Lezione 9

Nella lezione precedente, come esercizio, si sono calcolati alcuni valori caratteristici (media e varianza) sulle distribuzioni marginali delle due variabili X e Y.

Le stesse elaborazioni possono essere effettuate anche sulle distribuzioni condizionate e vedremo più avanti come i risultati ottenuti servano per verificare l'esistenza di possibili legami fra le variabili e per misurarne l'intensità.

VALORI CARATTERISTICI DELLE DISTRIBUZIONI CONDIZIONATE

Il generico **momento ordinario di ordine r** delle k distribuzioni della variabile quantitativa Y condizionata a X (che può essere di qualsiasi tipo) assume la forma seguente

$$m_{ry|c_j} = \frac{1}{n_{j.}} \sum_{l=1}^{h} d_l^r n_{jl} = \frac{1}{f_{j.}} \sum_{l=1}^{h} d_l^r f_{jl}$$

(per $r = 0, 1, 2, \ldots$) se la Y è discreta e la forma

$$m_{ry|c_j} = \frac{1}{n_{j.}} \sum_{l=1}^{h} \bar{d}_l^r n_{jl} = \frac{1}{f_{j.}} \sum_{l=1}^{h} \bar{d}_l^r f_{jl}$$

se la Y è continua e i suoi valori sono raggruppati in classi

Va sottolineato il fatto che le k distribuzioni della Y|x costituiscono le distribuzioni della Y all'interno di k gruppi distinti. Di conseguenza tutte le

considerazioni fatte a proposito di una popolazione suddivisa in gruppi valgono anche in questo caso.

Ricordando quanto detto nelle lezioni precedenti risulta quindi

- La media generale della Y corrisponde alla media ponderata delle medie delle distribuzioni condizionate (più brevemente, medie condizionate), dove i pesi corrispondono alle numerosità dei gruppi

- La varianza complessiva della Y corrisponde alla somma della varianza within e della varianza between, ossia alla somma della media delle varianze delle distribuzioni condizionate più la varianza delle medie delle distribuzioni condizionate

Esempio

Considerata la tabella

X\Y	-1 – 1	1 – 5	5 – 10	
1	0.0	0.1	0.1	0.2
3	0.1	0.3	0.4	0.8
	0.1	0.4	0.5	1.0

si determini:

a) la media e la varianza della X e della Y

b) le medie e le varianze delle due distribuzioni condizionate: Y|x=1 e Y|x=3

c) si verifichi che la media generale della Y corrisponde alla media delle due medie condizionate ponderate con le numerosità dei gruppi omogenei in X

d) si verifichi che la varianza della Y corrisponde alla somma della varianza within e della varianza between

a) I valori caratteristici della X e della Y si calcolano sulla base della loro distribuzione marginale, per cui risulta

$$\bar{x} = 1 \times 0.2 + 3 \times 0.8 = 2.6$$

$$m_{2x} = 1 \times 0.2 + 9 \times 0.8 = 7.4$$

$$s_x^2 = 7.4 - 2.6^2 = 0.64$$

$$\bar{y} = 3 \times 0.4 + 7.5 \times 0.5 = 4.95$$

$$m_{2y} = 9 \times 0.4 + 56.25 \times 0.5 = 31.725$$

$$s_y^2 = 31.725 - 4.95^2 = 7.2225$$

b) la media, il secondo momento ordinario e la varianza della Y|x=1 sono rispettivamente pari a

$$\bar{y}_1 = \frac{1}{0.2}(3 \times 0.1 + 7.5 \times 0.1) = 5.25$$

$$m_{2y|1} = \frac{1}{0.2}(3^2 \times 0.1 + 7.5^2 \times 0.1) = 32.625$$

$$s_{y|1}^2 = 32.625 - 5.25^2 = 5.0625$$

mentre la media, il secondo momento ordinario e la varianza della Y|x=3 sono rispettivamente pari a

$$\bar{y}_3 = \frac{1}{0.8}(3 \times 0.3 + 7.5 \times 0.4) = 4.875$$

$$m_{2y|3} = \frac{1}{0.8}(3^2 \times 0.3 + 7.5^2 \times 0.4) = 31.5$$

$$s_{y|3}^2 = 31.5 - 4.875^2 = 7.734375$$

c) La media ponderata delle medie delle distribuzioni condizionate è

$$\bar{y}_1 \times 0.2 + \bar{y}_3 \times 0.8 = 5.25 \times 0.2 + 4.875 \times 0.8 = 4.95 = \bar{y}$$

