Capitolo 6

Coppie di variabili aleatorie

In questo capitolo il concetto di variabile aleatoria viene generalizzato al caso di una coppia di variabili aleatorie: si mostra in particolare che in questo caso la caratterizzazione statistica completa avviene assegnando funzioni di due variabili, quali la CDF, la pdf o la DF congiunta (statistiche congiunte). Inoltre, le statistiche delle variabili aleatorie prese singolarmente (statistiche marginali) si possono ricavare univocamente una volta assegnate le statistiche congiunte. Un caso particolarmente semplice è quello delle variabili aleatorie indipendenti, per le quali le statistiche congiunte si fattorizzano semplicemente nel prodotto delle corrispondenti statistiche marginali. Particolarmente importante è il caso di una coppia di variabili aleatorie congiuntamente gaussiane, introdotto nell'esempio 6.2. Il capitolo si conclude con lo studio delle trasformazioni di coppie di variabili aleatorie e con l'introduzione del teorema fondamentale sulle trasformazioni di coppie di variabili aleatorie, che rappresenta la naturale estensione del teorema già studiato per le trasformazioni di una variabile aleatoria.

6.1 Introduzione

Nei capitoli precedenti abbiamo affrontato lo studio di *una singola* variabile aleatoria X, introducendo in particolare le funzioni (CDF, pdf o DF) che servono per la sua caratterizzazione statistica. Anche quando ci siamo occupati di trasformazioni di variabili aleatorie, del tipo $Y = g(X)$, mediante le quali a partire da una variabile aleatoria X si genera un'altra variabile aleatoria Y, ci siamo limitati a caratterizzare singolarmente Y (calcolandone ad esempio la CDF, la pdf, o la DF).

È chiaro che, se Y si ottiene a partire da X mediante una trasformazione $g(X)$, il legame esistente tra X ed Y è semplice, essendo un legame di tipo *deterministico*; se conosciamo il valore di X, il valore di Y è perfettamente determinato (il viceversa è vero solo se g è una funzione invertibile). Esistono tuttavia molti casi pratici in cui è possibile definire due variabili aleatorie su uno stesso esperimento di probabilità, ed il legame tra esse non è semplicemente deterministico. Ad

esempio, si consideri l'esperimento probabilistico che consiste nello scegliere una persona a caso in un insieme di persone, e si supponga che la variabile aleatoria X rappresenti l'altezza delle persona scelta, mentre la variabile aleatoria Y ne rappresenta il peso. È indubbio che esiste una dipendenza tra X ed Y, ma tale dipendenza non è espressa da una semplice relazione del tipo $Y = g(X)$, in quanto il legame tra peso ed altezza dipende da tutta una serie di altri parametri, quali costituzione fisica della persona, sesso, età, etc. D'altra parte appare abbastanza difficile, se non impossibile, individuare una formula esatta che descrive la relazione tra peso ed altezza e tenga conto di tutti i parametri del problema. È ragionevole invece descrivere in maniera approssimata tale relazione utilizzando le leggi della probabilità, il che costituisce l'oggetto del presente capitolo. Dovremo allora essere in grado di calcolare la probabilità che l'altezza di una persona sia compresa, diciamo, tra 180 e 190 cm, mentre il suo peso sia compreso tra 70 ed 80 kg. Tale problema è concettualmente simile a quello di descrivere congiuntamente due o più esperimenti aleatori (esperimento combinato) discusso in dettaglio nel § 2.4.

6.2 Funzione di distribuzione cumulativa (CDF) congiunta

Consideriamo due variabili aleatorie X ed Y costruite sullo stesso spazio di probabilità $(\Omega, \mathcal{S}, P)$. Sulla base delle conoscenze finora acquisite, siamo in grado di calcolare le probabilità che definiscono le CDF di X ed Y:

$$P(X \leq x) \triangleq F_X(x)\,;$$
$$P(Y \leq y) \triangleq F_Y(x)\,;$$

ma non sappiamo evidentemente calcolare la probabilità di eventi del tipo:

$$\{X \leq x\} \cap \{Y \leq y\} = \{X \leq x, Y \leq y\}\,,$$

che rappresentano la probabilità che i valori assunti dalla coppia (X, Y) appartengano alla regione del piano delimitata dalle rette di equazione $X = x$ ed $Y = y$ (regione in grigio in Fig. 6.2). Infatti questa probabilità *non* si può esprimere in termini di $F_X(x)$ e $F_Y(y)$.[1] Tale considerazione porta naturalmente all'introduzione di una misura della probabilità *congiunta* degli eventi $\{X \leq x\}$ e $\{Y \leq y\}$, rappresentata dalla funzione di distribuzione cumulativa (CDF) *congiunta* della coppia di variabili aleatorie (X, Y):

Definizione (CDF congiunta). Date due variabili aleatorie X ed Y costruite su uno stesso spazio di probabilità $(\Omega, \mathcal{S}, P)$, la loro CDF congiunta è:

$$F_{XY}(x, y) \triangleq P(X \leq x, Y \leq y)\,, \quad \forall (x, y) \in \overline{\mathbb{R}} \times \overline{\mathbb{R}}\,.$$

La CDF congiunta è chiaramente una funzione reale di due variabili reali, a valori in $[0, 1]$ (trattandosi di una probabilità). Essendo una funzione di due variabili, essa risulta più difficile da interpretare e manipolare matematicamente, rispetto alle CDF $F_X(x)$ ed $F_Y(y)$: le sue principali proprietà sono elencate nel paragrafo seguente. Notiamo che nel seguito, per brevità, utilizzeremo sia la notazione $F_{XY}(x, y)$, sia quella più sintetica $F(x, y)$.

