Rivista di Statistica Applicata Vol. 21 n. 2 Giugno 1988

ROSTRA

Rubrica a cura di Benito V. Frosini

Limiti di confidenza di una proporzione e di un tasso data l'osservazione di zero eventi

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

1. Premessa

Talora capita, soprattutto se la dimensione campionaria è ridotta, di concludere uno studio senza la osservazione di alcun evento (verificarsi di malattia o decesso o guarigione, risposta falsamente negativa o positiva ad un test diagnostico).

Potendo ragionevolmente escludere che l'osservazione è dovuta ad un artefatto, ad esempio per la scarsa affidabilità del sistema di rilevazione o per l'esiguità dell'intervallo di tempo fra l'osservazione e l'inizio dell'esposizione in studio, è necessario definire l'appropriato intervallo di confidenza per il parametro di interesse, ricorrendo ad un metodo esatto (Mood e Graybill, 1963).

Esso, come esemplificato, fornisce una soluzione diretta e semplice e costituisce pertanto un buon esercizio didattico che non si trova svolto nei testi di statistica a più larga diffussione; al contrario risulta che in quelli contenenti le appropriate tavole, con l'esclusione del volume di Fisher e Yates (1963), viene riportata una soluzione errata (Owen, 1962; Documenta Geigy, 1963; Bayer, 1966; Dixon e Massey, 1969).

2. Intervallo di confidenza esatto, a livello $(1-\alpha)$

A) Variabile binomiale

Assumiamo di avere osservato n soggetti per un eguale e congruo intervallo di tempo e di non avere rilevato alcun caso di malattia. Data l'osservazione è corretto assumere a limite inferiore dell'intervallo di confidenza per π il valore 0.

Infatti per tale valore di parametro la probabilità della osservazione fatta o di una più estrema (maggiore di zero) è rigorosamente eguale a 1.

Per quanto riguarda invece il limite superiore esso deve essere tale da soddisfare la seguente eguaglianza:

$$Pr(X = 0|n;\pi) = (1-\pi)^n = \alpha \tag{1}$$

da cui:

$$n \cdot Ln(1-\pi) = Ln(\alpha) \tag{2}$$

quindi:

$$(1-\pi) = exp(Ln(\alpha)/n) \tag{3}$$

e, infine:

$$\pi = 1 - exp(Ln(\alpha)/n) \tag{4}$$

Risulta così definito un intervallo di confidenza non centrale (Kendall e Stuart, 1976) a livello $(1-\alpha)$.

Inoltre, poiché la funzione $(1-exp(-y))$ è approssimabile con y per $y < 0.10$, se $n > 30$ e $\alpha = 0.05$ la (4) è approssimabile con la seguente:

$$\pi \approx 3/n \tag{5}$$

B) Variabile poissoniana

Assumiamo di avere osservato l'esperienza di una massa di N anni–persona (base dello studio) senza rilevare alcun caso di malattia.

La variabile casuale appropriata in questo caso ha distribuzione poissoniana con parametro $\mu = I \cdot N$, dove I denota l'incidenza vera della malattia nella popolazione studiata. L'intervallo di confidenza a livello $(1-\alpha)$ per I ha allora i seguenti estremi: 0 e μ/N, dove μ deve soddisfare la seguente:

$$Pr(X = 0|\mu) = \exp(-\mu) = \alpha \tag{6}$$

da cui:

$$-\mu = Ln(\alpha) \tag{7}$$

quindi:

$$I = -Ln(\alpha)/N \approx 3/N \tag{8}$$

dove l'approssimazione vale solo se $\alpha = 0.05$.

3. Tre esempi

A) Stima del rischio

Di 25 soggetti esposti a un determinato fattore nessuno contrae malattia entro 5 anni.

Quanto può valere il rischio vero al 95% di confidenza?

$$\pi = 1 - exp(Ln(0.05)/25) = 0.1129$$

Nelle tavole citate viene riportato il valore errato di 0.1372.

Ricorrendo alla approssimazione (5) si ottiene $\pi = 0.12$.

