

APPUNTI¹ SU: DISTRIBUZIONI CONGIUNTE, COVARIANZA E RETTA DI REGRESSIONE

1 Distribuzioni congiunte

Spesso nelle applicazioni si è portati a considerare più quantità aleatorie e a confrontarne il comportamento. Ciò motiva l'introduzione della nozione di *distribuzione congiunta*.

Sia, al solito, $(\Omega, \mathbb{P})$ uno spazio di probabilità, ovvero Ω è uno spazio campionario e $\mathbb{P} : \Omega \rightarrow [0, 1]$ è una (misura di) probabilità su Ω . Su Ω siano definite due variabili aleatorie (v.a.) discrete X e Y :

$$X, Y : \omega \in \Omega \mapsto X(\omega), Y(\omega) \in \mathbb{R}$$

Denotiamo con $E_X = \{x_1, \dots, x_k\}$ (o $E_X = \{x_1, x_2, \dots\}$) e con $E_Y = \{y_1, \dots, y_m\}$ (o $E_Y = \{y_1, y_2, \dots\}$) l'insieme dei valori ammissibili per X e Y rispettivamente. Ricordiamo brevemente che la *legge* o *distribuzione* di X e di Y sono date dalle funzioni p_X e p_Y , rispettivamente, che ad ogni valore ammissibile associano la probabilità con la quale possono verificarsi:

$$\begin{aligned} p_X(x) &= \mathbb{P}\{X = x\} & x \in E_X \\ p_Y(y) &= \mathbb{P}\{Y = y\} & y \in E_Y \end{aligned}$$

Si consideri ora l'applicazione

$$(X, Y) : \Omega \rightarrow \mathbb{R}^2, \quad \omega \mapsto (X(\omega), Y(\omega)).$$

L'immagine di tale applicazione è l'insieme delle coppie $(x, y) \in E_X \times E_Y$. Si definisce **legge** o **distribuzione congiunta** delle v.a. X e Y l'applicazione

$$p(x, y) = \mathbb{P}\{X = x, Y = y\}$$

che ad ogni coppia (ammissibile) $(x, y) \in E_X \times E_Y$ associa la probabilità con la quale tale coppia può essere assunta.

Vedremo nel seguito che tale nozione si rivela estremamente importante e che permette di risolvere numerosi problemi.

Osserviamo dapprima che la legge congiunta di X e Y consente di dedurre le leggi di X e di Y , dette **leggi** o **distribuzioni marginali**. Infatti, per la formula delle probabilità totali, fissato $x_i \in E_X$, si ha

$$p_X(x_i) = \mathbb{P}\{X = x_i\} = \sum_j \mathbb{P}\{X = x_i, Y = y_j\} = \sum_j p(x_i, y_j) \quad (1)$$

¹Queste note sono la traduzione di parte delle lezioni tenute dai Proff. Paolo Baldi e Nicole El Karoui all'Università Pierre et Marie Curie (Paris VI) nell'a.a. 1995/'96.

La legge (marginale) di X in x_i si calcola dunque a partire dalla legge congiunta, “congelando” la prima variabile, x_i , e “saturando” rispetto a tutti i possibili valori che può assumere la Y , ovvero sommando $p(x_i, y_j)$ su tutti i possibili y_j . Discorso analogo riguardo il calcolo della legge (marginale) di Y : fissato $y_j \in E_Y$,

$$p_Y(y_j) = \mathbb{P}\{Y = y_j\} = \sum_i p(x_i, y_j) \quad (2)$$

Esempio 1. Si consideri un’urna contenente 6 palline numerate da 1 a 6; se ne estraggano 2 *con reinserimento*: siano X_1 e X_2 i due numeri ottenuti. Si vuole calcolare la distribuzione congiunta di X_1 e X_2 .

I valori ammissibili per la coppia (X_1, X_2) sono dati dalle coppie (i, j) , al variare di i, j nell’insieme $\{1, \dots, 6\}$. Si tratta quindi di 36 possibili valori e, poiché ogni coppia è equiprobabile, ciascuna coppia ha probabilità $\frac{1}{36}$ di verificarsi. Quindi, $E_{(X_1, X_2)} = \{1, \dots, 6\}^2$. È possibile visualizzare la situazione dei due tiri come nella Figura 1, in cui $\bullet$ denotano i valori possibili.