d) La varianza delle medie delle distribuzioni condizionate è

$$s_b^2 = (5.25 - 4.95)^2 \times 0.2 + (4.875 - 4.95)^2 \times 0.8 = 0.0225$$

mentre la media delle varianze delle distribuzioni condizionate è

$$s_w^2 = 5.0625 \times 0.2 + 7.734375 \times 0.8 = 7.2$$

La loro somma risulta

$$s_b^2 + s_w^2 = 0.0225 + 7.2 = 7.2225 = s_y^2$$

Fino a questo momento si sono esaminati i valori caratteristici che possono essere calcolati sulle due distribuzioni marginali della X e della Y e sulle k distribuzioni della Y|x.

Esistono però altri importanti indici che vengono calcolati sulla distribuzione congiunta delle due variabili e che saranno l'argomento delle pagine successive.

MOMENTI MISTI DALL'ORIGINE (O MOMENTI MISTI ORDINARI)

Considerate due variabili quantitative X e Y, sulla loro distribuzione congiunta possono essere calcolati sia i momenti misti dall'origine sia i momenti misti centrali.

Dato che in seguito si utilizzeranno solo i momenti misti di ordine 1,1, ci si limiterà ad analizzare solo questo caso particolare.

Si definisce come **momento misto dall'origine** (o **ordinario**) **di ordine 1,1 la media del prodotto fra le due variabili**

Se si dispone della **sequenza** delle n coppie delle determinazioni delle due variabili, il momento misto ordinario (o dall'origine) di ordine 1,1, indicato con $m_{1,1}$, corrisponde quindi a

$$m_{1,1} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i$$

Esempio

Considerate 4 unità statistiche su cui le variabili X e Y hanno assunto i valori seguenti

(-1, 10) (0, 8) (3, 7) (2, 5)

calcolare il momento misto ordinario di ordine 1,1.

Si ottiene

$$m_{1,1} = \frac{1}{4}(-1 \times 10 + 0 \times 8 + 3 \times 7 + 2 \times 5) = 5.25$$

Il momento misto ordinario di ordine 1,1 si ottiene effettuando il prodotto delle intensità delle due variabili e calcolando poi la media di tali prodotti.

Se i dati raccolti sono organizzati in una **distribuzione di frequenza**, il momento misto ordinario (o dall'origine) di ordine 1,1 corrisponde a

$$m_{1,1} = \frac{1}{n}\sum_{j=1}^{k}\sum_{l=1}^{h} c_j d_l n_{jl} = \sum_{j=1}^{k}\sum_{l=1}^{h} c_j d_l f_{jl}$$

Va osservato che se una variabile è continua e i suoi valori sono raggruppati in classi, si usa sempre la medesima formula, sostituendo alla generica intensità il valore centrale della classe.

Esempi

1) Data la seguente distribuzione bivariata, se ne calcoli il momento misto ordinario di ordine 1,1

X\Y	1	2	3	
3	20	10	0	30
4	5	5	10	20
	25	15	10	50

In questo caso occorre tenere presente che a ogni prodotto fra le determinazioni della X e della Y va associata la frequenza corrispondente. Alla coppia di determinazioni (3, 1) è infatti associata una frequenza 20, alla coppia (3, 2) una frequenza 10, mentre la coppia (3,3) non si mai è osservata sulle unità statistiche esaminate.

Applicando la formula, il momento misto ordinario di ordine 1,1 risulta pari a

$$m_{1,1} = \frac{3 \times 1 \times 20 + 3 \times 2 \times 10 + 3 \times 3 \times 0 + 4 \times 1 \times 5 + 4 \times 2 \times 5 + 4 \times 3 \times 10}{50} = 6$$

2)

Considerata la distribuzione seguente

X\Y	-1 – 1	1 – 5	5 – 10	
0 – 2	0.0	0.1	0.1	0.2
2 – 4	0.1	0.3	0.4	0.8
	0.1	0.4	0.5	1.0

Il momento misto ordinario di ordine 1,1 corrisponde a

$$m_{1,1} = 1 \times 0 \times 0 + 1 \times 3 \times 0.1 + 1 \times 7.5 \times 0.1 + 3 \times 0 \times 0.1 + 3 \times 3 \times 0.3 + 3 \times 7.5 \times 0.4$$
$$= 12.75$$

MOMENTI MISTI CENTRALI

Il momento misto centrale di ordine 1,1, che viene indicato con il termine **covarianza**, corrisponde alla media aritmetica del prodotto degli scarti delle due variabili dalla propria media.

