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

---

1. Analisi della varianza a due criteri di classificazione: si illustri la scomposizione della devianza totale e la conseguente costruzione delle statistiche test necessarie alla verifica delle consuete ipotesi.

2. Una società che effettua indagini di mercato mediante l'invio di questionari postali vuole verificare se diversi incentivi associati ai questionari (Gadget, Concorso a premi o Raccolta Punti) influiscono significativamente sulla percentuale X di questionari ricevuti compilati. A tale scopo vengono selezionati casualmente tre campioni di numerosità, rispettivamente, 5 per le indagini con gadget, 6 per le indagini con concorso a premi e 4 per le indagini con raccolta punti. La seguente tabella riporta le percentuali di questionari ricevuti compilati:

Gadget	50	50	54	60	57	
Concorso a premi	65	63	55	66	50	52
Raccolta punti	47	47	50	48		

Sapendo che la *devianza totale* è $592,93$ dopo aver specificato il modello da utilizzare:

a) si verifichi (utilizzando un livello di significatività $\alpha = 0,05$) se la forma di incentivo associata al questionario influisce significativamente sulla percentuale media di questionari ricevuti compilati;

b) si verifichi, ad un livello di significatività $\alpha = 0,05$, la seguente ipotesi:

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

interpretando opportunamente il risultato ottenuto;

c) costruire l'intervallo di confidenza al 95% per la differenza: $(\alpha_2 - \alpha_3)$.

3. Per sei capoluoghi italiani si sono considerati i seguenti indicatori ambientali: X_1 = *chilometri di piste ciclabili nell'anno 2000 (per 100 kmq di superficie)*; X_2 = *mq di area verde per abitante (anno 2002);* X_3 = *numero di autovetture per kmq (anno 2002).* Di seguito è riportata la matrice dei dati standardizzati.

Comune	Z_1	Z_2	Z_3
Torino	1,979	0,136	1,394
Milano	0,217	-0,219	1,296
Firenze	-0,061	0,113	-0,227
Bologna	-0,143	1,971	-0,599
Campobasso	-1,116	-0,870	-1,270
Bari	-0,876	-1,131	-0,594

a) Applicare la procedura di analisi dei gruppi non gerarchica delle k-medie selezionando come centri iniziali i comuni di Milano e Bari;

b) descrivere la partizione ottenuta.

4. La seguente matrice C riporta i coefficienti di correlazione lineare tra quattro variabili originarie standardizzate (righe) e le quattro componenti principali da esse estratte (colonne):

$$C = \begin{bmatrix} 0,9124 & -0,0490 & 0,1593 & 0,3740 \\ -0,3310 & -0,8324 & -0,3786 & 0,2330 \\ 0,8651 & -0,2304 & -0,3306 & -0,2986 \\ -0,0355 & 0,8861 & -0,4328 & 0,1619 \end{bmatrix}.$$

L'applicazione dei consueti criteri per la scelta del numero di componenti principali da mantenere nell'analisi ha condotto a mantenere le prime due componenti principali.

a) Per le due unità $\mathbf{u}_1$ e $\mathbf{u}_2$ i valori standardizzati delle 4 variabili sono di seguito riportati:

unità	Z_1	Z_2	Z_3	Z_4
$\mathbf{u}_1$	0,6418	-0,1881	0,8116	-0,6823
$\mathbf{u}_2$	0,1245	0,9394	-0,1518	-0,3466

.

Si ricavino i punteggi delle due unità sulle componenti principali mantenute.

b) Si rappresentino graficamente le correlazioni tra le variabili originarie e le prime due componenti principali commentando opportunamente. Si riportino, nel medesimo grafico, i punti corrispondenti alle unità $\mathbf{u}_1$ e $\mathbf{u}_2$ interpretandone la posizione.