[1] Tranne nel caso particolare in cui gli eventi $\{X \leq x\}$ e $\{Y \leq y\}$ siano indipendenti, come vedremo nel seguito.

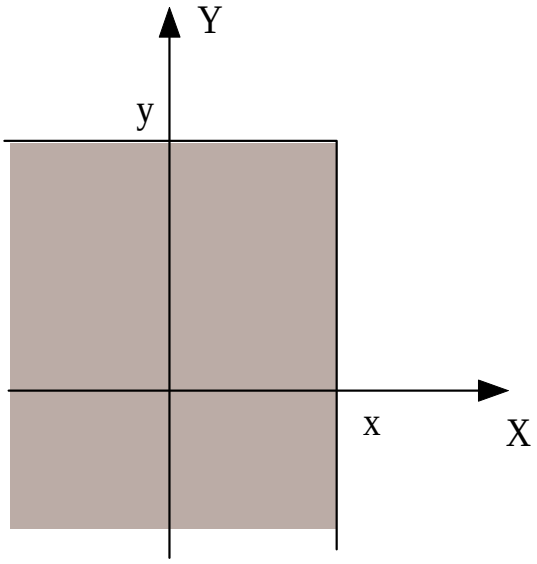


Fig. 6.1. L'evento $\{X \leq x, Y \leq y\}$ è costituito dai punti di Ω le cui immagini (X, Y) cadono nella regione in grigio.

6.2.1 Proprietà della CDF congiunta

La CDF congiunta $F(x, y)$ gode delle seguenti proprietà:

1. $F(-\infty, y) = 0$
   $F(x, -\infty) = 0$
   $F(+\infty, +\infty) = 1$

Prova. Per la prima identità , si ha:

$$F(-\infty, y) = P(X \leq -\infty, Y \leq y)\,,$$

ma $\{X \leq -\infty, Y \leq y\} \subseteq \{X \leq -\infty\} = \{X = -\infty\}$, per cui $P(X \leq -\infty, Y \leq y) \leq P(X = -\infty) = 0$, per la definizione di variabile aleatoria. Analogamente si procede per provare la seconda identità. Infine, per provare la terza si scrive semplicemente:

$$F(+\infty, +\infty) = P(X \leq +\infty, Y \leq +\infty) = P(\Omega) = 1\,.$$

□

2. $P(x_1 < X \leq x_2, Y \leq y) = F(x_2, y) - F(x_1, y)$;
   $P(X \leq x, y_1 < Y \leq y_2) = F(x, y_2) - F(x, y_1)$.

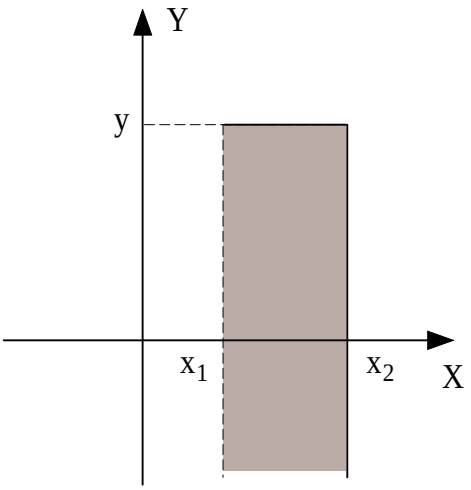


Fig. 6.2. L'evento $\{x_1 < X \leq x_2, Y \leq y\}$ è costituito dai punti di Ω le cui immagini (X, Y) cadono nella regione in grigio.

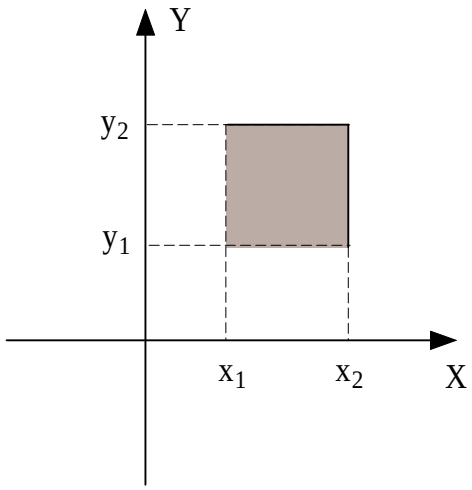


Fig. 6.3. L'evento $\{x_1 < X \leq x_2, y_1 < Y \leq y_2\}$ è costituito dai punti di Ω le cui immagini (X, Y) cadono nella regione in grigio.

