Statistica I

Unità J: covarianza e correlazione

Tommaso Rigon

Università Milano-Bicocca

Unità J

Argomenti affrontati

- Diagramma di dispersione
- Covarianza
- Coefficiente di correlazione (lineare)
- Matrice delle varianze e covarianze
- Matrice di correlazione

Riferimenti al libro di testo

- §7.1
- §7.6 — §7.7

Descrizione del problema

- Consideriamo **tre indicatori socio-economici** disponibili per $n = 47$ province svizzere di lingua francese. I dati sono storici e si riferiscono al 1888. Consideriamo:
- Una **misura di fertilità** (nati per donna), standardizzata in maniera tale che vari tra 0 e 100.
- Percentuale degli **occupati in agricoltura** sul totale degli occupati, interpretabile come un indicatore di urbanizzazione della provincia.
- Il logaritmo della percentuale della popolazione con un'**istruzione** superiore alla scuola primaria.
- Il problema che ci poniamo è di cercare di descrivere le **relazioni** esistenti tra i tre indicatori.

I dati grezzi

Fertilità

```
 [1] 80.2 83.1 92.5 85.8 76.9 76.1 83.8 92.4 82.4 82.9 87.1 64.1
[13] 66.9 68.9 61.7 68.3 71.7 55.7 54.3 65.1 65.5 65.0 56.6 72.5
[25] 57.4 74.2 72.0 60.5 58.3 65.4 75.5 69.3 77.3 70.5 79.4 65.0
[37] 92.2 79.3 70.4 65.7 72.7 64.4 77.6 67.6 35.0 44.7 42.8
```

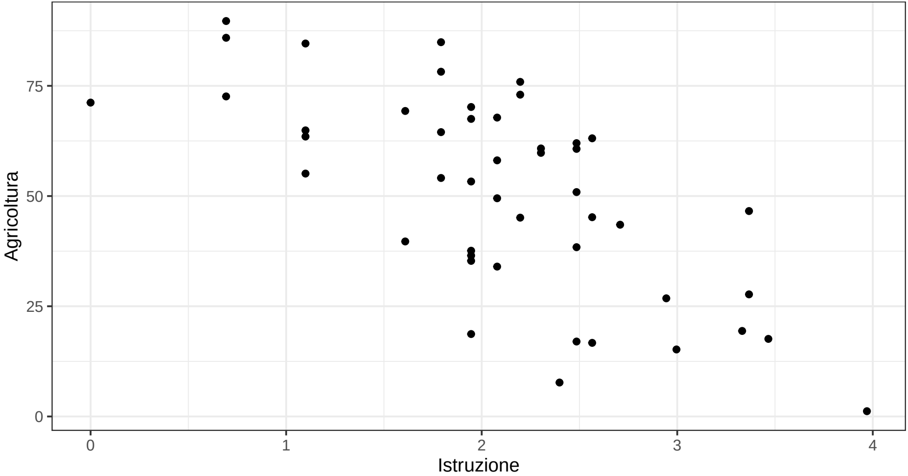
Agricoltura

```
 [1] 17.0 45.1 39.7 36.5 43.5 35.3 70.2 67.8 53.3 45.2 64.5 62.0
[13] 67.5 60.7 69.3 72.6 34.0 19.4 15.2 73.0 59.8 55.1 50.9 71.2
[25] 54.1 58.1 63.5 60.8 26.8 49.5 85.9 84.9 89.7 78.2 64.9 75.9
[37] 84.6 63.1 38.4  7.7 16.7 17.6 37.6 18.7  1.2 46.6 27.7
```