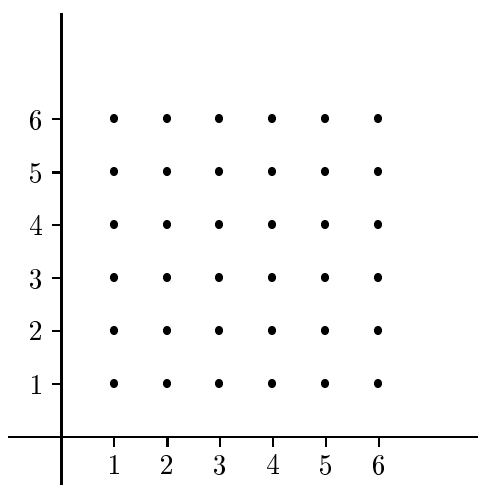


Figura 1 Insieme dei valori ammissibili per la coppia (X_1, X_2) .

Prese singolarmente, le due v.a. X_1 e X_2 possono assumere valori nell’insieme $\{1, \dots, 6\}$ e ciascun valore è assunto con probabilità $\frac{1}{6}$ (esercizio!). Ciò significa che le distribuzioni marginali di X_1 e X_2 sono *uniformi* nell’insieme $\{1, \dots, 6\}$.

Supponiamo ora che le due estrazioni siano effettuate *senza reinserimento*, e denotiamo con Y_1 e Y_2 il risultato della prima e della seconda estrazione rispettivamente. I valori ammissibili per la coppia (Y_1, Y_2) non sono più gli stessi perché, ad esempio, la coppia $(1, 1)$ non può più verificarsi. Infatti, i possibili risultati sono dati dalle coppie (i, j) tali che $i, j \in \{1, \dots, 6\}$ *ma con $i \neq j$* . Si tratta dunque di $30 (= 36 \text{ possibili coppie} - 6 \text{ coppie identiche})$

valori ammissibili, tutti equiprobabili, che dunque possono verificarsi ciascuno con probabilità $\frac{1}{30}$. Quindi, $E_{(Y_1, Y_2)} = \{(i, j) : i \neq j, i, j \in \{1, \dots, 6\}\}$. La legge congiunta di Y_1 e Y_2 è quindi data da

$$p(i, j) = \begin{cases} \frac{1}{30} & \text{se } i, j = 1, \dots, 6 \text{ e } i \neq j \\ 0 & \text{altrimenti} \end{cases}$$

La Figura 2 visualizza tutti i possibili risultati di questo secondo esperimento. Ancora una volta, i simboli $\bullet$ denotano tutti i possibili valori per la coppia (Y_1, Y_2) , che possono essere assunti con probabilità $\frac{1}{30}$.

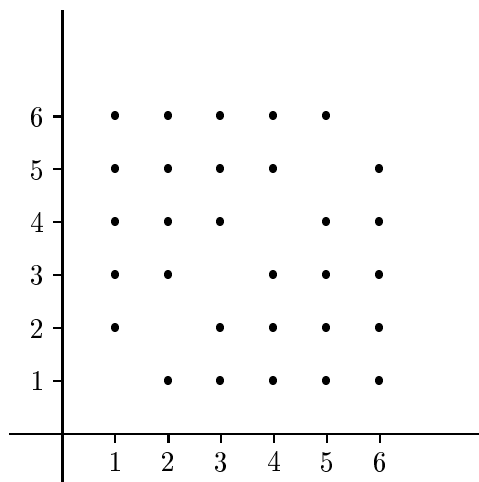


Figura 2 Valori ammissibili per (Y_1, Y_2) .

