

# Analisi delle corrispondenze

Giovanna Jona Lasinio  
giovanna.jonalasinio@uniroma1.it

Dipartimento di Scienze Statistiche - Università *Sapienza* Roma

- Oggetto dell'analisi delle corrispondenze (AC) sono insiemi di dati qualitativi (matrici faunistiche ad esempio).
- Il principio su cui si basa l'AC è lo stesso su cui si basa l'analisi in componenti principali. La PCA usa come misura di variabilità la correlazione o la covarianza, l'AC utilizza una distanza tra righe (o colonne) della matrice dei dati, basata sulla statistica  $\chi^2$ .
- Fine dell'analisi è spiegare perché la matrice dei dati si scosta da una situazione di omogeneità (righe o colonne proporzionali).

- **Matrice dei dati X**

Siti/Specie	$X_1$	...	$X_k$	Tot
$S_1$	$n_{11}$	...	$n_{1k}$	$n_{1.}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$S_p$	$n_{p1}$	...	$n_{pk}$	$n_{p.}$
Tot	$n_{.1}$	...	$n_{.k}$	$n$

- Possiamo estrarre informazione in diversi modi da questa matrice tenendo conto di diversi aspetti
  - a) **confrontiamo i siti tra loro** e per farlo dobbiamo tener conto della diversa numerosità nei siti
  - b) **confrontiamo le specie nei siti**, dobbiamo considerare l'effetto della diversa numerosità delle specie

## a) Matrice dei profili riga

Siti/Specie	$X_1$	...	$X_k$	Tot
$S_1$	$pr_{11} = \frac{n_{11}}{n_{1.}}$	...	$pr_{1k} = \frac{n_{1k}}{n_{1.}}$	1
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$S_p$	$pr_{p1} = \frac{n_{p1}}{n_{p.}}$	...	$pr_{pk} = \frac{n_{pk}}{n_{p.}}$	1
Tot	$pr_{.1} = \frac{n_{.1}}{n}$	...	$pr_{.k} = \frac{n_{.k}}{n}$	-

Questa matrice corrisponde ad una ponderazione delle righe di  $\mathbf{X}$  con l'abbondanza totale di individui nei siti campionati. Indicheremo con  $\mathbf{pr}_i$  la riga  $i$ -esima della matrice dei profili riga (il profilo riga  $i$ -esimo).

## b) Matrice dei profili colonna

Siti/Specie	$X_1$	...	$X_k$	Tot
$S_1$	$pc_{11} = \frac{n_{11}}{n_{.1}}$	...	$pc_{1k} = \frac{n_{1k}}{n_{.1}}$	$pc_{.1} = \frac{n_{1.}}{n}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$S_p$	$pc_{p1} = \frac{n_{p1}}{n_{.1}}$	...	$pc_{pk} = \frac{n_{pk}}{n_{.k}}$	$pc_{.p} = \frac{n_{p.}}{n}$
Tot	1	...	1	-

Questa matrice corrisponde ad una ponderazione delle colonne di  $\mathbf{X}$  con l'abbondanza totale di individui di ciascuna specie osservati nell'area di studio. Indicheremo con  $\mathbf{pc}_j$  la colonna  $j$ -esima della matrice dei profili colonna (il profilo colonna  $j$ -esimo)

**Come valutare correttamente la distanza,  
nello spazio delle specie, tra due specie o  
due siti?**



**Distanza del  $\chi^2$**

In particolare:

$$\begin{aligned}
 d(\mathbf{pr}_i, \mathbf{pr}_h) &= \sum_{j=1}^k \frac{1}{pr_{.j}} (pr_{ij} - pr_{hj})^2 \\
 &= \sum_{j=1}^k \frac{1}{\frac{n_{.j}}{n}} \left( \frac{n_{ij}}{n_{.i}} - \frac{n_{hj}}{n_{.h}} \right)^2
 \end{aligned} \tag{1}$$

$$\begin{aligned}
 d(\mathbf{pc}_j, \mathbf{pc}_h) &= \sum_{i=1}^p \frac{1}{pc_{.i}} (pc_{ij} - pc_{ih})^2 \\
 &= \sum_{i=1}^p \frac{1}{\frac{n_{i.}}{n}} \left( \frac{n_{ij}}{n_{j.}} - \frac{n_{ih}}{n_{h.}} \right)^2
 \end{aligned} \tag{2}$$

L'espressione (1) è la distanza tra due profili riga mentre l'espressione (2) è la distanza tra due profili colonna.

- A questo punto si può condurre una PCA sulla matrice dei profili riga o dei profili colonna.
- Otterremo di solito dei risultati più dispersi ma interpretabili nello stesso modo della PCA



## ESEMPI SU R