Istruzione

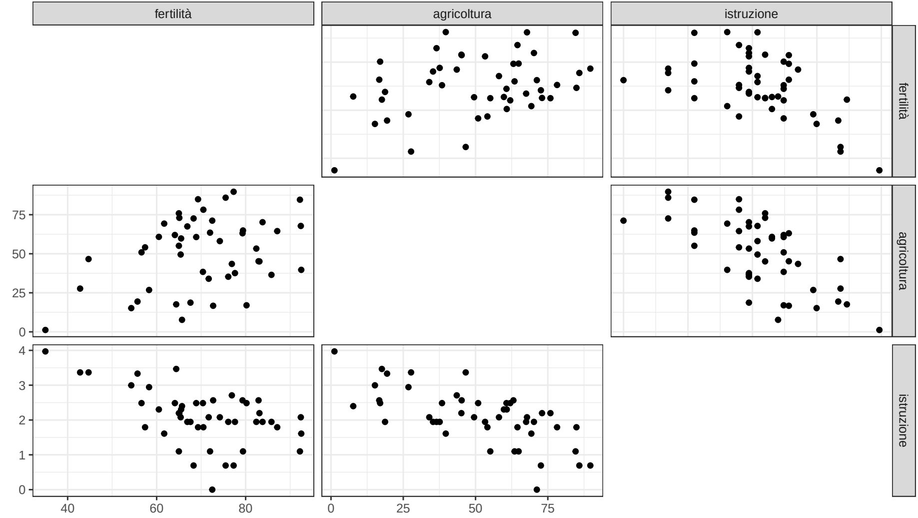
```
[1] 2.485 2.197 1.609 1.946 2.708 1.946 1.946 2.079 1.946 2.565
[11] 1.792 2.485 1.946 2.485 1.609 0.693 2.079 3.332 2.996 2.197
[21] 2.303 1.099 2.485 0.000 1.792 2.079 1.099 2.303 2.944 2.079
[31] 0.693 1.792 0.693 1.792 1.099 2.197 1.099 2.565 2.485 2.398
[41] 2.565 3.466 1.946 1.946 3.970 3.367 3.367
```

Il diagramma a dispersione



- I dati sono rappresentati come punti in un diagramma cartesiano.

Diagrammi a dispersione



Commenti ai grafici I

- La percentuale di occupati in `agricoltura` e `fertilità` sono **positivamente associati**.

- Province con una alta percentuale di occupati in agricoltura hanno anche una alta fertilità. Viceversa, basse percentuali di occupati in agricoltura si osservano in province con bassi livelli di fertilità.

- Esiste una **associazione negativa** tra `istruzione` e `fertilità`.

- Province con un alto livello di istruzione hanno una fertilità più bassa delle province con un basso livello di istruzione.

- Simili considerazioni possono essere fatte per la relazione tra le variabili `agricoltura` e `istruzione`, in cui si osserva una **associazione negativa**.

Commenti ai grafici II

- La relazione tra `agricoltura` e `fertilità` sembra **più debole** della relazione esistente tra `agricoltura` ed `istruzione`.

- Meno facile è valutare l'intensità delle relazioni intercorrenti tra `istruzione` e, rispettivamente, `agricoltura` e `fertilità`.

- La prima relazione (`istruzione` — `agricoltura`) sembra però in una qualche misura **più forte** della seconda (`istruzione` — `fertilità`).

- Per quantificare queste relazioni, abbiamo pertanto bisogno di un **indice** che sia in grado di identificare forza e direzioni delle associazioni tra variabili.

La covarianza

- Un indicatore che misura la forza della relazione tra due variabili è la **covarianza**.
- **Covarianza**. La covarianza delle coppie di dati $(x_1, y_1), \dots, (x_n, y_n)$ è

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

- Si noti che la covarianza è simmetrica, ovvero: $\text{cov}(x, y) = \text{cov}(y, x)$.
- La covarianza pertanto assume valori **positivi** se la maggior parte dei termini $(x_i - \bar{x})$ e $(y_i - \bar{y})$ sono concordi, ovvero se hanno lo stesso segno.
- La covarianza assume invece valori **negativi** se la maggior parte dei termini $(x_i - \bar{x})$ e $(y_i - \bar{y})$ sono discordi, ovvero se hanno segni diversi.
- Infine, la covarianza assume valori **prossimi a zero** se i termini $(x_i - \bar{x})$ e $(y_i - \bar{y})$ sono in ugual misura concordi e discordi.

Proprietà della covarianza

- **Proprietà**. La covarianza tra la variabile x e x stessa è pari alla varianza di x, ovvero

$$\mathrm{cov}(x, x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = \mathrm{var}(x) \geq 0.$$

  Poiché i termini $(x_i - \bar{x})$ e $(x_i - \bar{x})$ sono necessariamente sempre concordi, in questo caso la covarianza è grande e positiva.

- **Proprietà**. La covarianza tra la variabile x e $-x$ stessa è pari alla varianza di x cambiata di segno, ovvero

$$\mathrm{cov}(x, -x) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(-x_i + \bar{x}) = -\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 = -\mathrm{var}(x) \leq 0.$$

  Poiché i termini $(x_i - \bar{x})$ e $-(x_i - \bar{x})$ sono necessariamente sempre discordi, in questo caso la covarianza è grande e negativa.

