1 Varianza e covarianza: definizioni e proprietà

Sia (X, Y) un vettore aleatorio discreto di densità $p_{X,Y}$ con densità marginali p_X e p_Y. Poniamo $Im(X) = \{x_1, x_2, ...\}$ e $Im(Y) = \{y_1, y_2, ...\}$.

Supponiamo che X ammetta momento centrato di ordine 2. Questo particolare momento si chiama *Varianza* della variabile aleatoria X e si indica con $Var(X)$. Quindi

$$Var(X) = E[(X - E(X))^2]. \tag{1}$$

Proposizione 1.1. *Siano X ed Y variabili aleatorie che ammettono varianza.*

1 *Per ogni $c \in \mathbb{R}$ la variabile aleatoria cX ha varianza pari a*

$$Var(cX) = c^2 Var(X).$$

2 *Per ogni $c \in \mathbb{R}$ la variabile aleatoria $X + c$ ha varianza pari a*

$$Var(X + c) = Var(X).$$

3 *la variabile aleatoria $X + Y$ ha ha varianza pari a*

$$Var(X + Y) = Var(X) + Var(Y) + 2Cov(X, Y),$$

dove

$$Cov(X, Y) = E[(X - E(X))(Y - E(Y))]. \tag{2}$$

Proof. **1** Sia $c \in \mathbb{R}$. Calcoliamo la varianza di cX:

$$\begin{aligned} Var(cX) =& E[(cX - E(cX))^2] = E[(c(X - E(X)))^2] = \\ =& c^2 E[(X - E(X))^2] = c^2 Var(X). \end{aligned}$$

2 Sia $c \in \mathbb{R}$. Calcoliamo la varianza di $X + c$:

$$\begin{aligned} Var(X + c) =& E[((X + c) - E(X + c))^2] = E[(X + c - E(X) - c)^2] = \\ =& E[(X - E(X))^2] = Var(X). \end{aligned}$$

3 Calcoliamo la varianza di $X + Y$:

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

□

Esempio 1.2. *Se consideriamo due variabili aleatorie X ed Y tali che $X \sim Uni\{-10^{-10}, 10^{-10}\}$, $Y \sim Uni\{-10^{+10}, 10^{+10}\}$. Chiaramente $E(X) = E(Y) = 0$, e quindi $Var(X) = E(X^2) = 2 \times 10^{-20}$ $Var(Y) = E(Y^2) = 2 \times 10^{20}$*

Notiamo che la varianza può essere anche espressa come segue

$$Var(X) = E[X^2] - E(X)^2. \tag{3}$$

Infatti

$$\begin{aligned} E[(X - E(X))^2] =& E[X^2 - 2XE(X) + E(X)^2] = E[X^2] - 2E[XE(X)] + E[E(X)^2] = \\ =& E[X^2] - 2E(X)^2 + E(X)^2 = E[X^2] - E(X)^2. \end{aligned}$$

Analogamente per la covarianza vale la rappresentazione

$$Cov(X, Y) = E[XY] - E(X)E(Y). \tag{4}$$

Infatti

$$\begin{aligned} E[(X - E(X))(Y - E(Y)] = E[XY - XE(Y) - YE(X) + E(X)E(Y)] = \\ =& E[XY] - E(X)E(Y) - E(Y)E(X) + E(X)E(Y) = \\ =& E[XY] - E(X)E(Y). \end{aligned}$$

Osserviamo che dalla (4) deriva che *due variabili aleatorie indipendenti hanno covarianza nulla* (variabili aleatorie con varianza nulla si dicono *scorrelate*).
Infatti, se due variabili aleatorie sono indipendenti, allora la media del prodotto è il prodotto delle medie e dunque

$$Cov(X, Y) = E[XY] - E(X)E(Y) = E(X)E(Y) - E(X)E(Y) = 0.$$

È interessante notare che non vale il viceversa, ovvero *variabili aleatorie scorrelate non sono necessariamente indipendenti.* A tale proposito diamo il seguente

