COGNOME E NOME …… MATR. ……
ANALISI DEI DATI PER IL MARKETING – novembre 2008.

ESERCIZIO I
Si è applicata l'analisi delle componenti principali a 97 modelli di fotocamere digitali, considerando 7 variabili ed ottenendo la "**matrice di componenti**" riportata di seguito.

	Componente	
	1	2
Risoluz.Max	,797	,332
Display	,014	,926
Larghezza	,925	-,144
Altezza	,944	-,087
Profondità	,659	-,338
Peso	,977	-,081
Prezzo	,857	,280

- Si giudichi la validità dell'analisi con i diversi criteri.
- S'interpreti il significato delle componenti.
- Si disegni il "Grafico delle componenti" (prima parte del *biplot)* e si dica quali sono le variabili che influenzano maggiormente il prezzo delle fotocamere.

ESERCIZIO II
Si è estratto un campione casuale di 8 modelli di fotocamere dalla matrice dei dati utilizzata nell'esercizio I ed al medesimo si è applicata la *cluster analysis* con il metodo del legame completo e la distanza euclidea standardizzata.
Il "programma di agglomerazione" ottenuto con SPSS è riportato di seguito.

- Si costruisca il corrispondente dendrogramma, se ne proponga un taglio ragionevole e si scriva la corrispondente partizione.
- Si descrivano i passi della procedura di SPSS utilizzata **per estrarre il campione** e si dica quale scelta ha dovuto effettuare il ricercatore.

Programma di agglomerazione

	Cluster accorpati			Stadio di formazione del cluster		
Stadio	Cluster 1	Cluster 2	Coefficienti	Cluster 1	Cluster 2	Stadio successivo
1	1	8	1,232	0	0	3
2	3	4	1,594	0	0	4
3	1	2	1,825	1	0	4
4	1	3	1,930	3	2	6
5	5	6	2,989	0	0	6
6	1	5	4,340	4	5	7
7	1	7	6,204	6	0	0

ESERCIZIO III
In un sondaggio pilota per la valutazione della *customer satisfaction* su un campione casuale di 25 utenti d'un *call center* il coefficiente di correlazione tra il tempo di attesa (in minuti secondi) ed il grado di soddisfazione (espresso su una scala da 1 a 10) è risultato uguale a -0.45.

- Si verifichi la significatività della correlazione, scegliendo opportunamente il livello, e si dica a quale conclusione ragionevolmente si perviene.

DOMANDA FACOLTATIVA (rispondere solo dopo aver risolto gli altri esercizi)
Si calcoli a quanto dovrebbe ammontare la numerosità (**approssimata**) del campione affinché tale valore numerico del coefficiente di correlazione (-0.45) risulti significativo al livello dell'1 per mille.

TRACCIA DELLE RISOLUZIONI

ESERCIZIO I (Volume Zani – Cerioli, pp. 230-239; 247-250)

Per giudicare la validità **complessiva** dell'analisi occorre calcolare la percentuale di varianza totale spiegata dalle prime due componenti principali. Effettuando la somma dei quadrati per ciascuna colonna della "matrice di componenti" si ottengono i primi due autovalori: $\lambda_1 = 4.505$ e $\lambda_2 = 1.195$, che sono entrambi maggiori di 1. La percentuale di varianza totale spiegata, 81.431% è sensibilmente maggiore del livello di soglia: $0.95^7 = 69{,}83\%$, per cui l'analisi è complessivamente valida. I restanti autovalori riportati nella tabella di output di SPSS non possono essere calcolati dallo studente in base ai dati forniti nell'esercizio.

Varianza totale spiegata

Componente	Autovalori iniziali			Pesi dei fattori non ruotati		
	Totale	% di varianza	% cumulata	Totale	% di varianza	% cumulata
1	4,505	64,352	64,352	4,505	64,352	64,352
2	1,195	17,078	81,431	1,195	17,078	81,431
3	,631	9,011	90,441			
4	,359	5,134	95,575			
5	,174	2,488	98,063			
6	,100	1,431	99,494			
7	,035	,506	100,000			

