# Calcolo delle Probabilità e Statistica Matematica – A.A. 2009/10

## CAPITOLO 7 – Proprietà del valore atteso

### 7.1 Introduzione

### 7.2 Valore atteso di somme di variabili aleatorie

### 7.2 Covarianza, Varianza di una somma e correlazioni

# 7.1 Introduzione

Ricordiamo che il valore atteso di una variabile aleatoria discreta $X$ con densità $p(x)$ è

$$E[X] = \sum_x x\, p(x)$$

mentre se $X$ è (assolutamente) continua con densità $f(x)$ si ha

$$E[X] = \int_{-\infty}^{+\infty} x\, f(x)\, dx.$$

**Proposizione.** Se $X$ assume valori compresi tra $a$ e $b$, allora il valore atteso è compreso tra $a$ e $b$. In altre parole, se $P(a \le X \le b) = 1$ allora $a \le E[X] \le b$.

**Dimostrazione.** Poiché $p(x) = 0$ se $x$ non appartiene ad $[a, b]$, nel caso discreto si ha

$$E[X] = \sum_{x:p(x)>0} x\, p(x) \ge \sum_{x:p(x)>0} a\, p(x) = a \sum_{x:p(x)>0} p(x) = a.$$

Analogamente si ricava $E[X] \le b$. La dimostrazione nel caso continuo è analoga.

## 7.2 Valore atteso di somme di variabili aleatorie

Siano $X$ e $Y$ due variabili aleatorie e $g$ una funzione di due variabili.

**Proposizione.** Se $X$ e $Y$ hanno densità discreta congiunta $p(x, y)$, allora

$$E[g(X,Y)] = \sum_y \sum_x g(x,y)p(x,y).$$

Illustriamo un'importante applicazione di quanto su esposto. Se $E[X]$ e $E[Y]$ sono entrambe finite e poniamo $g(x, y) = x + y$, si ha

$$\begin{aligned} E[X+Y] &= \sum_x \sum_y (x+y)\, p(x,y) = \sum_x \sum_y x\, p(x,y) + \sum_x \sum_y y\, p(x,y) \\ &= \sum_x x\, p_X(x) + \sum_y y\, p_Y(y) = E[X] + E[Y]. \end{aligned}$$

Ragionando per induzione si dimostra che se $E[X_i]$ è finito per $i = 1, \dots, n$, allora

$$E[a_1 X_1 + \cdots + a_n X_n + b] = a_1 E[X_1] + \cdots + a_n E[X_n] + b.$$

**Esempio. La media campionaria.** Siano $X_1, \ldots, X_n$ variabili aleatorie indipendenti e identicamente distribuite con distribuzione $F$ e valore atteso $\mu$. Tale sequenza costituisce un *campione casuale* della distribuzione $F$. La variabile aleatoria

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

è detta media campionaria di $X_1, \ldots, X_n$. Calcolare $E[\overline{X}]$.

**Soluzione.** Dato che $E[X_i] = \mu$, per la proprietà di linearità si ha

$$E[\overline{X}] = E\left[\frac{1}{n} \sum_{i=1}^{n} X_i\right] = \frac{1}{n} E\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n} \sum_{i=1}^{n} E[X_i] = \mu.$$

Pertanto il valore atteso della media campionaria è $\mu$, la media della distribuzione. La media campionaria è spesso utilizzata per dare una stima del valore atteso $\mu$ della distribuzione, se questo è sconosciuto.

**Esempio.** Siano $X$ e $Y$ due variabili aleatorie tali che $X \geq Y$. In altre parole, per ogni esito dell'esperimento, il valore assunto da $X$ è maggiore o uguale del valore assunto da $Y$. Dato che $X - Y \geq 0$, si ha che $E[X - Y] \geq 0$, ossia $E[X] \geq E[Y]$.

**Esempio. La disuguaglianza di Boole.** Siano $A_1, \ldots, A_n$ degli eventi, e definiamo le loro variabili indicatrici

$$X_i = \begin{cases} 1 & \text{se } A_i \text{ si realizza,} \\ 0 & \text{altrimenti,} \end{cases} \qquad (i = 1, \ldots, n).$$

Poniamo $X = \displaystyle\sum_{i=1}^{n} X_i$, così $X$ è il numero di eventi $A_i$ che si realizzano. Infine sia

$$Y = \begin{cases} 1 & \text{se } X \geq 1, \\ 0 & \text{altrimenti,} \end{cases}$$

sicché $Y$ è uguale ad 1 se si realizza almeno uno degli eventi $A_i$ e vale 0 altrimenti.

