INTERVALLI DI CONFIDENZA NON PARAMETRICI PER L'AREA SOTTESA ALLA CURVA ROC

Gianfranco Adimari

1. INTRODUZIONE

In ambito sanitario, gli esami diagnostici vengono comunemente utilizzati con l'obiettivo di riconoscere, relativamente a una qualche patologia, i soggetti malati e quelli sani. Nella situazione più semplice, un esame diagnostico fornisce un risultato che può essere espresso come "positivo" o "negativo", con il risultato positivo che corrisponde a valori elevati (superiori ad un certo valore di soglia t fissato) o bassi (inferiori ad un certo valore di soglia t fissato) di una qualche variabile diagnostica. Naturalmente, l'esame non è infallibile, cosicché alcuni soggetti malati possono avere un esame negativo ("falsi negativi") ed alcuni soggetti sani possono avere un esame positivo ("falsi positivi"). Un test diagnostico è allora tanto più accurato quanto più è piccola la probabilità che esso produca dei "falsi".

Siano X e Y due variabili casuali indipendenti, con distribuzione continua, che descrivono, rispettivamente, la risposta di un test diagnostico su un soggetto sano e su un soggetto malato; si supponga che un risultato "positivo" corrisponda a un valore elevato del test. Il grafico che rappresenta la relazione tra 1 meno la probabilità di ottenere un "falso negativo" (la *sensibilità*) e la probabilità di ottenere un "falso positivo" (1 meno la *specificità*), per ogni possibile valore di soglia t, è la cosiddetta curva ROC (*Receiver Operating Characteristic Curve*) associata al test. Essa costituisce uno strumento importante per valutare la capacità discriminante del test e per mettere a confronto test alternativi. In particolare, l'area sottesa alla curva ROC rappresenta l'indice globale di accuratezza diagnostica più comunemente usato. Si può mostrare (Bamber, 1975) che il valore di tale area coincide con la quantità $\rho_0 = \Pr\{Y > X\}$.

Sia $X_1, X_2, \ldots, X_{n_x}$ un campione casuale semplice (c.c.s.) di dimensione n_x da X e $Y_1, Y_2, \ldots, Y_{n_y}$ un c.c.s. di dimensione n_y da Y. Si consideri il problema della costruzione di intervalli di confidenza per ρ_0, in ambito non parametrico. Tale problema potrebbe essere risolto, in linea di principio, ricorrendo alla tecnica basata sulla funzione di verosimiglianza empirica (si veda Owen, 2001, come riferimento generale) che si è affermata nell'ultimo decennio come valida alternativa al

bootstrap e permette di ottenere, in molte situazioni, regioni di confidenza con buone proprietà teoriche, sufficientemente accurate anche quando le dimensioni campionarie sono medio-piccole.

Sia F_{q_x} la generica distribuzione multinomiale sul campione $X_1,\ldots,X_{n_x}$, che assegna probabilità q_{x_h} all'osservazione X_h. Analogamente, sia F_{q_y} la distribuzione multinomiale sul campione $Y_1,\ldots,Y_{n_y}$, che assegna probabilità q_{y_k} all'osservazione Y_k. Poiché $\rho_0 = \Pr\{Y > X\} = E[\Pr\{Y > X|X\}] = E[S_y(X)]$, dove $S_y(\cdot)$ indica la funzione di sopravvivenza di Y, la funzione di verosimiglianza empirica per ρ_0, calcolata nel punto ρ, è definita come

$$\mathcal{L}(\rho) = \max_{q_{x_1},\ldots,q_{x_{n_x}},\, q_{y_1},\ldots,q_{y_{n_y}}} \prod_{h=1}^{n_x}\prod_{k=1}^{n_y} q_{x_h} q_{y_k} \,, \tag{1}$$

dove il massimo deve essere ottenuto sotto i vincoli

$$\sum_{h=1}^{n_x} q_{x_h} = 1 \;,\; \sum_{k=1}^{n_y} q_{y_k} = 1 \text{ e } \sum_{h=1}^{n_x}\sum_{k=1}^{n_y} I(Y_k > X_h) q_{y_k} q_{x_h} = \rho.$$