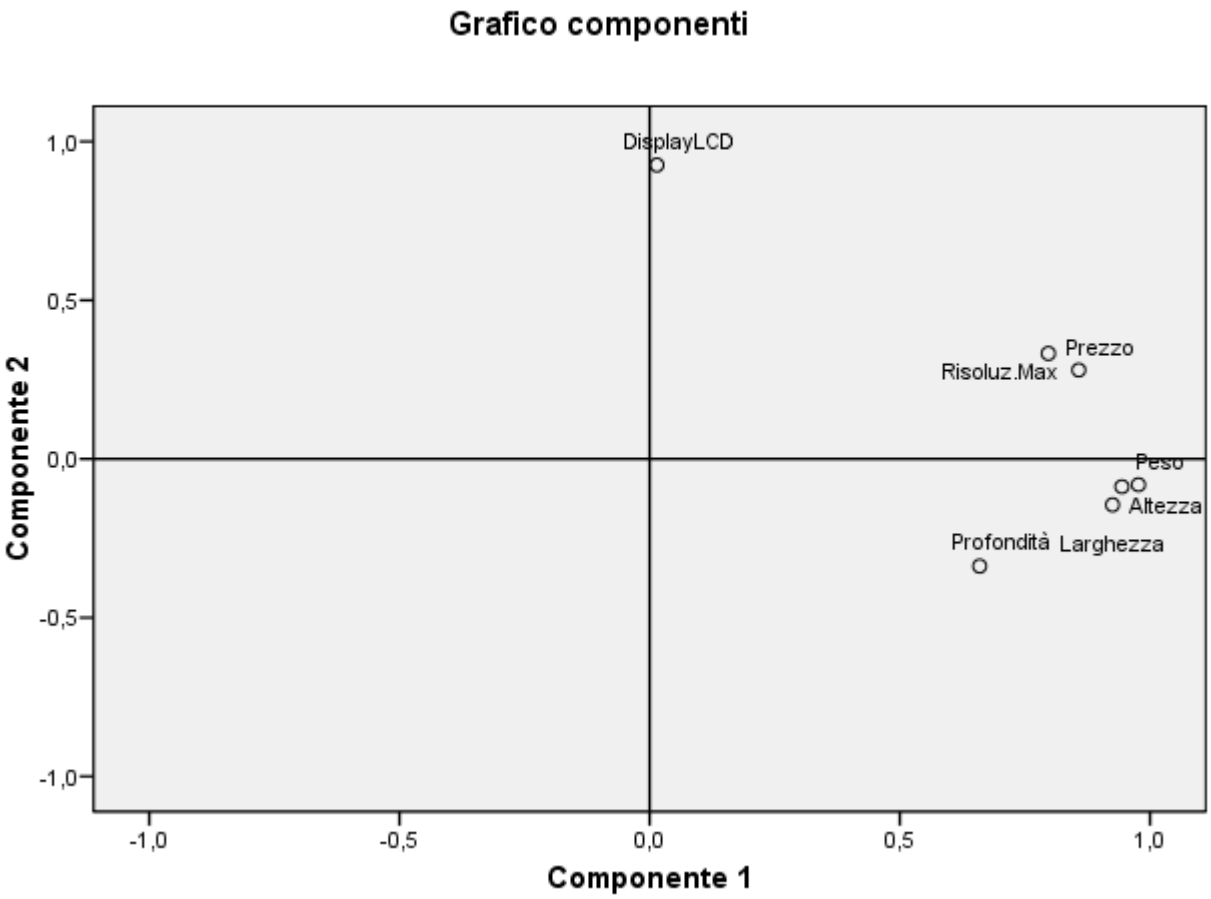
La validità **con riferimento a ciascuna variabile** è fornita dalla corrispondente comunalità, che si ottiene calcolando la somma dei quadrati per riga della "matrice di componenti". La variabile "profondità" è spiegata meno bene rispetto alle altre variabili, ma comunque più del 50%.

Comunalità

	Iniziale	Estrazione
Risoluz.Max	1,000	,745
Display	1,000	,858
Larghezza	1,000	,876
Altezza	1,000	,898
Profondità	1,000	,549
Peso	1,000	,961
Prezzo	1,000	,813

b) la prima componente è un indicatore sintetico di tutte le variabili, ad eccezione del Display. La seconda componente è espressione del solo Display.

c) Il grafico delle componenti si ottiene riportando in un diagramma cartesiano (avente la prima CP sull'asse delle ascisse e la seconda CP sull'asse delle ordinate) per ciascuna variabile il punto corrispondente ai valori dei due rispettivi coefficienti di correlazione che compaiono nella "matrice di componenti.
La variabile che influenza maggiormente il prezzo delle fotocamere è la risoluzione massima, il cui vettore forma un angolo molto piccolo con il vettore del prezzo, per cui presenta una correlazione assai elevata con detta variabile.