La covarianza: formula per il calcolo

- La covarianza ammette una seconda definizione, spesso utilizzata in pratica perché **semplice da calcolare**.

- **Covarianza**. La covarianza delle coppie di dati $(x_1, y_1), \dots, (x_n, y_n)$ è

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

- La dimostrazione si ottiene facilmente come segue

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

Il secondo termine è pari a zero, essendo la somma degli scarti dalla media. Pertanto

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

Esempio di calcolo della covarianza

- **Dati** $x_1, \dots, x_4$: `1, 3, 2, 5`.
- **Dati** $y_1, \dots, y_4$: `4, 9, 5, 6`.
- Le medie aritmetiche (**momenti primi**) dei dati sono $\bar{x} = 2.75$ e $\bar{y} = 6$.
- La media dei prodotti (**momento misto**) è pari a

$$\frac{1}{n}\sum_{i=1}^{n} x_i y_i = \frac{1 \times 4 + 3 \times 9 + 2 \times 5 + 5 \times 6}{4} = 17.75.$$

- Pertanto la covarianza è pari a

$$\text{cov}(x, y) = \left(\frac{1}{n}\sum_{i=1}^{n} x_i y_i\right) - \bar{x}\bar{y} = 17.75 - 6 \times 2.75 = 1.25.$$

- **Esercizio**. Si ottenga la covarianza utilizzando la prima definizione.

Covarianza: trasformazioni lineari

- **Proprietà**. Se consideriamo i dati trasformati $v_1, \dots, v_n$ e $w_1, \dots, w_n$, tali che

$$v_i = a_x + b_x x_i, \qquad w_i = a_y + b_y y_i, \qquad i = 1, \dots, n,$$

dove $a_x, b_x, a_y, b_y \in \mathbb{R}$ sono quattro numeri reali, allora

$$\text{cov}(v, w) = b_x b_y \text{cov}(x, y).$$

- La relazione precedente permette di calcolare agevolmente la covarianza tra le v_i e le w_i senza dover calcolare le v_i stesse.

- La dimostrazione segue dalle proprietà della media e delle sommatorie, infatti

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n (v_i - \bar{v})(w_i - \bar{w}) &= \frac{1}{n}\sum_{i=1}^n (a_x + b_x x_i - a_x - b_x \bar{x})(a_y + b_y y_i - a_y - b_y \bar{y}) \\ &= \frac{1}{n}\sum_{i=1}^n b_x b_y (x_i - \bar{x})(y_i - \bar{y}) = b_x b_y \frac{1}{n}\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).\end{aligned}$$

La varianza della somma di due variabili

- La covarianza risulta utile anche quando vogliamo ottenere la varianza di una nuova variabile che sia la somma di altre due variabili.

- **Proprietà**. Siano $x_1, \dots, x_n$ e $y_1, \dots, y_n$ due insiemi di dati e siano $w_1, \dots, w_n$ dei dati trasformati tali che
$$w_i = x_i + y_i, \qquad i = 1, \dots, n.$$
Allora vale che
$$\text{var}(w) = \text{var}(x) + \text{var}(y) + 2\text{cov}(x, y).$$

- La dimostrazione anche in questo caso è immediata:

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}(w_i - \bar{w})^2 &= \frac{1}{n}\sum_{i=1}^{n}(x_i + y_i - (\bar{x} + \bar{y}))^2 = \frac{1}{n}\sum_{i=1}^{n}((x_i - \bar{x}) + (y_i - \bar{y}))^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 + \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 + \frac{2}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).\end{aligned}$$

La matrice delle varianze e covarianze

- Nel caso in esame, troviamo che

$$
\begin{aligned}
\text{cov}(\texttt{fertilità}, \texttt{agricoltura}) &= 98.0, \\
\text{cov}(\texttt{fertilità}, \texttt{istruzione}) &= -5.1, \\
\text{cov}(\texttt{agricoltura}, \texttt{istruzione}) &= -11.9.
\end{aligned}
$$

- Tipicamente, le varianze e le covarianze di tutte le coppie di variabili vengono organizzate in una matrice, chiamata **matrice delle varianze e covarianze**.

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

- In tale matrice, l'elemento in posizione (i, j) rappresenta la covarianza tra la variabile i-esima e la variabile j-esima.