Usando le formule (1) e (2), è possibile calcolare le distribuzioni di Y_1 e di Y_2 come distribuzioni marginali della loro legge congiunta. La somma in (1) per $x_i = 1, 2, 3, 4, 5, 6$ contiene 5 termini uguali a $\frac{1}{30}$. Ad esempio, per $x_i = 3$,

$$p_{Y_1}(3) = p(3, 1) + p(3, 2) + p(3, 4) + p(3, 5) + p(3, 6) = \frac{1}{6}$$

dunque la legge di Y_1 è la legge uniforme su $\{1, \dots, 6\}$. Usando la (2), si verifica facilmente (esercizio!) che anche Y_2 ha distribuzione uniforme su $\{1, \dots, 6\}$.

Tale risultato non è sorprendente per Y_1 (la probabilità che alla prima estrazione si ottenga la pallina numero i è ovviamente $\frac{1}{6}$, per ogni $i = 1, 2, 3, 4, 5, 6$) ma lo è un po' meno per Y_2 ...

Le leggi marginali sono dunque identiche, sia che le estrazioni vengano effettuate *con reinserimento* sia che le palline vengano estratte *senza reinserimento*.

L'esempio precedente mostra che è possibile avere due distribuzioni congiunte diverse aventi le medesime distribuzioni marginali. Ciò significa che

non è possibile, in generale, stabilire quale sia il comportamento congiunto di due v.a. noto il comportamento di ognuna di esse prese singolarmente. Al contrario, la distribuzione congiunta consente *sempre* di determinare le due distribuzioni marginali, a partire dalle (1) e (2).

L'esempio che segue mostra un particolare utilizzo della distribuzione congiunta.

Esempio 2. Da un'urna contenente 6 palline numerate da 1 a 6 vengono effettuate due estrazioni senza reinserimento. Qual è la probabilità che i risultati delle due estrazioni differiscano al più di 2? E se le estrazioni fossero state effettuate con reinserimento?

Usando le notazioni dell'Esempio 1, si tratta di calcolare $\mathbb{P}\{|Y_1 - Y_2| \leq 2\}$. Tale probabilità si può riscrivere come $\mathbb{P}\{(Y_1, Y_2) \in A\}$, dove A è la regione del piano $\{(x_1, x_2); |x_1 - x_2| \leq 2\}$. Dunque, la probabilità richiesta è data da

$$\sum_{(i,j) \in A} p(i, j) \quad (3)$$

dove p denota la distribuzione congiunta di (Y_1, Y_2) , calcolata nell'Esempio 1.

La regione A è data dalla striscia compresa tra le rette di equazione $x = y + 2$ e $x = y - 2$, disegnate nella Figura 3.

In A si trovano 18 valori ammissibili per la coppia (Y_1, Y_2) e poiché ciascuna coppia ha probabilità $\frac{1}{30}$, la somma in (3) vale $\frac{18}{30} = \frac{3}{5}$.

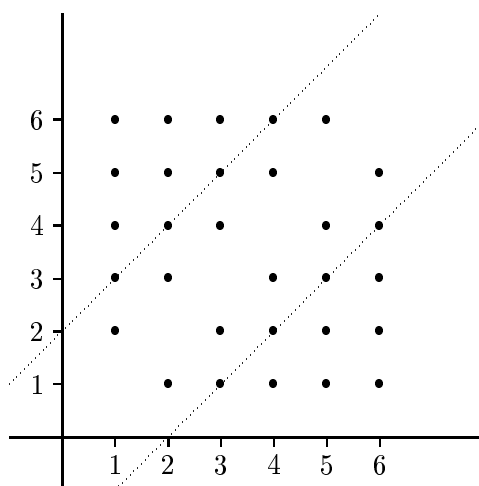


Figura 3 L'insieme A è dato dall'intersezione della striscia compresa tra le due rette disegnate in figura e l'insieme dei valori ammissibili.

Se le estrazioni fossero state effettuate con reinserimento, il procedimento sopra descritto sarebbe comunque valido, ma in tal caso A conterrebbe 24 (= 18 + 6) valori possibili, ciascuno verificabile con probabilità $\frac{1}{36}$. Ciò significa che, sotto tale ipotesi, la probabilità dell'evento in questione sarebbe $\frac{2}{3}$.

Un'altra situazione nella quale la nozione di distribuzione congiunta è cruciale è la seguente: se X e Y denotano due v.a., di legge congiunta p , qual è la legge di $X + Y$?