ESERCIZIO II

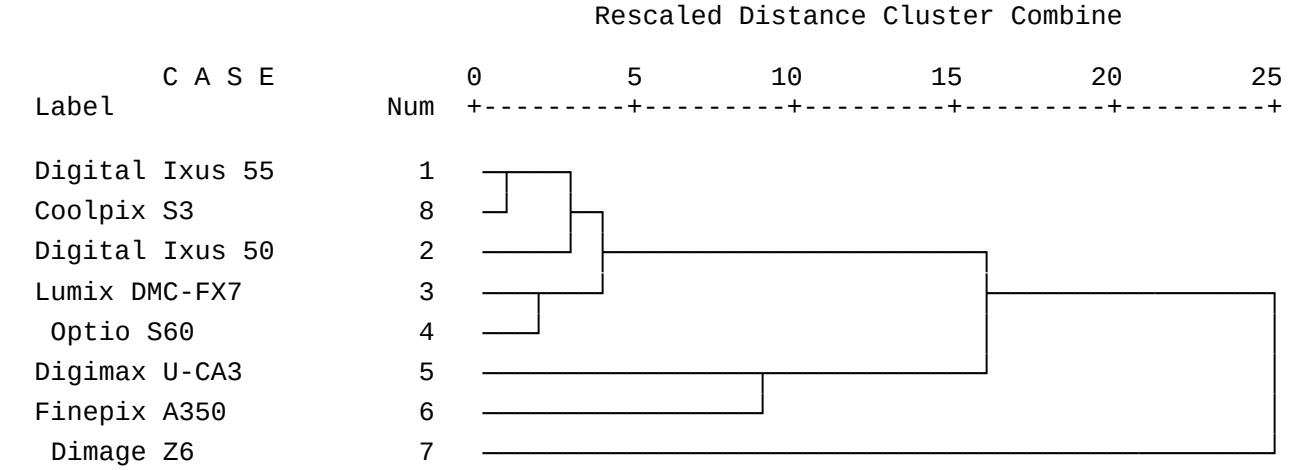
a) Il dendrogramma fornito da SPSS (con distanze riscalate nell'intervallo 0 -25) è riportato di seguito.
Lo studente poteva costruire il dendrogramma anche in funzione dei livelli di distanza originari (coefficienti) riportati nella tabella "Programma di agglomerazione" dell'esercizio.
Il "taglio" più opportuno è dopo il quinto stadio, che corrisponde ad una distanza originaria di circa 3 e ad una distanza riscalata di circa 13. La partizione individuata è la seguente:
(1, 8, 2, 3, 4) (5, 6) (7)
che presenta un gruppo di 5 elementi, un altro gruppo di 2 elementi ed un singolo modello di fotocamera, che può interpretarsi come un *outlier*.

b) I passi della procedura di SPSS per estrarre un campione casuale sono i seguenti (si veda il testo Zani - Cerioli, pag. 33):
dati – seleziona casi – campione casuale di casi – dimensione del campione – approssimativamente … % di tutti i casi.
Il ricercatore ha dovuto quindi scegliere la percentuale di casi, e cioè circa l'8% delle 97 unità di partenza.

▽

* * * * * * H I E R A R C H I C A L C L U S T E R A N A L Y S I S * * * * * *

Dendrogram using Complete Linkage



ESERCIZIO III

Utilizzando la formula (2.4) riportata a pag. 46 del testo Zani – Cerioli, si ottiene:
t_r = -2.417.
Consideriamo detto valore in modulo.
Sulla tavola della variabile aleatoria T di Student, per *g* = 23, si legge:
t(0.05) = 2.069
t(0.02) = 2.500
Il valore campionario t_r si colloca in questo intervallo, per cui il *p-value* corrispondente è compreso tra il 2% e il 5%.
Al livello del 5% è possibile rifiutare l'ipotesi nulla di assenza di correlazione, mentre al livello del 2% (e *a fortiori* al livello dell'1% o dell'1 per mille) non si può rifiutare H_0. Il campione non consente quindi di pervenire ad una conclusione univoca, poiché questa dipende dalla scelta del livello di significatività nell'ambito dei valori di più comune impiego. Si suggerisce di **aumentare la numerosità del campione** per ottenere una risposta più chiara.

DOMANDA FACOLTATIVA

La soluzione dell'esercizio III consente *in primis* di affermare che un coefficiente di correlazione uguale a -0.45 risulterà significativo al livello dell'1 per mille se la numerosità del campione sarà **maggiore di 25**.
Per trovare il valore della numerosità occorre calcolare il valore di *n* tale che:

$$\left| \frac{-0.45}{\sqrt{1-(-0.45)^2}} \sqrt{n-2} \right| > t(0.001)$$

Ma *t(0.001)* è a sua volta funzione di *g* = *n -2*, per cui l'equazione precedente contiene due incognite. Si può allora procedere per tentativi, considerando i valori di *g* riportati sulla tavola.

Per n = 42 e g = 40 si ottiene:

$| t_r | = 3.187 < 3.551$.

Pertanto, una numerosità campionaria uguale a 42 non è sufficiente per rifiutare al livello dell'1 per mille l'ipotesi nulla di assenza di correlazione per un coefficiente uguale a -0,45 .
Per n = 62 e g = 60 si ottiene:

$| t_r | = 3,903 > 3.460$

Una numerosità campionaria **maggiore o uguale a 62** consente senza dubbio di rifiutare l'ipotesi nulla al livello dell'1 per mille.
Volendo ottenere una risposta più precisa si potrebbe procedere per interpolazione lineare tra i valori riportati sulla tavola in corrispondenza di g = 40 e g = 60.
Ad esempio, per g = 50 si ottiene $t(0.001)$ = 3.506 e quindi:

$| t_r | = 3,563 > 3.506$.

Pertanto, già una numerosità campionaria maggiore o uguale a 52 consente di rifiutare l'ipotesi nulla al livello dell'1 per mille.