

Data Science

Statistica Descrittiva

Covarianza → tendenza che hanno due variabili (X, Y) a variare insieme, ovvero a covariare

Tipi di relazione

Lineare → si avvicina alla forza di una retta

↳ All'aumentare/diminuire di X aumenta/diminuisce Y

↳ direttamente proporzionale
inversamente proporzionale

Non Lineare → andamento curvilineo (parabola / iperbole)

↳ livelli Bassi e Alti di X corrispondono livelli bassi di Y , mentre livelli intermedi di X corrispondono livelli alti di Y

Forza di Relazione

Direzione → positiva → all'aumentare di X aumenta anche Y

↳ direttamente proporzionale

→ negativa → all'aumentare di X diminuisce Y

↳ inversamente proporzionale

Entità → Forza della relazione esistente tra due variabili

↳ Più i punteggi sono raggruppati attorno ad una retta, tanto più forte è la relazione tra due variabili

↳ disparsi in maniera uniforme → non esiste alcuna relazione

Abbiamo bisogno di un indice che sia in grado di identificare forza e direzioni delle associazioni tra variabili.

Covarianza → indicatore che misura la forza di relazione tra due variabili

$$\text{cov}(X, Y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$$

- Valori positivi → se la maggior parte dei termini $(x_i - \bar{x})$ e $(y_i - \bar{y})$ sono concordi
- Valori negativi → se la maggior parte dei termini $(x_i - \bar{x})$ e $(y_i - \bar{y})$ sono discordi
- Prossimi a 0 → se i termini $(x_i - \bar{x})$ e $(y_i - \bar{y})$ sono in ugual misura concordi e discordi

Proprieta'

- La covarianza tra una variabile x e se stessa è pari alla varianza di x
$$\text{cov}(x, x) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(x_i - \bar{x}) \longrightarrow \text{sempre concordi} \longrightarrow \text{cov grande e positivo}$$

$$\text{Cov}(x, -x) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(-x_i + \bar{x}) = -\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2 = -\text{Var}(x) \leq 0$$

↓
 covarianza
 grande

↑
 invertendo i
 segni perche'
 - x

↳ tercini discordi

Matrici di Covarianza

- Nella diagonale ci sono le varianze $\text{cov}(x, x) = \text{var}(x)$

	fertilità	agricoltura	istruzione
fertilità	152.7	98.0	-5.1
agricoltura	98.0	504.8	-11.9
istruzione	-5.1	-11.9	0.6

- $$\bullet \text{Cov}(y, x) = \text{cov}(x, y)$$

↳ La matrice è simmetrica

Coefficiente di correlazione r di Pearson

→ Per abberuare se la covarianza e' piccola o grande dobbiamo confrontarla con il prodotto degli scarti quadratici medi (σ)

- La covarianza solitamente viene rappresentata nella sua forma nonnormalizzata
Cov. Nonnormalizzata \rightarrow Correlazione

Correlazione → può assumere valori da -1 a 1 ($[-1, 1]$)

- $-1 \rightarrow$ correlazione perfetta negativa
 - $1 \rightarrow$ correlazione perfetta positiva
 - $0 \rightarrow$ nessuna relazione

- La correlazione ci dice se c'è una relazione sistematica, ma non se una causa l'altra

Coefficiente di correlazione di Pearson → misura la corr. tra variabili a intervalli o a rapporti equivalenti.

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$

Assume values $[-1, 1]$

Matrice di Correlazione

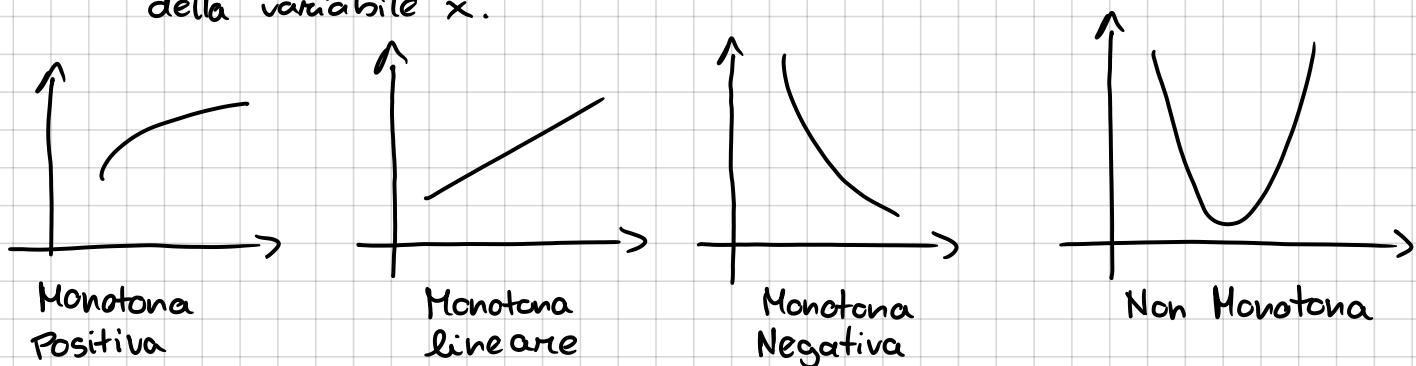
	fertilità	agricoltura	istruzione
fertilità	1	0.35	-0.52
agricoltura	0.35	1	-0.68
istruzione	-0.52	-0.68	1