Qui $I(E)$ è la funzione indicatrice dell'evento E. La funzione $\mathcal{L}(\cdot)$ raggiunge il suo massimo assoluto in corrispondenza del valore

$$\hat{\rho} = U = \frac{1}{n_x n_y}\sum_{h=1}^{n_x}\sum_{k=1}^{n_y} I(Y_k > X_h),$$

dove U è la ben nota statistica di Mann-Whitney, stimatore non parametrico per ρ_0. Sfortunatamente, a differenza di quanto accade usualmente, non si riesce, in questo caso, ad ottenere un'espressione esplicita per $\mathcal{L}(\rho)$ via moltiplicatori di Lagrange. Questo inconveniente rende, di fatto, inutilizzabile tale approccio, a causa dell'elevata complessità che caratterizza il calcolo della funzione $\mathcal{L}(\cdot)$.

Si può però aggirare l'ostacolo e costruire, in maniera relativamente semplice, una pseudoverosimiglianza per ρ_0, alternativa alla (1), combinando la funzione di verosimiglianza empirica per il funzionale media e gli pseudo-valori jackknife derivati dalla statistica di Mann-Whitney. Tale approccio, ispirato da un'idea di Jing *et al.* (2005), è l'oggetto di discussione di questo lavoro. Esso è descritto nel paragrafo 2, nel quale si mostra, in particolare, che per la suddetta pseudo-verosimiglianza, indicata con $L(\rho)$, vale un risultato asintotico analogo a quello fornito dal teorema di Wilks nel caso parametrico. La funzione $L(\rho)$ si può dunque usare, allo stesso modo in cui si usa la funzione di verosimiglianza parametrica, per ottenere intervalli di confidenza, con livelli di copertura asintoticamente esatti, per l'area sottesa alla curva ROC. Alcuni risultati di uno studio di simulazione, effettuato per valutare l'accuratezza nel finito degli intervalli di confidenza prodotti dal

metodo proposto, sono riportati nel paragrafo 3. Il paragrafo 4 contiene alcune considerazioni conclusive.

2. L'APPROCCIO PROPOSTO

Si considerino gli pseudo-valori $V_i = nU - (n-1)U_{-i}$, $i = 1, 2, ..., n$, dove $n = n_y + n_x$ e U_{-i} indica la statistica di Mann-Whitney ottenuta eliminando dal campione *pooled* la i-esima generica osservazione. In particolare, nel seguito si userà, quando opportuno, anche la notazione U_{-y_k} per indicare la statistica di Mann-Whitney ottenuta eliminando la k-esima osservazione del campione da Y, oppure U_{-x_h} se l'osservazione eliminata è la h-esima del campione da X. Posto

$$A = \sum_{h=1}^{n_x} \sum_{k=1}^{n_y} I(Y_k > X_h) ,$$

si ha $U = A/(n_x n_y)$ e

$$U_{-y_k} = \frac{A - \sum_{h=1}^{n_x} I(Y_k > X_h)}{n_x (n_y - 1)} , \qquad U_{-x_h} = \frac{A - \sum_{k=1}^{n_y} I(Y_k > X_h)}{(n_x - 1) n_y} \tag{2}$$

Quindi, $\sum_{k=1}^{n_y} U_{-y_k} = A/n_x$, $\sum_{h=1}^{n_x} U_{-x_h} = A/n_y$ e $(1/n) \sum_{i=1}^{n} U_{-i} = U$; cosicché

$$\frac{1}{n} \sum_{i=1}^{n} V_i = nU - \frac{n-1}{n} \sum_{i=1}^{n} U_{-i} = U .$$

Risulta, dunque, che U è lo stimatore jackknife per ρ_0, nel senso che risulta $U = (1/n) \sum_{i=1}^{n} V_i$. Inoltre, le variabili V_i hanno media ρ_0, cioè $\mathrm{E}\{V_i\} = \rho_0$, $i = 1, ..., n$. Una funzione di pseudoverosimiglianza per ρ_0 può allora essere ottenuta combinando la funzione di verosimiglianza empirica per il funzionale media (Owen, 1988) e gli pseudo-valori jackknife V_i derivati dalla statistica di Mann-Whitney.