Essendo chiaramente $X \geq Y$, dall'esempio precedente si deduce che $E[X] \geq E[Y]$.
Dato che

$$E[X] = \sum_{i=1}^{n} E[X_i] = \sum_{i=1}^{n} P(X_i = 1) = \sum_{i=1}^{n} P(A_i)$$

e che

$$E[Y] = P\{\text{si realizza almeno uno degli } A_i\} = P\left(\bigcup_{i=1}^{n} A_i\right),$$

si ottiene infine la seguente relazione, nota come *disuguaglianza di Boole*:

$$P\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} P(A_i).$$

**Esempio. Valore atteso di una variabile aleatoria binomiale.** Sia $X$ una variabile aleatoria binomiale di parametri $n$ e $p$. Ricordando che questa variabile rappresenta il numero di successi in $n$ prove indipendenti, dove ogni prova ha successo con probabilità $p$, si ha che

$$X = X_1 + X_2 + \cdots + X_n,$$

dove

$$X_i = \begin{cases} 1 & \text{se l'}i\text{-esima prova è un successo,} \\ 0 & \text{se l'}i\text{-esima prova è un insuccesso.} \end{cases}$$

La variabile aleatoria di Bernoulli $X_i$ ha valore atteso $E[X_i] = 0 \cdot (1-p) + 1 \cdot p = p$. Di conseguenza risulta

$$E[X] = E[X_1] + E[X_2] + \cdots + E[X_n] = n\,p.$$

**Esempio. Valore atteso di una variabile aleatoria ipergeometrica.** Si scelgano a caso $n$ biglie da un'urna contenente $N$ biglie delle quali $m$ sono bianche. Determinare il numero atteso di biglie bianche selezionate.

**Soluzione.** Il numero $X$ di biglie bianche selezionate si può rappresentare come

$$X = X_1 + X_2 + \cdots + X_m,$$

dove

$$X_i = \begin{cases} 1 & \text{se è stata scelta l'}i\text{-esima biglia bianca,} \\ 0 & \text{altrimenti.} \end{cases}$$

Essendo

$$E[X_i] = P(X_i = 1) = P\{\text{è stata scelta l'}i\text{-esima biglia bianca }\} = \frac{\binom{1}{1}\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N},$$

si ottiene infine

$$E[X] = E[X_1] + \cdots + E[X_m] = \frac{m\,n}{N}.$$

Tale risultato può essere ottenuto anche utilizzando una rappresentazione alternativa:

$$X = Y_1 + \cdots + Y_n,$$

dove le variabili di Bernoulli

$$Y_i = \begin{cases} 1 & \text{se l'}i\text{-esima biglia selezionata è bianca,} \\ 0 & \text{altrimenti,} \end{cases} \qquad (i = 1, 2, \ldots, n).$$

non sono indipendenti, ma sono identicamente distribuite. Notiamo infatti che

$$P(Y_i = 1 \,|\, X = k) = \frac{\binom{n-1}{k-1}}{\binom{n}{k}} = \frac{\dfrac{(n-1)!}{(k-1)!\,(n-k)!}}{\dfrac{n!}{k!\,(n-k)!}} = \frac{k}{n},$$

dove a denominatore c'è il numero $\binom{n}{k}$ di sequenze booleane di lunghezza $n$ con $k$ elementi pari a 1 e con $n-k$ elementi pari a 0, mentre a numeratore è presente il numero $\binom{n-1}{k-1}$ di sequenze booleane di lunghezza $n$ aventi 1 nel posto $i$-esimo e $k-1$ elementi pari a 1 nei rimanenti $n-1$ posti.

Notiamo che la probabilità

$$P(Y_i = 1 \,|\, X = k) = \frac{k}{n}$$

si può ottenere anche in modo diretto, tenendo presente che se si devono collocare $k$ cifre pari a 1 e $n-k$ cifre pari a 0 in una sequenza di lunghezza $n$, ci sono $k$ casi favorevoli su un totale di $n$ affinché nel posto $i$-esimo della sequenza sia collocata una cifra pari a 1. Ne segue che la distribuzione di $Y_i$ non dipende da $i$, essendo

$$P(Y_i = 1) = \sum_{k=0}^{n} P(Y_i = 1 \,|\, X = k)\, P(X = k) = \sum_{k=0}^{n} \frac{k}{n}\, P(X = k) = \frac{1}{n}\, E(X).$$

Si ricava pertanto

$$E(X) = n\, P(Y_i = 1) = n\, P(Y_1 = 1) = n\, \frac{m}{N}.$$

**Esempio. Valore atteso del numero di accoppiamenti.** $N$ persone lanciano il loro cappello nel centro di una stanza. I cappelli vengono mescolati, e ogni persona ne prende uno a caso. Determinare il valore atteso del numero di persone che selezionano il proprio cappello.

**Soluzione.** Indicando con $X$ il numero di accoppiamenti, possiamo calcolare $E[X]$ scrivendo $X = X_1 + X_2 + \cdots + X_N$, dove

$$X_i = \begin{cases} 1 & \text{se la persona } i \text{ prende il suo cappello,} \\ 0 & \text{altrimenti.} \end{cases}$$

Dato che, per ogni $i$, l'$i$-esima persona può scegliere ugualmente uno degli $N$ cappelli, si ha

$$E[X_i] = P(X_i = 1) = \frac{1}{N},$$

da cui

$$E[X] = E[X_1] + \cdots + E[X_N] = \frac{1}{N} \cdot N = 1.$$

Quindi, in media, esattamente una persona seleziona il proprio cappello.