Si ha

$$\begin{aligned}\mathbb{P}\{X + Y = t\} &= \sum_j \mathbb{P}\{X + Y = t, Y = y_j\} \\ &= \sum_j \mathbb{P}\{X = t - y_j, Y = y_j\} = \sum_j p(t - y_j, y_j) = \sum_{(x_i, x_j) : x_i + y_j = t} p(x_i, y_j)\end{aligned}$$

Tale formula permette di calcolare esplicitamente la distribuzione della v.a. $X + Y$ ed ha una grande importanza teorica perché mostra che la legge di $X + Y$ non dipende dalle marginali ma solo dalla distribuzione congiunta: se X' e Y' sono due v.a. che hanno stessa legge (congiunta) p allora la distribuzione di $X' + Y'$ è la stessa di $X + Y$.

2 Covarianza e correlazione

Siano X e Y due v.a. Si vuole determinare una formula per la varianza di $X + Y$. Si ha

$$\begin{aligned}\text{Var}(X + Y) &= \mathbb{E}[(X + Y - \mathbb{E}[X + Y])^2] = \\ &= \mathbb{E}[(X - \mathbb{E}[X]) + (Y - \mathbb{E}[Y])]^2 = \\ &= \mathbb{E}[(X - \mathbb{E}[X])^2] + \mathbb{E}[(Y - \mathbb{E}[Y])^2] + 2\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \\ &= \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)\end{aligned}\tag{4}$$

dove

$$\text{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]\tag{5}$$

La quantità $\text{Cov}(X, Y)$ prende il nome di **covarianza di X e Y** . Un'altra espressione per la covarianza (spesso più utile nei calcoli) è la seguente:

$$\begin{aligned}\text{Cov}(X, Y) &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] = \\ &= \mathbb{E}[XY - X\mathbb{E}[Y] - Y\mathbb{E}[X] + \mathbb{E}[X]\mathbb{E}[Y]] = \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]\end{aligned}\tag{6}$$

dove, ricordiamo,

$$\mathbb{E}[XY] = \sum_{i,j} x_i x_j p(x_i, x_j).$$

Sappiamo che se X e Y sono indipendenti allora $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$, dunque

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] = 0\tag{7}$$

In tal caso, $\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y)$. Per il caso di m v.a. segue facilmente (esercizio!)

$$\text{Var}(X_1 + \cdots + X_m) = \sum_{i=1}^m \text{Var}(X_i) + \sum_{i \neq j} \text{Cov}(X_i, X_j) \quad (8)$$

e se $X_1, \dots, X_m$ sono indipendenti, allora

$$\text{Var}(X_1 + \cdots + X_m) = \text{Var}(X_1) + \cdots + \text{Var}(X_m) \quad (9)$$

Dunque, la varianza della somma di v.a. indipendenti è data dalla somma delle varianze.

Cerchiamo ora di capire il significato della covarianza di due v.a. Abbiamo appena visto che la covarianza di due v.a. indipendenti è zero. Sfortunatamente, l'inverso è falso (se così non fosse, infatti, si potrebbe usare tale risultato per una verifica, tutto sommato non difficile dal punto di vista computazionale, dell'indipendenza di due v.a.). Mostriamo qui di seguito un controesempio.

Esempio 3. Sia X una v.a. che può assumere i valori $\{-1, 0, +1\}$, con distribuzione

$$p_X(0) = \alpha, \quad p_X(-1) = p_X(+1) = \beta$$

con $\alpha, \beta > 0$, $\alpha + 2\beta = 1$. Si noti che $\mathbb{E}[X] = 0$. Sia ora Y definita nel modo seguente: $Y = 1$ se $X \neq 0$, $Y = 0$ se $X = 0$. Mostriamo che $\text{Cov}(X, Y) = 0$. Infatti, è immediato verificare che $XY = X$ (esercizio!), dunque $\mathbb{E}[XY] = \mathbb{E}[X] = 0$. Allora,

$$\text{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X] \mathbb{E}[Y] = 0.$$