B) Stima di incidenza

Da una massa a rischio di 784 anni–persona non sono stati generati casi di malattia nel corso della osservazione.

Quanto può valere l'incidenza vera al 95% di confidenza?

$$I = -Ln(0.05)/784 = 3.82 \text{ per } 1000 \text{ anni–persona}$$

In questo caso la (8) fornisce una soluzione soddisfacente (3.83 per 1000 anni–persona).

Il lettore non si faccia fuorviare dalla similitudine delle formule (5) e (8).

Nel primo caso argomento è il rischio o probabilità che un soggetto della popolazione in studio contragga malattia in un definito intervallo di tempo mentre nel secondo è l'incidenza o forza istantanea di morbosità che si esercita sulla popolazione in studio.

Nel primo caso n è il numero di soggetti inizialmente a rischio, seguiti per un eguale periodo di tempo, tutti potenziali casi incidenti nel periodo di osservazione.

Nel secondo caso N è una massa–tempo espressa in anni–persona, somma dei diversi periodi a rischio individuali osservati. Da N non è deducibile il numero potenziale di casi incidenti.

C) Stima della sensibilità di un test diagnostico

Un test diagnostico, somministrato a 17 malati, è risultato positivo in tutti i soggetti.

Quale può essere la sensibilità vera del test al 95% di confidenza?

La sensibilità di un test diagnostico è pari al complemento a 1 della proporzione di falsi negativi, per la cui stima intervallare si può applicare la (4). Il limite superiore dell'intervallo di confidenza per la sensibilità è 1.00 e il limite inferiore si ricava facendo il complemento a 1 della (4) e quindi:

$$\pi = exp(Ln(0.05)/17) = 0.8384$$

Per $n > 30$ e $\alpha = 0.05$ si può ricorrere alla seguente approssimazione:

$$\pi \approx (n - 3)/n$$

4. Conclusione

Sembra appropriato trattare estesamente il calcolo dell'intervallo esatto di confidenza nei casi considerati sia per la semplicità dell'approccio che consentono e la efficacia didattica sia perché tali situazioni si presentano nella realtà.

Ad esempio, la osservazione di zero decessi (eventi) può capitare frequentemente nella analisi di dati di mortalità disaggregati per piccole aree geografiche (Province, USL, Comuni). In questo contesto nasce anche il problema di come introdurre tali stime nel calcolo di una "media ponderata" che in genere utilizza come peso il reciproco della varianza di stima.

Una soluzione è offerta dalle procedure Bayesiane e neo–Bayesiane; alternativamente si possono utilizzare i risultati sopra esposti, ricorrendo a un sistema di pesi basato sul reciproco dell'ampiezza dell'intervallo di confidenza. In questo caso però va tenuta nel giusto conto anche la non centralità degli intervalli di confidenza ottenuti.

Riferimenti bibliografici

Bayer W.H., 1966, *Handbook of tables for probability and statistics*, CRC Publ., Cleveland, 219–237.

Dixon W.J., Massey F.J., 1969, *Introduction to statistical analysis*, McGraw–Hill, New York, 501–504.

Documenta Geigy, 963, *Tables scientifiques*, J.R. Geigy, S.A. Basel, 85–103.

Fisher R.A., Yates F., 1963, *Statistical tables for biological agricultural and medical research*, Oliver–Boyd, Edinburgh, 65.

Kendall M., Stuart A., 1976, *The advanced theory of statistics*, C. Griffin & Co. Ltd, London, 113–116.

Mood A.M., Graybill F.A., 1963, *Introduction to the theory of statistics*, McGraw–Hill, New York, 256–262.
Owen D.B., 1962, *Handbook of statistical tables*, Pergamon Press, London, 274–285.

Summary

Confidence limits for proportions and rates when zero events are observed.

The exact method to calculate confidence limits for proportions and rates when zero events are observed is described.
In this case the approach is very simple and can be usefully employed for teaching the concept of confidence interval.
In spite of its underlined simpliciy many handbooks of statistical tables report incorrect results.