- Nella diagonale ci sono le varianze, poiché $\text{cov}(x, x) = \text{var}(x)$. Inoltre, poichè $\text{cov}(x, y) = \text{cov}(y, x)$, la matrice è **simmetrica**.

Grande quanto?

- L'esempio illustra uno dei problemi connessi con l'utilizzo della covarianza.
- L'**interpretazione del segno** non pone nessuno problema. Le covarianze riportate ci indicano una relazione tendenzialmente crescente tra `fertilità` ed `agricoltura` ed una relazione tendenzialmente decrescente tra queste due variabili e l'istruzione.
- Però, ad esempio, che $\text{cov}(\texttt{fertilità}, \texttt{agricoltura})$ risulti uguale a 98.0 indica un legame debole o forte tra le due variabili?
- Per rispondere alla domanda avremmo bisogno di conoscere un **estremo superiore**, possibilmente con una chiara interpretazione, per il valore assoluto della covarianza.

Minimo e massimo della covarianza

- Il minimo ed il massimo valore che la covarianza può assumere sono noti.
- In particolare, il valore assoluto della covarianza non è mai superiore al prodotto degli scarti quadratici medi. Questo aiuta moltissimo la sua interpretazione.
- **Proprietà**. Siano $x_1, \dots, x_n$ e $y_1, \dots, y_n$ due insiemi di dati, allora

$$-\text{sqm}(x)\text{sqm}(y) \leq \text{cov}(x, y) \leq \text{sqm}(x)\text{sqm}(y).$$

Di conseguenza si ottiene che

$$|\text{cov}(x, y)| \leq \text{sqm}(x)\text{sqm}(y).$$

- Nell'esempio precedente, questo significa che

$$\text{cov}(\texttt{fertilità}, \texttt{agricoltura}) = 98.0 \leq \sqrt{152.7}\sqrt{504.8} = 277.64,$$

ovvero circa la frazione $0.35 = 98.0/277.64$ del valore massimo.

Dimostrazione

- Per due variabili x ed y poniamo $\sigma_x = \text{sqm}(x)$ e $\sigma_y = \text{sqm}(y)$. Allora

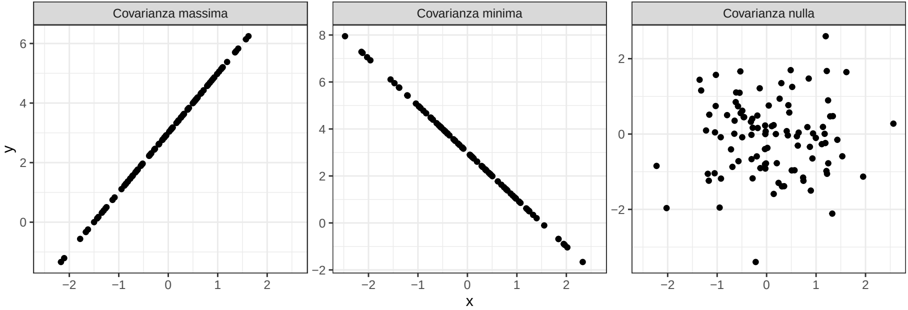
$$
\begin{aligned}
\text{var}\left(\frac{x}{\sigma_x} + \frac{y}{\sigma_y}\right) &= \text{var}\left(\frac{x}{\sigma_x}\right) + \text{var}\left(\frac{y}{\sigma_y}\right) + 2\text{cov}\left(\frac{x}{\sigma_x}, \frac{y}{\sigma_y}\right) \\
&= \frac{1}{\sigma_x^2}\text{var}(x) + \frac{1}{\sigma_y^2}\text{var}(y) + \frac{2}{\sigma_x\sigma_y}\text{cov}(x, y) \\
&= 2\left(1 + \frac{1}{\sigma_x\sigma_y}\text{cov}(x, y)\right).
\end{aligned}
$$

- Trattandosi di una varianza, allora otteniamo che

$$
2\left(1 + \frac{\text{cov}(x, y)}{\sigma_x\sigma_y}\right) \geq 0 \quad \iff \quad \frac{\text{cov}(x, y)}{\sigma_x\sigma_y} \geq -1 \quad \iff \quad \text{cov}(x, y) \geq -\sigma_x\sigma_y.
$$

- Procedendo in maniera analoga, la seconda diseguaglianza di ottiene considerando $\text{var}\,(x/\sigma_x - y/\sigma_y)$, da cui segue (dopo un po' di conti...) che $\text{cov}(x, y) \leq \sigma_x\sigma_y$.