Prova. Per la prima relazione, si ha (Fig. 6.2)

$$\{X \leq x_1, Y \leq y\} \cup \{x_1 < X \leq x_2, Y \leq y\} = \{X \leq x_2, Y \leq y\},$$

e gli eventi a primo membro sono mutuamente esclusivi, per cui:

$$P(X \leq x_1, Y \leq y) + P(x_1 < X \leq x_2, Y \leq y) = P(X \leq x_2, Y \leq y),$$

ovvero:

$$F(x_1, y) + P(x_1 < X \leq x_2, Y \leq y) = F(x_2, y),$$

da cui l'asserto. Analogamente si procede per provare la seconda relazione. □

3. $$\boxed{P(x_1 < X \leq x_2, y_1 < Y \leq y_2) = F(x_2, y_2) - F(x_1, y_2) - F(x_2, y_1) + F(x_1, y_1).}$$

Prova. Si noti che si ha (Fig. 6.3):

$$\{x_1 < X \leq x_2, Y \leq y_2\} = \{x_1 < X \leq x_2, Y \leq y_1\} \cup \{x_1 < X \leq x_2, y_1 < Y \leq y_2\},$$

e gli eventi a secondo membro sono mutuamente esclusivi, per cui:

$$P(x_1 < X \leq x_2, Y \leq y_2) = P(x_1 < X \leq x_2, Y \leq y_1) + P(x_1 < X \leq x_2, y_1 < Y \leq y_2),$$

ovvero:

$$F(x_2, y_2) - F(x_1, y_2) = F(x_2, y_1) - F(x_1, y_1) + P(x_1 < X \leq x_2, y_1 < Y \leq y_2),$$

da cui l'asserto. □

Per ricordare mnemonicamente tale proprietà, osserviamo che la CDF compare con il segno positivo nelle coppie "concordi" (x_1, y_1) e (x_2, y_2), mentre compare con il segno negativo nelle coppie "discordi" (x_1, y_2) ed (x_2, y_1).

6.3 Funzione densità di probabilità (pdf) congiunta

A partire dalla CDF congiunta, è naturale definire la funzione densità di probabilità (pdf) *congiunta* di una coppia di variabili aleatorie (X, Y):

Definizione (pdf congiunta). Date due variabili aleatorie X ed Y con CDF congiunta $F_{XY}(x,y)$, la loro pdf congiunta è:

$$f_{XY}(x,y) \triangleq \frac{\partial^2}{\partial x \partial y} F_{XY}(x,y)\,. \tag{6.1}$$

Notiamo che nella definizione di pdf congiunta compare la derivata mista (rispetto a x ed y) della funzione di due variabili $F_{XY}(x,y)$; poiché la pdf è unica, assumeremo che tale derivata mista non dipenda dall'ordine di derivazione, ovvero che la funzione $F_{XY}(x,y)$ soddisfi la seguente *condizione di Schwartz* per lo scambio dell'ordine di derivazione: le derivate miste di $F_{XY}(x,y)$ rispetto ad x ed y devono esistere ed essere continue.[2] Notiamo che nel seguito utilizzeremo per la pdf congiunta sia la notazione $f_{XY}(x,y)$, sia quella più snella $f(x,y)$.

6.3.1 Proprietà della pdf congiunta

Mentre, sulla base della definizione, la pdf congiunta si ottiene a partire dalla CDF congiunta per semplice derivazione, la seguente relazione consente di calcolare la CDF congiunta a partire dalla pdf congiunta per integrazione:

$$F(x,y) = \int_{-\infty}^{x}\int_{-\infty}^{y} f(u,v)\,du\,dv\,. \tag{6.2}$$

Prova. Integrando membro a membro la (6.1), si ha:

$$\int_{-\infty}^{x}\int_{-\infty}^{y} f(u,v)\,du\,dv = \int_{-\infty}^{x}\int_{-\infty}^{y} \frac{\partial^2}{\partial u \partial v} F(u,v)\,du\,dv\,,$$

ed il secondo membro si scrive:

$$\begin{aligned}
&\int_{-\infty}^{x} \frac{\partial}{\partial u}\left[\int_{-\infty}^{y} \frac{\partial}{\partial v} F(u,v)\,dv\right] du = \int_{-\infty}^{x} \frac{\partial}{\partial u}\left[F(u,v)\right]_{v=-\infty}^{v=y} du \\
&= \int_{-\infty}^{x} \frac{\partial}{\partial u}\left[F(u,y) - \underbrace{F(u,-\infty)}_{=0}\right] du \\
&= \left[F(u,y)\right]_{u=-\infty}^{u=x} = F(x,y)\,,
\end{aligned}$$

per cui resta provato l'asserto. □

Dalla (6.2), ponendo $x = y = +\infty$, e ricordando che $F(+\infty,+\infty) = 1$ si ricava:

$$\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f(u,v)\,du\,dv = 1\,,$$

che rappresenta la cosiddetta *condizione di normalizzazione* della pdf, che va interpretata geometricamente nel senso che il volume compreso tra la superficie di equazione $z = f(x,y)$ ed il piano $z = 0$ è unitario.