Esempio 1.3. *Diamo un esempio di due variabili aleatorie scorrelate ma non indipendenti.*
Sia $Z \sim Unif\{0, \frac{\pi}{2}, \pi\}$. Consideriamo le variabili aleatorie

$$\begin{aligned} X =& \sin Z; \\ Y =& \cos Z. \end{aligned}$$

X ed Y sono scorrelate.
Infatti

$$
\begin{aligned}
E(X) &= \sin 0 \times \frac{1}{3} + \sin \frac{\pi}{2} \times \frac{1}{3} + \sin \pi \times \frac{1}{3} = \frac{1}{3}; \\
E(Y) &= \cos 0 \times \frac{1}{3} + \cos \frac{\pi}{2} \times \frac{1}{3} + \cos \pi \times \frac{1}{3} = 0; \\
E(XY) &= \sin 0 \cos 0 \times \frac{1}{3} + \sin \frac{\pi}{2} \cos \frac{\pi}{2} \times \frac{1}{3} + \sin \pi \cos \pi \times \frac{1}{3} = 0,
\end{aligned}
$$

e quindi

$$Cov(X,Y) = E(XY) - E(X)E(Y) = 0.$$

X ed Y non sono indipendenti.
Infatti $P(X = 1, Y = 1) = P(\sin Z = 1, \cos Z = 1) = P(Z = \frac{\pi}{2}, Z = 0) = 0$, *mentre* $P(X = 1) = P(Z = \frac{\pi}{2}) = \frac{1}{3}$ *e* $P(Y = 1) = P(Z = 0) = \frac{1}{3}$, *e dunque*

$$0 = p_{XY}(1,1) \neq p_X(1)p_Y(1) = \frac{1}{9}.$$

2 Calcolo della varianza delle Variabili Aleatorie notevoli

Per il calcolo utilizzeremo la rappresentazione (3).

- **Variabile Aleatoria di Bernoulli di parametro** p

  Sia $X \sim B(p)$. Ricordiamo che la sua densità $p(1) = p$, $p(0) = 1 - p$.

  $$E(X^2) = 1 \times p + 0 \times (1 - p) = p.$$

  Quindi

  $$Var(X) = E(X^2) - E(X)^2 = p - p^2 = p(1 - p).$$

- **Variabile Aleatoria Binomiale di parametri** n, p

  Sia $X \sim B(n, p)$. Ricordiamo che

  $$p_X(k) = \binom{n}{k} p^k (1 - p)^{n-k}, \quad k \in \{0, ..., n\}.$$

Calcoliamo il momento secondo

$$\begin{aligned}E(X^2) = &\sum_{k=0}^{n} k^2 \binom{n}{k} p^k (1-p)^{n-k} = \sum_{k=1}^{n} k^2 \frac{n!}{(n-k)!k!} p^k (1-p)^{n-k} = \\ =&np \sum_{k=1}^{n} k \binom{n-1}{k-1} p^{k-1} (1-p)^{n-k} = \\ =&np \sum_{j=0}^{n-1} (j+1) \binom{n-1}{j} p^j (1-p)^{n-1-j} = \\ =&np[(n-1)p+1] = n^2p^2 - np^2 + np.\end{aligned}$$

Quindi

$$Var(X) = E(X^2) - E(X)^2 = n^2p^2 - np^2 + np - n^2p^2 = np(1-p).$$

Utilizziamo ora iterativamente la proprietà 3 della varianza. Sappiamo che $X = X_1 + ... + X_n \sim B(n,p)$ con $X_1, ..., X_n$ indipendenti e tali che $X_k \sim B(p)$. Allora
poiché sono indipendenti, la varianza della somma è somma delle varianze
e quindi

$$Var(X) = Var(X_1) + ... + Var(X_n) = np(1-p).$$

- **Variabile Aleatoria Ipergeometrica di parametri** n, r, b

Sia X una variabile aleatoria ipergeometrica, definita da n prove, r elementi di tipo 1 (successo) b elementi di tipo 2 (insuccesso). Come per il calcolo del valor medio rappresentiamo X come somma di n variabili aleatorie (non indipendenti!!!) $X_1, ..., X_n$, essendo X_k la variabile aleatoria che vale 1 se si è ottenuto un successo all k-sima prova. Sappiamo che $X_k \sim B(\frac{r}{r+b})$ $k \in \{0, ..., n\}$ e che $X = X_1 + ... + X_n$. Quindi dalla proprietà 3 si ottiene:

$$Var(X) = Var(X_1 + ... + X_n) = \sum_i Var(X_i) + 2 \sum_{i<j} Cov(X_i, X_j).$$

Ora si può dimostrare che ciascuna coppia di variabili aleatorie (X_i, X_j) ha la stessa legge, e quindi tutti i tertmini della seconda sommatoria nella precedente uguaglianza sono uguali. Quindi

$$Var(X) = Var(X_1 + ... + X_n) = nVar(X_1) + n(n-1)Cov(X_1, X_2).$$

Ora sappiamo che $Var(X_1) = \frac{rb}{(r+b)^2}$. Per calcolare $Cov(X_1, X_2)$ usando la (4) è necessario calcolare $E(X_1X_2)$. Ora

$$X_1X_2 = \begin{cases} 1 & \text{se si hanno 2 successi sulle prime due estrazioni;} \\ 0 & \text{altrimenti .} \end{cases}$$