**Esempio.** Al passaggio di uno stormo di $n$ anatre, vi sono $n$ cacciatori che sparano all'istante, e ognuno mira ad un bersaglio a caso, indipendentemente dagli altri. Se ogni cacciatore colpisce il suo bersaglio con probabilità $p$, calcolare il numero atteso di anatre che non sono colpite.

**Soluzione.** Sia $X_i = 1$ se l'$i$-esima anatra non è colpita, e 0 altrimenti, per $i = 1, 2, \ldots, n$. Il numero atteso richiesto è dato da

$$E[X_1 + \cdots + X_n] = E[X_1] + \cdots + E[X_n].$$

Per calcolare $E[X_i] = P\{X_i = 1\}$, osserviamo che ognuno dei cacciatori colpirà, indipendentemente, l'$i$-esima anatra con probabilità $p/n$, sicché

$$P\{X_i = 1\} = \left(1 - \frac{p}{n}\right)^n \qquad \text{e quindi} \qquad E[X] = n\left(1 - \frac{p}{n}\right)^n.$$

**Nota.** La frazione attesa di anatre che non sono colpite è

$$E\left[\frac{X}{n}\right] = \frac{E[X]}{n} = \left(1 - \frac{p}{n}\right)^n \to e^{-p} \quad \text{per} \quad n \to \infty.$$

**Esempio. Algoritmo quick-sort.** Supponiamo di disporre di un insieme di $n$ valori distinti $x_1, x_2, \ldots, x_n$ e di volerli ordinare in modo crescente (sort). Una procedura efficace al riguardo è l'algoritmo di quick-sort, definito come segue. Se $n = 2$ l'algoritmo confronta due valori e li mette in ordine appropriato. Se $n > 2$, uno degli elementi è scelto a caso, ad esempio $x_i$, e si confrontano gli altri valori con $x_i$. I valori inferiori a $x_i$ vengono messi tra parentesi graffe alla sinistra di $x_i$, mentre quelli superiori a $x_i$ vengono messi alla destra di $x_i$. L'algoritmo si ripete per i valori interni alle parentesi graffe e continua finché i valori non sono tutti ordinati.
Supponiamo ad esempio di voler ordinare i seguenti 10 numeri distinti

$$5, 9, 3, 10, 11, 14, 8, 4, 17, 6.$$

Iniziamo scegliendo un numero a caso, ad esempio 10 (ogni numero può essere scelto con probabilità $1/10$). Confrontando gli altri valori con tale numero si ottiene

$$\{5, 9, 3, 8, 4, 6\}, 10, \{11, 14, 17\}$$

Rivolgiamo ora l'attenzione al sottoinsieme $\{5, 9, 3, 8, 4, 6\}$ e scegliamo a caso uno dei suoi valori, ad esempio il 6. Confrontando ognuno dei valori tra parentesi con 6 e mettendo a sinistra di esso i valori più piccoli di 6 e a destra quelli più grandi di 6 si ottiene

$$\{5, 3, 4\}, 6, \{9, 8\}, 10, \{11, 14, 17\}.$$

Considerando ora la parentesi a sinistra e scegliendo 4 per i successivi confronti, si giunge a

$$\{3\}, 4, \{5\}, 6, \{9, 8\}, 10, \{11, 14, 17\}.$$

Si continua finché non ci sono più sottoinsiemi che contengono più di un elemento.

Sia $X$ il numero di confronti che servono all'algoritmo di quick-sort per ordinare $n$ numeri distinti. Si ha allora che $E[X]$ è una misura dell'efficienza dell'algoritmo. Per calcolare $E[X]$, esprimiamo $X$ come somma di altre variabili aleatorie

$$X = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} I(i, j),$$

dove per $1 \leq i < j \leq n$, $I(i,j)$ è uguale ad 1 se $i$ e $j$ sono prima o poi confrontati direttamente mentre è uguale a 0 altrimenti. Quindi si ha

$$
\begin{aligned}
E[X] = E[\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} I(i,j)] &= \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} E[I(i,j)] \\
&= \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} P\{i \, \mathrm{e} \, j \, \text{sono confrontati tra loro}\}.
\end{aligned}
$$

Per determinare la probabilità che $i$ e $j$ non siano mai confrontati, osserviamo che i valori $\{i, i+1, \ldots, j-1, j\}$ sono inizialmente nella stessa parentesi e vi rimarranno se il numero scelto per il primo confronto non è compreso tra $i$ e $j$. Ad esempio se il numero da confrontare con gli altri è strettamente maggiore di $j$, allora tutti i valori $\{i, i+1, \ldots, j-1, j\}$ andranno in una parentesi a sinistra del numero scelto, mentre se esso è strettamente inferiore a $i$, allora tali valori andranno messi in una parentesi a destra.