Si pensi a $V_1, V_2, ..., V_n$ come a un c.c.s. da una variabile V con media ρ_0. Sia F_p la distribuzione multinomiale su $V_1, ..., V_n$, che assegna probabilità p_i alla osservazione V_i. La funzione di verosimiglianza empirica per la media di V, valutata nel punto ρ, è definite come

$$L(\rho) = \max_{p_1, ..., p_n} \prod_{i=1}^{n} p_i \text{ , sotto i vincoli } \sum_{i=1}^{n} p_i = 1 \text{ e } \sum_{i=1}^{n} V_i p_i = \rho .$$

Siano $V_{(1)}$ e $V_{(n)}$ il più piccolo ed il più grande tra gli pseudo-valori, rispettiva-

mente. Quando $\rho \in (V_{(1)}, V_{(n)})$, ricorrendo al metodo dei moltiplicatori di Lagrange, si ottiene l'espressione esplicita

$$L(\rho) = \prod_{i=1}^{n} \frac{1}{n\{1+\lambda(V_i - \rho)\}} ,$$

dove $\lambda = \lambda(\rho) \in \Re$ è l'unica soluzione dell'equazione

$$\frac{1}{n}\sum_{i=1}^{n} \frac{V_i - \rho}{1+\lambda(V_i - \rho)} = 0 . \tag{3}$$

Anche la funzione $L(\rho)$ raggiunge il suo massimo assoluto quando $\rho = U$ e, in tale punto, vale n^{-n}. Perciò, la funzione log-rapporto di verosimiglianza empirica per ρ_0 (la media di V) è

$$l(\rho) = \log\{n^n L(\rho)\} = -\sum_{i=1}^{n} \log\{1+\lambda(V_i - \rho)\} , \tag{4}$$

quando $\rho \in (V_{(1)}, V_{(n)})$. Se $\rho \notin (V_{(1)}, V_{(n)})$, si ha $L(\rho) = 0$ e, equivalentemente, $l(\rho) = -\infty$.

In realtà, le variabili $V_1, ..., V_n$ non sono né identicamente distribuite né indipendenti. Ciò non di meno, un risultato analogo a quello usuale (relativo alla situazione di c.c.s.) sulla distribuzione asintotica della statistica di verosimiglianza empirica per la media, può essere ottenuto, nel caso in questione, con argomentazioni *ad hoc*. Tale risultato, fornito dal teorema 1, è analogo a quello sancito dal teorema di Wilks nel caso parametrico e giustifica l'uso del termine pseudo-verosimiglianza per la funzione $L(\cdot)$. Il teorema 1 è preceduto dal lemma 1 e dal lemma 2 che riportano alcuni risultati preliminari. Il lemma 1, in particolare, fornisce il risultato classico sulla distribuzione asintotica della statistica di Mann-Whitney (si veda, per esempio, Serfling, 1980, capitolo 5).

Lemma 1. Siano $\varsigma_x(\rho_0) = \Pr\{Y_1 > X, Y_2 > X\} - \rho_0^2$ e $\varsigma_y(\rho_0) = \Pr\{Y > X_1, Y > X_2\} - \rho_0^2$. Si assuma che sia $\varsigma_x(\rho_0) > 0$ o $\varsigma_y(\rho_0) > 0$ e che $n_y/n_x \to \kappa$, con $0 < \kappa < +\infty$, quando $n \to \infty$. Allora, quando $n \to \infty$,

$$\sqrt{n}(U - \rho_0) \xrightarrow{d} N(0, \omega_0)$$

con $\omega_0 = (\kappa + 1)\{\kappa\varsigma_x(\rho_0) + \varsigma_y(\rho_0)\}/\kappa$.