Quindi $p_{X_1X_2}(1) = \frac{\binom{r}{2}\binom{b}{0}}{\binom{r+b}{2}} = \frac{r(r-1)}{(r+b)(r+b-1)}$, e dunque

$$\begin{aligned} Cov(X_1, X_2) =& \frac{r(r-1)}{(r+b)(r+b-1)} - \frac{r^2}{(r+b)^2} = \\ =& \frac{r}{(r+b)}\left(\frac{r-1}{(r+b-1)} - \frac{r}{(r+b)}\right) = -\frac{rb}{(r+b)(r+b-1)}. \end{aligned}$$

Finalmente, sostituendo si ottiene

$$\begin{aligned} Var(X) =& Var(X_1 + ... + X_n) = n\frac{rb}{(r+b)^2} - n(n-1)\frac{rb}{(r+b)(r+b-1)} = \\ =& \frac{nrb}{(r+b)}\left(\frac{1}{(r+b)} - (n-1)\frac{1}{(r+b-1)}\right) = \frac{nrb}{(r+b)^2}\frac{r+b-n}{(r+b-1)}. \end{aligned}$$

- **Variabili aleatorie Geometrica modificata e Geometrica**

Sia X una variabile aleatoria geometrica modificata di parametro p. Ricordiamo che

$$p_X(k) = p(1-p)^{k-1},\ k \geq 1.$$

Per calcolare il momento di ordine 2 impostiamo una equazione come segue

$$\begin{aligned} E(X^2) =& \sum_{k\geq 1} k^2 p(1-p)^{k-1} = \sum_{k\geq 1}(k-1+1)^2 p(1-p)^{k-1} = \\ =& \sum_{k\geq 1}(k-1)^2 p(1-p)^{k-1} + 2\sum_{k\geq 1}(k-1)p(1-p)^{k-1} + \sum_{k\geq 1} p(1-p)^{k-1} = \\ =& (1-p)\sum_{h\geq 1} h^2 p(1-p)^{h-1} + 2(1-p)\sum_{h\geq 1} hp(1-p)^{h-1} + 1 = \\ =& (1-p)E(X^2) + 2(1-p)E(X) + 1, \end{aligned}$$

ovvero $E(X^2)$ è soluzione dell' equazione

$$E(X^2) = (1-p)E(X^2) + 2\frac{1-p}{p} + 1$$

cioé

$$E(X^2) = 2\frac{1-p}{p^2} + \frac{1}{p}$$

e quindi

$$Var(X) = E(X^2) - E(X)^2 = 2\frac{1-p}{p^2} + \frac{1}{p} - \frac{1}{p^2} = \frac{1-p}{p^2}.$$

Se consideriamo la variabile aleatoria $Y = X - 1$, allora Y è una variabile aleatoria geometrica di parametro p e, dalla proprietà 2 della varianza si ha $Var(Y) = Var(X) = \frac{1-p}{p^2}$.

- **Variabile aleatoria di Poisson**

  Sia $X \sim Po(\lambda)$.

$$\begin{aligned} E(X^2) =& \sum_{k\geq 0} k^2 \frac{\lambda^k}{k!} e^{-\lambda} = \sum_{k\geq 1} k^2 \frac{\lambda^k}{k!} e^{-\lambda} = \\ =& \sum_{k\geq 1} k \frac{\lambda^k}{(k-1)!} e^{-\lambda} = \lambda \sum_{h\geq 0} (h+1) \frac{\lambda^h}{h!} e^{-\lambda} = \\ =& \lambda \sum_{h\geq 0} h \frac{\lambda^h}{h!} e^{-\lambda} + \lambda \sum_{h\geq 0} \frac{\lambda^h}{h!} e^{-\lambda} = \lambda^2 + \lambda \end{aligned}$$

  e quindi

$$Var(X) = E(X^2) - E(X)^2 = \lambda^2 + \lambda - \lambda^2 = \lambda.$$

3 Covarianza e coefficiente di correlazione lineare

A differenza della varianza, che è sempre non negativa, la covarianza può assumere valori negativi. Più precisamente, se

$$Cov(X,Y) \begin{cases} > 0 & \text{allora } X \text{ e } Y \text{ si dicono positivamente correlate;} \\ = 0 & \text{allora } X \text{ e } Y \text{ si dicono scorrelate;} \\ < 0 & \text{allora } X \text{ e } Y \text{ si dicono negativamente correlate.} \end{cases}$$