Come interpretare la pdf congiunta? Una possibile interpretazione è fornita dalla seguente relazione:

$$f(x,y)\,dx\,dy = P(x < X \leq x + dx, y < Y \leq y + dy) \tag{6.3}$$

[2] Salvo nel caso in cui la pdf presenti un impulso nel punto (x,y), caso che peraltro non considereremo mai in pratica

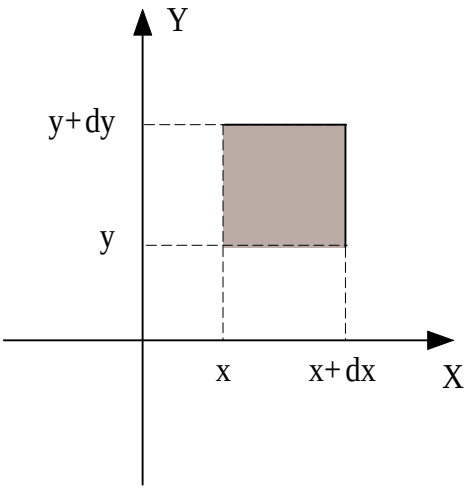


Fig. 6.4. L'evento $\{x < X \leq x + dx, y < Y \leq y + dy\}$ è costituito dai punti di Ω le cui immagini (X, Y) cadono nel "rettangolino" di area $dx\,dy$ (regione in grigio).

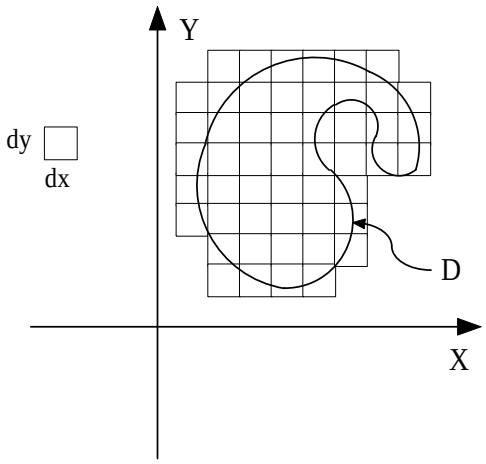


Fig. 6.5. La probabilità che la coppia (X, Y) appartenga al dominio D si può esprimere come somma di sovrapposizione (al limite, come integrale) delle probabilità che la coppia (X, Y) appartenga a "rettangolini" di area infinitesima che ricoprono il dominio D.

cioè $f(x, y)$ rappresenta la probabilità che la coppia di variabili aleatorie (X, Y) appartenga ad un "rettangolino" di lati infinitesimi, divisa per l'area $dx\,dy$ del rettangolino (Fig. 6.4). Questo risultato giustifica, anche nel caso bidimensionale, la denominazione di *densità* di probabilità, e prova anche implicitamente che $f(x, y) \geq 0$.

Prova. Applicando la proprietà 3 della CDF vista in precedenza, si ha:

$$P(x < X \leq x + dx, y < Y \leq y + dy) = F(x, y) + F(x + dx, y + dy) - F(x, y + dy) - F(x + dx, y)\,,$$

che possiamo riscrivere anche come:

$$P(x < X \leq x + dx, y < Y \leq y + dy) = [F(x + dx, y + dy) - F(x, y + dy)] - [F(x + dx, y) - F(x, y)]\,,$$

da cui, dividendo e moltiplicando per $dx\,dy$ e sfruttando la definizione di derivata parziale come limite del rapporto incrementale rispetto alla variabile d'interesse (con l'altra variabile fissa), si ha:

$$\begin{aligned}
&P(x < X \leq x + dx, y < Y \leq y + dy) = \\
&= \frac{1}{dy}\left\{\frac{F(x + dx, y + dy) - F(x, y + dy)}{dx} - \frac{F(x + dx, y) - F(x, y)}{dx}\right\} dx\,dy = \\
&= \frac{1}{dy}\left(\frac{\partial F(x, y + dy)}{\partial x} - \frac{\partial F(x, y)}{\partial x}\right) dx\,dy = \\
&= \frac{\partial^2 F(x, y)}{\partial y \partial x}\,dx\,dy\,,
\end{aligned}$$

da cui, ricordando l'assunzione che la derivata mista rispetto ad x ed y non dipende dall'ordine di derivazione, si ha l'asserto. □

Più in generale, se D è un dominio qualsiasi di $\mathbb{R}^2$, posso vederlo come la sovrapposizione di

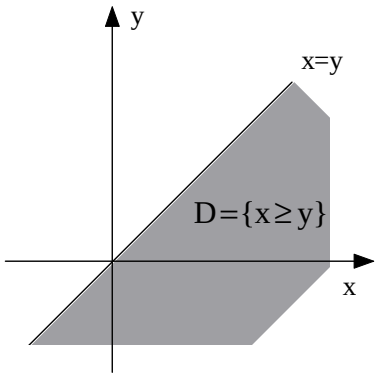


Fig. 6.6. La probabilità che $X \geq Y$ si ottiene integrando la pdf congiunta sul dominio $D = \{x \geq y\}$ (regione in grigio).

