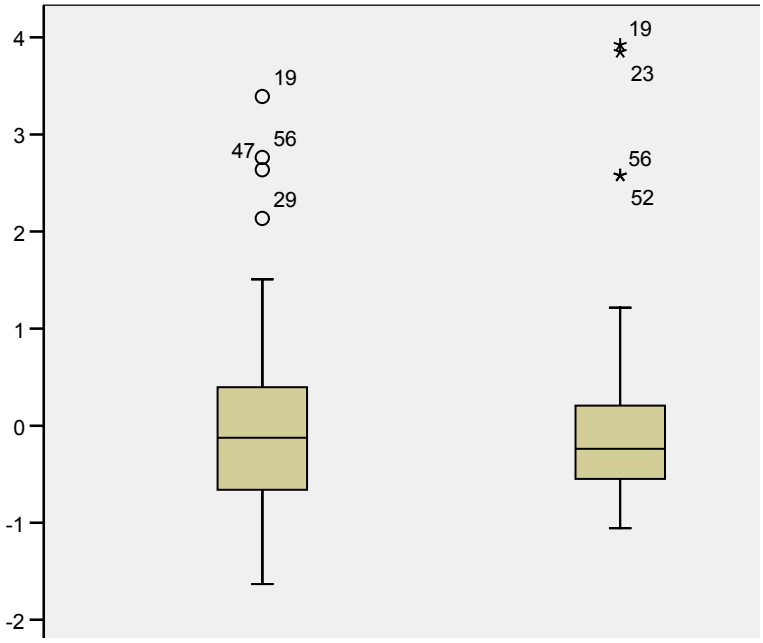
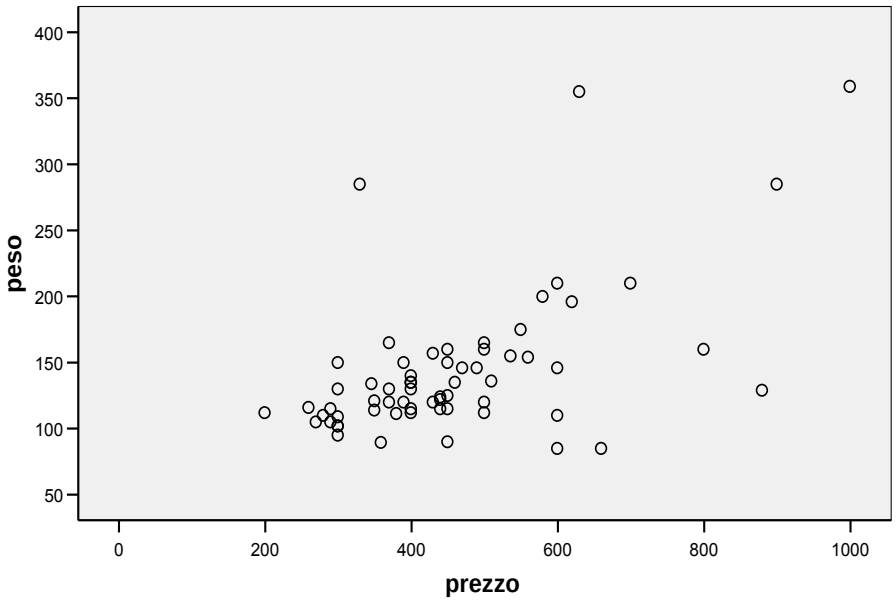
COGNOME E NOME …… MATR. ……

ANALISI DEI DATI PER IL MARKETING – febbraio 2008 - I

1) Per un insieme di modelli di *smartphone* si sono costruiti i *boxplot* degli scostamenti standardizzati del prezzo e del peso ed il diagramma di dispersione delle due variabili.

- Si dica quali informazioni si possono trarre dall'esame congiunto di questi grafici.
- Si descriva la tipologia dei valori anomali che compaiono nel diagramma di dispersione e si illustri la strategia per il trattamento degli stessi ai fini delle analisi multivariate.





2) Al medesimo data set considerato nell'esercizio I, dopo aver effettuato un'operazione di *data cleaning*, si è applicata l'analisi delle componenti principali, considerando 57 modelli e 6 variabili standardizzate. Conoscendo la "matrice di componenti" riportata di seguito:

- si valuti la bontà del modello e si discuta la scelta del numero di componenti;
- si interpreti il significato delle componenti estratte;
- si disegni la parte del *biplot* tracciabile in base ai dati forniti dall'esercizio.

Matrice di componenti(a)

	Componente	
	1	2
prezzo	,847	-,102
peso	,937	-,092
lunghezza	,890	,014
altezza	,811	-,055
dimensione schermo	,828	-,035
durata standby	,242	,969

Metodo estrazione: analisi componenti principali.
a 2 componenti estratti

3) Si è estratto un campione casuale dalla matrice dei dati dell'esercizio 2, applicando a tale campione la *cluster analysis* col metodo del legame completo e la distanza euclidea standardizzata. In base alla tabella seguente:

- si costruisca il corrispondente dendrogramma;
- si proponga un "taglio" ragionevole dello stesso e si scriva la corrispondente partizione.