Tutti i valori $\{i, i+1, \ldots, j-1, j\}$ rimarranno quindi all'interno della stessa parentesi finché uno di essi è scelto per effettuare i confronti. Il valore scelto viene poi confrontato con gli altri compresi tra $i$ e $j$. Se non è né $i$, né $j$, allora, dopo il confronto, $i$ andrà in una parentesi alla sua sinistra e $j$ in una parentesi alla sua destra. D'altro lato, se il valore scelto nell'insieme $\{i, i+1, \ldots, j-1, j\}$ è uguale a $i$ o $j$, ci sarà un confronto diretto tra $i$ e $j$.

Supponendo allora che il valore scelto sia uno dei $j-i+1$ valori tra $i$ e $j$, la probabilità che si tratti di $i$ o di $j$ è pari a $2/(j-i+1)$. Si può quindi concludere che

$$P\{i \,\mathrm{e}\, j \,\text{sono confrontati tra loro}\} = \frac{2}{j-i+1},$$

da cui si ricava

$$E[X] = \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{2}{j-i+1}.$$

Per ottenere l'ordine di grandezza di $E[X]$ per $n$ grande, approssimiamo le somme con degli integrali. Ora

$$\begin{aligned}\sum_{j=i+1}^{n} \frac{2}{j-i+1} &\approx \int_{i+1}^{n} \frac{2}{x-i+1}\mathrm{d}x = 2\log(x-i+1)\big|_{i+1}^{n} \\ &= 2\log(n-i+1) - 2\log(2) \approx 2\log(n-i+1).\end{aligned}$$

Quindi

$$\begin{aligned}E[x] &\approx \sum_{i=1}^{n-1} 2\log(n-i+1) \approx 2\int_{1}^{n-1} \log(n-x+1)\mathrm{d}x = 2\int_{2}^{n} \log(y)\mathrm{d}y \\ &= 2(y\log(y) - y)\big|_{2}^{n} \approx 2n\log(n).\end{aligned}$$

Pertanto, per $n$ grande, l'algoritmo di quick-sort richiede, in media, approssimativamente $2n\, log(n)$ confronti per ordinare $n$ valori distinti.

## 7.3 Covarianza, Varianza di una somma e correlazioni

**Proposizione.** Se $X$ e $Y$ sono indipendenti, ed $h$ e $g$ sono due funzioni, allora

$$E[g(X)h(Y)] = E[g(X)]E[h(Y)]$$

**Dimostrazione.** Supponiamo che $X$ ed $Y$ siano variabili discrete con distribuzione congiunta $p(x, y)$. Allora

$$\begin{aligned} E[g(X)h(Y)] &= \sum_x \sum_y g(x)h(y)p(x,y) = \sum_x \sum_y g(x)h(y)p_X(x)p_Y(y) \\ &= \sum_x g(x)p_X(x) \sum_y h(y)p_Y(y) = E[g(X)]E[h(Y)]. \end{aligned}$$

Come il valore atteso e la varianza di una singola variabile forniscono delle informazioni sulla variabile aleatoria, così la covarianza tra due variabili aleatorie fornisce un'informazione sulla relazione tra le due variabili.

**Definizione.** La covarianza tra $X$ ed $Y$, indicata con $Cov(X,Y)$, è definita da

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

Sviluppando il membro a destra della precedente definizione, vediamo che

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

Se $Cov(X,Y) > 0$ le variabili aleatorie $X$ ed $Y$ si dicono *positivamente correlate*; in tal caso entrambe tendono ad assumere valore maggiore, o minore, della propria media. Al contrario, se $Cov(X,Y) < 0$ le variabili aleatorie $X$ ed $Y$ si dicono *negativamente correlate*; in tal caso se una tende ad assumere valore maggiore della propria media, l'altra tende ad assumere valore minore della propria media. Se $Cov(X,Y) = 0$ le variabili aleatorie $X$ ed $Y$ si dicono *scorrelate*, ovvero *non correlate*.

Se $X$ e $Y$ sono variabili indipendenti allora si ha che $Cov(X,Y)=0$, invero usando $g(x)=x$ e $h(y)=y$ nella Proposizione precedente, si ha $E[XY]=E[X]\,E[Y]$ e quindi $Cov(X,Y)=0$.

Il viceversa è falso. Un semplice esempio di due variabili dipendenti $X$ e $Y$ la cui covarianza è zero si ottiene supponendo che $X$ e $Y$ siano tali che

$$P(X=0)=P(X=1)=P(X=-1)=\frac{1}{3}, \qquad Y=\begin{cases} 0 & \text{se } X\neq 0, \\ 1 & \text{se } X=0. \end{cases}$$

Essendo $XY=0$ si ha che $E[XY]=0$. Inoltre, $E[X]=0$ e quindi

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

Tuttavia chiaramente $X$ e $Y$ non sono indipendenti, come si può ricavare dalla tabella seguente:

| $x\backslash y$ | $0$ | $1$ | $p_X(x)$ |
|---|---|---|---|
| $-1$ | $1/3$ | $0$ | $1/3$ |
| $0$ | $0$ | $1/3$ | $1/3$ |
| $1$ | $1/3$ | $0$ | $1/3$ |
| $p_Y(y)$ | $2/3$ | $1/3$ | $1$ |

(ad esempio, $p(0,0)=0\neq p_X(0)p_Y(0)=\frac{1}{3}\cdot\frac{1}{3}$).

