# Analisi Esplorativa

Aldo Solari aldo.solari@unimib.it

Pagina del corso: https://aldosolari.github.io/AE

Analisi Esplorativa è il secondo modulo dell'insegnamento **Analisi Statistica Multivariata**. Il modulo si propone di fornire un'introduzione ai principali metodi statistici per l'analisi di dati multidimensionali al fine di identificare strutture che consentano di ridurne la complessità preservando l'informazione originariamente presente nelle misurazioni.

## 0.1 L'analisi multivariata

L'analisi multivariata (o multidimensionale) riguarda l'*analisi congiunta di più variabili* misurate sul medesimo insieme di unità statistiche.

In qualche caso ha senso l'analisi delle singole variabili raccolte, molto più spesso le variabili sono legate in modo tale che solo un'analisi congiunta di esse permette di rilevare pienamente la struttura dei dati

Le tecniche per l'analisi di dati multivariati possono avere una natura *descrittiva/esplorativa* oppure *inferenziale*. Per gli scopi di questo corso, ci occuperemo principalmente delle tecniche descrittive/esplorative, lasciando gli aspetti inferenziali a corsi più avanzati.

Fra i molteplici obiettivi dell'analisi multivariata, considereremo:

1. Esplorazione di dati multidimensionali (*exploratory analysis*)
2. Riduzione della dimensionalità dei dati (*dimensionality reduction*)
   - Analisi delle componenti principali (*principal component analysis*)
   - Analisi fattoriale (*factor analysis*)
3. Raggruppamento delle unità statistiche (*cluster analysis*)

- $k$-medie (*k-means*)
- analisi dei gruppi gerarchica (*hierarchical clustering*)

Nella nomenclatura dell'apprendimento automatico o *machine learning*, questi temi vanno sotto il nome di *unsupervised learning.*

Significa che l'apprendimento non è guidato da una variabile risposta, come invece accade nei problemi di *supervised learning*

| | *Output* discreto | *Output* continuo |
|---|---|---|
| *Supervised learning* | Classificazione | Regressione |
| *Unsupervised learning* | Raggruppamento | Riduzione dimensionalità |

## 0.2 Riduzione della dimensionalità

$$\underset{n\times p}{X} \mapsto \underset{n\times q}{Y}$$

- *Input*: matrice $\underset{n\times p}{X}$ con $p$ variabili quantitative.
- *Output*: matrice $\underset{n\times q}{Y}$ con $q < p$ variabili quantitative.
- *Obiettivo*: Ridurre la dimensione perdendo meno informazione possibile.

**Esempio 0.2.1** (Dati `heptathlon`). *L'eptathlon è una specialità dell'atletica leggera che contempla $p = 7$ gare di discipline diverse: 100 metri ostacoli, salto in alto, getto del peso, 200 metri piani, salto in lungo, tiro del giavellotto e 800 metri piani. I dati che abbiamo a disposizione riguardano i risultati di $n = 25$ atlete alle Olimpiadi di Seul del 1988:*