Lemma 2. Sotto le condizioni del lemma 1, quando $n \to \infty$, si ha:

(*i*) $\max_{i=1,...,n} |V_i| = O(1)$;

(*ii*) $\Pr\{V_{(1)} < \rho_0\} \to 1$ e $\Pr\{V_{(n)} > \rho_0\} \to 1$.

Dimostrazione. (*i*) Si può scrivere $V_i = U_{-i} + n(U - U_{-i})$, cosicché $|V_i| \leq |U_{-i}| + n|U - U_{-i}|$.

Inoltre, in base alla (2), risulta

$$U - U_{-i} = \frac{A}{n_x n_y} - \frac{A - j}{n_x (n_y - 1)} = \frac{n_y j - A}{n_x n_y (n_y - 1)} = \frac{j / n_x - U}{n_y - 1} \ ,$$

oppure

$$U - U_{-i} = \frac{A}{n_x n_y} - \frac{A - j}{(n_x - 1) n_y} = \frac{n_x j - A}{n_x n_y (n_x - 1)} = \frac{j / n_y - U}{n_x - 1} \ ,$$

per qualche opportuno intero j tale che $0 \leq j \leq n_x$ o $0 \leq j \leq n_y$, rispettivamente, a seconda che la statistica U_{-i} sia ottenuta eliminando un'osservazione del campione da Y o del campione da X. Pertanto, tenendo presente che $0 \leq U \leq 1$ (e che lo stesso vale per U_{-i}), si ha

$$|U - U_{-i}| \leq \max\{1/(n_x - 1), 1/(n_y - 1)\} \ ,$$

da cui segue l'asserto. (*ii*) Siano $U_{-i}^{\min}$ e $U_{-i}^{\max}$ il più piccolo ed il più grande dei valori U_{-i}, $i = 1, ..., n$. Si ha, evidentemente, $V_{(1)} = nU - (n - 1)\ U_{-i}^{\max}$ e $V_{(n)} = nU - (n - 1)\ U_{-i}^{\min}$. Quindi,

$$V_{(1)} - U_{-i}^{\min} = n(U - U_{-i}^{\max}) + (U_{-i}^{\max} - U_{-i}^{\min}) \leq n(U - U_{-i}^{\max}),$$

$$V_{(n)} - U_{-i}^{\max} = n(U - U_{-i}^{\min}) - (U_{-i}^{\max} - U_{-i}^{\min}) \geq n(U - U_{-i}^{\min}),$$

valendo il segno di uguaglianza, in entrambi i casi, esclusivamente se $U_{-i}^{\max} = U_{-i}^{\min} = U$. In definitiva, dunque, risulta essere $V_{(1)} < U_{-i}^{\min}$ e $V_{(n)} > U_{-i}^{\max}$, a meno che non sia $V_{(1)} = U_{-i}^{\min} = U_{-i}^{\max} = V_{(n)} = U$. È facile convincersi che quest'ultima circostanza si verifica, per ogni fissata coppia di valori ($n_x > 1$, $n_y > 1$) per le dimensioni campionarie, solo se $U = 0$ o $U = 1$, cioè solo quando i due campioni, da Y e da X, risultano completamente "separati" nel campione *pooled* ordinato. In tutti gli altri casi, infatti, l'osservazione campionaria è tale che almeno una determinazione della variabile X è più grande di almeno una determinazione della variabile Y e almeno una determinazione di Y è più grande di almeno una determinazione di X. E in questi casi i numeratori nelle espressioni per le statistiche U_{-y_k} (e U_{-x_h}) date nella (2), non possono risultare tutti uguali al variare dell'osservazione Y_k (X_h) eliminata.