Minimo e massimo della covarianza



- La covarianza è massima, ovvero $\text{cov}(x, y) = \text{sqm}(x)\text{sqm}(y)$, quando i punti sono allineati lungo una **retta crescente**.
- La covarianza è minima, ovvero $\text{cov}(x, y) = -\text{sqm}(x)\text{sqm}(y)$, quando i punti sono allineati lungo una **retta decrescente**.
- La covarianza è nulla, ovvero $\text{cov}(x, y) = 0$, quando i punti sono dispersi.

Il coefficiente di correlazione

- Per affermare se la covarianza è piccola o grande dobbiamo quindi confrontarla con il prodotto degli scarti quadratici medi.

- Di conseguenza, solitamente la covarianza viene presentata direttamente nella sua forma **normalizzata**, chiamata correlazione.

- **Coefficiente di correlazione (lineare)**. La correlazione delle coppie di dati $(x_1, y_1), \dots, (x_n, y_n)$ è

$$\rho = \text{cor}(x, y) = \frac{\text{cov}(x, y)}{\text{sqm}(x)\text{sqm}(y)} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\sigma_x}\right)\left(\frac{y_i - \bar{y}}{\sigma_y}\right),$$

dove $\sigma_x = \text{sqm}(x)$ e $\sigma_y = \text{sqm}(y)$.

- Il coefficiente di correlazione è quindi pari alla **covarianza** dei **dati standardizzati**.

La matrice di correlazione

- Nel caso in esame, troviamo che

$$
\begin{aligned}
\text{cor}(\texttt{fertilità}, \texttt{agricoltura}) &= 0.35,\\
\text{cor}(\texttt{fertilità}, \texttt{istruzione}) &= -0.52,\\
\text{cor}(\texttt{agricoltura}, \texttt{istruzione}) &= -0.68.
\end{aligned}
$$

- Anche le correlazioni vengono tipicamente organizzate in una matrice, chiamata **matrice di correlazione**.

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

- In tale matrice, l'elemento in posizione (i, j) rappresenta la correlazione tra la variabile i-esima e la variabile j-esima.

- **Esercizio**. La diagonale è pari 1, dato che $\text{cor}(x, x) = 1$. Si verifichi quest'ultima equazione.

Interpretazione della correlazione

- **Proprietà**. Siano $x_1, \dots, x_n$ e $y_1, \dots, y_n$ due insiemi di dati, allora

$$-1 \le \mathrm{cor}(x, y) \le 1.$$

Questa proprietà segue dalle proprietà della covarianza.

- Se $\mathrm{cor}(x, y) < 0$ allora i dati indicano una **associazione negativa** tra le due variabili (al crescere di una l'altra decresce). Se $\mathrm{cor}(x, y) = -1$ allora i dati sono perfettamente allineati lungo una retta descrescente.

- Se $\mathrm{cor}(x, y) = 0$ (in pratica se $\mathrm{cor}(x, y) \approx 0$), allora non esiste una relazione lineare tra le due variabili.

- Se $\mathrm{cor}(x, y) > 0$ allora i dati indicano una **associazione positiva** tra le due variabili (al crescere di una, cresce anche l'altra). Se $\mathrm{cor}(x, y) = 1$ allora i dati sono perfettamente allineati lungo una retta crescente.

La correlazione misura relazioni lineari

- Per ragioni che diventeranno esplicite nell'Unità K, 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 $\text{cor}(x, y) < 1$.

- **Esempio**. Si considerino i dati $x_1, \dots, x_5$ pari a $-2, -1, \dots, 2$ e si consideri

$$y_i = e^{x_i}, \qquad i = 1, \dots, 5.$$

  Nonostante la relazione tra le variabili x ed y sia monotona, $\text{cor}(x, y) = 0.89 < 1$.

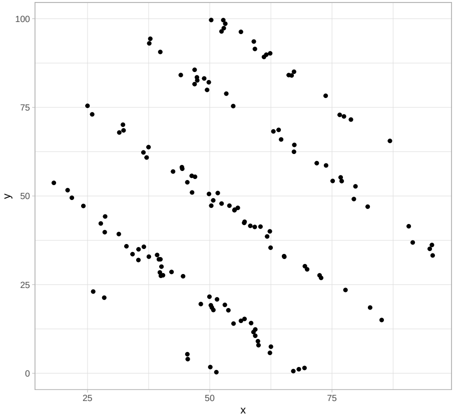
- Il fatto che $\text{cor}(x, y) = 0$ non permette di escludere la presenza di relazioni non-monotone nei dati.