| | hurdles | highjump | shot | run200m | longjump | javelin | run800m |
|---|---|---|---|---|---|---|---|
| Joyner-Kersee (USA) | 12.69 | 1.86 | 15.80 | 22.56 | 7.27 | 45.66 | 128.51 |
| John (GDR) | 12.85 | 1.80 | 16.23 | 23.65 | 6.71 | 42.56 | 126.12 |
| Behmer (GDR) | 13.20 | 1.83 | 14.20 | 23.10 | 6.68 | 44.54 | 124.20 |
| Sablovskaite (URS) | 13.61 | 1.80 | 15.23 | 23.92 | 6.25 | 42.78 | 132.24 |
| Choubenkova (URS) | 13.51 | 1.74 | 14.76 | 23.93 | 6.32 | 47.46 | 127.90 |
| Schulz (GDR) | 13.75 | 1.83 | 13.50 | 24.65 | 6.33 | 42.82 | 125.79 |
| Fleming (AUS) | 13.38 | 1.80 | 12.88 | 23.59 | 6.37 | 40.28 | 132.54 |
| Greiner (USA) | 13.55 | 1.80 | 14.13 | 24.48 | 6.47 | 38.00 | 133.65 |
| Lajbnerova (CZE) | 13.63 | 1.83 | 14.28 | 24.86 | 6.11 | 42.20 | 136.05 |
| Bouraga (URS) | 13.25 | 1.77 | 12.62 | 23.59 | 6.28 | 39.06 | 134.74 |
| Wijnsma (HOL) | 13.75 | 1.86 | 13.01 | 25.03 | 6.34 | 37.86 | 131.49 |
| Dimitrova (BUL) | 13.24 | 1.80 | 12.88 | 23.59 | 6.37 | 40.28 | 132.54 |
| Scheider (SWI) | 13.85 | 1.86 | 11.58 | 24.87 | 6.05 | 47.50 | 134.93 |
| Braun (FRG) | 13.71 | 1.83 | 13.16 | 24.78 | 6.12 | 44.58 | 142.82 |
| Ruotsalainen (FIN) | 13.79 | 1.80 | 12.32 | 24.61 | 6.08 | 45.44 | 137.06 |
| Yuping (CHN) | 13.93 | 1.86 | 14.21 | 25.00 | 6.40 | 38.60 | 146.67 |
| Hagger (GB) | 13.47 | 1.80 | 12.75 | 25.47 | 6.34 | 35.76 | 138.48 |
| Brown (USA) | 14.07 | 1.83 | 12.69 | 24.83 | 6.13 | 44.34 | 146.43 |
| Mulliner (GB) | 14.39 | 1.71 | 12.68 | 24.92 | 6.10 | 37.76 | 138.02 |
| Hautenauve (BEL) | 14.04 | 1.77 | 11.81 | 25.61 | 5.99 | 35.68 | 133.90 |
| Kytola (FIN) | 14.31 | 1.77 | 11.66 | 25.69 | 5.75 | 39.48 | 133.35 |
| Geremias (BRA) | 14.23 | 1.71 | 12.95 | 25.50 | 5.50 | 39.64 | 144.02 |
| Hui-Ing (TAI) | 14.85 | 1.68 | 10.00 | 25.23 | 5.47 | 39.14 | 137.30 |
| Jeong-Mi (KOR) | 14.53 | 1.71 | 10.83 | 26.61 | 5.50 | 39.26 | 139.17 |
| Launa (PNG) | 16.42 | 1.50 | 11.78 | 26.16 | 4.88 | 46.38 | 163.43 |

*L'obiettivo è determinare un punteggio da attribuire a ciascun atleta che sintetizzi la prestazione nelle sette gare al fine di ottenere la classifica finale, ovvero ridurre la dimensionalità da $p = 7$ a $q = 1$:*

$$\underset{25\times 7}{X} \mapsto \underset{25\times 1}{y}$$

**Esempio 0.2.2** (Dati `face`). *Una immagine in bianco e nero può essere rappresentata come una matrice di dati, dove l'intensità di grigio di ogni pixel viene rappresentata nella corrispondente elemento della matrice. I colori più chiari sono associati valori più alti, colori più scuri sono associati valori più bassi nell'intervallo $[0, 1]$: si veda la Tabella 0.2.2 riferita alla Figura 0.2.2.*

Figura 0.1: Immagine originale (una matrice $X$ di dimensione $243 \times 22$) e immagine compressa. Fonte: materiale didattico di Marloes Maathuis.

| r/c | ... | 110 | 111 | 112 | 113 | 114 | ... |
|---|---|---|---|---|---|---|---|
| ... | ... | ... | ... | ... | ... | ... | |
| 110 | ... | 0.96 | 0.93 | 0.92 | 0.93 | 0.90 | ... |
| 111 | ... | 0.97 | 0.96 | 0.95 | 0.95 | 0.93 | ... |
| 112 | ... | 0.95 | 0.96 | 0.94 | 0.93 | 0.90 | ... |
| 113 | ... | 0.87 | 0.90 | 0.90 | 0.87 | 0.82 | ... |
| 114 | ... | 0.85 | 0.86 | 0.87 | 0.85 | 0.82 | ... |
| ... | ... | ... | ... | ... | ... | ... | ... |