Programma di agglomerazione

Stadio	Cluster accorpati		Coefficienti	Stadio di formazione del cluster		Stadio successivo
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	3	6	1,253	0	0	2
2	3	4	1,546	1	0	8
3	2	10	2,060	0	0	6
4	1	5	2,085	0	0	7
5	8	9	2,220	0	0	6
6	2	8	2,839	3	5	8
7	1	7	3,330	4	0	9
8	2	3	3,884	6	2	9
9	1	2	4,793	7	8	10
10	1	11	5,882	9	0	0

Traccia di risoluzione

ESERCIZIO I

Il *boxplot* degli scostamenti standardizzati del prezzo segnala che la distribuzione è approssimativamente simmetrica, con valore mediano prossimo a zero e con 4 valori anomali eccezionalmente grandi. Il *boxplot* degli scostamenti standardizzati del peso mostra invece una distribuzione con lieve asimmetria positiva, con variabilità minore di quella del prezzo e con 4 valori anomali. In particolare, i modelli di *smartphone* 19 e 56 risultano *outliers* per entrambe le variabili, cioè sono molto più costosi e pesanti rispetto agli altri. Tali modelli corrispondono ai due punti che si collocano in alto e a destra nel successivo diagramma di dispersione. Inoltre, in tale diagramma di dispersione gli altri due punti che si collocano in alto, sopra la nuvola principale, sono anomali per il peso, ma non per il prezzo, mentre il punto che si trova a destra (maggiore di 800) e verso il basso è un valore anomalo per il prezzo, ma non per il peso.

Per le strategie di trattamento dei valori anomali, si legga il testo: Zani – Cerioli, 2007, pp.82-83, adattando il commento al caso in esame.

Esercizio II

Effettuando i calcoli si ottiene:

λ_1 = 3.791, percentuale di varianza spiegata = 63,18%

λ_2 = 0.962, percentuale di varianza spiegata = 16,04%

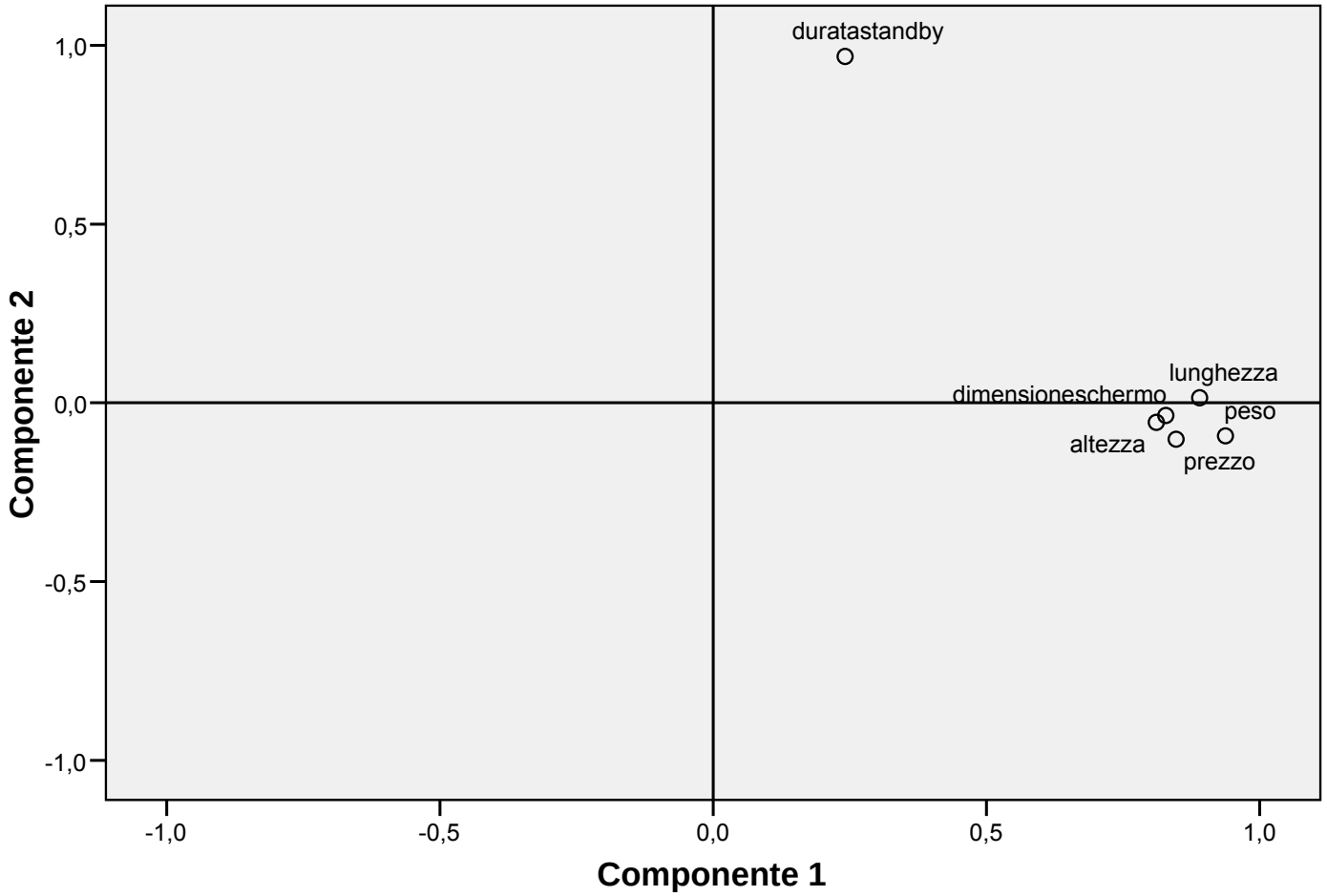
percentuale cumulata di varianza spiegata = 79,22%