Il fatto che questo momento abbia un nome specifico sottolinea l'importanza che gli si attribuisce in statistica.

Le formule per calcolare la covarianza variano, come sempre, a seconda di come sono organizzati i dati raccolti.

Nel caso di una **sequenza** di n coppie di valori, la covarianza fra X e Y, di solito indicata con i simboli s_{xy} oppure $\bar{m}_{1,1}$, corrisponde a

$$s_{xy} = \bar{m}_{1,1} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})$$

Se i dati raccolti sono organizzati in una **distribuzione di frequenza**, la covarianza corrisponde a

$$s_{xy} = \bar{m}_{1,1} = \frac{1}{n}\sum_{j=1}^{k}\sum_{l=1}^{h}(c_j - \bar{x})(d_l - \bar{y})n_{jl} = \sum_{j=1}^{k}\sum_{l=1}^{h}(c_j - \bar{x})(d_l - \bar{y})f_{jl}$$

Anche in questo caso, se una variabile è continua e data per classi di valori, si usa sempre la medesima formula, sostituendo alla generica intensità il valore centrale della classe.

Prima di vedere qualche esempio numerico è opportuno dimostrare una prima proprietà della covarianza, che consente di calcolarla in modo molto più semplice

PRIMA PROPRIETÀ

La covarianza fra due variabili X e Y corrisponde alla media del prodotto delle due variabili meno il prodotto delle loro medie. In simboli

$$\bar{m}_{1,1} = s_{xy} = m_{1,1} - \bar{x}\bar{y}$$

Per effettuare questa dimostrazione è sufficiente sviluppare la formula originaria della covarianza

DIMOSTRAZIONE

Date *n* coppie di osservazioni relative alle due variabili quantitative X e Y, la covarianza è data da

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

Distribuendo la somma si ottiene

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \frac{1}{n}\sum_{i=1}^{n}\bar{x}y_i - \frac{1}{n}\sum_{i=1}^{n} x_i\bar{y} + \frac{1}{n}\sum_{i=1}^{n}\bar{x}\bar{y}$$

ed anche

$$s_{xy} = \frac{1}{n}\sum_{i=1}^{n} x_i y_i - \bar{x}\frac{1}{n}\sum_{i=1}^{n} y_i - \bar{y}\frac{1}{n}\sum_{i=1}^{n} x_i + \frac{1}{n}\sum_{i=1}^{n} \bar{x}\bar{y}$$

Il primo termine, nel rettangolo azzurro, corrisponde al momento misto ordinario di ordine 1,1.

Il termine nel rettangolo giallo corrisponde al prodotto di $\bar{x}$ per la media della Y e, in modo analogo, la quantità riportata nel rettangolo arancione è ancora pari al prodotto delle medie $\bar{x}\bar{y}$.

Anche l'ultimo termine, nel rettangolo viola, è pari a $\bar{x}\bar{y}$ perché è la media della costante $\bar{x}\bar{y}$.

In conclusione, si ha

$$s_{xy} = m_{1,1} - \bar{x}\bar{y} - \bar{x}\bar{y} + \bar{x}\bar{y} = m_{1,1} - \bar{x}\bar{y}$$

Questa ottenuta è la formula più semplice da utilizzare per il calcolo della covarianza

Esempi

1) Considerata la sequenza delle coppie (x_i, y_i) relative a due variabili discrete X e Y

(-1, 10) (0, 8) (3, 7) (2, 5)

calcolarne la covarianza

Le medie delle due variabili assumono i valori

$$\bar{x} = \frac{-1 + 3 + 2}{4} = 1$$

$$\bar{y} = \frac{10 + 8 + 7 + 5}{4} = 7.5$$

mentre il momento misto ordinario di ordine 1,1 è

$$m_{1,1} = \frac{-1 \times 10 + 3 \times 7 + 2 \times 5}{4} = 5.25$$

per cui la covarianza è pari a

$$s_{xy} = 5.25 - 1 \times 7.5 = -2.25$$

2)

