Quaderni di Statistica e Matematica
applicata alle Scienze Economico-Sociali
VOLUME I — N. 2 — OTTOBRE 1978

LA DISTANZA DELLE ALTERNATIVE NELLE VERIFICHE DI IPOTESI RIGUARDANTI LE DISTRIBUZIONI MULTINOMIALI *

di Michele ZENGA **

1. *Introduzione e sommario*

In un precedente lavoro [1] si è mostrata l'opportunità di misurare in una multinomiale a k classi la distanza di una alternativa, caratterizzata dalle probabilità $(p_1, p_2, \ldots, p_k)$, dalle probabilità $(\pi_1, \pi_2, \ldots, \pi_k)$ dell'ipotesi di nullità, utilizzando la funzione $\triangle = \max \left\{ \frac{|p_i - \pi_i|}{\pi_i} ; \ i = 1, 2, \ldots, k \right\}$.

In [1] si è anche proceduto alla determinazione delle numerosità campionarie necessarie, impiegando il consueto test basato sulla statistica X^2 di Karl-Pearson, per avere valori piccoli di α (probabilità di errori di prima specie) e di max β (β è la probabilità di errori di seconda specie; max β è riferito alle alternative con $\triangle \geqslant \delta$, essendo $\delta > 0$). Una caratteristica singolare riscontrata dall'esame di dette numerosità campionarie è che esse aumentano enormemente all'aumentare di k.

In questo lavoro si effettueranno dei confronti fra $\triangle$ e la consueta funzione di distanza $H^2 = \Sigma \frac{(p_i - \pi_i)^2}{\pi_i}$, e fra $\triangle$ e la funzione $\Gamma = \Sigma \, | p_i - \pi_i |$. In particolare si mostrerà come varia H^2 sul luogo dei punti ω_δ dello spazio parametrico Ω in cui $\triangle = \delta$, nonché come varia Γ su ω_δ. Si ricaveranno anche le variazioni di $\triangle$ sul luogo dei punti ω_η dello spazio parametrico in cui $H = \eta$, nonché sul luogo dei punti ω_γ in cui $\Gamma = \gamma$. Nel presente lavoro si verificherà inoltre se il test basato su X^2 è adeguato quando si vogliono caratterizzare le alternative con la funzione $\triangle$. Partendo dalla constatazione che su ω_δ la differenza fra max H^2 e min H^2 cresce al crescere di k, si riuscirà a spiegare anche il perché del considerevole aumento, all'aumentare di k, delle numerosità campionarie riscontrate in [1]. Ciò sarà sufficiente per ritenere inadeguato il test basato su X^2 nel caso in cui sia necessario caratterizzare le alternative, invece che con la consueta funzione H^2, tramite la funzione $\triangle$. Il lavoro porta così alla conclusione che in questo ultimo caso sarà forse meglio basare il test su una statistica analoga a $\triangle$ come è ad esempio la statistica

$$Z = \max \left\{ \frac{| n_i - n\pi_i |}{\sqrt{n\pi_i}} ; \ i = 1, 2, \ldots, k \right\};$$

* Lavoro sostenuto finanziariamente anche con il supporto del ministero della Pubblica Istruzione.

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

dove n è la numerosità campionaria ed n_i indica la frequenza osservata nella classe i^{ma} della multinomiale.

2. *Funzioni di distanza*

Si abbia una multinomiale a k classi con probabilità $(p_1, p_2, \ldots, p_k)$ e con numero di prove n. Lo spazio parametrico Ω può considerarsi un sottospazio dello spazio euclideo k-dimensionale E^k le cui coordinate $(p_1, \ldots, p_k)$ sono tali che $\sum_1^k p_i = 1, p_i \geqslant 0 \; (i = 1, 2, \ldots, k)$.