Sia ora $\varepsilon_0 > 0$ un valore reale fissato. Si supponga che, quando $n \to \infty$, la $\Pr\{U_{-i}^{\min} < \rho_0 + \varepsilon_0\}$ converga al valore $1 - \beta$, dove β è un qualche reale stretta-

mente positivo. Allora, per ogni ε tale che $0 < \varepsilon \leq \varepsilon_0$, $\lim_{n\to\infty} \Pr\{U_{-i}^{\min} < \rho_0 + \varepsilon\} \leq 1 - \beta$, visto che l'evento $\{U_{-i}^{\min} < \rho_0 + \varepsilon\}$ implica l'evento $\{U_{-i}^{\min} < \rho_0 + \varepsilon_0\}$. Allora, si avrebbe $\lim_{n\to\infty} \Pr\{U_{-i}^{\min} \geq \rho_0 + \varepsilon\} \geq \beta$ e, poiché l'evento $\{U_{-i}^{\min} \geq \rho_0 + \varepsilon\}$ implica l'evento $\{U \geq \rho_0 + \varepsilon\}$, risulterebbe $\lim_{n\to\infty} \Pr\{U \geq \rho_0 + \varepsilon\} \geq \beta > 0$. Verrebbe pertanto negata la consistenza dello stimatore U. Deve dunque essere necessariamente $\lim_{n\to\infty} \Pr\{U_{-i}^{\min} \leq \rho_0\} = 1$. In maniera analoga si può mostrare che $\lim_{n\to\infty} \Pr\{U_{-i}^{\max} \geq \rho_0\} = 1$.

Per completare la dimostrazione occorre mostrare che $\Pr\{0 < U < 1\} \to 1$ quando $n \to \infty$. Ma ciò segue dalla consistenza della statistica U, dovendo essere, sotto le ipotesi fatte, $0 < \rho_0 < 1$.

Teorema 1. Sotto le condizioni del lemma 1, $-2l(\rho_0) \xrightarrow{d} \chi_1^2$.

Dimostrazione. In base al risultato (*ii*) del lemma 2, si ha che $\Pr\{V_{(1)} < \rho_0 < V_{(n)}\} \to 1$ quando $n \to \infty$. Pertanto, la quantità $l(\rho_0)$ è finita con probabilità che tende a 1 quando $n \to \infty$.

Da un'applicazione del teorema di Dini, la funzione $\lambda(\cdot)$ definita dall'equazione (3) è continua in un intorno di U e risulta

$$\left.\frac{d\lambda(\rho)}{d\rho}\right|_{\rho=U} = -\frac{1}{\hat{\omega}},$$

dove $\hat{\omega} = n^{-1}\sum_{i=1}^{n}(V_i - U)^2$ è lo stimatore jackknife per la varianza asintotica di $\sqrt{n}U$. Si tratta di stimatore consistente, cioè $\hat{\omega} = \omega_0 + o_p(1)$. Sia $\lambda_0 = \lambda(\rho_0)$. Lo sviluppo in serie di Taylor della funzione $\lambda(\cdot)$ in un intorno di U, arrestato al primo ordine, permette di ottenere

$$\lambda_0 = \frac{U - \rho_0}{\hat{\omega}} + o_p(n^{-1/2}). \tag{5}$$

Perciò, in base a quanto sancito dal lemma 1 ed alla consistenza di $\hat{\omega}$, si ha $\lambda_0 = O_p(n^{-1/2})$ e, di conseguenza, tenendo conto anche del risultato (*i*) del lemma 2,

$$|\lambda_0| \max_{1\leq i\leq n} |V_i - \rho_0| = O_p(n^{-1/2}).$$

Lo sviluppo di McLaurin

$$\log(1+z) = z - \frac{1}{2}z^2 + \frac{z^3}{3(1+\bar{z})^3}, \quad |\bar{z}| \le |z|,$$

usato nella espressione di $l(\rho_0)$, cioè nella (4) calcolata in $\rho = \rho_0$, porta poi a

$$l(\rho_0) = -\lambda_0 \sum_{i=1}^{n}(V_i - \rho_0) + (\lambda_0^2/2)\sum_{i=1}^{n}(V_i - \rho_0)^2 + o_p(1).$$

Quindi, tenendo conto della (5) e del fatto che

$$\frac{1}{n}\sum_{i=1}^{n}(V_i - \rho_0)^2 = \frac{1}{n}\sum_{i=1}^{n}(V_i - U)^2 + (U - \rho_0)^2 = \hat{\omega} + O_p(n^{-1}),$$

si ha

$$-2l(\rho_0) = \frac{n(U-\rho_0)^2}{\hat{\omega}} + o_p(1).$$

Il risultato segue dalla normalità asintotica di $\sqrt{n}(U - \rho_0)$ e dalla consistenza di $\hat{\omega}$.

