QUALCHE OSSERVAZIONE SULLA STATISTICA GOODNESS-OF-FIT DI K.PEARSON

G.Pederzoli
Libero Istituto Universitario Cattaneo
Castellanza

1. Introduzione

L'articolo originale di K. Pearson, pubblicato nel 1900 con un titolo piuttosto lungo, ha avuto un notevole successo. Si tratta, come tutti sanno, di un test statistico per la verifica di un'ipotesi semplice multinominale. Questa statistica, universalmente detta "goodness-of-fit", è approssimativamente distribuita secondo una chi-quadrato quando la numerosità del campione è grande.

L'estensione ad ipotesi composite contiene, come afferma Lehmann (1959), un errore nel numero dei gradi di libertà della distribuzione limite. La soluzione corretta per il caso generale è stata trovata più tardi da Fisher (1924) che tra le altre cose si è occupato della dipendenza di questa distribuzione dei metodi impegnati nella stima dei parametri.

Tra i tanti lavori sulla statistica di Pearson ricordiamo quello di Neyman (1949) su i test chi-quadrato con alternative limitate. Una panoramica di questi metodi è contenuta nel libro di Cramér (1946), negli articoli di Cochran (1952, 1954) e nel capitolo 30 del volume di Kendall and Stuart (1978) dove sono anche fornite altre fonti bibliografiche.

Per quanto riguarda la distribuzione delle forme quadrate e relative proprietà si veda Mathai and Provost (1992).

2. La statistica di Pearson

Si consideri la legge di probabilità multinominale

$$f(x_1,\ldots,x_k) = \frac{n!}{x!\ldots x_k!} p_1^{x_1}\ldots p_k^{x_k}$$

per $0 < p_j < 1$ con $x_j = 0,1,\ldots,n$, dove $j = 1,\ldots,k$, $p_1 + \ldots + p_k = 1$, $x_1 + \ldots + x_k = n$ e $f(x_1,\ldots,x_k) = 0$ altrimenti. Ne segue che $f(x_1,\ldots,x_k)$ è una funzione di probabilità discreta di $(k-1)$ variabili con parametri $n, p_1,\ldots,p_k$. In particolare si ha che $E(x_j) = np_j$, $Var(x_j) = np_j(1-p_j)$, $Cov(x_i,x_j) = -p_ip_j$, $i \neq j$ per $i,j = 1,\ldots,k$.

La statistica goodness-of-fit di Pearson, che indicheremo con U^2, è data da

$$U^2 = \sum \frac{(x_j - np_j)^2}{np_j}$$

Per la verifica dell'ipotesi $H_0 : p_j = p_j^0$, $j = 1,\ldots,k$ questa viene rifiutata qualora la discrepanza tra il vettore osservato $(x_1,\ldots,x_k)$ e il corrispondente vettore atteso sotto H_0, vale a dire $(np_1^0,\ldots,np_k^0)$ risulti ampia. Per una definizione assiomatica di U^2 quale misura di discrepanza si veda l'articolo di Kaufman, Mathai and Rathie (1972). Il test impiegato sfrutta la proprietà che U^2 è approssimativamente distribuito secondo una chi-quadrato come $(k-1)$ gradi di libertà quando n è grande.

Quando $k = 2$ il risultato è ovvio in quanto per $x_2 = n - x_1$ e $p_2 = 1 - p_1$ si ottiene

$$U^2 = \sum_{j=1}^{2} \frac{(x_j - np_j)^2}{np_j} = \frac{(x_1 - np_1)^2}{np_1(1-p_1)} = u^2$$

dove $u = \frac{x_1 - np_1}{\sqrt{np_1(1-p_1)}}$, vale a dire una variabile x_1 standardizzata. Per il teorema del limite centrale si ha che $u \to N(0,1)$ quando $n \to \infty$ e perciò $u^2 \to X_1^2$ con una buona approssimazione per $n \geq 20$, $np_1 \geq 5$, $n(1-p_1) \geq 5$.

