# *SUGGERIMENTI PER L'IMPIEGO DEL TEST CHI-QUADRATO* *

di *Michele Zenga* **

## 1. *Introduzione e sommario*

Una delle critiche più frequentemente fatte al test chi-quadrato è che per n (numerosità campionaria) grande, differenze dalla ipotesi nulla anche di lieve entità portano quasi sempre all'accettazione della ipotesi alternativa (il test sarebbe troppo sensibile).

In questa nota, dopo aver mostrato l'opportunità di misurare la distanza delle probabilità $p_i$ $(i = 1, 2, \ldots, k)$ di una alternativa dalle probabilità $\pi_i$ della ipotesi nulla, facendo riferimento alla funzione $D = \max \left\{ \frac{|p_i - \pi_i|}{\pi_i}; i = 1, 2, \ldots . k \right\}$, si è proceduto alla determinazione delle numerosità campionarie necessarie per avere valori piccoli di $\alpha$ (probabilità errore di prima specie) e di max $\beta$ ($\beta$ è la probabilità di un errore di seconda specie; max $\beta$ è riferito alle alternative con $D \geqslant \delta$, essendo $\delta > 0$).

Le numerosità campionarie così ottenute sono molto elevate ed indicano che il test chi-quadrato è meno sensibile di quanto si ritiene e che quindi le critiche di cui sopra non sembrano molto fondate.

Da ultimo, si è mostrato come utilizzare la statistica $X^2$ di Karl-Pearson e la funzione D per ottenere un test con errori proporzionali, cioè un test che soddisfa la condizione $\frac{\max \beta}{\alpha} = \nu$, essendo $\nu$ un valore positivo prefissato.

## 2. *Distanza dell'alternativa*

Si abbia una multinomiale a k classi con probabilità $p_1, p_2, \ldots, p_k$ e con numero di prove n.

Si supponga di voler verificare l'ipotesi $H_0 : (p_1 = \pi_1)$ e $(p_2 = \pi_2)$

---

* Lavoro presentato alla XXIX riunione scientifica della Società Italiana di Statistica, Bologna 20-22 marzo 1978.

** Libera Università degli Studi di Trento - Facoltà di Economia e Commercio.

e .... e ($p_k = \pi_k$), contro l'alternativa $H_1$ : ($p_1 \neq \pi_1$) o ($p_2 \neq \pi_2$) o .... o ($p_k \neq \pi_k$).

Come è noto, tale verifica si esegue ricorrendo alla statistica $X^2 = \Sigma \frac{(n_i - n\pi_i)^2}{n\pi_i}$, essendo $n_i$ il numero delle volte in cui nelle n prove si è verificata la classe $i^{ma}$, ovviamente $\Sigma n_i = n$. Sotto Ho la statistica $X^2$ tende a distribuirsi, per n grande, secondo una v.c. chi-quadrato centrale con $k - 1$ gradi di libertà. Sotto una qualsiasi alternativa ($p_1$, $p_2$, ...., $p_k$) la variabile $X^2$ tende a distribuirsi, per n grande, secondo una v.c. chi-quadrato non centrale con $k - 1$ gradi di libertà e parametro di non centralità $\lambda^2 = n \Sigma \frac{(p_i - \pi_i)^2}{\pi_i}$.

Alcuni studiosi sostengono [1] che per valori elevati di n, anche ai più bassi livelli di significatività, il test basato su $X^2$ porta a respingere l'ipotesi Ho anche per alternative assai prossime ad Ho. Queste affermazioni sono basate, più che sull'effettivo calcolo della funzione di potenza, sulla proprietà della consistenza del test $X^2$. La consistenza è però una proprietà asintotica per cui non è possibile sapere da quali valori di n si possa ritenere trascurabile la probabilità $\beta$.

Si mostrerà in seguito che, se fra l'alternativa ed Ho vi è una piccola distanza e si vogliono avere valori di $\alpha$ e di $\beta$ piccoli, è necessario impiegare valori molto elevati di n.