Ma, X e Y *non* sono indipendenti:

$$0 = \mathbb{P}(X = 0, Y = 1) \neq \mathbb{P}(X = 0) \mathbb{P}(Y = 1) = \alpha \cdot 2\beta.$$

Se $\text{Cov}(X, Y) = 0$, diremo che X e Y sono **non correlate**. In ogni caso, ci si può servire della covarianza per dare una *misura dell'indipendenza* di due v.a.: valori della covarianza vicini a zero significano che le due v.a. sono *quasi indipendenti*; viceversa, valori grandi (in valore assoluto) della covarianza indicano la presenza di una forte dipendenza tra le due v.a. Il prossimo esempio mostra la relazione che lega la covarianza *positiva* e *negativa* con il concetto di “dipendenza” tra due v.a.

Esempio 4. (Retta di regressione) Siano X e Y due v.a. sullo stesso spazio di probabilità (Ω, P) . Si vuole trovare la migliore approssimazione possibile della Y tramite una funzione lineare della X . In altre parole, si cercano due numeri $a, b \in \mathbb{R}$ tali che $aX + b$ approssima Y nel miglior modo possibile.

Un problema di questo tipo appare, ad esempio, quando si vuole stimare una quantità Y che non è osservabile direttamente. Si consideri, a titolo di esempio, il caso in cui Y è un segnale e X è un'osservazione “disturbata” (vale a dire, X è uguale a Y più un errore dovuto all'osservazione).

Il primo problema consiste nel definire un criterio di ottimalità: che significato si vuole dare a “migliore approssimazione possibile”? Ovviamente, occorre partire dalla distanza tra $aX + b$ e Y , ovvero da $|aX + b - Y|$. Poichè però lavorare con i moduli può essere oneroso in termini di calcoli, si preferisce considerare $(aX + b - Y)^2$. Ora, tale quantità è aleatoria e il problema che è stato posto è “deterministico”: i numeri a e b che si cercano *non* sono aleatori. Per tale ragione, si considera il valore atteso di $(aX + b - Y)^2$. Dunque, il criterio di ottimalità scelto è il seguente: si vogliono determinare due numeri a e b tali da minimizzare

$$(a, b) \mapsto \mathbb{E}[(aX + b - Y)^2]$$

(va detto che esistono altri criteri di ottimizzazione, che qui non saranno presi in considerazione). Denotiamo, per semplicità, $\mu = \mathbb{E}(X)$, $\nu = \mathbb{E}(Y)$. Allora,

$$\begin{aligned} \mathbb{E}[(aX + b - Y)^2] &= \mathbb{E}[(a(X - \mu) + (b + a\mu) - (Y - \nu) - \nu)^2] \\ &= \mathbb{E}[(a(X - \mu) - (Y - \nu)) + (b + a\mu - \nu)]^2 \\ &= \mathbb{E}[(a(X - \mu) - (Y - \nu))^2] + (b + a\mu - \nu)^2 \\ &\quad + 2(b + a\mu - \nu)\mathbb{E}[a(X - \mu) - (Y - \nu)] \\ &= I_1 + I_2 + I_3 \end{aligned}$$

Osserviamo che il termine I_3 è nullo perché

$$\begin{aligned} \mathbb{E}[a(X - \mu) - (Y - \nu)] &= \mathbb{E}[a(X - \mu)] - \mathbb{E}[Y - \nu] \\ &= a(\mathbb{E}(X) - \mu) - (\mathbb{E}(Y) - \nu) = 0 \end{aligned}$$

Cominciamo con il cercare il valore di a che minimizza I_1 :

$$\begin{aligned} I_1 &= \mathbb{E}[(a(X - \mu) - (Y - \nu))^2] = \\ &= a^2\mathbb{E}[(X - \mu)^2] + \mathbb{E}[(Y - \nu)^2] - 2a\mathbb{E}[(X - \mu)(Y - \nu)] = \\ &= a^2\text{Var}(X) - 2a\text{Cov}(X, Y) + \text{Var}(Y). \end{aligned}$$