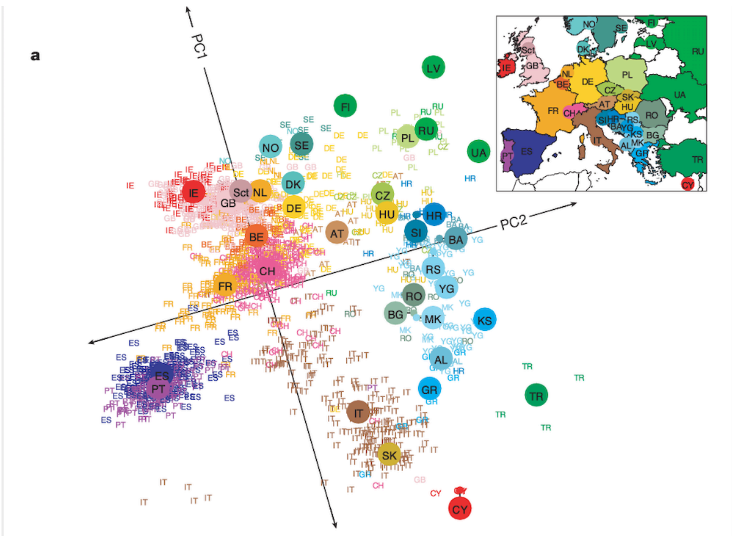
*L'obiettivo dell'analisi è la compressione dell'immagine per ridurne le dimensioni senza perdere troppa informazione. L'immagine compressa che vedete in Figura 0.2.2 si ottiene con* $\underset{n\times q}{Y}\,\underset{q\times p}{V'} + \underset{n\times 1}{1}\,\underset{1\times p}{\bar{x}'}$ *e* $q = 10$. *In termini di numeri utilizzati per descrivere ciascuna immagine, abbiamo:*

- $\underset{243\times 220}{X}$*:* $243 \times 220 = 53460$ *numeri*
- $\underset{243\times 10}{Y}, \underset{220\times 10}{V}, \underset{220\times 1}{\bar{x}}$*:* $243 \times 10 + 220 \times 10 + 220 = 4850$ *numeri*

**Esempio 0.2.3** (I geni europei rispecchiano la geografia europea?)**.** *Il lavoro descritto in Novembre et al. (2008) utilizza l'analisi delle componenti principali per lo studio della struttura genetica delle popolazioni. A tal fine, hanno raccolto un campione di* $n \approx 1300$ *persone contenente le misurazioni su* $p \approx 200.000$ *SNP (polimorfismi a singolo nucleotide).*

*Il risultato è sorprendente: le mappe genetiche e geopolitiche dell'Europa si sovrappongono in misura notevole: si veda la Figura 0.2.3. Tuttavia, il risultato dovrebbe essere interpretato con cautela.*

Figura 0.2: Fonte: Novembre et al. (2008), Figure 1




## 0.3 Raggruppamento delle unità statistiche

$$\underset{n\times p}{X} \mapsto \underset{n\times 1}{y}$$

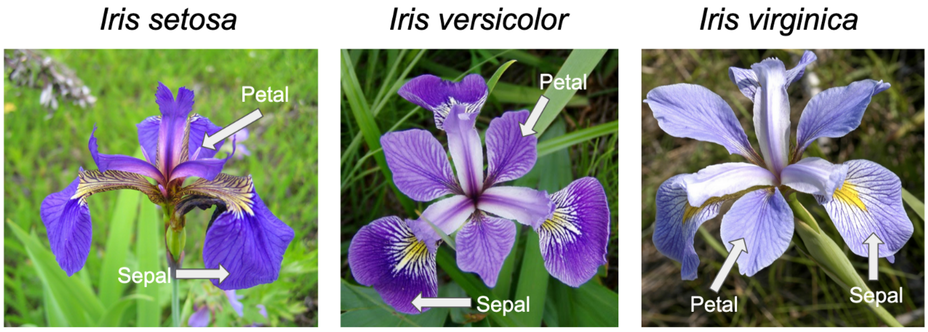
- *Input*: matrice $\underset{n\times p}{X}$ con $p$ variabili quantitative e/o qualitative

- *Output* vettore $\underset{n\times 1}{y} = \begin{bmatrix} y_1 \\ \cdots \\ y_i \\ \cdots \\ y_n \end{bmatrix}$ con $y_i \in \{G_1, G_2, \ldots, G_k\}$
  dove $G_1, G_2, \ldots, G_k$ rappresenta il primo, ..., il $k$-simo gruppo

- *Obiettivo*: Formare $k$ gruppi omogenei al loro interno e disomogenei tra di loro

**Esempio 0.3.1** (I dati `iris`). *Il dati* `iris` *sono stati analizzati da Ronald Fisher nel 1936. Il dataset consiste in $n = 150$ fiori di genere Iris (dalla parola greca iris che significa arcobaleno) misurate da Edgar Anderson e classificate secondo tre specie: Iris setosa, Iris virginica e Iris versicolor: si veda la Figura 0.3.1.*

Figura 0.3: Iris setosa, Iris virginica e Iris versicolor. Fonte: Wikipedia.