Nota

- La covarianza e la correlazione misurano esclusivamente relazioni lineari. Questo ha importanti conseguenze
- Se la relazione tra x ed y è monotona ma non lineare, allora $\text{cor}(x,y) < 1$

Monotona → Una relazione è monotona quando all'aumentare di una variabile, l'altra si muove esclusivamente in una direzione. Non importa come lo fai, l'importante è che non cambia direzione.

Lineare → Una relazione è lineare quando il tasso di cambiamento è costante. Rappresentata graficamente è una retta. L'aumento della variabile y è sempre proporzionale all'aumento della variabile x .



Caratteristica Numerica di una distribuzione di dati

Statistica → indica ogni funzione dei dati

I dati quantitativi possono essere divisi in:

Dati discreti → hanno valori in un insieme numerabile es. numero di cervelli

Dati continui → hanno valori in un intervallo di numeri es. lunghezza

Caratteristiche di un insieme di dati

Centro → tendenza centrale dei dati

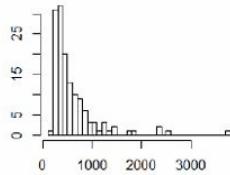
Dissusione → L'estensione dei dati nel range dei valori

Forca → Somma di un grafico, tipo istogramma.

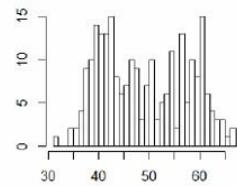
↪ Prevalentemente 2:

- Simetria (skewness) → i dati sono simmetrici rispetto al centro
 - ↪ Assimmetrica a destra (right skewed) → se la coda destra della distribuzione appare allungata rispetto al centro dei dati
(Analogia → Asimmetrica a sx)

- Curtosi → Indica la somiglianza rispetto alla distribuzione normale per quanto riguarda la forma del picco
 - ↪ Platocurtica → è più piatta rispetto alla normale
 - ↪ Leptocurtica → è più allungata rispetto alla normale
 - ↪ Mesocurtica → non si discosta dalla normale



(a) Distribuzione asimmetrica a destra



(b) Distribuzione simmetrica

Figura 4.1: Esempi di alcune distribuzioni di dati

Cluster → dati raggruppati attorno ad alcuni valori → ammassi (cluster)

Gap → intervalli in cui i dati sono quasi assenti → vuoti (gap)

Outliers → Osservazioni estremhe che provocano dei problemi, perché influenzano statistiche e parametri in modo eccessivo

↪ Cause:

- ↪ Non appartengono all'insieme dei dati che si vogliono analizzare
 - ↪ è opportuno elimerarli prima dell'analisi
- ↪ errore tipografico → viene corretto e reinserito
- ↪ Indicano una tendenza non prevista del fenomeno da analizzare → comportamento nuovo

Misure delle caratteristiche del centro

$$\text{Media Simple} \quad \bar{x} = \frac{\sum_{i=1}^N x_i}{N}$$

- sensibile a dati estremi

Mediana Simple → si devono ordinare i dati in maniera crescente.

- ↳ N dispari → la mediana è il valore centrale
- ↳ N pari → la media fra i due valori centrali
- ↳ resistente ai valori estremi
- ↳ necessita di ordinamento dei dati

Media Troncata → Si elicuina una frazione $0 < p < s$ delle osservazioni dagli estremi della lista ordinata e poi si calcola la media sui rimanenti.

- ↳ resistente ai valori estremi
- ↳ richiede ordinamento

↳ es. media troncata al 20% abbiamo 9 valori.

$$9 \text{ valori} \cdot 20\% = 1.8 \text{ (eccesso 2)}$$

- togliamo 2 valori dalla coda sx e 2 valori dalla coda dx
- 5 valori rimanenti con cui fare la media

Statistica ordinata

Statistica ordinata → Statistica detta dopo aver ordinato i dati in ordine

$$\text{crescente } x_1 \leq x_2 \leq \dots \leq x_N$$

→ dà informazioni sulla distribuzione dei dati e su dove essi sono più o meno concentrati.