Per misurare la distanza di un generico punto $(p_1, p_2, \ldots, p_k)$ da un prefissato punto di Ω, le cui coordinate si indicheranno con $(\pi_1, \pi_2, \ldots, \pi_k)$, esistono diversi criteri. Il più impiegato consiste nel calcolare la media quadratica ponderata delle differenze relative assolute $\zeta_i = \dfrac{|p_i - \pi_i|}{\pi_i}$ con pesi π_i, data dall'espressione

$$H = \sqrt{\Sigma \left(\frac{p_i - \pi_i}{\pi_i}\right)^2 \cdot \pi_i} = \sqrt{\Sigma \frac{(p_i - \pi_i)^2}{\pi_i}} \tag{1}$$

Un altro tipo di media delle differenze relative assolute potrebbe essere la media aritmetica ponderata data dall'espressione

$$\Gamma = \Sigma \, \zeta_i \cdot \pi_i = \Sigma \, |p_i - \pi_i| \, . \tag{2}$$

Un criterio del tutto diverso dalla (1) e dalla (2) è stato introdotto in [1] ed è dato dalla funzione

$$\triangle = \max \{ \zeta_i ; \; i = 1, 2, \ldots, k \} \, . \tag{3}$$

Dal punto di vista formale tutti e tre i criteri sono ben definiti. La (1) e la (2), esendo delle medie, non tengono conto dei valori che possono essere assunti dalle singole ζ_i in corrispondenza ad un valore prefissato, rispettivamente, di H e di Γ. Vedremo in seguito che le singole ζ_i possono assumere valori di gran lunga più elevati dei valori prefissati di H e di Γ.
La (3) effettua un controllo delle singole deviazioni delle p_i da π_i. Il significato della (3) è immediato. Infatti, se $\triangle \geqslant \delta$ significa che almeno una differenza relativa assoluta è maggiore o uguale a δ. Se $\triangle = \delta$ significa che vi è almeno una ζ_i uguale a δ e che nessuna è maggiore di δ. Se $\triangle < \delta$ significa che tutte le differenze relative assolute sono inferiori a δ.

3. *Variazioni di* H *e di* Γ *su* ω_δ

Si indichi con $\zeta_{(1)}$ e con $\zeta_{(k)}$ rispettivamente la più piccola e la più grande fra le ζ_i. Dalle proprietà delle medie potenziate si deduce che

$$\zeta_{(1)} \leqslant \Gamma \leqslant H \leqslant \zeta_{(k)} = \triangle \tag{4}$$

Si considerino ora i punti dello spazio parametrico giacenti su ω_δ. Dalla (4) si ricava che su ω_δ si ha l'uguaglianza $\Delta = \delta$ e che vale la relazione $\Gamma \leqslant H \leqslant \delta$.

Per studiare le variazioni di Γ e H su ω_δ faremo, per semplicità, l'ipotesi che sotto H_0 si abbia

$$\pi_1 = \pi_2 = \ldots = \pi_k = \frac{1}{k}\,.$$

Cominciamo ad esaminare le variazioni di Γ su ω_δ. Dalla (2) si ricava la scomposizione che segue:

$$\Gamma = \Sigma\,|p_i - 1/k| = \sum_{(p_i \geqslant 1/k)} (p_i - 1/k) + \sum_{(p_i < 1/k)} (1/k - p_i)\,.$$

Dovendo essere $\sum_1^k (p_i - 1/k) = 0$, si avrà

$$\sum_{(p_i \geqslant 1/k)} (p_i - 1/k) = - \sum_{(p_i < 1/k)} (p_i - 1/k)\,.$$

Per comodità di scrittura si ponga ora:

$$\Gamma_1 = \sum_{(p_i \geqslant 1/k)} (p_i - 1/k) \quad \text{e} \quad \Gamma_2 = \sum_{(p_i < 1/k)} (1/k - p_i)\,.$$

Risulta pertanto $\Gamma_1 = \Gamma_2$ ed inoltre $\Gamma = \Gamma_1 + \Gamma_2$. (5)
Pertanto massimizzare (minimizzare) Γ equivale a massimizzare (minimizzare) Γ_1 oppure Γ_2.