Il minimo di questo trinomio è raggiunto nel punto

$$a^* = \frac{\text{Cov}(X, Y)}{\text{Var}(X)} \tag{10}$$

Cerchiamo ora il valore di b che minimizza I_2 , calcolato sul valore di a appena trovato:

$$b^* = \nu - a^* \mu = \mathbb{E}(Y) - a^* \mathbb{E}(X)$$

Si definisce **retta di regressione** la retta di equazione

$$y = a^* x + b^*$$

dove a^* e b^* sono i numeri appena determinati. Quindi, l'equazione della retta di regressione è

$$y = \frac{\text{Cov}(X, Y)}{\text{Var}(X)}(x - \mathbb{E}[X]) + \mathbb{E}[Y]$$

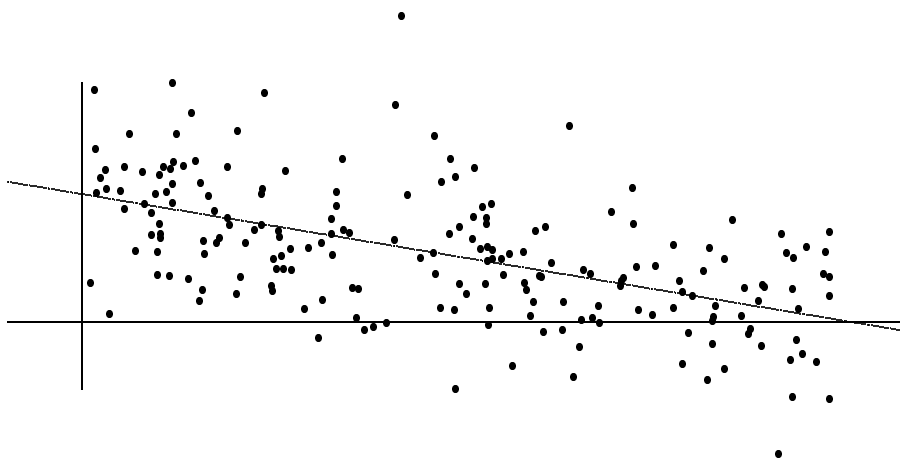


Figura 4 Si suppone, qui e nel grafico successivo, che i punti disegnati in figura rappresentino i valori ammissibili per la coppia (X, Y) , ognuno dei quali può essere assunto con la stessa probabilità. Qui la retta di regressione ha un coefficiente angolare negativo e la correlazione è negativa. Si noti che Y ha tendenza ad assumere valori più piccoli quando X è grande.

Si noti che la retta di regressione non è “simmetrica”, ovvero scambiando i ruoli ad X e Y cambia anche la retta di regressione. Inoltre, il coefficiente angolare della retta di regressione ha lo stesso segno della covarianza (perché $\text{Var}(X)$ è sempre positiva). Ciò indica che, se la covarianza è positiva, Y tende ad assumere valori grandi in corrispondenza di valori grandi della X . Tale proprietà è detta di *dipendenza positiva*. Al contrario, se la covarianza è negativa, Y tende ad assumere valori piccoli in corrispondenza di valori grandi della X e in tal caso si dice che c'è *dipendenza negativa* tra le due v.a. Ad esempio, se si volesse creare un modello che legghi la relazione tra il peso X e l'altezza Y di un individuo preso a caso da una popolazione, ci si aspetta una covarianza positiva tra queste due v.a. perché individui alti hanno la tendenza ad essere più pesanti (ciò non toglie che possano esistere individui alti e leggeri oppure bassi e molto pesanti, eventi che, in

generale, hanno probabilità di verificarsi molto bassa). Ci si attende una covarianza negativa quando si considerino quantità aleatorie che hanno un effetto “antitetico” l’una sull’altra.