In base a quanto affermato dal teorema 1, la pseudo-verosimiglianza $L(\rho)$ può essere utilizzata, nella maniera usuale, per ottenere intervalli di confidenza per (o risolvere problemi di verifica d'ipotesi su) l'area ρ_0 sottesa alla curva ROC. In particolare, se c_γ è tale che $\Pr\{\chi_1^2 > c_\gamma\} = \gamma$, l'insieme

$$\{\rho : -2l(\rho) \le c_\gamma\}$$

costituisce un intervallo di confidenza approssimato per ρ_0, con livello di copertura nominale $1 - \gamma$.

La funzione $L(\rho)$, nella sua versione normalizzata, rappresenta una approssimazione della versione normalizzata della funzione di verosimiglianza empirica $\mathcal{L}(\rho)$, almeno in prossimità del punto di massimo U. Questo approccio ripropone, quindi, alcune proprietà della tecnica "pura" basata sulla verosimiglianza empirica e presenta alcuni vantaggi rispetto a metodi alternativi quali quello classico basato sull'approssimazione normale per la distribuzione della statistica U. In particolare, la costruzione di intervalli di confidenza mediante $L(\rho)$ non richiede la stima della varianza asintotica di alcuna statistica. Inoltre, gli intervalli corrispondenti hanno forma determinata automaticamente dai dati che non è soggetta a vincoli predeterminati di simmetria.

3. ALCUNI RISULTATI DI SIMULAZIONE

Per valutare l'accuratezza degli intervalli di confidenza prodotti dal metodo descritto, è stato effettuato uno studio di simulazione. Le tavole 1 e 2 forniscono i livelli di copertura empirici degli intervalli di confidenza per ρ_0, costruiti mediante la funzione $L(\rho)$, ottenuta a partire da $n_x + n_y$ pseudo-valori dalla statistica di Mann-Whitney. I livelli empirici riportati si riferiscono ad alcuni valori del livello nominale $1 - \gamma$ e delle dimensioni n_x e n_y dei campioni. A scopo comparativo, sono forniti anche i livelli empirici degli intervalli ottenuti mediante il metodo classico basato sull'approssimazione normale per la distribuzione della statistica U.

Ogni esperimento di simulazione è basato su 5000 replicazioni. Nel caso dei risultati riportati nella tavola 1, i valori per le variabili Y e X sono stati generati, rispettivamente, da una $N(\mu, 4)$ e da una $N(0, 1)$. Sono stati considerati tre diversi valori di μ, a cui corrispondono i valori 0.55, 0.75 e 0.95 per ρ_0. Per quanto riguarda invece i risultati riportati nella tavola 2, i valori per le variabili Y e X sono stati generati da una distribuzione gamma con parametro di forma α e parametro di scala 2, $Ga(\alpha, 2)$, e da una $Ga(1, 1)$, rispettivamente. Anche in questo caso, ai valori scelti per α corrispondono i valori 0.55, 0.75 e 0.95 per ρ_0.

Dall'analisi dei risultati forniti dalle tavole, gli intervalli di confidenza basati sulla pseudoverosimiglianza $L(\rho)$ appaiono sufficientemente accurati anche quando le dimensioni campionarie sono medio-piccole e, come ci si poteva aspettare, tendono ad essere più accurati degli intervalli ottenuti con il metodo classico basato sull'approssimazione normale. Naturalmente, le dimensioni campionarie che assicurano un livello di accuratezza sufficientemente elevato dipendono dal vero valore ρ_0 e sono tanto più grandi quanto più tale valore si avvicina a 1. Risultati di simulazione analoghi a quelli riportati (e relative simili conclusioni) si ottengono per valori piccoli di ρ_0 ($0 < \rho_0 < 1/2$), anche se tali valori risultano di scarso interesse nell'analisi ROC.