*Le quattro variabili considerate sono la lunghezza e la larghezza del sepalo e del petalo: si veda la seguente Tabella.*

Figura 0.4: Alcune osservazioni dei dati `iris`.

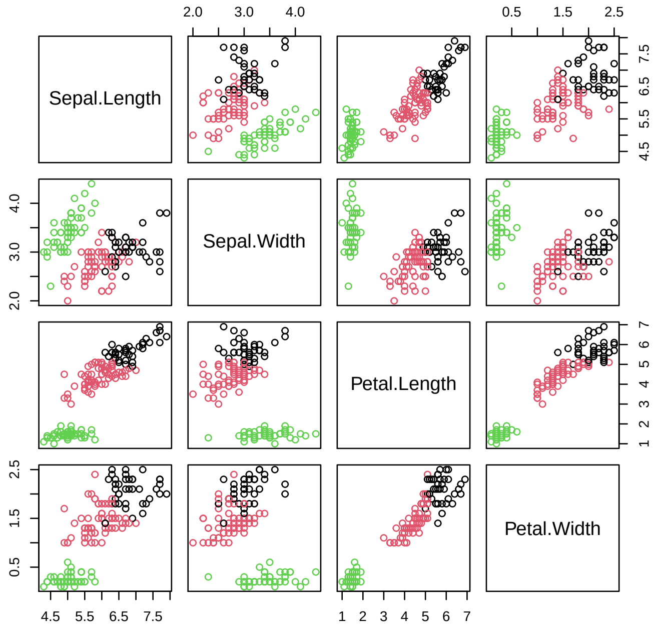
**Table I**

| *Iris setosa* | | | | *Iris versicolor* | | | | *Iris virginica* | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Sepal length | Sepal width | Petal length | Petal width | Sepal length | Sepal width | Petal length | Petal width | Sepal length | Sepal width | Petal length | Petal width |
| 5·1 | 3·5 | 1·4 | 0·2 | 7·0 | 3·2 | 4·7 | 1·4 | 6·3 | 3·3 | 6·0 | 2·5 |
| 4·9 | 3·0 | 1·4 | 0·2 | 6·4 | 3·2 | 4·5 | 1·5 | 5·8 | 2·7 | 5·1 | 1·9 |
| 4·7 | 3·2 | 1·3 | 0·2 | 6·9 | 3·1 | 4·9 | 1·5 | 7·1 | 3·0 | 5·9 | 2·1 |
| 4·6 | 3·1 | 1·5 | 0·2 | 5·5 | 2·3 | 4·0 | 1·3 | 6·3 | 2·9 | 5·6 | 1·8 |
| 5·0 | 3·6 | 1·4 | 0·2 | 6·5 | 2·8 | 4·6 | 1·5 | 6·5 | 3·0 | 5·8 | 2·2 |
| 5·4 | 3·9 | 1·7 | 0·4 | 5·7 | 2·8 | 4·5 | 1·3 | 7·6 | 3·0 | 6·6 | 2·1 |
| 4·6 | 3·4 | 1·4 | 0·3 | 6·3 | 3·3 | 4·7 | 1·6 | 4·9 | 2·5 | 4·5 | 1·7 |
| 5·0 | 3·4 | 1·5 | 0·2 | 4·9 | 2·4 | 3·3 | 1·0 | 7·3 | 2·9 | 6·3 | 1·8 |
| 4·4 | 2·9 | 1·4 | 0·2 | 6·6 | 2·9 | 4·6 | 1·3 | 6·7 | 2·5 | 5·8 | 1·8 |

*L'analisi di raggruppamento fornisce circa il* 90% *di osservazioni classificate correttamente, come descritto dalla tabella 0.3.1 e rappresentato con la Figura 0.3.1.*

| | setosa | versicolor | virginica |
|---|---|---|---|
| gruppo A | 0 | 2 | 36 |
| grippo B | 0 | 48 | 14 |
| gruppo C | 50 | 0 | 0 |

Figura 0.5: Raggruppamento dei dati `iris`.