Il livello della soglia di accettabilità è: $0.95^6 \times 100 = 73{,}5\%$, per cui considerando entrambe le componenti principali il modello risulta soddisfacente. Occorre osservare, tuttavia, che la seconda CP presenta un autovalore minore di 1 (anche se di poco), per cui in base al secondo criterio di scelta del numero di componenti essa dovrebbe venire esclusa. Così facendo, però, la percentuale di varianza spiegata sarebbe molto al di sotto della soglia. In questo caso, dunque, i diversi criteri non forniscono una risposta univoca sul numero di CP da adottare (vedi: Zani – Cerioli, 2007, p. 238).

La prima CP è un indicatore sintetico del prezzo e delle caratteristiche fisiche degli *smartphone* (presenta correlazioni elevate con le prime 5 variabili), mentre la seconda CP è espressione della sola durata in standby.

La parte del *biplot* disegnabile in base ai dati disponibili è riportata di seguito.

Grafico componenti

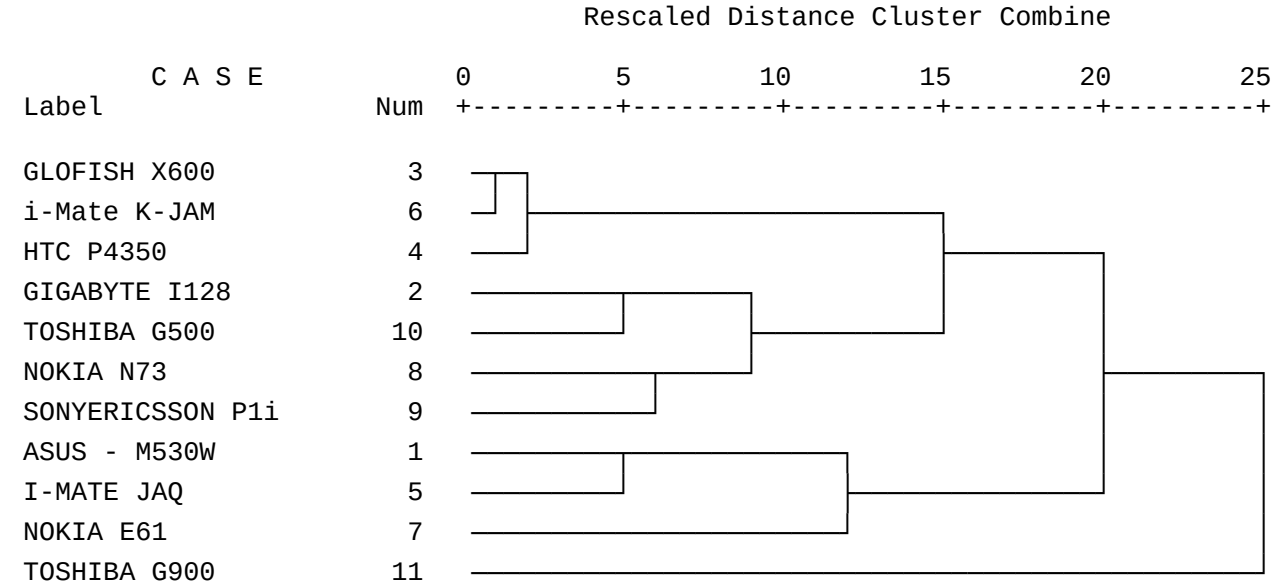


ESERCIZIO III

▽

* * * * * * 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



Il taglio più ragionevole è dopo il passo 8 nella tabella (distanza riscalata tra 15 e 20 nel dendroagramma), che individua due gruppi (segmenti del mercato) di 7 e 3 modelli ed un *outlier* (il modello n. 11).
La partizione corrispondente è:
(3, 6, 4, 2, 10, 8, 9) (1, 5, 7) (11).