Data la seguente distribuzione bivariata

X\Y	1	2	3	
3	20	10	0	30
4	5	5	10	20
	25	15	10	50

se ne calcoli la covarianza

La media della variabile X è pari a

$$\bar{x} = \frac{3 \times 30 + 4 \times 20}{50} = 3.4$$

La media della variabile Y è pari a

$$\bar{y} = \frac{1 \times 25 + 2 \times 15 + 3 \times 10}{50} = 1.7$$

Il momento misto ordinario di ordine 1,1 è dato da

$$m_{1,1} = \frac{3 \times 1 \times 20 + 3 \times 2 \times 10 + 4 \times 1 \times 5 + 4 \times 2 \times 5 + 4 \times 3 \times 10}{50} = 6.0$$

e quindi la covarianza risulta pari a

$$s_{xy} = 6 - 3.4 \times 1.7 = 0.22$$

3)

Data la seguente distribuzione bivariata

X\Y	-1 – 1	1 – 5	5 – 10	
0 – 2	0.0	0.1	0.1	0.2
2 – 4	0.1	0.3	0.4	0.8
	0.1	0.4	0.5	1.0

se ne calcoli la covarianza

Risulta

$$\bar{x} = 1 \times 0.2 + 33.4$$

$\bar{x} = 1 \times 0.2 + 3 \times 0.8 = 2.6$

$\bar{y} = 0 \times 0.1 + 3 \times 0.4 + 7.5 \times 0.5 = 4.95$

$m_{1,1} = 1 \times 3 \times 0.1 + 1 \times 7.5 \times 0.1 + 3 \times 3 \times 0.3 + 3 \times 7.5 \times 0.4 = 12.75$

$s_{xy} = 12.75 - 2.6 \times 4.95 = -0.12$

Come si nota anche dai risultati degli esempi, la covarianza può assumere un valore minore, maggiore o uguale a zero e il suo segno indica il tipo di relazione fra le variabili considerate.

Se la relazione è diretta (al crescere dei valori di X anche Y tende a crescere) le variabili si dicono **concordi.** Gli scarti $(X - \bar{x})$ e $(Y - \bar{y})$ avranno

tendenzialmente lo stesso segno: a scarti negativi della X tenderanno a essere associati scarti negativi della Y e a scarti positivi della X tenderanno a essere associati scarti positivi della Y. Il prodotto di tali scarti avrà segno positivo nella maggior parte dei casi.
Di conseguenza la covarianza, che corrisponde alla media dei prodotti di tali scarti, avrà segno positivo.

Se due variabili sono concordi la loro covarianza assume un valore maggiore di zero

Se la relazione è inversa (al crescere dei valori di X la Y tende a decrescere, e viceversa) le variabili si dicono **discordi.** Gli scarti $(X - \bar{x})$ e $(Y - \bar{y})$ avranno tendenzialmente segno opposto: a scarti negativi della X tenderanno a essere associati scarti positivi della Y e a scarti positivi della X tenderanno a essere associati scarti negativi della Y. Il prodotto di tali scarti avrà quindi segno negativo nella maggior parte dei casi.
Di conseguenza la covarianza, che corrisponde alla media dei prodotti di tali scarti, avrà segno negativo.

Se due variabili sono discordi la loro covarianza assume un valore minore di zero

SECONDA PROPRIETÀ

La covarianza è invariante rispetto a eventuali traslazioni, ma non lo è rispetto a cambiamenti di scala.