**Esempio 0.3.2** (Dati `movielens`). *I dati che abbiamo a disposizione riguardano la valutazione (*rating*) (da 0.5 a 5) attributo a $n = 9125$ film da parte di $p = 671$ utenti tra il 09 gennaio 1995 e il 16 ottobre 2016 . L'esempio che segue considerera $n = 50$ film e $p = 139$ utenti.*

*I dati sono rappresentati nella Tabella 0.3.2. Uno delle sfide da affrontare è il problema dei valori mancanti (missing values). Cosa fare quando il nostro dataset presenta dei buchi?*

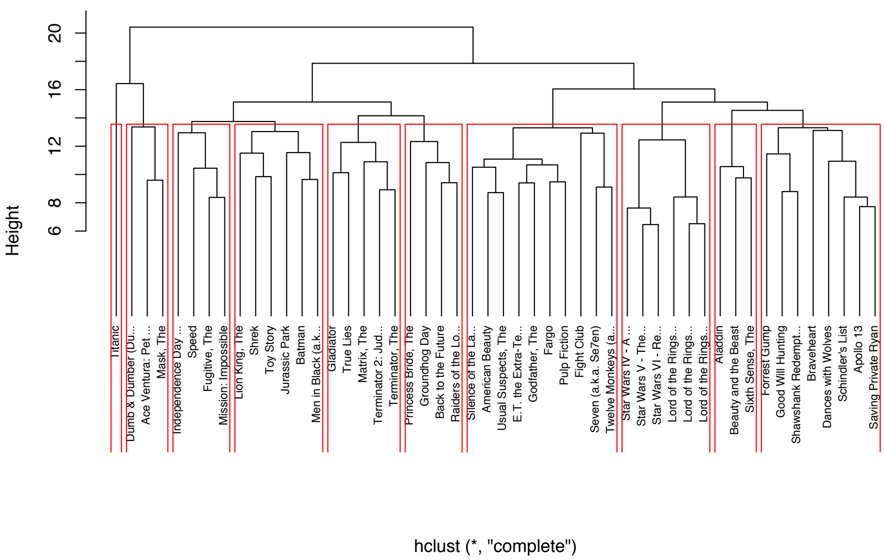
*Una volta affrontato il problema dei dati mancanti, si può procedere raggruppando i film in gruppi omogenei al loro interno e disomogenei tra di loro rispetto al* rating *che hanno ottenuto dagli utenti. Ad esempio, se decidiamo di raggruppare i $n = 50$ film in $k = 10$ gruppi A, B, C, D, E, F, G, H, I, L otteniamo i raggruppamenti descritti dalla Figura 0.3.2.*

$$\underset{50\times 139}{X} \mapsto \underset{50\times 1}{y} = \begin{bmatrix} \text{B} \\ \text{A} \\ \cdots \\ \text{A} \\ \cdots \\ \text{C} \\ \text{D} \end{bmatrix}$$

Tabella 0.1: Valutazioni di alcuni utenti su alcuni film.