4. NOTE CONCLUSIVE

La funzione $L(\rho)$, discussa in questo lavoro, costituisce una funzione di pseudo-verosimiglianza per l'area ρ_0 sottesa alla curva ROC. Per essa vale un risultato asintotico analogo a quello fornito dal teorema di Wilks nel caso parametrico; ciò giustifica il termine pseudo-verosimiglianza ad essa attribuito e ne consente l'uso, in una maniera standard, per ottenere intervalli di confidenza non parametrici, con livelli di copertura asintoticamente corretti.

TAVOLA 1

Livelli di copertura empirici degli intervalli di confidenza per ρ_0 *ottenuti mediante la pseudo-verosimiglianza* $L(\rho)$ (PV) *e l'approssimazione normale* (AN). *Dati generati da distribuzioni normali*

			$1-\gamma$		
			0.99	0.95	0.90
$\rho_0 = 0.55$	$n_x = 10$, $n_y = 15$	PV	0.993	0.957	0.911
		AN	0.973	0.924	0.872
	$n_x = 35$, $n_y = 50$	PV	0.991	0.951	0.899
		AN	0.985	0.941	0.887
$\rho_0 = 0.75$	$n_x = 10$, $n_y = 15$	PV	0.982	0.956	0.907
		AN	0.954	0.908	0.868
	$n_x = 15$, $n_y = 20$	PV	0.988	0.955	0.912
		AN	0.963	0.919	0.872
	$n_x = 35$, $n_y = 50$	PV	0.989	0.950	0.899
		AN	0.980	0.935	0.882
$\rho_0 = 0.95$	$n_x = 35$, $n_y = 50$	PV	0.945	0.904	0.864
		AN	0.915	0.884	0.851
	$n_x = 55$, $n_y = 70$	PV	0.967	0.929	0.889
		AN	0.937	0.891	0.848
	$n_x = 100$, $n_y = 125$	PV	0.987	0.943	0.891
		AN	0.962	0.916	0.869

TAVOLA 2

Livelli di copertura empirici degli intervalli di confidenza per ρ_0 *ottenuti mediante la pseudo-verosimiglianza* $L(\rho)$ (PV) *e l'approssimazione normale* (AN). *Dati generati da distribuzioni gamma*

			$1-\gamma$		
			0.99	0.95	0.90
$\rho_0 = 0.55$	$n_x = 10$, $n_y = 15$	PV	0.991	0.961	0.914
		AN	0.975	0.931	0.882
	$n_x = 35$, $n_y = 50$	PV	0.992	0.952	0.904
		AN	0.987	0.943	0.895
$\rho_0 = 0.75$	$n_x = 10$, $n_y = 15$	PV	0.979	0.938	0.892
		AN	0.952	0.908	0.862
	$n_x = 15$, $n_y = 20$	PV	0.984	0.952	0.904
		AN	0.967	0.923	0.876
	$n_x = 35$, $n_y = 50$	PV	0.993	0.955	0.904
		AN	0.982	0.941	0.892
$\rho_0 = 0.95$	$n_x = 35$, $n_y = 50$	PV	0.973	0.929	0.876
		AN	0.942	0.899	0.852
	$n_x = 55$, $n_y = 70$	PV	0.982	0.945	0.900
		AN	0.959	0.922	0.876
	$n_x = 100$, $n_y = 125$	PV	0.990	0.951	0.902
		AN	0.975	0.934	0.889

La versione normalizzata della funzione $L(\rho)$ rappresenta un'approssimazione della versione normalizzata della funzione di verosimiglianza empirica $\mathcal{L}(\rho)$ per ρ_0, di cui ripropone alcune proprietà. Le due funzioni non sono però la stessa cosa e ci si aspetta che possano essere anche molto diverse quando le dimensioni campionarie sono piccole, in particolare in regioni distanti dal punto di massimo U. Uno studio specifico delle differenze tra $L(\rho)$ e $\mathcal{L}(\rho)$ potrebbe rivelarsi quindi molto utile, eventualmente per individuare tecniche di aggiustamento che permettano di migliorare il comportamento di $L(\rho)$ come surrogato di $\mathcal{L}(\rho)$.