Non potendo però disporre nella pratica di valori tanto elevati di n, si impone l'esigenza di considerare le alternative aventi almeno una certa distanza da Ho e di trascurare le alternative più vicine. Trascurare le alternative molto prossime ad Ho non sembra possa portare a conseguenze pratiche, perché al ricercatore interessa scoprire se l'alternativa si discosti o meno di una certa entità da Ho.

Inoltre, si tenga presente che una certa distanza fra Ho e l'alternativa è necessaria anche per motivi strettamente tecnici, in quanto anche per valori elevati di n, se fra una alternativa ed Ho non vi è una certa distanza, le distribuzioni di $X^2$ sotto le due ipotesi tendono a confondersi ed è allora difficile discriminare fra le stesse.

Vi sono diversi modi per valutare la distanza di una alternativa da Ho. Solitamente [2] si ricorre al parametro di non centralità $\lambda^2 = n \Sigma \frac{(p_i - \pi_i)^2}{\pi_i}$. Ciò è forse dovuto al fatto che la distribuzione di $X^2$ dipende solo da $(k - 1)$ e da $\lambda^2$.

A fronte di questi vantaggi, $\lambda^2$ presenta alcuni inconvenienti. Innanzitutto $\lambda^2$ fa dipendere la distanza di una alternativa da Ho dalla numerosità campionaria n. Questo inconveniente è facilmente elimina-

bile riferendosi ad $\eta^2 = \Sigma \frac{(p_i - \pi_i)^2}{\pi_i}$. Ma anche $\eta^2$ presenta ancora degli inconvenienti in quanto il ricercatore non riesce a valutare che cosa voglia significare un certo valore di $\eta^2$, essendo lo stesso, generalmente, più abituato a ragionare in termini di differenze percentuali fra $p_i$ e $\pi_i$, piuttosto che sulle differenze al quadrato fra $p_i$ e $\pi_i$ rapportate a $\pi_i$.

Sembra, pertanto, più utile fornire un procedimento, per misurare la distanza, basato sulle differenze percentuali assolute $\frac{|p_i - \pi_i|}{\pi_i}$.

Ad esempio, si potrebbe considerare la seguente funzione degli scarti percentuali assoluti: $D = \max \left\{ \frac{|p_i - \pi_i|}{\pi_i} ; \; i = 1, 2, \ldots, k \right\}$.

Al ricercatore, potrebbero interessare solo quelle alternative per le quali $D \geqslant \delta$, essendo $\delta > 0$. D'ora in avanti si indicherà con $\Omega_\delta$ il sottospazio dello spazio parametrico caratterizzato da $D \geqslant \delta$. Affermare che $D \geqslant \delta$ è equivalente ad affermare che almeno una differenza percentuale assoluta $\frac{|p_i - \pi_i|}{\pi_i}$ sia maggiore od uguale a $\delta$.

Ovviamente, è possibile far riferimento anche ad altre funzioni delle differenze percentuali assolute per misurare la distanza fra una alternativa ed Ho. Ad esempio, si potrebbe considerare la media aritmetica delle percentuali assolute $\frac{1}{k} \Sigma \frac{|p_i - \pi_i|}{\pi_i}$ e considerare le alternative per le quali detta media sia maggiore o uguale a v, essendo $v > 0$. Come anche, si potrebbe far riferimento ad una funzione delle differenze assolute. In questo lavoro verrà considerata la funzione:

$$D = \max \left\{ \frac{|p_i - \pi_i|}{\pi_i} ; \; 1, 2, \ldots, k \right\}.$$

Sembra ora utile avere qualche idea sulla forma della porzione dello spazio parametrico caratterizzato da $D \leqslant \delta$.