Inoltre *la covarianza è una forma bilineare simmetrica*, ovvero tale che $Cov(X,Y) = Cov(Y,X)$ ed inoltre per ogni $a, b \in \mathbb{R}$,

$$Cov(aX + bY, Z) = aCov(X,Z) + bCov(Y,Z).$$

La covarianza è una sorta di misura di dipendenza tra variabili aleatorie. La dipendenza è tanto più grande quanto più è grande in valore assoluto, la covarianza. Questa indicazione di dipendenza però non è molto precisa, dipendendo dal range delle variabili aleatorie coinvolte. Un indicatore più affidabile è il coefficiente di correlazione lineare. Per definirlo introduciamo la

Definizione 3.1. *Sia X dotata di varianza. Si chiama **deviazione standard di** X (e si indica con σ_X) la quantità $\sigma_X = \sqrt{Var(X)}$*

Definizione 3.2. *Siano X ed Y dotate di varianza. Si chiama **coefficiente di correlazione lineare** (e si indica con ρ_{XY}) la quantità*

$$\rho_{XY} = \frac{Cov(X,Y)}{\sigma_X \sigma_Y}.$$

Come vedremo nella seguente proposizione, il coefficiente di correlazione lineare fornisce una indicazione molto più precisa della dipendenza tra variabili aleatorie.

Proposizione 3.3. *Siano X ed Y dotate di varianza. Allora*

1 $\rho_{XY} \in [-1, 1]$;

2

$$\rho_{XY} = \begin{cases} 0 & \text{se } X \text{ e } Y \text{ sono indipendenti;} \\ 1 & \text{se } Y = aX + b \text{ e } a > 0; \\ -1 & \text{se } Y = aX + b \text{ e } a < 0. \end{cases}$$

Proof. **1** Siano U e V due variabili aleatorie dotate di momento di ordine 2. Allora

$$E(|UV|)^2 \leq E(U^2)E(V^2). \tag{5}$$

Infatti $0 \leq E((\theta U + V)^2)$ per ogni $\theta \in \mathbb{R}$ e sviluppando il valore medio si ha $0 \leq \theta^2 E(U^2) + 2\theta E(UV) + E(V^2)$ per ogni $\theta \in \mathbb{R}$ ovvero $\Delta = E(UV)^2 - E(U^2)E(V^2) \leq 0$.

Poniamo

$$U = X - E(X);$$
$$V = Y - E(Y)$$

e sostituiamo nella (5):

$$E((X - E(X))(Y - E(Y)))^2 \leq E((X - E(X))^2)E((Y - E(Y))^2),$$

ovvero

$$|Cov(X, Y)| \leq \sqrt{Var(X)Var(Y)},$$

e quindi

$$|\rho_{XY}| \leq 1,$$

da cui l' asserto.

2 Se X ed Y sono indipendenti, allora $Cov(X, Y) = 0$, e dunque anche $\rho_{XY} = 0$.

Sia ora $Y = aX + b$. Allora

$$\begin{aligned} Cov(X, Y) =& Cov(X, aX + b) = aVar(X) + Cov(X, b) = aVar(X); \\ Var(Y) =& Var(aX + b) = a^2Var(X). \end{aligned}$$

Di conseguenza

$$\rho_{XY} = \frac{Cov(X, Y)}{\sigma_X\sigma_Y} = \frac{aVar(X)}{\sigma_X|a|\sigma_X} = \frac{a}{|a|} = \begin{cases} 1 & \text{se } a > 0; \\ -1 & \text{se } a < 0. \end{cases}$$

□

Esempio 3.4. *Sia data un' urna contenente b biglie bianche ed r biglie rosse. Si effettuano 2 estrazioni senza reimmissione. Sia*

$$X_i = \begin{cases} 1 & \textit{se esce una pallina rossa alla i-sima estrazione;} \\ 0 & \textit{se esce una pallina bianca alla i-sima estrazione .} \end{cases} \quad i = 1, 2.$$

Calcolare $\rho_{X_1X_2}$.

Dal paragrafo precedente sappiamo che

$$\begin{aligned} Cov(X_1, X_2) =& -\frac{rb}{(r + b)^2(r + b - 1)}, \\ Var(X_i) =& \frac{rb}{(r + b)^2}, \quad i = 1, 2. \end{aligned}$$

Quindi, applicando la definizione, si ottiene

$$\rho_{X_1X_2} = -\frac{\frac{rb}{(r+b)^2(r+b-1)}}{\frac{rb}{(r+b)^2}} = -\frac{1}{r + b - 1}.$$

<u>Esercizi x casa</u>: 1.34, 1.54 Abundo.