Un'ultima considerazione riguarda una possibile estensione dell'approccio presentato. In talune situazioni, piuttosto che l'intera area sottesa alla curva ROC, è preferibile usare, come indice di accuratezza diagnostica, una porzione dell'area

stessa; per esempio, quella corrispondente a valori piccoli, inferiori a una certa soglia u, della probabilità associata a un "falso positivo" (si veda Dodd e Pepe, 2003). In questo caso, per l'area parziale τ_0, sottesa alla curva ROC nella regione $(0, u)$, vale la relazione $\tau_0 = \Pr\{Y > X, X > t\}$, dove $t = S_x^{-1}(u)$ e $S_x(\cdot)$ indica la funzione di sopravvivenza di X. In alcune circostanze è ragionevole assumere che il quantile t sia noto. Allora, uno stimatore non parametrico per τ_0 è dato da

$$\hat{\tau} = \frac{1}{n_x n_y} \sum_{b=1}^{n_x} \sum_{k=1}^{n_y} I(Y_k > X_b, X_b > t)$$

e gli pseudo-valori jackknife derivati da tale statistica possono essere utilizzati, come descritto nel lavoro, per ottenere una pseudo-verosimiglianza per τ_0.

Dipartimento di Scienze Statistiche GIANFRANCO ADIMARI
Università di Padova

RIFERIMENTI BIBLIOGRAFICI

D. BAMBER, (1975), *The area above the ordinal dominance graph and the area below the receiver operating graph*, "Journal of Mathematical Psychology", 12, pp. 387-415.

L.E. DODD, M.S. PEPE, (2003), *Partial AUC estimation and regression*, "Biometrics", pp. 614-623.

B.Y. JING, J. YUAN, W. ZHOU (2005), *Empirical likelihood for non-degenerate U-statistics*, "Prooceedings ASMDA (Applied Stochastic Models and Data Analysis)", pp. 793-803.

A.B. OWEN, (1988), *Empirical likelihood ratio confidence intervals for a single functional*, "Biometrika", 75, pp. 237-249.

A.B. OWEN, (2001), *Empirical likelihood*, Chapman and Hall, London.

R.J. SERFLING, (1980), *Approximation theorems of mathematical statistics*, Wiley, New York.

RIASSUNTO

Intervalli di confidenza non parametrici per l'area sottesa alla curva ROC

Seguendo un'idea di Jing *et al.* (2005), in questo lavoro si combinano la funzione di verosimiglianza empirica per il funzionale media e gli pseudo-valori jackknife derivati dalla statistica di Mann-Whitney per due campioni. Ciò permette di ottenere una funzione di pseudo-verosimiglianza $L(\rho)$ per l'area ρ_0 sottesa alla curva ROC. Si dimostra che vale un risultato asintotico analogo al teorema di Wilks, cosicché $L(\rho)$ può essere usata, nella maniera usuale, per ottenere intervalli di confidenza approssimati per ρ_0. Vengono inoltre forniti alcuni risultati di simulazione che mostrano l'utilità del metodo proposto.

SUMMARY

Nonparametric confidence intervals for the area under the ROC curve

Following an idea by Jing *et al.* (2005), this paper combines the empirical likelihood for the mean functional with jackknife pseudo-values obtained from the Mann-Witney two-sample statistic. This leads to a pseudo-likelihood $L(\rho)$ for the area ρ_0 under the ROC curve. A Wilks-type theorem is proved, so that $L(\rho)$ can be used in a standard way to obtain approximate confidence intervals for ρ_0. Some simulation results are given, in order to show the usefulness of the proposed method.