3. La distribuzione Limite.

Esistono diversi metodi per dimostrare che in generale $U^2 \to X^2_{k-1}$ quando $n \to \infty$. Scopo di questo lavoro è di proporre una derivazione particolarmente semplice di questo risultato importante che non richiede l'inversione della matrice di covarianza.

La versione multivariata del teorema del limite centrale implica che

$$Q = (x_1 - np_1, \ldots, x_{k-1} - np_{k-1}) \begin{bmatrix} np_1(1-p_1) & -np_1p_2 \cdots & -np_1p_{k-1} \\ \vdots & & \vdots \\ -np_{k-1}p_1 & \cdots & np_{k-1}(1-p_{k-1}) \end{bmatrix}^{-1} \begin{bmatrix} x_1 - np_1 \\ \vdots \\ x_{k-1} - np_{k-1} \end{bmatrix}$$

tende a X^2_{k-1} al crescere di n, dove $x_k = n - x_1 - \ldots - x_{k-1}$ e $p_k = 1 - p_1 - \ldots - p_{k-1}$. Per mostrare che Q non è altro che la statistica di Pearson definiamo una matrice diagonale D i cui elementi sono $\sqrt{p_j}$, $j = 1, \ldots, k-1$, vale a dire tale che $D = diag\left(\sqrt{p_j}, \ldots, \sqrt{p_{k-1}}\right)$. Allora l'inversa della matrice di covarianza può scriversi come segue

$$\begin{bmatrix} np_1(1-p_1) & -np_1p_2 \cdots & -np_1p_{k-1} \\ \vdots & & \vdots \\ -np_{k-1}p_1 & & np_{k-1}(1-p_{k-1}) \end{bmatrix}^{-1} =$$

$$= (\sqrt{n}D)^{-1} \begin{bmatrix} 1-p_1 & -\sqrt{p_1p_2} \cdots & -\sqrt{p_1p_{k-1}} \\ & & \\ -\sqrt{p_{k-1}p_1} & \cdots & 1-p_{k-1} \end{bmatrix} (\sqrt{n}D)^{-1}$$

[Facciamo presente che il metodo proposto da Rao (1973) suggerisce di impiegare direttamente l'inversa della matrice $Diag\,(p_1, \ldots, p_{k-1}) - pp'$ dove $p = (p_1, \ldots, p_{k-1})'$].

Cambiando $(\sqrt{n}D)^{-1}$ con i vettori della forma quadratica si ottiene

$$Q = \left(\frac{x_1 - np_1}{\sqrt{np_1}}, \ldots . \frac{x_{k-1} - np_{k-1}}{\sqrt{np_{k-1}}}\right) \begin{bmatrix} 1-p_1 & -\sqrt{p_1p_2}\cdots & -\sqrt{p_1p_{k-1}} \\ & & \\ -\sqrt{p_{k-1}p_1} & \cdots & 1-p_{k-1} \end{bmatrix}^{-1} \begin{bmatrix} \frac{x_1-np_1}{\sqrt{np_1}} \\ \vdots \\ \frac{x_{k-1}-np_{k-1}}{\sqrt{np_{k-1}}} \end{bmatrix}$$

scrivendo

$$\left(I - CC^t\right)^{-1} = \begin{bmatrix} 1-p_1 & -\sqrt{p_1p_2}\cdots & -\sqrt{p_1p_{k-1}} \\ & & \\ -\sqrt{p_{k-1}p_1} & \cdots & 1-p_{k-1} \end{bmatrix}^{-1}$$

con

$$C = \begin{bmatrix} \sqrt{p_1} & 0 & \cdots & 0 \\ 0 & \sqrt{p_2} & \cdots & 0 \\ \vdots & & & \vdots \\ 0 & 0 & \cdots & \sqrt{p_{k-1}} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \vdots \\ 1 \end{bmatrix}$$

