Università degli Studi di Milano-Bicocca - Facoltà di Economia
Esame di Analisi dei Dati I
29 giugno 2006
NB: Commentare sempre i risultati ottenuti

---

1. Data una matrice $\underset{(n\times p)}{X}$ individui-caratteri, si illustri analiticamente come ricavare la matrice di correlazione tra le p variabili.

2. Con l'obiettivo di valutare l'influenza dello *stato civile* (single, sposati, divorziati e vedovi) sullo stipendio annuo lordo degli uomini occupati di almeno 25 anni della Lombardia, da ciascuna delle 4 popolazioni viene estratto un campione di numerosità n_j $(j = 1, \ldots, 4)$ e viene rilevato lo stipendio annuo lordo X in centinaia di euro. Indicati con:

   - X_{ji} lo stipendio annuo lordo dell'i-mo intervistato della popolazione j;
   - $\overline{X}_{j\cdot}$ lo stipendio medio aritmetico degli n_j individui del campione estratto dalla j-ma popolazione

   la seguente tabella riporta alcune informazioni di sintesi sulle osservazioni campionarie:

Popolazione	j	n_j	$\overline{X}_{j\cdot}$	$S_j^2 = \frac{1}{n_j-1}\sum_{i=1}^{n_j}\left(X_{ji} - \overline{X}_{j\cdot}\right)^2$
Single	1	200	246,2	72,77
Sposati	2	350	278,5	132,25
Divorziati	3	120	266	98,41
Vedovi	4	60	281,1	108,12

   a) si calcolino le stime per gli effetti specifici $\alpha_j = \mu_{j\cdot} - \mu_{\cdot\cdot}$ $j = 1, 2, 3, 4$ imputabili ai livelli del fattore *stato civile* e si commenti uno di essi;
   b) si verifichi (utilizzando un livello di significatività $\alpha = 0,05$) se lo stato civile influisce significativamente sul reddito medio annuo lordo;
   c) si verifichi, ad un livello di significatività $\alpha = 0,05$, la seguente ipotesi:

$$H_0 : \frac{\mu_1 + \mu_2}{2} = \mu_3 \qquad \text{contro} \qquad H_1 : \frac{\mu_1 + \mu_2}{2} \neq \mu_3$$

   interpretando opportunamente il risultato ottenuto;
   d) si costruisca l'intervallo di confidenza al 98% per il seguente contrasto lineare:

$$L = \mu_1 - \frac{\mu_2 + \mu_3 + \mu_4}{3}$$

3. Per sei capoluoghi di provincia della Lombardia si sono considerati i seguenti indicatori ambientali urbani (espressi in Kg per abitante) relativi alla raccolta di rifiuti urbani nell'anno 2003 (Fonte Istat):

   $X_1 =$ raccolta differenziata;
   $X_2 =$ raccolta indifferenziata in cassonetti/recipienti;
   $X_3 =$ raccolta indifferenziata ingombranti (poltrone, divani, mobili, elettrodomestici ecc...);
   $X_4 =$ raccolta indifferenziata da spazzamento stradale;
   $X_5 =$ raccolta indifferenziata altro.

   I dati sono riportati nella seguente tabella:

Provincia	X_1	X_2	X_3	X_4	X_5
Varese	188,8	290,6	46,1	27,6	7,2
Milano	163,0	30,8	11,2	45,8	303,6
Bergamo	218,0	287,0	13,3	38,4	14,9
Brescia	286,6	432,7	1,4	17,9	10,4
Pavia	146,4	349,0	10,4	7,6	63,5
Lodi	171,2	368,8	19,5	11,1	8,0

   La matrice D delle *distanze city-block* calcolata sui dati standardizzati è:

$$D = \begin{bmatrix} 0 & 9,149 & 3,836 & 7,128 & 5,871 & 4,076 \\ 9,149 & 0 & 6,572 & 11,256 & 7,896 & 8,670 \\ 3,836 & 6,572 & 0 & 4,983 & 4,901 & 4,116 \\ 7,128 & 11,256 & 4,983 & 0 & 5,555 & 4,787 \\ 5,871 & 7,896 & 4,901 & 5,555 & 0 & 2,106 \\ 4,076 & 8,670 & 4,116 & 4,787 & 2,106 & 0 \end{bmatrix}.$$

   a) Tracciare il dendrogramma riferito ai sei capoluoghi di provincia avvalendosi del *metodo del legame completo*;
   b) suggerire una opportuna partizione, giustificando la scelta;
   c) descrivere la partizione individuata al punto precedente;
   d) si dica se e come varierebbe la successione delle partizioni ottenuta al punto a se invece di considerare le distanze city-block si utilizzassero i logaritmi delle stesse.

4. Si illustrino i criteri per stabilire quali e quante componenti principali mantenere in un'analisi realizzata su n unità statistiche sulle quali sono state rilevate p variabili.