Per effettuare questa dimostrazione occorre innanzitutto considerare due trasformazioni lineari W e Z delle variabili X e Y aventi covarianza s_{xy} *e poi scrivere la formula della covarianza fra W e Z andando ad effettuare le sostituzioni opportune all'interno di tale formula*

DIMOSTRAZIONE

Date due variabili X e Y con covarianza si considerino le variabili trasformate

$$W = a + bX$$

$$Z = a' + b'Y$$

Per definizione, la loro covarianza è data da

$$s_{wz} = \frac{1}{n}\sum_{i=1}^{n}[(w_i - \bar{w})(z_i - \bar{z})]$$

Dalla relazione esistente fra W e X risulta

$w_i = a + bx_i$

$\bar{w} = a + b\bar{x}$

mentre dalla relazione esistente fra Z e Y si ha

$z_i = a' + b'y_i$

$\bar{z} = a' + b'\bar{y}$

Andando a effettuare queste sostituzioni nella formula della covarianza fra W e Z si ottiene

$$s_{wz} = \frac{1}{n}\sum_{i=1}^{n}[(a + bx_i - a - b\bar{x})(a' + b'y_i - a' - b'\bar{y})] =$$

$$= \frac{1}{n}\sum_{i=1}^{n}[(bx_i - b\bar{x})(b'y_i - b'\bar{y})] = b\,b'\frac{1}{n}\sum_{i=1}^{n}[(x_i - \bar{x})(y_i - \bar{y})]$$

dove

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

per definizione, è la covarianza fra le variabili X e Y.

Si ottiene quindi

$$s_{wz} = bb's_{xy}$$

Esempio

Considerate due variabili X e Y con covarianza $s_{xy} = -2$ si calcoli la covarianza delle variabili trasformate

$$W = -\frac{1}{2} + 0.5X$$

$$Z = 3 - 0.1Y$$

In questo caso i parametri b e b' sono rispettivamente pari a

$b = 0.5$

$b' = -0.1$

Quindi la covarianza fra W e Z risulta

$$s_{wz} = 0.5 \times (-0.1) \times (-2) = 0.1$$

COEFFICIENTE DI CORRELAZIONE LINEARE

Date due variabili quantitative X e Y si considerino le corrispondenti variabili scarto standardizzato

$$U = \frac{X - \bar{x}}{s_x}$$

$$V = \frac{Y - \bar{y}}{s_y}$$

La covarianza fra U e V è un caso particolare della dimostrazione precedente dove

$b = \frac{1}{s_x}$

$b' = \frac{1}{s_y}$

per cui si ottiene

$$s_{uv} = \frac{1}{s_x}\frac{1}{s_y}s_{xy}$$

La covarianza fra le due variabili standardizzate U e V è il cosiddetto **coefficiente di correlazione lineare** calcolato per le variabili X e Y, che viene indicato con il simbolo r_{xy}

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

Come sarà chiarito in seguito, questo è uno degli indici statistici più frequentemente calcolato quando le due variabili di interesse sono entrambe quantitative.

Per il momento si può solo affermare che per due variabili X e Y concordi r_{xy} assume un valore positivo, mentre se le variabili sono discordi l'indice assume un valore negativo (le deviazioni standard che compaiono al denominatore sono infatti sempre positive e la covarianza è minore di zero per variabili discordi e maggiore di zero per variabili concordi).

Nella formula del coefficiente di correlazione lineare compaiono indici già noti, per cui può essere calcolato senza difficoltà.

ESERCIZI

1) Considerate le seguenti coppie di valori relative a due variabili X e Y

(-2, 0) (0, 1) (1, 1) (2, 3) (3, 5)

si calcoli il coefficiente di correlazione lineare r_{xy} fra le due variabili e si stabilisca se sono concordi o discordi.

Per la variabile X risulta

$\bar{x} = 0.8$

$m_{2x} = 3.6$

$s_x^2 = 2.96$

Per la variabile Y risulta

$\bar{y} = 2.0$

$m_{2y} = 7.2$

$s_y^2 = 3.2$

Il momento misto dall'origine di ordine 1,1 risulta

$$m_{1,1} = \frac{1 \times 1 + 2 \times 3 + 3 \times 5}{5} = 4.4$$

per cui la covarianza è

$$s_{xy} = m_{1,1} - \bar{x}\bar{y} = 4.4 - 0.8 \times 2 = 2.8$$

Il coefficiente di correlazione lineare è quindi pari a

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}} = \frac{2.8}{\sqrt{2.96 \times 3.2}} \approx 0.9098$$

Le variabili X e Y sono concordi

2)

Considerata la seguente distribuzione bivariata

X\Y	– 1	0	1	
-1	2	3	5	10
0	3	2	0	5
	5	5	5	15

calcolare r_{xy} e si stabilisca se X e Y sono concordi o discordi.