ciò implica che $\|CC^t\| \leq \|C\| \, \|C^t\| = \|C\|^2 < 1$ dove $\|C\|$ denota una norma di C. Basta poi prendere $\|C\|^2 =$ il più grande dei $p_j < 1$. Ne segue che può esprimersi come una serie convergente di potenza, vale a dire

$$\left(I - CC^t\right) = I + \left(CC^t\right) + \left(CC^t\right)^2 + \cdots$$

Impiegheremo proprio questo sviluppo in serie per evitare l'inversione della matrice delle covarianze, ed al tempo stesso calcolare il valore esatto delle statistiche. Sostituendo nell'espressione di Q il primo termine risulta uguale a

$$\sum_{j=1}^{k-1} \frac{\left(x_j - np_j\right)^2}{np_j}$$

Si consideri poi

$$\left(\frac{x_1 - np_1}{\sqrt{np_1}}, \cdots, \frac{x_{k-1} - np_{k-1}}{\sqrt{np_{k-1}}}\right) C = \frac{(x_1 - np_1) + \cdots + (x_{k-1} - np_{k-1})}{\sqrt{n}}.$$

Il secondo termine in Q è dato da

$$\frac{1}{n}\left[(x_1 - np_1) + \cdots + (x_{k-1} - np_{k-1})\right]^2 = \frac{(x_k - np_k)^2}{n}$$

Tuttavia $(CC^t)^2 = CC^tCC^t = C(C^tC)C^t = (p_1 + \cdots + p_{k-1})CC^t = (1 - p_k)CC^t$ e di conseguenza $(CC^t)^r = (1 - p_k)^{r-1} CC^t$.

Pertanto, sommando i termini di una serie geometrica di ragione $(1 - p_k)$, si ha che

$$\sum_{r=2}^{\infty} (CC^t)^r = \left[\sum_{r=1}^{\infty} (1 - p_k)\right] CC^t = \left(\frac{1 - p_k}{p_k}\right) CC^t$$

Ne consegue che la somma di tutti i termini in Q, dal secondo in poi, risulta la seguente:

$$\frac{(x_k - np_k)^2}{n}\left[1 + \frac{(1 - p_k)}{p_k}\right] = \frac{(x_k - np_k)^2}{np_k}.$$

e cioè la statistica di Pearson per tutti i valori di k, e il risultato risulta dimostrato.

References

Cochran, W.G. (1952). The χ^2 test of goodness of fit, *Ann. Maht. Stat.*, 2, 315-345

Cochran, W.G. (1954). Some methods for strengthening the common χ^2 tests, *Biometrics*, 10, 417-451

Cramér, H. (1946). *Mathematical Methods of Statistics,* Princeton University Press, Princeton

Fisher, R.A. (1924). The conditions under which chi square measures the discrepancy between observation and hypothesis, *J. Roy Stat. Soc.*, 87, 442-450

Kaufman, H., Mathai, A.M. and Rathie, P.N. (1972). A mathematical foundation for Pearson's chi square goodness-of-fit statistic. *Sankhya. Series A*, 34, 441-442.

Kendall, M.G. and Stuart, A. (1978). *The Advanced Theory of Statistics*, Charles Griffin and Co., fourth edition, London

Lehmann, E.L. (1959). *Testing Statistical Hypotheses,* J.Wiley & Sons, Inc., New York

Mathai, A.M. and Provost, S.B. (1992). *Quadratic Forms i Random Variables: Theory and Applications*, Marcel Dekker, New York.

Neymann, J. (1949). Contribution to the theory of the χ^2 test, *Proc. Berkeley Symposium on Mathematical Statistics and Probability,* University of California Press, Berkeley, 239-273

Pearson, K. (1900). On a criterion that a given system of derivations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen in random sampling, *Phil. Mag.*, Ser. 5, 50, 157-172

Rao, C.R. (1973). *Linear Statistical Inference and Its Applications*, Wiley, New York.