Università degli Studi di Milano-Bicocca - Facoltà di Economia
Esame di Analisi dei Dati (modulo A)
4 febbraio 2004
NB: Commentare sempre i risultati ottenuti

---

1. Si illustri il modello lineare associato all'analisi della varianza ad un criterio di classificazione ed il relativo contesto di applicazione, specificando il significato dei parametri del modello stesso.

2. Un'azienda incaricata di gestire l'assistenza per la riparazione dei telefoni cellulari di due differenti marche (*I* e *II*) si serve di tre centri di assistenza (A, B e C). Si vuole verificare se il tempo medio di riparazione (in giorni) dipende dalla marca del telefono cellulare guasto e/o dal centro in cui viene eseguita la riparazione. A tale scopo viene estratto, dalla banca dati di tutte le riparazioni eseguite, un campione di 4 telefoni cellulari per ogni associazione tra centro di riparazione e marca. La seguente tabella riporta i tempi medi di riparazione in giorni:

	Marca		
Centro ass.	*I*	*II*	$\overline{X}_{j..}$
A	26,25	21,75	$\mathbf{24,000}$
B	36,00	34,50	$\mathbf{35,250}$
C	29,25	30,00	$\mathbf{29,625}$
$\overline{X}_{.k.}$	**30,50**	**28,75**	**29,625**

Sapendo che la *devianza nei gruppi* è 411,25 dopo aver specificato le ipotesi necessarie, si verifichi, ad un livello di significatività $\alpha = 0,05$, se:

a) il centro di assistenza influenza significativamente il tempo medio di riparazione;

b) la marca del telefono cellulare influenza significativamente il tempo medio di riparazione;

c) vi è *interazione* tra centro di assistenza e marca del telefono cellulare.

3. Si illustrino le caratteristiche generali dei metodi non gerarchici aggregativi e si presentino i criteri per fissare il numero k di gruppi della partizione e i relativi centri iniziali.

4. Data una matrice X di ordine *(n,p)* riferita alla rilevazione di p variabili quantitative su n unità statistiche, si definisca la prima componente principale e se ne determini la varianza.

CONTINUA SUL RETRO

5. Per nove fastfood del New Jersey, sono state rilevate 4 variabili: X_1 = *numero di impiegati a tempo indeterminato*; X_2 = *numero di impiegati a tempo determinato;* X_3 = *paga media oraria (dollari);* X_4 = *aumento medio della paga oraria dopo un anno dall'assunzione (dollari).* Di seguito è riportata la matrice dei dati:

Fastfood	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$	$\mathbf{X_4}$	**Fastfood**	$\mathbf{X_1}$	$\mathbf{X_2}$	$\mathbf{X_3}$	$\mathbf{X_4}$
A	2,50	2,50	4,75	0,29	**F**	6,50	7,00	4,60	0,27
B	12,50	11,50	4,25	0,29	**G**	3,00	2,00	4,85	0,35
C	7,00	6,50	4,50	0,24	**H**	12,00	12,00	4,35	0,27
D	13,00	11,00	4,50	0,30	**I**	6,00	6,50	4,70	0,23
E	2,00	3,00	5,00	0,31					

La matrice delle *distanze city-block* tra i nove fastfood, calcolata sui dati standardizzati, considerati nello stesso ordine di presentazione della precedente tabella, è:

$$D=\begin{bmatrix} 0 & 7,071 & 4,736 & 6,241 & 1,946 & 3,433 & 2,448 & 7,225 & 3,900 \\ & 0 & 5,249 & 1,655 & 8,747 & 4,803 & 9,277 & 1,282 & 6,668 \\ & & 0 & 4,419 & 6,412 & 1,573 & 6,942 & 4,238 & 1,419 \\ & & & 0 & 7,334 & 3,973 & 7,864 & 2,051 & 5,838 \\ & & & & 0 & 5,109 & 2,343 & 8,901 & 5,575 \\ & & & & & 0 & 5,639 & 3,792 & 1,865 \\ & & & & & & 0 & 9,431 & 6,105 \\ & & & & & & & 0 & 5,657 \\ & & & & & & & & 0 \end{bmatrix}$$

A partire dalla matrice D si è applicato il *metodo gerarchico del legame singolo* (si veda il dendrogramma) fino ad ottenere la partizione P={(D,H,B);(F,I,C);(E,A);G} e la corrispondente matrice D* delle distanze aggiornata:

$$D^*=\left|\begin{array}{cccc|c} (D,H,B) & (F,I,C) & (E,A) & G & \\ 0 & 3,792 & 6,241 & 7,864 & (D,H,B) \\ & 0 & 3,433 & 5,639 & (F,I,C) \\ & & 0 & 2,343 & (E,A) \\ & & & 0 & G \end{array}\right.$$

a) Completare il dendrogramma;

b) suggerire una opportuna partizione, giustificando la scelta;

c) descrivere la partizione individuata al punto precedente.