Minimo di Γ su ω_δ

Per studiare l'andamento di Γ su ω_δ bisogna tener presente che almeno uno scarto relativo $|p_i - 1/k| \,/\, \frac{1}{k}$ è uguale a δ. Dal momento che siamo alla ricerca del minimo di Γ, si ipotizzerà di avere, senza perdere in generalità, un solo scarto relativo pari a δ e inoltre si ipotizzerà che esso sia positivo.

Ne consegue che $\min\limits_{\omega_\delta} \Gamma_1 = \frac{1}{k}\,\delta$ e che si raggiunge se non vi sono altri scarti positivi.

Dall'uguaglianza $\Gamma_1 = \Gamma_2$ consegue che anche $\min\limits_{\omega_\delta} \Gamma_2 = \frac{1}{k}\,\delta$ e quindi

$$\min_{\omega_\delta} \Gamma = \frac{2}{k}\,\delta\,. \tag{6}$$

Si noti che l'uguaglianza $\Gamma_2 = \sum_{(p_i < 1/k)} (1/k - p_i) = \frac{1}{k}\delta$ si può conseguire in diversi modi. Ad esempio, si può avere una sola differenza $(1/k - p_i)$ uguale a $\frac{1}{k}\delta$ oppure avere $(k-1)$ scarti $(1/k - p_i) \geqslant 0$ la cui somma è $\frac{1}{k}\delta$. Si noti inoltre che $\min_{\omega_\delta} \Gamma$ è inversamente proporzionale al numero k delle classi della multinomiale.

Massimo di Γ su ω_δ

Bisogna distinguere fra k pari e k dispari. Se k è pari su ω_δ è possibile ottenere che tutti gli scarti relativi siano uguali al massimo consentito δ. Ciò si realizza con k/2 scarti relativi positivi e k/2 scarti relativi negativi. Consegue che per k pari $\max_{\omega_\delta} \Gamma = \delta$. Per k dispari non si possono avere k scarti relativi assoluti uguali a δ. Dal vincolo $\Gamma_1 = \Gamma_2$ e dal vincolo $\frac{|p_i - 1/k|}{\frac{1}{k}} \leqslant \delta$ deriva che $\max_{\omega_\delta} \Gamma_1 = \frac{k-1}{2}\frac{\delta}{k}$ e che si può raggiungere se vi sono almeno $\frac{k-1}{2}$ scarti positivi. Supponiamo per assurdo che $\Gamma_1 > \frac{k-1}{2} \cdot \frac{\delta}{k}$. Ciò si può realizzare se vi sono almeno $\frac{k-1}{2} + 1$ scarti positivi. Ne consegue che devono esservi allora al più $\frac{k-1}{2}$ scarti negativi. Quand'anche ciascuno di questi raggiungesse in valore assoluto il massimo consentito δ, la loro somma sarebbe inferiore a Γ_1, ma ciò non è possibile in quanto $\Gamma_1 = \Gamma_2$.

In definitiva $\max_{\omega_\delta} \Gamma_1 = \frac{k-1}{2} \cdot \frac{\delta}{k}$ e quindi $\max_{\omega_\delta} \Gamma = \frac{k-1}{k}\delta$. (7)

Deriva da tutto quanto precede che su ω_δ la funzione Γ oscilla da un valore minimo di $\delta \cdot \frac{2}{k}$ ad un valore massimo pari a δ o $\delta\frac{k-1}{k}$. Si noti come all'aumentare di k la differenza fra $\max_{\omega_\delta} \Gamma$ e $\min_{\omega_\delta} \Gamma$ aumenti.

Si analizzerà ora come varia H^2 su ω_δ. Nel caso in cui $\pi_i = 1/k$ $(i = 1, 2, \ldots, k)$