| | U8 | U15 | U17 | U19 | U20 | U21 | U22 | U23 | U26 |
|---|---|---|---|---|---|---|---|---|---|
| Ace Ventura | | 2.0 | | 3.0 | 1.0 | 3.0 | | 2.0 | 0.5 |
| Aladdin | | 0.5 | | 3.0 | 3.5 | | 2.0 | 4.0 | |
| American Beauty | 4.5 | 4.0 | 4.5 | | | | 4.0 | 3.5 | 4.0 |
| Apollo 13 | | 3.0 | | 3.0 | 3.0 | | | 3.5 | |
| Back to the Future | 4.0 | 5.0 | 4.5 | 5.0 | 3.5 | 4.0 | 4.0 | 4.5 | |
| Batman | | 4.0 | | 4.0 | 4.0 | 3.0 | 4.5 | 3.5 | |
| Beauty and the Beast | | | | 5.0 | 4.0 | 3.0 | | 4.5 | |
| Braveheart | 4.0 | 3.0 | | 3.0 | 2.0 | | | 3.5 | |
| Dances with Wolves | | 3.0 | 3.0 | 3.0 | 2.0 | 4.0 | | 2.5 | |
| Dumb & Dumber | | 3.5 | | 3.0 | 1.0 | | 2.5 | | |
| E.T. | | 4.0 | | 5.0 | 1.5 | 3.0 | 2.5 | 5.0 | |
| Fargo | | 5.0 | 3.5 | 5.0 | 2.0 | | | 4.5 | 3.5 |
| Fight Club | 4.0 | 5.0 | 5.0 | | 0.5 | | 4.0 | 3.5 | 4.0 |
| Forrest Gump | 4.0 | 1.0 | 2.5 | 5.0 | 2.0 | 4.0 | 3.5 | 4.5 | 4.5 |
| Fugitive, The | 4.5 | 5.0 | | 4.0 | 4.5 | 3.0 | 4.5 | 3.5 | 3.5 |
| Gladiator | 5.0 | 2.0 | 4.0 | | | | 3.0 | 4.0 | 2.5 |
| Godfather, The | 5.0 | 5.0 | 5.0 | 5.0 | 2.0 | 4.0 | 4.0 | 5.0 | 4.0 |
| Good Will Hunting | 4.0 | 4.0 | 4.0 | | | | | 3.5 | |
| ⋮ | | | | | | | | | |

Figura 0.6: Raggruppamento gerarchico dei dati `movielens`.



## 0.4 Libri di testo

- Johnson and Wichern (2015)
- Everitt and Hothorn (2011)

# Indice

# Capitolo 1

# La matrice dei dati

## 1.1 La matrice $X$

$$\underset{n\times p}{X} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2j} & \cdots & x_{2p} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix}$$

- Media per la $j$-sima variabile

$$\bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}, \qquad j = 1, \ldots, p$$

- Varianza per la $j$-sima variabile

$$s_j^2 = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)^2, \qquad j = 1, \ldots, p$$

- Covarianza tra la $j$-sima e la $k$-sima variabile

$$s_{jk} = \frac{1}{n}\sum_{i=1}^{n}(x_{ij} - \bar{x}_j)(x_{ik} - \bar{x}_k), \qquad j = 1, \ldots, p,\ k = 1, \ldots, p$$

  Si noti che $s_{jk} = s_{kj}$ e che $s_{jj} = s_j^2$

- Correlazione tra la $j$-sima e la $k$-sima variabile

$$r_{jk} = \frac{s_{jk}}{\sqrt{s_{jj}}\sqrt{s_{kk}}}, \qquad j = 1, \ldots, p,\ k = 1, \ldots, p$$

  Si noti che $-1 \leq r_{jk} \leq 1$

## 1.2 Vettore delle medie, matrice di varianze/covarianze e di correlazione

- Vettore delle medie

$$\underset{p\times 1}{\bar{x}} = \begin{bmatrix} \bar{x}_1 \\ \cdots \\ \bar{x}_j \\ \cdots \\ \bar{x}_p \end{bmatrix}$$

- Matrice di varianze/covarianze

$$\underset{p\times p}{S} = \begin{bmatrix} s_{11} & s_{12} & \cdots & s_{1j} & \cdots & s_{1p} \\ s_{21} & s_{22} & \cdots & s_{2j} & \cdots & s_{2p} \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ s_{j1} & s_{j2} & \cdots & s_{jj} & \cdots & s_{jp} \\ \vdots & \vdots & \cdots & \vdots & \ddots & \cdots \\ s_{p1} & s_{p2} & \cdots & s_{pj} & \cdots & s_{pp} \end{bmatrix}$$

- Matrice di correlazione

$$\underset{p\times p}{R} = \begin{bmatrix} 1 & r_{12} & \cdots & r_{1j} & \cdots & r_{1p} \\ r_{21} & 1 & \cdots & r_{2j} & \cdots & r_{2p} \\ \vdots & \vdots & \ddots & \vdots & & \vdots \\ r_{j1} & r_{j2} & \cdots & 1 & \cdots & r_{jp} \\ \vdots & \vdots & \cdots & \vdots & \ddots & \cdots \\ r_{p1} & r_{p2} & \cdots & r_{pj} & \cdots & 1 \end{bmatrix}$$