"rettangolini" di area infinitesima (Fig. 6.5), e quindi scrivere:

$$P[(X,Y) \in D] = \iint_D f_{XY}(x,y)\,dx\,dy\,,$$

per cui la pdf congiunta integrata su un qualunque dominio D restituisce la probabilità che la coppia di variabili aleatorie (X,Y) appartenga al dominio. Come si vede, la pdf congiunta è un potente strumento per il calcolo di probabilità relative alla coppia di variabili aleatorie (X,Y): l'unica difficoltà si può incontrare nella risoluzione dell'integrale doppio nella (6.3.1), specialmente se il dominio D ha una forma complicata.

► *Esempio* 6.1. Sia (X,Y) una coppia di variabili aleatorie con pdf congiunta $f(x,y)$: applichiamo i concetti precedenti per calcolare $P(X \geq Y)$. Evidentemente, il dominio D da considerare in questo caso è quello definito da $D = \{(x,y) \in \mathbb{R}^2 \text{ tali che } x \geq y\}$, raffigurato in Fig. 6.3.1.

Tale dominio si può riguardare come *normale*[3] sia rispetto all'asse x che all'asse y, per cui la probabilità cercata si può calcolare nei due modi equivalenti:

$$P(X \geq Y) = \int_{-\infty}^{\infty} dx \int_{-\infty}^{x} f(x,y)\,dy = \int_{-\infty}^{\infty} dy \int_{y}^{\infty} f(x,y)\,dx\,.$$

◄

6.4 Funzione di distribuzione di probabilità (DF) congiunta

Nel caso in cui le variabili aleatorie X ed Y siano entrambi discrete, anziché descriverle in termini di CDF o pdf congiunta, risulta più semplice fornire la loro descrizione congiunta attraverso l'introduzione della funzione di distribuzione di probabilità (DF) *congiunta*:

[3] Ricordiamo che un dominio D si dice normale rispetto all'asse x se si può esprimere come $D = \{a \leq x \leq b, \alpha(x) \leq y \leq \beta(x)\}$, dove $\alpha(x)$ e $\beta(x)$ sono opportune funzioni di x (al limite costanti).

Definizione (DF congiunta). Date due variabili aleatorie X ed Y discrete costruite su uno stesso spazio di probabilità $(\Omega, \mathcal{S}, P)$, a valori in $\mathcal{X}$ e $\mathcal{Y}$, rispettivamente, la loro DF congiunta è:

$$p_{XY}(x,y) = P(X = x, Y = y)\,, \tag{6.4}$$

dove $(x,y) \in \mathcal{X} \times \mathcal{Y}$.

Concludiamo osservando che per caratterizzare statisticamente una coppia di variabili aleatorie è necessaria la conoscenza della CDF congiunta, della pdf congiunta, o della DF congiunta (nel caso discreto).

6.5 Statistiche congiunte e marginali

Per una coppia di variabili aleatorie $(X.Y)$, le CDF, pdf e DF congiunte sono dette statistiche *congiunte*, mentre quelle delle singole variabili aleatorie sono dette statistiche *marginali*. Si pone allora il seguente problema: abbiamo visto che non è possibile "ricavare" le statistiche congiunte da quelle marginali. Ci chiediamo se sia possibile il viceversa: in effetti vedremo che è possibile ricavare le statistiche marginali da quelle congiunte. Per le CDF si ha, infatti,

$$\begin{aligned} F_X(x) &= F_{XY}(x, +\infty)\,, \\ F_Y(y) &= F_{XY}(+\infty, y)\,; \end{aligned}$$

mentre per le pdf

$$\begin{aligned} f_X(x) &= \int_{-\infty}^{\infty} f_{XY}(x,y)\,dy\,, \\ f_Y(y) &= \int_{-\infty}^{\infty} f_{XY}(x,y)\,dx\,; \end{aligned}$$

ed infine per le DF:

$$\begin{aligned} p_X(x) &= \sum_{y \in \mathcal{Y}} p_{XY}(x,y)\,, \\ p_Y(y) &= \sum_{x \in \mathcal{X}} p_{XY}(x,y)\,. \end{aligned}$$

