17. Test chi-quadrato per tabelle di contingenza

Per applicare il test chi-quadrato (χ^2) nella forma generalizzata per tabelle di contingenza, che è quella qui impiegata, è necessario che ciascun elemento del campione in esame possa essere classificato per una caratteristica in un numero R di classi, e per una seconda caratteristica in un numero C di classi, in modo tale che i dati possano essere organizzati in una tabella di R righe per C colonne, comprendente quindi un totale di $N = R \cdot C$ celle.

Essendo allora O il valore osservato in una data cella e A il valore atteso per la stessa cella, il test χ^2 viene calcolato come somma dei rapporti $(O - A)^2 / A$ per tutte le celle della tabella, cioè come

$$\chi^2 = \Sigma (O - A)^2 / A$$

Nel caso delle tabelle 1 x 2 e 2 x 2 viene raccomandata la correzione di Yates per la continuità, applicando la quale il test χ^2 viene calcolato come somma dei rapporti $(|O - A| - 1/2)^2 / A$ per tutte le celle della tabella, cioè come

$$\chi^2 = \Sigma (|O - A| - 1/2)^2 / A$$

Il valore di O per ciascuna cella è noto, essendo come detto O il valore osservato. Il valore di A, cioè il valore atteso, non è noto, ma può essere specificato qualora si operi alla luce di una ben definita ipotesi riguardante i dati. Nel caso particolare dell'ipotesi "non vi è differenza fra le frequenze osservate", cioè di quella che gli statistici chiamano la "ipotesi nulla", e che viene qui impiegata, il valore atteso A può essere stimato, e assume un valore pari al prodotto del totale della riga per il totale della colonna cui la cella appartiene diviso per il totale n dei casi osservato, ovvero

$$A = (\text{totale della riga}) \cdot (\text{totale della colonna}) / n$$

Per illustrare le modalità di sviluppo dei calcoli si consideri il caso più semplice, quello della seguente tabella 2 x 2

O_1	Of_2
O_3	Of_4

in il valore osservato per ciascuna delle quattro celle è indicato rispettivamente con O_1, O_2, O_3 e O_4.

Indicando allora:

→ il totale della riga 1 con R_1 ($R_1 = O_1 + O_2$);
→ il totale della riga 2 con R_2 ($R_2 = O_3 + O_4$);
→ il totale della colonna 1 con C_1 ($C_1 = O_1 + O_3$);
→ il totale della colonna 2 con C_2 ($C_2 = O_2 + O_4$);
→ il totale dei casi con n ($n = O_1 + O_2 + O_3 + O_4$);

i quattro valori di A attesi

A_1	A_2
A_3	A_4

corrispondenti ai valori di *f* osservati saranno per definizione uguali rispettivamente a

$$A_1 = C_1 \cdot R_1 / n \qquad A_2 = C_2 \cdot R_1 / n$$
$$A_3 = C_1 \cdot R_2 / n \qquad A_4 = C_2 \cdot R_2 / n$$

essendo ancora ovviamente $A_1 + A_2 + A_3 + A_4 = n$, mentre il valore di χ^2 sarà pari a

$$\chi^2 = (O_1 - A_1)^2 / A_1 + (O_2 - A_2)^2 / A_2 + (O_3 - A_3)^2 / A_3 + (O_4 - A_4)^2 / A_4$$

con $(R - 1) \cdot (C - 1)$ gradi di libertà.

E' quindi facile estendere i calcoli dalla tabella 2 x 2 a tabelle di qualsiasi estensione.

Il valore di *p* corrispondente alla statistica χ^2 rappresenta la probabilità di osservare per caso una differenza tra frequenze osservate e frequenze attese della grandezza di quella effettivamente osservato: se tale probabilità è sufficientemente piccola, si conclude per una differenza significativa di incidenza nei diversi gruppi del fattore in esame.