Università degli Studi di Milano-Bicocca - Facoltà di Economia
Esame di Analisi dei Dati (modulo A)
30 giugno 2005
NB: Commentare sempre i risultati ottenuti

---

1. Sia X un carattere quantitativo presente su quattro popolazioni distinte. É noto che X ha distribuzione normale di valore atteso μ_j $(j = 1, 2, 3, 4)$ e varianza σ^2 identica nelle quattro popolazioni.
   Si è interessati al contrasto lineare:

   $$L = -0,5\mu_1 + \mu_2 - \mu_3 + 0,5\mu_4.$$

   A tale scopo si seleziona da ciascuna delle quattro popolazioni un campione casuale di numerosità n_j $(j = 1, 2, 3, 4)$: nella fattispecie $n_1 = 5$, $n_2 = 7$, $n_3 = 6$ e $n_4 = 5$.

   a) Si consideri lo stimatore *naturale* per L e se ne determini il valore atteso e la varianza;

   b) si ricavi la distribuzione dello stimatore individuato al punto precedente;

   c) si costruisca la statistica test per verificare l'ipotesi nulla $H_0 : L = 0$ contro l'alternativa $H_1 : L \neq 0$ ad un livello di significatività α.

2. Si sono considerate, per alcune marche di autoveicoli, le immatricolazioni di autovetture con alimentazione a benzina (X_1) e diesel (X_2) dell'anno 2004. Di seguito è riportata la matrice dei dati standardizzati.

Marca	Z_1	Z_2
CITROEN	1,329	0,735
OPEL	0,870	1,552
TOYOTA	0,752	0,461
SMART	−0,939	−1,327
HYUNDAI	−0,978	−0,894
NISSAN	−1,035	−0,527

   a) Applicare la procedura di analisi dei gruppi non gerarchica delle k-medie selezionando come centri iniziali le marche OPEL e HYUNDAI;

   b) descrivere la partizione ottenuta calcolando opportuni indici basati sulla scomposizione della devianza totale in devianza nei gruppi e fra i gruppi.

3. Si illustrino le fasi comuni ai metodi gerarchici aggregativi, soffermandosi successivamente sulle diverse definizioni di distanze fra gruppi.

4. Su un campione di 150 single sono state rilevate le seguenti variabili: X_1 = *numero di pranzi/cene fuori casa nei fine settimana (ultimi tre mesi)*; X_2 = *numero di fine settimana trascorsi fuori dal comune di residenza (ultimi tre mesi);* X_3 = *numero di vacanze di almeno 10 giorni (ultimo anno);* X_4 = *spesa media mensile per l'abitazione di residenza.* Di seguito è riportata la matrice $\mathbf{C}$ di correlazione tra le variabili originarie e le componenti principali:

$$\mathbf{C} = \begin{bmatrix} -0,5121 & 0,7503 & 0,3898 & -0,1511 \\ -0,1759 & 0,9233 & -0,2740 & 0,2037 \\ -0,8882 & -0,1830 & -0,3902 & -0,1588 \\ 0,8222 & 0,4672 & -0,2374 & -0,2221 \end{bmatrix}$$

   a) Si determini, mediante opportuni criteri, il numero di componenti principali da mantenere nell'analisi;

   b) si interpretino le componenti principali selezionate al punto precedente;

   c) si calcoli e si commenti il coefficiente di correlazione lineare tra X_1 e X_2;

   d) si rappresentino graficamente le correlazioni tra le variabili originarie e le componenti principali mantenute commentando opportunamente.