Prova. La dimostrazione per le CDF è banale. Infatti, poiché:

$$\{X \leq x\} = \{X \leq x\} \cap \Omega = \{X \leq x\} \cap \{Y \leq +\infty\},$$

allora si ha $F_X(x) = F_{XY}(x, +\infty)$, e scambiando i ruoli di X ed Y si ottiene anche la seconda relazione.
Per le pdf, si consideri la relazione (6.2), e si derivi rispetto ad x, applicando il teorema fondamentale del calcolo integrale:

$$\frac{\partial F_{XY}(x,y)}{\partial x} = \int_{-\infty}^{y} f_{XY}(x,v)\, dv\,.$$

Ponendo nella precedente $y = +\infty$, si ha $F_{XY}(x, +\infty) = F_X(x)$ e quindi:

$$\frac{d}{dx} F_X(x) = f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,v)\, dv\,.$$

cioè l'asserto. La seconda relazione per le pdf si ottiene con ragionamento analogo, scambiando i ruoli di X ed Y.
Infine, per quanto riguarda le DF, il ragionamento è semplice. Infatti, si ha:

$$\{X = x\} = \cup_{y \in \mathcal{Y}} \{X = x\} \cap \{Y = y\},$$

da cui si ha l'asserto, essendo gli eventi a secondo membro mutuamente esclusivi. La seconda relazione per le DF si ottiene banalmente scambiando i ruoli di X ed Y. □

► *Esempio* 6.2. Una coppia di variabili aleatorie (X, Y) si dicono congiuntamente gaussiane, e si denotano con $(X, Y) \sim \mathrm{N}(\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho)$, se la loro pdf congiunta ha le seguente espressione:

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\, e^{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2} - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]}, \tag{6.5}$$

raffigurata in Fig. 6.7. Si noti il caratteristico andamento a campana della distribuzione gaussiana, che si manifesta anche nel caso bidimensionale. Le curve di livello della funzione $f_{XY}(x,y)$, ovvero le curve ottenute dall'intersezione della superficie di Fig. 6.7 con piani orizzontali di equazione $z =$ costante, sono ellissi di equazione (vedi equazione (6.5))

$$\frac{(x-\mu_X)^2}{\sigma_X^2} - 2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\,\sigma_Y} + \frac{(y-\mu_Y)^2}{\sigma_Y^2} = \text{costante}$$

e sono raffigurate in Fig. 6.8 e Fig. 6.9, per due diverse scelte del parametro ρ. Gli assi maggiori e minori di tali ellissi sono inclinati rispetto all'asse x di due angoli α_1 ed α_2 (che differiscono di $\pi/2$) e che si ottengono dalla seguente equazione trigonometrica:

$$\tan(2\alpha) = \frac{2\,\rho\,\sigma_X\,\sigma_Y}{\sigma_X^2 - \sigma_Y^2}\,. \tag{6.6}$$

Notiamo che la pdf congiunta di una coppia di variabili aleatorie congiuntamente gaussiane dipende da 5 parametri, dei quali μ_X e μ_Y possono assumere valori arbitrari, σ_X e σ_Y sono non negativi, ed infine ρ deve assumere valori tali che $|\rho| \leq 1$ (osserviamo in realtà che la (6.5) perde di significato per $|\rho| = 1$). Notiamo che la distribuzione assume il valore massimo in (μ_X, μ_Y), che σ_X e σ_Y rappresentano l'estensione della campana lungo X e Y, rispettivamente, mentre ρ governa la "strettezza" degli ellissi: si confrontino le Figg. 6.8 e 6.9 dove si riportano le curve di livello per $\rho = 0.5$ e $\rho = 0.9$. Per $|\rho| \to 1$, gli ellissi degenerano in segmenti e la pdf congiunta tende a concentrarsi sempre più su una retta obliqua. Vedremo nel prossimo capitolo il significato di ρ e della condizione $|\rho| = 1$, mentre il significato degli altri parametri sarà chiarito nel corso di questo stesso esempio.

A partire dalle statistiche congiunte, applicando le relazioni tra pdf congiunte e marginali, è possibile determinare le statistiche marginali di X ed Y. Procediamo per X (per Y i calcoli sono simili); dobbiamo calcolare