Lo spazio parametrico $\Omega$ può considerarsi un sottospazio dello spazio euclideo K-dimensionale $E^k$ i cui punti $(p_1, p_2, \ldots, p_k)$ sono tali che: $\Sigma p_i = 1$, $p_i \geqslant 0$ $(i = 1, 2, \ldots k)$. Sia $(\pi_1, \ldots, \pi_k)$ un punto prefissato di $\Omega$ e si definisca la funzione.

$$D = \max \left\{ \frac{|p_i - \pi_i|}{\pi_i} ; \; i = 1, 2, \ldots, k \right\}, \qquad (1)$$

essendo $(p_1, p_2, \ldots, p_k)$ un punto arbitrario di $\Omega$.

Fissato un numero reale $\delta > 0$ consideriamo il luogo dei punti che soddisfano la condizione: $\max \left\{ \frac{|p_i - \pi_i|}{\pi_i} ; i = 1, 2, \ldots, k \right\} = \delta$. (2)

Si controlla subito che trattasi di una ipersuperficie di dimensione $(k-1)$, e precisamente di un poliedro le cui facce sono di dimensione $(k-2)$ e sono caratterizzate dall'avere una coordinata $p_i$ ($i = 1, 2, \ldots, k$) costante. In particolare le facce sono a due a due parallele, poiché giacciono su iperpiani di equazioni: $p_i = \pi_i + \delta\pi_i$ e $p_i = \pi_i - \delta\pi_i$.

Non sempre però le facce sono 2k, come si riscontrerà negli esempi che seguono. Tuttavia, come verrà mostrato in seguito, ciò non ha nessuna conseguenza operativa nell'impiego del test. Per esemplificare consideriamo il caso $k = 3$ con $\pi_1 = 0{,}4$, $\pi_2 = 0{,}3$ e $\pi_3 = 0{,}3$ e con $\delta = 0{,}2$.

I punti dello spazio parametrico in cui $D \leqslant \delta$ sono caratterizzati dal contemporaneo realizzarsi delle seguenti relazioni:

$$0{,}32 \leqslant p_1 \leqslant 0{,}48$$
$$0{,}24 \leqslant p_2 \leqslant 0{,}36$$
$$0{,}64 \leqslant p_1 + p_2 \leqslant 0{,}76$$

Conseguentemente, il contemporaneo verificarsi delle relazioni precedenti dà luogo in $\Omega$ all'esagono indicato nel grafico 1.

Non sempre però, per $k = 3$, il luogo dei punti che si ottiene ponendo $D = \delta$ dà luogo ad un esagono.

Per mostrare ciò e senza perdere in generalità si supponga $\pi_1 \geqslant \pi_2 \geqslant \pi_3$, il che implica $\pi_1 \geqslant 1/3$. Si ricava facilmente che se $1/3 \leqslant \pi_1 < 1/2$ si ha un esagono e se $\pi_1 \geqslant 1/2$ si ha un quadrilatero, come è indicato nel grafico 2 in cui si è supposto $\pi_1 = 0{,}6$, $\pi_2 = 0{,}2$ e $\pi_3 = 0{,}2$ e $\delta = 0{,}20$.

## 3. *Riformulazione del problema ed alternativa più sfavorevole.*

È possibile riformulare il problema di verifica di ipotesi riguardante la multinomiale ipotizzando che il sottospazio parametrico dell'ipotesi alternativa $H_1$ sia $\Omega_\delta$.

Per il problema di verifica di ipotesi di cui sopra è possibile far ancora riferimento alla statistica $X^2$.

Si supponga che il valore critico della variabile test sia c. Per n sufficientemente elevato il valore di $\alpha$ è approssimato bene dalla probabilità che la variabile chi-quadrato con $k-1$ gradi di libertà, che si indica con $X^2_{(k-1)}$, sia maggiore o uguale a c. Analogamente per la alternativa $(p_1, p_2, \ldots, p_k) \in \Omega_\delta$ e sempre per n grande il valore di $\beta$