**Esempio.** Nel lancio di una moneta ripetuto 3 volte, sia $X$ il numero di volte che esce testa e sia $Y$ il numero di variazioni, ossia quanti lanci danno risultati diversi dal lancio precedente. Calcolare $Cov(X,Y)$.

**Soluzione.** Abbiamo già ricavato la densità congiunta di $(X,Y)$, da cui si trae che $X$ e $Y$ non sono indipendenti; inoltre si ricava:

| $x\backslash y$ | 0 | 1 | 2 | $p_X(x)$ |
|---|---|---|---|---|
| 0 | 1/8 | 0 | 0 | 1/8 |
| 1 | 0 | 1/4 | 1/8 | 3/8 |
| 2 | 0 | 1/4 | 1/8 | 3/8 |
| 3 | 1/8 | 0 | 0 | 1/8 |
| $p_Y(y)$ | 1/4 | 1/2 | 1/4 | 1 |

$$E(X) = 0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = \frac{3}{2}$$

$$E(Y) = 0 \cdot \frac{1}{4} + 1 \cdot \frac{1}{2} + 2 \cdot \frac{1}{4} = 1$$

$$E(XY) = 1 \cdot 1 \cdot \frac{1}{4} + 1 \cdot 2\,\frac{1}{8} + 2 \cdot 1 \cdot \frac{1}{4} + 2 \cdot 2 \cdot \frac{1}{8} = \frac{3}{2}$$

e pertanto: $Cov(X,Y) = E(XY) - E(X)\,E(Y) = \dfrac{3}{2} - \dfrac{3}{2} \cdot 1 = 0.$

**Proposizione.** La covarianza possiede le seguenti proprietà:

(i) $Cov(X,Y) = Cov(Y,X)$,

(ii) $Cov(X,X) = \text{Var}(X)$,

(iii) $Cov(aX,Y) = a\,Cov(X,Y)$,

(iv) $Cov(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j) = \sum_{i=1}^{n}\sum_{j=1}^{m} Cov(X_i, Y_j)$.

**Dimostrazione.** La (i) e la (ii) seguono immediatamente dalla definizione di covarianza. Per la proprietà di linearità del valore medio si ha la (iii):

$$Cov(aX,Y) = E[(aX - E(aX))(Y - E(Y))] = a\,Cov(X,Y).$$

Per dimostrare la (iv), ossia la proprietà di additività della covarianza, poniamo $\mu_i = E[X_i]$ e $v_j = E[Y_j]$. Allora

$$E\left[\sum_{i=1}^{n} X_i\right] = \sum_{i=1}^{n} \mu_i, \qquad E\left[\sum_{j=1}^{m} Y_j\right] = \sum_{j=1}^{m} v_j.$$

Inoltre

$$
\begin{aligned}
Cov\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j\right) &= E\left[\left(\sum_{i=1}^{n} X_i - \sum_{i=1}^{n} \mu_i\right)\left(\sum_{j=1}^{m} Y_j - \sum_{j=1}^{m} v_j\right)\right] \\
&= E\left[\sum_{i=1}^{n}(X_i - \mu_i)\sum_{j=1}^{m}(Y_j - v_j)\right] \\
&= E\left[\sum_{i=1}^{n}\sum_{j=1}^{m}(X_i - \mu_i)(Y_j - v_j)\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m} E[(X_i - \mu_i)(Y_j - v_j)],
\end{aligned}
$$

dove l'ultima uguaglianza segue in quanto il valore atteso di una somma di variabili aleatorie è uguale alla somma dei valori attesi.

**Proposizione.**

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \mathrm{Var}(X_i) + 2\sum_{i<j} Cov(X_i, X_j).$$

**Dimostrazione.** Dalla precedente proposizione, posto $Y_j = X_j$, $j = 1, \ldots, n$, si ha

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = Cov\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{n} Cov(X_i, X_j)$$

$$= \sum_{i=1}^{n} \mathrm{Var}(X_i) + \sum_{i\neq j} Cov(X_i, X_j).$$

Dato che ogni coppia di indici $(i, j)$, $i \neq j$, appare due volte nella doppia sommatoria, la proposizione segue immediatamente. (In altri termini, la *matrice di covarianza* $\{Cov(X_i, X_j)\}_{i,j}$ è simmetrica, con le varianze sulla diagonale principale).