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)\, dy\,.$$

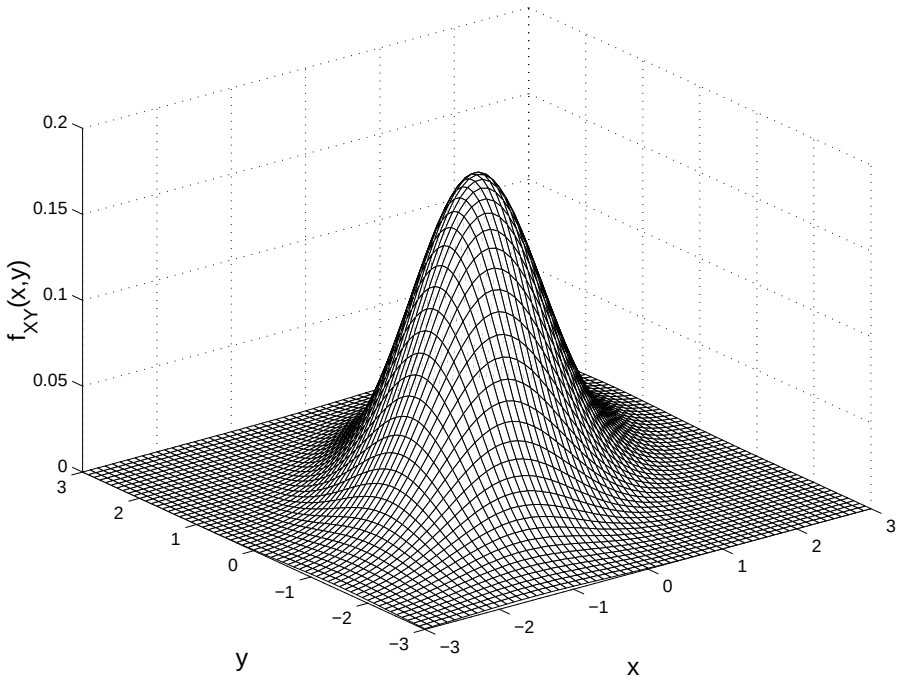


Fig. 6.7. La pdf $f_{XY}(x,y)$ di due variabili aleatorie congiuntamente gaussiane $(X,Y) \sim \mathrm{N}(0,0,1,1,0.5)$.

Manipoliamo la pdf congiunta per scriverla in una forma che consenta la semplice risoluzione dell'integrale. Si ha

$$\begin{aligned} f_{XY}(x,y) &= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\, e^{-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_X)^2}{\sigma_X^2}-2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}+\frac{(y-\mu_Y)^2}{\sigma_Y^2}\right]} \\ &= \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\rho^2}}\, e^{-\frac{1}{2(1-\rho^2)}\frac{(x-\mu_X)^2}{\sigma_X^2}}\; e^{-\frac{1}{2(1-\rho^2)}\left[\frac{(y-\mu_Y)^2}{\sigma_Y^2}-2\rho\frac{(x-\mu_X)(y-\mu_Y)}{\sigma_X\sigma_Y}\right]}. \end{aligned}$$

Aggiungiamo e sottraiamo la quantità $\frac{\rho^2(x-\mu_X)^2}{\sigma_X^2}$ nell'argomento del secondo esponenziale, così da far comparire un quadrato perfetto. Dopo alcune manipolazioni algebriche si ha:

$$f_{XY}(x,y) = \left[\frac{1}{\sigma_X\sqrt{2\pi}}\, e^{-\frac{1}{2\sigma_X^2}(x-\mu_X)^2}\right]\left[\frac{1}{\sigma_Y\sqrt{1-\rho^2}\sqrt{2\pi}}\, e^{-\frac{1}{2\sigma_Y^2(1-\rho^2)}\left[y-\mu_Y-\rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)\right]^2}\right].$$

Osserviamo che il primo fattore (tra parentesi quadre) rappresenta la pdf di una variabile aleatoria $X \sim \mathrm{N}(\mu_X,\sigma_X)$; per quanto riguarda il secondo, per un fissato valore di x, è facile verificare che esso rappresenta la pdf di una variabile aleatoria $Y \sim \mathrm{N}(\mu_Y+\rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X),\sigma_Y\sqrt{1-\rho^2})$, vale a dire con media $\mu_Y+\rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)$ e deviazione standard $\sigma_Y\sqrt{1-\rho^2}$.

Se adesso integriamo la pdf congiunta rispetto ad y per ottenere la pdf marginale $f_X(x)$, osserviamo che il secondo fattore, essendo una pdf per ogni valore di x, ha integrale rispetto ad y unitario. Pertanto, si ha semplicemente:

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x,y)\,dy = \frac{1}{\sigma_X\sqrt{2\pi}}\, e^{-\frac{1}{2\sigma_X^2}(x-\mu_X)^2},$$

e quindi è evidente che $X \sim \mathrm{N}(\mu_X,\sigma_X)$, cioè X è *marginalmente* gaussiana, con media μ_X e deviazione standard σ_X.

Ovviamente saremmo potuti giungere ad una decomposizione simmetrica operando rispetto ad y anziché rispetto ad x, per cui con analogo ragionamento si conclude che:

$$f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)\,dx = \frac{1}{\sigma_Y\sqrt{2\pi}}\, e^{-\frac{1}{2\sigma_Y^2}(y-\mu_Y)^2},$$

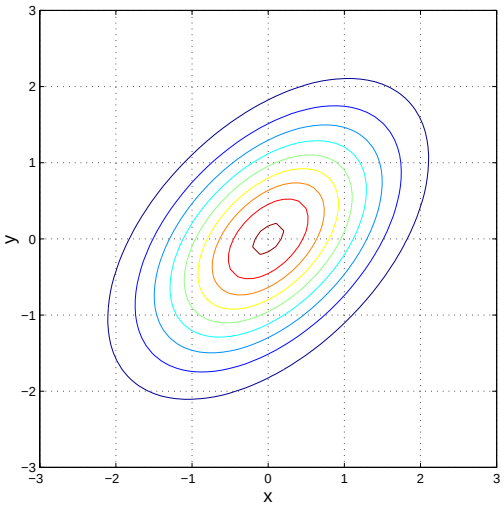


Fig. 6.8. Curve di livello della pdf $f_{XY}(x,y)$ di due variabili aleatorie congiuntamente gaussiane $(X,Y) \sim \mathrm{N}(0,0,1,1,\rho)$, per $\rho = 0.5$.