Per la variabile X risulta

$\bar{x} = -0.\overline{6}$

$m_{2x} = 0.\overline{6}$

$s_x^2 = 0.\overline{2}$

Per la variabile Y risulta

$\bar{y} = 0$

$m_{2y} = 0.\overline{6}$

$s_y^2 = 0.\overline{6}$

Il momento misto dall'origine di ordine 1,1 risulta

$$m_{1,1} = \frac{-1 \times (-1) \times 2 - 1 \times 1 \times 5}{15} = -0.2$$

per cui la covarianza è

$s_{xy} = m_{1,1} - \bar{x}\bar{y} = -0.2 - (0.\overline{6}) \times 0 = -0.2$

Il coefficiente di correlazione lineare è quindi pari a

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{s_{xy}}{\sqrt{s_x^2 s_y^2}} = \frac{-0.2}{\sqrt{0.\overline{2} \times 0.\overline{6}}} \approx -0.5196$$

Le variabili sono quindi discordi

PRIMA PROPRIETÀ

Il coefficiente di correlazione lineare è un indice adimensionale, ossia un numero puro, il cui campo di variazione è [-1, +1], ossia

$$-1 \leq r_{xy} \leq 1$$

Questa dimostrazione utilizza la disuguaglianza dovuta a Cauchy-Swartz, secondo cui vale la relazione seguente

$$\left(\sum_{i=1}^{n} v_i z_i\right)^2 \leq \left(\sum_{i=1}^{n} v_i^2\right)\left(\sum_{i=1}^{n} z_i^2\right)$$

dove

$$v_1, v_2, \dots, v_n$$

e

$$z_1, z_2, \dots, z_n$$

sono due ennuple di numeri reali.

Una volta scritta la disuguaglianza occorre sostituire al generico termine v_i l'i-esimo valore della variabile scarto per la X, ossia $(x_i - \bar{x})$ e al generico termine z_i l'i-esimo valore della variabile scarto per la Y, ossia $(y_i - \bar{y})$.

Moltiplicando tutti i termini della disuguaglianza per la costante $\frac{1}{n^2}$ e sviluppando tutti i quadrati, si ottiene la dimostrazione

DIMOSTRAZIONE

Considerata la disuguaglianza

$$\left(\sum_{i=1}^{n} v_i z_i\right)^2 \leq \left(\sum_{i=1}^{n} v_i^2\right)\left(\sum_{i=1}^{n} z_i^2\right)$$

sia

$$v_i = x_i - \bar{x}$$

$$z_i = y_i - \bar{y}$$

Effettuando le sostituzioni si ha

$$\left[\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right]^2 \leq \left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]\left[\sum_{i=1}^{n}(y_i-\bar{y})^2\right]$$

Moltiplicando tutti i termini della disuguaglianza per la costante

$$\frac{1}{n^2}$$

si ottiene

$$\frac{1}{n^2}\left[\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right]^2 \leq \frac{1}{n^2}\left[\sum_{i=1}^{n}(x_i-\bar{x})^2\right]\left[\sum_{i=1}^{n}(y_i-\bar{y})^2\right]$$

che corrisponde anche a

$$\left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})\right]^2 \leq \left[\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2\right]\left[\frac{1}{n}\sum_{i=1}^{n}(y_i-\bar{y})^2\right]$$

Il termine a sinistra del segno di disuguaglianza è la covarianza al quadrato

Il primo termine a destra del segno di disuguaglianza è la varianza di X

Il secondo termine a destra del segno di disuguaglianza è la varianza di Y

Risulta quindi verificata la seguente disuguaglianza

$$s_{xy}^2 \leq s_x^2 s_y^2$$

per cui il quadrato della covarianza risulta sempre minore o tutt'al più uguale al prodotto delle varianze delle due variabili.
Calcolando la radice quadrata dei due termini della disuguaglianza precedente risulta

$$-s_x s_y \leq s_{xy} \leq s_x s_y$$

ed infine, dividendo tutti i termini per $s_x s_y$ si ottiene

$$-1 \leq \frac{s_{xy}}{s_x s_y} = r_{xy} \leq 1$$

SECONDA PROPRIETÀ

Il coefficiente di correlazione lineare è invariante rispetto a trasformazioni lineari delle due variabili, a parte il segno