Notiamo che dalla proposizione precedente si ha, per $a_1, \ldots, a_n$ costanti arbitrarie,

$$\mathrm{Var}\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 \mathrm{Var}(X_i) + 2\sum_{i<j} a_i a_j Cov(X_i, X_j).$$

Da ciò si trae che

$$\mathrm{Var}\,(X - Y) = \mathrm{Var}\,(X) + \mathrm{Var}\,(Y) - 2Cov(X, Y).$$

Se $X_1, \ldots, X_n$ sono a due a due scorrelate, cioè se $Cov(X_i, X_j) = 0$ per ogni $i \neq j$, risulta:

$$Var\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} Var(X_i).$$

**Esempio.** Siano $X_1, \ldots, X_n$ variabili indipendenti ed identicamente distribuite con valore atteso $\mu$ e varianza $\sigma^2$, e sia $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ la media campionaria. Le variabili $X_i - \overline{X}$ sono chiamate deviazioni, mentre $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$ è chiamata varianza campionaria. Determinare (a) $Var[\overline{X}]$ e (b) $E[S^2]$.

**Soluzione.** Per l'indipendenza delle variabili $X_1, \ldots, X_n$ si ha

$$\text{(a)} \qquad Var(\overline{X}) = \left(\frac{1}{n}\right)^2 Var\left(\sum_{i=1}^{n} X_i\right) = \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} Var(X_i) = \frac{\sigma^2}{n}.$$

(b) Consideriamo la seguente identità algebrica

$$
\begin{aligned}
(n-1)S^2 &= \sum_{i=1}^{n}(X_i - \mu + \mu - \overline{X})^2 \\
&= \sum_{i=1}^{n}(X_i-\mu)^2 + \sum_{i=1}^{n}(\overline{X}-\mu)^2 - 2(\overline{X}-\mu)\sum_{i=1}^{n}(X_i-\mu) \\
&= \sum_{i=1}^{n}(X_i-\mu)^2 + n(\overline{X}-\mu)^2 - 2(\overline{X}-\mu)n(\overline{X}-\mu) \\
&= \sum_{i=1}^{n}(X_i-\mu)^2 - n(\overline{X}-\mu)^2
\end{aligned}
$$

Prendendo i valori attesi di entrambi i membri dell'uguaglianza precedente si ottiene

$$
(n-1)E[S^2] = \sum_{i=1}^{n} E[(X_i-\mu)^2] - nE[(\overline{X}-\mu)^2] = n\sigma^2 - nVar(\overline{X}) = (n-1)\sigma^2,
$$

dove si è utilizzato il fatto che $E[\overline{X}] = \mu$ ed il punto (a) nell'uguaglianza finale. Dividendo per $n-1$ si ottiene che $E[S^2] = \sigma^2$.

**Esempio.** Calcolare la varianza di una variabile aleatoria binomiale $X$ di parametri $n$ e $p$.

**Soluzione.** Dato che una tale variabile rappresenta il numero di successi in $n$ prove indipendenti quando il successo in una prova ha probabilità $p$, possiamo scrivere

$$X = X_1 + \cdots + X_n,$$

dove le $X_i$ sono variabili di Bernoulli indipendenti tali che

$$X_i = \begin{cases} 1 & \text{se l'}i\text{-esima prova è un successo,} \\ 0 & \text{altrimenti.} \end{cases}$$

Si ottiene pertanto

$$\mathrm{Var}(X) = \mathrm{Var}(X_1) + \cdots + \mathrm{Var}(X_n).$$

Essendo

$$\mathrm{Var}(X_i) = E[X_i^2] - (E[X_i])^2 = p - p^2 = p\,(1-p),$$

si ha infine

$$\mathrm{Var}(X) = n\,p\,(1-p).$$

**Esempio. Varianza del numero di accoppiamenti.** Calcolare la varianza di $X$, il numero di persone che seleziona il proprio cappello tra $N$.

**Soluzione** Utilizzando la rappresentazione di $X$ usata in un esempio precedente, si ha che $X = X_1 + X_2 + \cdots + X_N$, dove

$$X_i = \begin{cases} 1 & \text{se la persona } i \text{ prende il suo cappello,} \\ 0 & \text{altrimenti.} \end{cases}$$

Si ha poi

$$\mathrm{Var}(X) = \sum_{i=1}^{N} \mathrm{Var}(X_i) + 2 \sum_{i<j} Cov(X_i, X_j).$$

Essendo $P(X_i = 1) = 1/N$, si ha

$$E[X_i] = \frac{1}{N}, \qquad \mathrm{Var}(X_i) = \frac{1}{N}\left(1 - \frac{1}{N}\right) = \frac{N-1}{N^2}.$$

Inoltre

$$X_i X_j = \begin{cases} 1 & \text{se l'}i\text{-esima e la }j\text{-esima persona selezionano il loro cappello,} \\ 0 & \text{altrimenti.} \end{cases}$$

Pertanto si ha

$$E[X_i X_j] = \sum_{u=0}^{1} \sum_{v=0}^{1} u\, v\, p(u,v) = p(1,1) = P(X_i = 1)\, P(X_j = 1 | X_i = 1) = \frac{1}{N}\, \frac{1}{N-1}$$

da cui segue

$$Cov(X_i, X_j) = \frac{1}{N(N-1)} - \left(\frac{1}{N}\right)^2 = \frac{N-(N-1)}{N^2(N-1)} = \frac{1}{N^2(N-1)}$$

e quindi

$$\mathrm{Var}(X) = \frac{N-1}{N} + 2\binom{N}{2}\frac{1}{N^2(N-1)} = \frac{N-1}{N} + \frac{1}{N} = 1$$

In conclusione sia media che varianza del numero di accoppiamenti sono pari a 1.