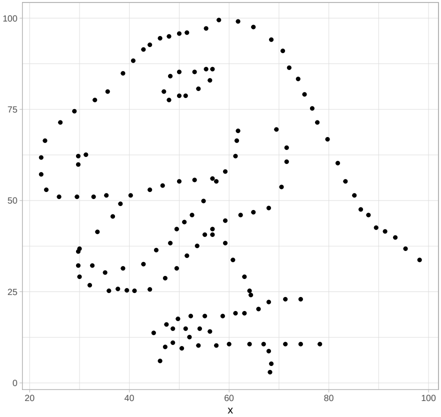
- **Esempio**. Si considerino i dati $x_1, \dots, x_5$ pari a $-2, -1, \dots, 2$ e si consideri

$$y_i = x_i^2, \qquad i = 1, \dots, 5.$$

  Nonostante ci sia una relazione ben precisa tra le variabili x ed y, $\text{cor}(x, y) = 0$.

Insiemi di dati con correlazione nulla





- Entrambi questi insiemi di dati hanno **correlazione nulla** ($\rho = 0$).

La correlazione e trasformazioni non lineari

- Nonostante quanto visto nella slide precedente, la covarianza tra una variabile ed una sua trasformazione monotona crescente (decrescente) è sempre positiva (negativa).

- **Proprietà**. Sia $x_1, \dots, x_n$ un insieme di dati e si considerino i dati trasformati

$$y_i = g(x_i),$$

per $i = 1, \dots, n$, dove $g(x)$ è una funzione **monotona crescente**. Allora:

$$\text{cov}\{x, g(x)\} \geq 0 \qquad \text{e quindi} \qquad \text{cor}\{x, g(x)\} \geq 0.$$

- Se invece $g(x)$ una è funzione **monotona decrescente**, allora:

$$\text{cov}\{x, g(x)\} \leq 0 \qquad \text{e quindi} \qquad \text{cor}\{x, g(x)\} \leq 0.$$

Dimostrazione I

- Sia $g(x)$ una funzione monotona **crescente**. Il caso di funzione monotona **decrescente** è lasciato come esercizio. Anzitutto, notiamo che

$$
\begin{aligned}
\text{cov}\{x, g(x)\} &= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\{g(x_i) - g(\bar{x}) + g(\bar{x}) - \bar{y}\} \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\{g(x_i) - g(\bar{x})\} + \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\{g(\bar{x}) - \bar{y}\} \\
&= \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\{g(x_i) - g(\bar{x})\}.
\end{aligned}
$$

- Per definizione di funzione monotona crescente, abbiamo che

$$
\text{se } x_i \geq \bar{x}, \qquad \text{allora} \qquad g(x_i) \geq g(\bar{x}),
$$

mentre

$$
\text{se } x_i \leq \bar{x}, \qquad \text{allora} \qquad g(x_i) \leq g(\bar{x}).
$$

Dimostrazione II

- Ovviamente, di conseguenza:

$$\text{se } x_i - \bar{x} \geq 0, \qquad \text{allora} \qquad g(x_i) - g(\bar{x}) \geq 0,$$

mentre

$$\text{se } x_i - \bar{x} \leq 0, \qquad \text{allora} \qquad g(x_i) - g(\bar{x}) \leq 0.$$

- Questi risultati implicano in particolare che:

$$(x_i - \bar{x})\{g(x_i) - g(\bar{x})\} \geq 0,$$

dato che i segni della quantità $x_i - \bar{x}$ e della quantità $g(x_i) - g(\bar{x})$ sono **concordi**.

- Quindi, questi risultati implicano che

$$\text{cov}\{x, g(x)\} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})\{g(x_i) - g(\bar{x})\} \geq 0,$$

essendo la somma di termini non-negativi.

Ulteriori osservazioni, modelli lineari

- In questa unità abbiamo presentato covarianza e correlazione come degli indici in grado di misurare l'associazione tra due variabili numeriche.
- In particolare, ci siamo posti in maniera **simmetrica** rispetto alle variabili.
- Tuttavia tali indici si possono anche "scoprire" analizzando il problema da un punto di vista diverso.
- Infatti, covarianza e correlazione sono strettamente collegate ai **modelli di regressione lineare**, ovvero l'argomento della prossima unità.