Rivista di Statistica Applicata Vol. 20 n. 3 settembre 1987

FORUM

Rubrica a cura di Gabriella Serio

Alcune considerazioni sulla varianza del rapporto standardizzato di mortalità (RSM)

Piergiorgio Duca, Istituto di Biometria e Statistica Medica, Università degli Studi – Milano

1. Premessa

La presente nota è stata suggerita dal constatare che, in alcuni testi elementari di statistica per epidemiologi (vedi ad esempio H.A. Kahn, 1983), non è dato il giusto rilievo né al fatto che l'errore standard di RSM sotto ipotesi nulla differisce da quello di RSM sotto ipotesi alternativa, né alla derivazione, peraltro elementare, di tali errori standard, né alle implicazioni statistiche dell'uso dell'errore standard di RSM sotto ipotesi alternativa per saggiare l'ipotesi nulla.

2. Il problema

Negli studi epidemiologici di mortalità è usuale, per una definita popolazione, stimare il Rapporto Standardizzato di Mortalità:

$$\widehat{RSM} = \frac{\text{numero di decessi osservati}}{\text{numero di decessi attesi}}$$

Al denominatore compare il numero il numero di decessi attesi nel campione in studio qualora la mortalità strato specifica della popolazione eletta a standard, che si assume non affetta da errore di stima, coincidesse con quella della popolazione campionata.

Gli strati sono definiti in base al valore di una variabile di confondimento di cui è necessario controllare l'effetto (ad esempio l'età).

La quantità che compare al denominatore varia con la composizione del campione in studio ma, dato un definito campione, è una costante.

RSM è così il rapporto tra una variabile casuale, che si assume distribuita secondo una funzione di probabilità di Poisson, e una costante che, sotto ipotesi nulla ($H_0 : RSM = 1$), ne è il valore atteso.

La statistica appropriata per saggiare H_0, quando l'elevato numero degli eventi attesi (dicamo 20 o più) consente l'approssimazione gaussiana, è:

$$X^2 = (O - A)^2/A \tag{1}$$

dove:

O = numero di eventi osservati

A = numero di eventi attesi sotto H_0

La (1) ha distribuzione chi–quadrato con un grado di libertà. Per definizione: $\widehat{RSM} = (O/A)$ e sotto $H_0 : Var(O) = A$. Pertanto:

$$Var(\widehat{RSM}) = Var(O/A) = 1/A^2 \cdot A = \widehat{RSM}/O \tag{2}$$

Nel caso in cui si rifiuti H_0, permane l'interesse alla stima intervallare di $\widehat{RSM}$, ma $Var(O)$ è ignota.

Sotto ipotesi alternativa ($H_1 : RSM \neq 1$) la migliore stima di $Var(O)$, assumendo la distribuzione di Poisson, è $Var(O) = O$. Da cui:

$$\widehat{Var}(\widehat{RSM}) = \widehat{Var}(O/A) = 1/A^2 \cdot O = \widehat{RSM}^2/O \tag{3}$$

E' da rilevare che in questo caso si ha solo una stima dell'$ES(\widehat{RSM})$ e non è quindi rigorosamente corretto usare tale quantità, unitamente al valore appropriato della deviata gaussiana standardizzata, per stimare l'intervallo di confidenza di RSM.

A maggior ragione è improprio usare il valore della (3) per saggiare H_0, in quanto la statistica:

$$X'^2 = (O - A)^2/O \tag{4}$$

è distribuita secondo chi–quadrato solo asintoticamente. Si rimanda alla appendice chi volesse avere una dimostrazione di ciò unitamente a una descrizione di come tale statistica si comporta per alcuni valori di A.

La (2) e la (3) danno risultati identici solo per $\widehat{RSM} = 1$; la prima equazione dà un valore maggiore della seconda se $\widehat{RSM} < 1$, minore se $\widehat{RSM} > 1$.

Un modo per stimare l'intervallo di confidenza quando si sia rifiutata H_0, esente dalle critiche sopra esposte, è il seguente.

Consideriamo la seguente equazione nell'incognita E:

$$(O - E)^2/E = z^2 \tag{5}$$

Le due soluzioni che consentono di stimare gli estremi dell'intervallo di confidenza di RSM a livello $(1 - \alpha)$ sono:

$$RSM_S = E_S/A = [O + \frac{1}{2} z_{\alpha/2} \cdot (z_{\alpha/2} + \sqrt{(z^2_{\alpha/2} + 4 \cdot O)})]/A \tag{6}$$

$$RSM_I = E_I/A = [O + \frac{1}{2} z_{\alpha/2} \cdot (z_{\alpha/2} + \sqrt{(z^2_{\alpha/2} + 4 \cdot O)})]/A \tag{7}$$

Applicando la (3) l'ampiezza dell'intervallo di confidenza di RSM che se ne ricava vale:

$$2 \cdot z_{\alpha/2} \cdot \sqrt{O}/A \tag{8}$$

mentre dalle (6) e (7) si ricava che l'ampiezza dell'intervallo così calcolato vale:

$$z_{\alpha/2} \cdot \sqrt{(z^2_{\alpha/2} + 4 \cdot O)}/A \tag{9}$$

Come si vede quanto maggiore è O rispetto a $z^2_{\alpha/2}$ tanto più i due risultati tendono a coincidere.

3. Esemplificazione

Consideriamo ora, a titolo esemplificativo, le seguenti applicazioni.

