsi positivi è bassa
- b. La percentuale di falsi positivi è elevata
- c. La percentuale di falsi negativi è elevata
- d. La percentuale di falsi negativi è superiore a quella dei falsi positivi.

5) In un test diagnostico, la sensibilità e la specificità:

- a. Sono inversamente proporzionali: quando un test ha elevata sensibilità, ha sempre anche una ridotta specificità, e viceversa
- b. Sono direttamente proporzionali: tutti i test con sensibilità elevata hanno anche elevata specificità e viceversa
- c. Possono assumere valori tra loro indipendenti, per cui esistono test molto sensibili e molto specifici, test molto sensibili e poco specifici, test poco sensibili e molto specifici ed anche test poco sensibili e poco specifici
- d. Hanno valori additivi: sommando la sensibilità e la specificità si ottiene sempre una probabilità del 100%.

6) Per la malattia X, letale se non curata in tempo, e con prevalenza pari a 1 caso ogni 100 mila abitanti, è disponibile un test diagnostico estremamente accurato, con sensibilità pari al 99% e specificità del 98%: quindi, se un soggetto risulta positivo al test, qual è la probabilità che sia veramente ammalato?

- a. 99%
- b. Dipende dall'accuratezza del test, che non si può calcolare esattamente con i dati forniti
- c. Inferiore a 1 su 1000
- d. 98%.

RIASSUNTO

La statistica bayesiana, chiamata così dal nome dello scienziato britannico Thomas Bayes, che l'ha usata per primo nel XVII secolo, è una tecnica di analisi ampiamente utilizzata in molti rami delle scienze applicate, ma viene citata ancora raramente nella letteratura medica. Il lavoro si pone l'obiettivo di presentare una trattazione elementare dei metodi statistici bayesiani, introducendone gli strumenti matematici di base (principalmente, la probabilità condizionale), e poi applicandoli alla valutazione delle caratteristiche dei test diagnostici.

DICHIARAZIONE DI CONFLITTO DI INTERESSI

Gli Autori dichiarano di non avere conflitto di interessi.

BIBLIOGRAFIA

1. Albert JH, Rossman AJ. Workshop statistics: discovery with data, a Bayesian approach. Emeryville: Key College Publishing, 2001.
2. Berry DA. Statistics: a Bayesian perspective. Belmont: Duxbury Press, 1996.