### Esempio

| | | | | |
|---|---|---|---|---|
| Variabile 1 (prezzo in Dollari per libro): | 42 | 52 | 48 | 58 |
| Variabile 2 (numero di libri venduti): | 4 | 5 | 4 | 3 |

$$\underset{4\times 2}{X} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \\ x_{41} & x_{42} \end{bmatrix} = \begin{bmatrix} 42 & 4 \\ 52 & 5 \\ 48 & 4 \\ 58 & 3 \end{bmatrix}$$

$$\underset{2\times 1}{\bar{x}} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix} = \begin{bmatrix} 50 \\ 4 \end{bmatrix}$$

$$\underset{2\times 2}{S} = \begin{bmatrix} s_{11} & s_{12} \\ s_{21} & s_{22} \end{bmatrix} = \begin{bmatrix} 34 & -1.5 \\ -1.5 & 0.5 \end{bmatrix}$$

$$\underset{2\times 2}{R} = \begin{bmatrix} 1 & r_{12} \\ r_{21} & 1 \end{bmatrix} = \begin{bmatrix} 1 & -0.36 \\ -0.36 & 1 \end{bmatrix}$$

## 1.3 Diagramma di dispersione

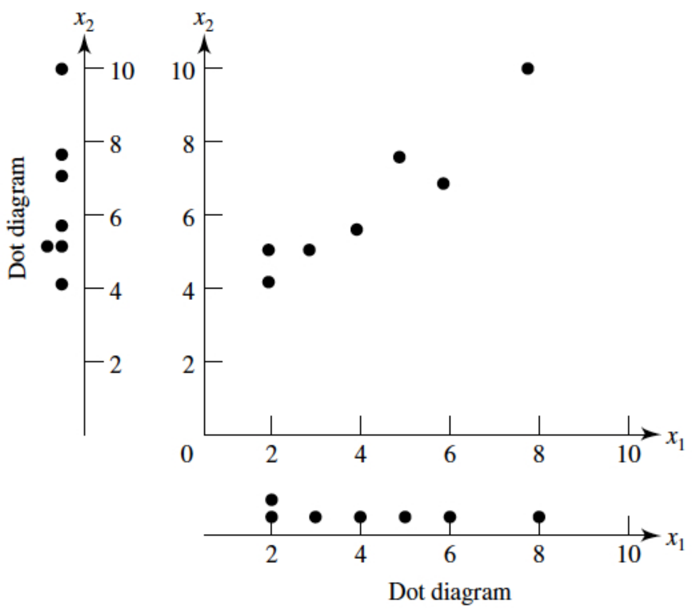
| | $u_1$ | $u_2$ | $u_3$ | $u_4$ | $u_5$ | $u_6$ | $u_7$ |
|---|---|---|---|---|---|---|---|
| $x_1$ | 3 | 4 | 2 | 6 | 8 | 2 | 5 |
| $x_2$ | 5 | 5.5 | 4 | 7 | 10 | 5 | 7.5 |

Medie: $\bar{x}_1 = 4.2$, $\bar{x}_2 = 6.2$

Varianze: $s_{11} = 4.2$, $s_{22} = 0.56$

Covarianza: $s_{12} = 3.70$

Correlazione: $r_{12} = 0.95$



Cosa succede se mescolo a caso (permutazione) i valori della prima riga della tabella?

| | $u_1$ | $u_2$ | $u_3$ | $u_4$ | $u_5$ | $u_6$ | $u_7$ |
|---|---|---|---|---|---|---|---|
| $x_1$ | 5 | 4 | 6 | 2 | 2 | 8 | 3 |
| $x_2$ | 5 | 5.5 | 4 | 7 | 10 | 5 | 7.5 |

Medie: $\bar{x}_1 = 4.2$, $\bar{x}_2 = 6.2$

Varianze: $s_{11} = 4.20$, $s_{22} = 0.56$

Covarianza $s_{12} = -3.01$

Correlazione $r_{12} = -0.78$

# Bibliografia


Everitt, B., & Hothorn, T. (2011). *An introduction to applied multivariate analysis with r*. Springer Science & Business Media.

Johnson, R. A., & Wichern, D. (2015). Applied multivariate statistical analysis. *Statistics*, *6215*(10), 10.

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A. R., Auton, A., ... others (2008). Genes mirror geography within europe. *Nature*, *456*(7218), 98–101.