In Italia nel 1979 i tassi di mortalità (maschi) per cardiopatia ischemica (codici $ICD - IX$: 410 – 414) specifici per classi quinquennali di età fra 35 e 59 anni, espressi per 100.000 persone anno, furono i seguenti:

19 41 85 168 313

In Lombardia si sono verificati, nello stesso intervallo di età, 1887 decessi fra i maschi.

Considerato che la popolazione lombarda è in media più giovane di quella nazionale, per confrontare la mortalità della regione Lombardia con quella nazionale è opportuno "aggiustare" il confronto per la diversa struttura per età.

Nel 1979 i maschi residenti in Lombardia, distribuiti per classe quinquennale di età, risultarono essere (in migliaia):

348 317 304 276 203

Applicando i tassi nazionali alla popolazione lombarda si ottiene come numero di decessi atteso (A) la quantità 1554.

Si può saggiare H_0 calcolando la statistica seguente:

$$(1887 - 1554)^2/1554 = 71.36$$

Il valore osservato consente di rifiutare H_0.

Si rilevi che in questo caso la statistica X'^2 vale 58.76, valore ancora altamente significativo, se interpretato come chi–quadrato, ma molto differente da quello "corretto" precedentemente calcolato.

RSM stimato risulta eguale a $(1887/1554) = 1.21$ e il suo ES, sotto ipotesi alternativa, vale:

$$(1887/1554)/\sqrt{1887} = 0.028$$

L'intervallo di confidenza al 95%, applicando la (5), ha come estremi i valori:

$$1.16 - 1.27$$

e ricorrendo alla (3) lo stesso intervallo vale $1.15 - 1.26$.

In una USL della regione Lombardia si sono verificati 34 decessi contro un numero atteso pari a 23.

In questa situazione il valore della statistica test per H_0, con correzione per la continuità, è pari a:

$$(33.5 - 23)^2/23 = 4.79$$

Il risultato consente di rifiutare H_0 a livello di significatività del 5%. Da notare che in questo caso $X'^2 = 3.24$ non consentirebbe invece tale rifiuto al 5%.

$\widehat{ES}(\widehat{RSM})$ vale, applicando la (3):

$$(34/23)/\sqrt{34} = 0.308$$

Utilizzando tale valore si ricavano i seguenti limiti di confidenza al 95%:

$$0.87 - 2.08$$

Ricorrendo alle (6) e (7) i limiti di confidenza di RSM al 95% risultano invece:

$$1.06 - 2.07$$

Come si vede per valori di O non troppo elevati, ma pur sempre nei limiti della approssimabilità della distribuzione poissoniana con quella gaussiana, si possono avere risultati anche sostanzialmente diversi.

Appendice

Si assuma di estrarre da una popolazione gaussiana, con media e varianza eguali ad A, un numero infinito di campioni di dimensione 1 e di calcolare le statistiche:

$$X^2 = (x - A)^2/A \tag{10}$$

$$X'^2 = (x - A)^2/x \tag{11}$$

La statistica X^2 in (10) ha distribuzione chi–quadrato con un grado di libertà, mentre X'^2 in (11) ha una distribuzione approssimabile con la distribuzione chi–quadrato con un grado di libertà solo quando A sia molto elevato.

La bontà di tale approssimazione per $A = 400$, ad esempio, si evince dalla tab. I, che riporta le probabilità cumulative calcolate in relazione ad alcuni centili di particolare interesse della distribuzione chi–quadrato.

Tali probabilità sono quelle corrispondenti a un deviata gaussiana standardizzata (z) che sta nella seguente relazione con la quantità $\sqrt{X'^2} = X'$:

$$z = X'/\sqrt{4A} \cdot (X' \pm \sqrt{(X'^2 + 4A)}) \tag{12}$$

Questa espressione si ricava risolvendo in x la (11):

$$x = \frac{1}{2} \cdot (2A + X'^2 \pm X'\sqrt{(4A + X'^2)}) \tag{13}$$

e esprimendo x come deviata gaussiana standardizzata:

$$z = (x - A)/\sqrt{A} \tag{14}$$

Per A tendente all'infinito dalla (12) si ricava direttamente la seguente:

$$z = X' \qquad \text{da cui} \qquad X'^2 = z^2 \tag{15}$$

Resta così dimostrato che la variabile X'^2 tende asintoticamente ad una distribuzione chi–quadrato con un grado di libertà.

Tab. I – Corrispondenza fra centili della distribuzione di $X^2 = (O-A)^2/A$ e centili della distribuzione di $X'^2 = (O-A)^2/O$, per diversi valori di A. (*)

u	Pr(X² >u)	Pr(X'² >u)				
		A=10^	A=10	A=20	A=100	A=400
.102	.750	.761	.749	.749	.753	.749
.455	.500	.541	.497	.497	.500	.500
1.32	.250	.214	.252	.251	.251	.250
2.71	.100	.094	.119	.109	.101	.101
3.84	.050	.074	.078	.065	.053	.050
6.63	.010	.029	.042	.026	.014	.011

(*) Nella colonna intestata con "A=10^" sono riportate le probabilità esatte calcolate secondo la distribuzione di Poisson; nelle altre colonne invece sono riportate le probabilità che si ricaverebbero approssimando con la distribuzione gaussiana una distribuzione poissoniana di parametro A.

Riferimenti bibliografici

Kahn H.A., 1983, *An Introduction to Epidemiologic Methods*, Oxford University Press, 72–78