Quantile Simple → di ordine p , $0 < p < 1 \rightarrow q_p$

↳ Valore che indica il $100 \cdot p\%$ dei dati e cui nome di q_p
 q_p è maggiore di $100 \cdot p\%$ dei dati

$$p = 0,5 \quad q_{0,5} \Rightarrow 100 \cdot 0,5\% \text{ dati} < q_{0,5} \rightarrow \text{mediana}$$

$$p = 0,25 \quad q_{0,25} \Rightarrow 100 \cdot 0,25\% \text{ dati} < q_{0,25}$$

$$p = 0,75 \quad q_{0,75} \Rightarrow 100 \cdot 0,75\% \text{ dati} < q_{0,75}$$

Varianza Campionaria Standard

→ Media Campionaria

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

- molto sensibile a valori estremi

Range Interquantile (IQR)

Il range interquantile è definito come $IQR = q_{0,75} - q_{0,25}$

- Poco sensibile a valori estremi
- Necessita di ordinamento
- Coinvolge solo il 50% dei dati



Deviazione Assoluta dalla Media (MAD)

$MAD = c \cdot \text{mediana delle deviazioni assolute della media}$

$$\downarrow \\ MAD = c \cdot \text{median}(|x_1 - \bar{x}|, |x_2 - \bar{x}|, \dots, |x_n - \bar{x}|)$$

- molto robusto
- poco conosciuto

Misura della Simmetria

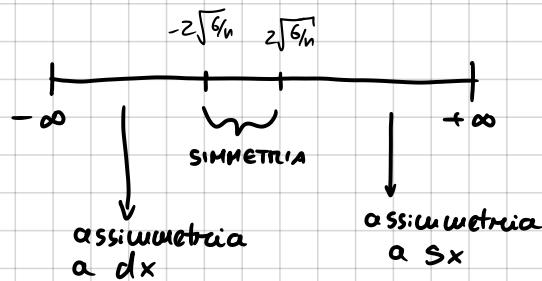
Misura della simmetria

→ Il segno di g_1 indica la direzione dell'assimetria, mentre la grandezza $|g_1|$ dice se l'assimetria è statisticamente rilevante o no.

$$g_1 = \frac{1}{n} \frac{\sum_{i=1}^n (x_i - \bar{x})^3}{S^3} \quad \begin{matrix} \bar{x} \rightarrow \text{media} \\ S \rightarrow \text{dev. std.} \end{matrix}$$

$$|g_1| > 2 \sqrt{6/n}$$

g_1 è nell'intervallo $]-\infty, +\infty[$

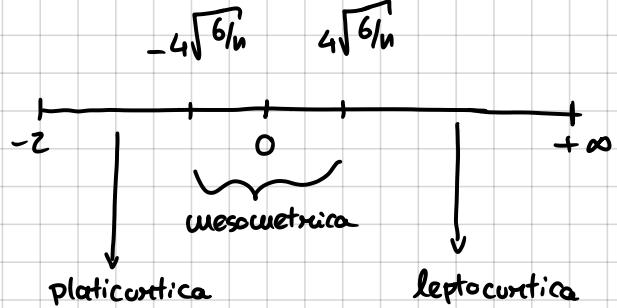


Misura della Cointosi

$$|g_2| > 4 \sqrt{6/n}$$

$$g_2 = \frac{1}{n} \frac{\sum_{i=1}^n (x_i - \bar{x})^4}{S^4} - 3$$

g_2 è nell'intervallo $]-2, +\infty[$



Mesocointica → distribuzione normale.

Leptocointica → distribuzione più appuntita della normale (code più pesanti).

Platicointica → distribuzione più piatta della normale, con code più leggere.

| cardini \rightarrow i cardini di un insieme sono definiti dopo aver ordinato in cuadro crescente gli elementi

- Cardine inferiore $\rightarrow h_L$

\hookrightarrow elemento in posizione $L = \lfloor (n+3)/2 \rfloor / 2$

\hookrightarrow con (x_1, \dots, x_n)

$$h_L = x_L$$

- Se L non è intero, allora il cardine inferiore è la media dei due valori adiacenti ad L .

- Cardine superiore $\rightarrow h_U$

\hookrightarrow elemento in posizione $U = n+1 - L$

| Cinque Numeri di Sintesi (5 NS)

$x_1 \rightarrow$ minimo

$x_N \rightarrow$ massimo

$$\begin{aligned} 5\text{NS} &= (x_1, h_L, \tilde{x}, h_U, x_N) && \text{mediana} \\ 5\text{NS} &= (x_1, q_{0.25}, q_{0.5}, q_{0.75}, x_N) \end{aligned}$$

La rappresentazione dei 5NS è il boxplot.

Outliers

\hookrightarrow Outliers Potenziale \rightarrow osservazione che si trova ad una distanza dal centro 1.5 volte rispetto all'ampiezza dell'intervallo $[h_L, h_U]$

$$[h_L - 1.5(h_U - h_L), h_U + 1.5(h_U - h_L)]$$

\hookrightarrow Outliers Sospetto \rightarrow osservazione che si trova ad una distanza dal centro superiore a 3 volte rispetto l'ampiezza dell'intervallo $[h_L, h_U]$

$$[h_L - 3(h_U - h_L), h_U + 3(h_U - h_L)]$$