Questa dimostrazione si basa sulle proprietà delle trasformazioni lineari dimostrate per la covarianza e per la deviazione standard. Considerate le variabili originarie X e Y, si considerano due loro trasformazioni lineari e si determina il valore della deviazione standard di tali variabili trasformate e della loro covarianza

DIMOSTRAZIONE

Date le variabili X e Y si considerano due loro trasformazioni lineari

$$W = a + bX$$

$$Z = a' + b'Y$$

In base alle proprietà della covarianza risulta

$$s_{wz} = bb's_{xy}$$

mentre per le proprietà della deviazione standard si ha

$$s_w = |b|s_x$$
$$s_z = |b'|s_y$$

Il coefficiente di correlazione lineare fra W e Z corrisponde quindi a

$$r_{wz} = \frac{s_{wz}}{s_w s_z} = \frac{bb's_{xy}}{|b|s_x|b'|s_y} = \frac{bb'}{|b||b'|}\frac{s_{xy}}{s_x s_y} = segno(bb')r_{xy}$$

Il coefficiente di correlazione lineare calcolato per W e Z ha quindi lo stesso valore del coefficiente di correlazione lineare calcolato per X e Y e il suo segno dipende dal segno del prodotto delle due costanti b e b'

ESEMPI

1) Date due variabili X e Y con coefficiente di correlazione lineare $r_{xy} = 0.71$ si determini il coefficiente di correlazione lineare delle variabili trasformate

$$W = -0.3 + 0.5X$$
$$Z = 3 - 8Y$$
$$r_{wz} = segno[0.5 \times (-8)]r_{xy} = -r_{xy} = -0.71$$

2) Date due variabili X e Y con coefficiente di correlazione lineare $r_{xy} = -0.94$ si determini il coefficiente di correlazione lineare delle variabili trasformate

$$W = -8 - 0.5X$$

$$Z = 3 - 2.5Y$$

$$r_{wz} = segno[(-0.5) \times (-2.5)] r_{xy} = r_{xy} = -0.94$$

TERZA PROPRIETÀ

Se fra due variabili quantitative X e Y esiste una relazione lineare diretta, il loro coefficiente di correlazione lineare risulta necessariamente uguale e a $+1$.
Se fra due variabili quantitative X e Y esiste una relazione lineare inversa, il loro coefficiente di correlazione lineare risulta necessariamente uguale e a -1.

Per effettuare questa dimostrazione è necessario considerare la variabile X e una sua trasformata lineare Y, in modo che fra X e Y sussista una relazione lineare che sarà diretta o inversa a seconda del segno del parametro che moltiplica la X.
A questo punto occorre andare a determinare quanto vale il coefficiente di correlazione lineare fra X e Y, andando a calcolare la covarianza fra X e Y e ricordando la proprietà della deviazione standard di una trasformazione lineare

DIMOSTRAZIONE

Considerata una variabile X e la sua trasformazione lineare

$$Y = a + bX$$

il loro coefficiente di correlazione lineare è

$$r_{xy} = \frac{s_{xy}}{s_x s_y}$$

dove

$$s_y = |b| s_x$$

per cui risulta

$$r_{xy} = \frac{s_{xy}}{|b| s_x^2}$$

Per definizione, la covarianza fra X e Y che compare al numeratore è data da

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

Tenendo presente la relazione lineare esistente fra X e Y si ha

$y_i = a + b x_i$

$\bar{y} = a + b\bar{x}$

Sostituendo queste due uguaglianze nella formula della covarianza fra X e Y si ottiene

$$s_{xy} = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(a + b x_i - a - b\bar{x}) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})(b x_i - b\bar{x}) =$$

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

Sostituendo questo risultato nell'espressione di r_{xy} si ottiene

$$r_{xy} = \frac{bs_x^2}{|b|s_x^2} = \frac{b}{|b|} = \begin{cases} -1 & \text{per } b < 0 \\ +1 & \text{per } b > 0 \end{cases}$$

ESEMPIO

Data una variabile X e una sua combinazione lineare

$$Y = 3 - 8.5X$$

determinare il valore coefficiente di correlazione lineare fra X e Y

Dalla dimostrazione precedente risulta

$$r_{xy} = \frac{-8{,}5}{|-8.5|} = -1$$