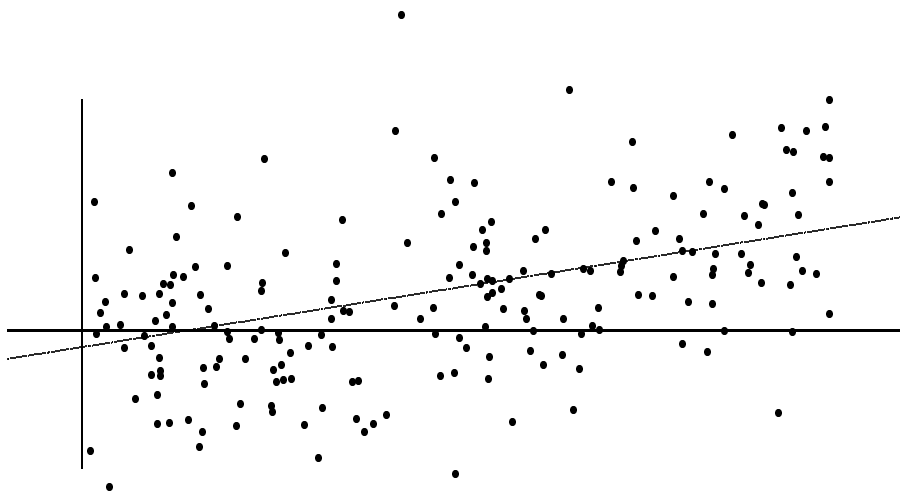


Figura 5 Qui in corrispondenza di valori grandi per X si osservano valori piuttosto grandi anche per Y . La covarianza è positiva.

Abbiamo appena visto come la covarianza possa essere utilizzata per misurare la dipendenza di due v.a. In realtà ciò non è del tutto vero perché la covarianza presenta un difetto:

$$\text{Cov}(aX, bY) = ab\text{Cov}(X, Y)$$

Ciò significa che la covarianza tra due v.a. è strettamente legata all’unità di misura: la covarianza cambia se si cambia l’unità di misura con la quale le v.a. X e Y sono state misurate. Per tale ragione, si introduce la seguente quantità:

$$\rho_{X,Y} = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)\text{Var}(Y)}} \quad (11)$$

$\rho_{X,Y}$ è detto **coefficiente di correlazione di X e Y** e non dipende dai cambiamenti di misurazione, cioè (esercizio!)

$$\rho_{aX,bY} = \rho_{X,Y}$$

per ogni $a > 0, b > 0$ (se $ab < 0$, il coefficiente di correlazione cambia segno: esercizio!).

La proposizione che segue assicura che

$$-1 \leq \rho_{X,Y} \leq 1$$

Proposizione 1. *Vale la seguente disuguaglianza, detta **disuguaglianza di Cauchy-Schwartz**:*

$$\mathbb{E}[XY]^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2]$$

Dimostrazione. poiché il valore atteso di una v.a. non negativa è sempre non negativo, per ogni $\vartheta \in \mathbb{R}$ si ha

$$0 \leq \mathbb{E}[(\vartheta X + Y)^2] = \vartheta^2 \mathbb{E}[X^2] + 2\vartheta \mathbb{E}[XY] + \mathbb{E}[Y^2]$$

Tale disuguaglianza è vera per ogni $\vartheta \in \mathbb{R}$: il trinomio $\vartheta \rightarrow \vartheta^2 \mathbb{E}[X^2] + 2\vartheta \mathbb{E}[XY] + \mathbb{E}[Y^2]$ non può assumere valori negativi e dunque il suo discriminante deve necessariamente essere minore di 0, il che significa

$$\mathbb{E}[XY]^2 - \mathbb{E}[X^2]\mathbb{E}[Y^2] \leq 0$$

□

La Proposizione 1 applicata alle v.a. $X - \mathbb{E}[X]$ et $Y - \mathbb{E}[Y]$ dà

$$\begin{aligned} \text{Cov}(X, Y)^2 &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])]^2 \leq \\ &\leq \mathbb{E}[(X - \mathbb{E}[X])^2] \mathbb{E}[(Y - \mathbb{E}[Y])^2] = \text{Var}(X)\text{Var}(Y) \end{aligned}$$

e quindi $\rho_{X,Y}^2 \leq 1$.

Come abbiamo visto, la condizione di non correlazione è in realtà molto più debole di quella di indipendenza, ma è anche più facile da verificare per cui è abbastanza usata nella pratica come forma debole di indipendenza.