La correlazione tra due variabili aleatorie $X$ e $Y$, indicata con $\rho(X,Y)$, è definita, se $Var(X)$ e $Var(Y)$ non sono nulle, dal coefficiente di correlazione

$$\rho(X,Y) = \frac{Cov(X,Y)}{\sqrt{\mathrm{Var}(X)\,\mathrm{Var}(Y)}}.$$

**Proposizione.** Risulta

$$-1 \leq \rho(X,Y) \leq 1.$$

**Dimostrazione.** Denotando con $\sigma_X^2$ e $\sigma_Y^2$ le varianze di $X$ e $Y$, si ha

$$0 \leq \mathrm{Var}\left(\frac{X}{\sigma_X} + \frac{Y}{\sigma_Y}\right) = \frac{\mathrm{Var}(X)}{\sigma_X^2} + \frac{\mathrm{Var}(Y)}{\sigma_Y^2} + \frac{2Cov(X,Y)}{\sigma_X\,\sigma_Y} = 2\,[1+\rho(X,Y)]$$

da cui segue $\rho(X,Y) \geq -1$. D'altronde risulta

$$0 \leq \mathrm{Var}\left(\frac{X}{\sigma_X} - \frac{Y}{\sigma_Y}\right) = \frac{\mathrm{Var}(X)}{\sigma_X^2} + \frac{\mathrm{Var}(Y)}{\sigma_Y^2} - \frac{2Cov(X,Y)}{\sigma_X\,\sigma_Y} = 2\,[1-\rho(X,Y)]$$

e quindi si ha anche $\rho(X,Y) \leq 1$.

Ricordiamo che se $\mathrm{Var}(Z) = 0$ allora $Z$ è costante con probabilità 1. Quindi, dalla precedente dimostrazione segue che se $\rho(X,Y) = 1$ allora $Y = aX + b$, con $a = \sigma_Y/\sigma_X > 0$. Similmente, se $\rho(X,Y) = -1$ allora $Y = aX + b$, con $a = -\sigma_Y/\sigma_X < 0$. Vale anche il viceversa, cioè se $Y = aX + b$ allora $\rho(X,Y) = \pm 1$, a seconda del segno di $a$. Infatti, poiché $Cov(X, aX + b) = a\,\mathrm{Var}(X)$, si ha

$$\rho(X,Y) = \rho(X, aX+b) = \frac{a\,\mathrm{Var}(X)}{\sqrt{\mathrm{Var}(X)\,a^2\,\mathrm{Var}(X)}} = \frac{a}{|a|} = \begin{cases} 1, & a > 0 \\ -1, & a < 0. \end{cases}$$

ll coefficiente di correlazione è una misura del grado di linearità tra $X$ e $Y$. Un valore di $\rho(X,Y)$ vicino a $+1$ o a $-1$ indica un alto livello di linearità tra $X$ e $Y$, mentre un valore vicino a 0 indica un'assenza di tale linearità. Un valore positivo di $\rho(X,Y)$ indica che $Y$ tende a crescere quando $X$ cresce, mentre un valore negativo indica che $Y$ tende a decrescere quando $X$ cresce. Se $\rho(X,Y) = 0$ allora $X$ e $Y$ sono scorrelate.

Dalla definizione segue che il coefficiente di correlazione $\rho(X,Y)$ è adimensionale.

**Esercizio** Nel lancio di un dado ripetuto $n$ volte, sia $X$ il numero di volte che esce 6 e $Y$ il numero di volte che esce 5. Calcolare il coefficiente di correlazione di $(X, Y)$.

**Soluzione.** Per $i, j = 1, 2, \ldots, n$ poniamo

$$X_i = \begin{cases} 1 & \text{se esce 6 al lancio } i\text{-esimo} \\ 0 & \text{altrimenti,} \end{cases} \qquad Y_j = \begin{cases} 1 & \text{se esce 5 al lancio } j\text{-esimo} \\ 0 & \text{altrimenti,} \end{cases}$$

da cui si segue che $X = \sum_{i=1}^{n} X_i$ e $Y = \sum_{j=1}^{n} Y_j$. Risulta

$$E(X_i) = E(Y_j) = \frac{1}{6}, \qquad E(X_i Y_j) = \begin{cases} 0 & \text{se } i = j, \\ \frac{1}{36} & \text{altrimenti,} \end{cases} \qquad \mathrm{Var}(X_i) = \mathrm{Var}(Y_j) = \frac{5}{36},$$

e pertanto

$$\mathrm{Cov}(X_i, Y_j) = E(X_i Y_j) - E(X_i)E(Y_j) = \begin{cases} -\frac{1}{36} & \text{se } i = j, \\ 0 & \text{altrimenti,} \end{cases}$$

$$\mathrm{Cov}(X, Y) = \mathrm{Cov}\left(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} Y_j\right) = \sum_{i=1}^{n} \sum_{j=1}^{n} \mathrm{Cov}(X_i, Y_j) = -\frac{n}{36},$$

e, poiché $X$ e $Y$ sono variabili binomiali,

$$\mathrm{Var}(X) = \mathrm{Var}(Y) = n\,\frac{5}{36}.$$

Infine segue:

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = \frac{-n/36}{n\,5/36} = -\frac{1}{5}.$$

Si noti che $\rho(X,Y)$ non dipende da $n$, a differenza di $\mathrm{Cov}(X,Y)$.

**Esempio** Siano $X$ e $Y$ tali che $p(0,0) = p(1,1) = p$ e $p(1,0) = p(0,1) = \frac{1}{2} - p$.
(i) Determinare i valori ammissibili di $p$.
(ii) Stabilire per quali valori di $p$ si ha che $X$ e $Y$ sono indipendenti.
(iii) Calcolare $\rho(X,Y)$ e studiarlo al variare di $p$.
**Soluzione.** Dalle condizioni $p(x,y) \geq 0\ \forall x,y$ segue $p \geq 0$ e $\frac{1}{2} - p \geq 0$, mentre $\sum_x \sum_y p(x,y) = 1$ per ogni $p \in \mathbb{R}$; quindi i valori ammissibili di $p$ sono

$$0 \leq p \leq \frac{1}{2}.$$

(ii) Dalla seguente tabella segue che $p(x,y) = p_X(x)\,p_Y(y)$ per ogni $x$ e $y$ se e solo se $p = 1/4$.

| $x\backslash y$ | 0 | 1 | $p_X(x)$ |
|---|---|---|---|
| 0 | $p$ | $1/2-p$ | $1/2$ |
| 1 | $1/2-p$ | $p$ | $1/2$ |
| $p_Y(y)$ | $1/2$ | $1/2$ | 1 |

Poiché $X$ e $Y$ sono entrambe variabili aleatorie di Bernoulli di parametro $1/2$, si ha

$$E[X] = E[Y] = \frac{1}{2}, \qquad \mathrm{Var}(X) = \mathrm{Var}(Y) = \frac{1}{2}\left(1-\frac{1}{2}\right) = \frac{1}{4},$$

ed inoltre $E[X\,Y] = \sum_{x=0}^{1}\sum_{y=0}^{1} x\,y\,p(x,y) = p(1,1) = p$, da cui segue

$$Cov(X,Y) = E[XY] - E[X]\,E[Y] = p - \frac{1}{4}.$$

Il coefficiente di correlazione è quindi dato da

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}} = \frac{p-\frac{1}{4}}{\frac{1}{4}} = 4\,p - 1, \qquad 0 \le p \le \frac{1}{2}.$$

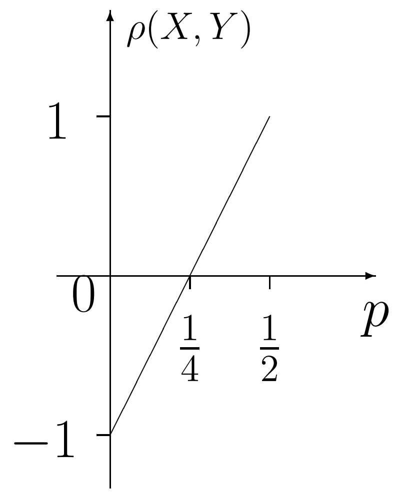
Grafico di $\rho(X,Y) = 4\,p - 1$, $0 \le p \le \frac{1}{2}$;

risulta:

$\rho(X,Y) = -1$ per $p = 0$,

$\rho(X,Y) = 0$ per $p = 1/4$,

$\rho(X,Y) = 1$ per $p = 1/2$.




Per $p = 0$ si ha:

$P(Y = 1 - X) = p(0,1) + p(1,0) = 1$

ed infatti in tal caso $\rho(X,Y) = -1$.

| $x\backslash y$ | 0 | 1 | $p_X(x)$ |
|---|---|---|---|
| 0 | 0 | 1/2 | 1/2 |
| 1 | 1/2 | 0 | 1/2 |
| $p_Y(y)$ | 1/2 | 1/2 | 1 |

Per $p = 1/2$ si ha:

$P(Y = X) = p(0,0) + p(1,1) = 1$

ed infatti in tal caso $\rho(X,Y) = 1$.

| $x\backslash y$ | 0 | 1 | $p_X(x)$ |
|---|---|---|---|
| 0 | 1/2 | 0 | 1/2 |
| 1 | 0 | 1/2 | 1/2 |
| $p_Y(y)$ | 1/2 | 1/2 | 1 |