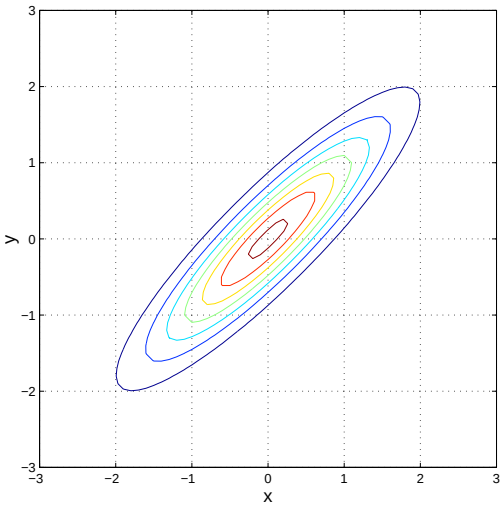


Fig. 6.9. Curve di livello della pdf $f_{XY}(x,y)$ di due variabili aleatorie congiuntamente gaussiane $(X,Y) \sim \mathrm{N}(0,0,1,1,\rho)$, per $\rho = 0.9$.

e quindi è evidente che $Y \sim \mathrm{N}(\mu_Y, \sigma_Y)$, cioè Y è *marginalmente* gaussiana, con media μ_Y e deviazione standard σ_Y. In conclusione: se $(X,Y) \sim \mathrm{N}(\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho)$ sono congiuntamente gaussiane, X ed Y sono marginalmente gaussiane, e si ha $X \sim \mathrm{N}(\mu_X, \sigma_X)$ e $Y \sim \mathrm{N}(\mu_Y, \sigma_Y)$. Il viceversa non è sempre vero: è possibile costruire esempi di variabili aleatorie marginalmente gaussiane la cui pdf congiunta non sia gaussiana (si veda [3, Es. 6-1]). Notiamo infine che in questo modo abbiamo anche interpretato 4 dei 5 parametri che compaiono nella pdf congiunta di due variabili aleatorie congiuntamente gaussiane, e che in sostanza sono quelli caratteristici delle statistiche marginali delle variabili aleatorie X ed Y; in effetti il parametro ρ è l'unico parametro che descrive la relazione di interdipendenza tra le due variabili aleatorie congiuntamente gaussiane. Come già osservato, il suo significato sarà chiarito nel prossimo capitolo. ◀

6.6 Coppie di variabili aleatorie indipendenti

Un caso particolarmente semplice da affrontare è quello in cui le variabili aleatorie X ed Y risultano *indipendenti*:

Definizione (coppie di variabili aleatorie indipendenti). Due variabili aleatorie X ed Y si dicono indipendenti se

$$F_{XY}(x,y) = F_X(x)\, F_Y(y)\,, \quad \forall (x,y) \in \mathbb{R}^2\,. \tag{6.7}$$

Come si vede, così come nella teoria della probabilità elementare l'indipendenza tra eventi si può esprimere come *fattorizzazione* della probabilità congiunta, ovvero $P(AB) = P(A)\,P(B)$, così per le variabili aleatorie l'indipendenza si può esprimere come fattorizzazione della CDF congiunta nel prodotto delle CDF marginali. Questo è chiaro, in quanto la CDF rappresenta in ultima analisi la probabilità di una collezione di eventi dello spazio di probabilità.

Nel caso di variabili aleatorie discrete, la definizione di indipendenza si può dare direttamente in termini di probabilità:

$$P(X = x, Y = y) = P(X = x)\, P(Y = y)\,,$$

il che equivale a dire, in termini di DF, che:

$$p_{XY}(x,y) = p_X(x)\, p_Y(y)\,.$$

6.6.1 Proprietà delle variabili aleatorie indipendenti

Enunciamo e dimostriamo alcune semplici proprietà delle coppie di variabili aleatorie indipendenti, che sono diretta conseguenza della definizione (6.7):

1. Se X ed Y sono indipendenti, allora $f_{XY}(x,y) = f_X(x) f_Y(y)$, $\forall (x,y) \in \mathbb{R}^2$ (fattorizzazione della pdf congiunta).

   Prova. Si ottiene immediatamente derivando la definizione (6.7). □

2. Se X ed Y sono indipendenti, allora gli eventi $\{X \in I_1\}$ e $\{Y \in I_2\}$ sono indipendenti, $\forall I_1, I_2 \subseteq \mathbb{R}$.

   Prova. Si ha:

   $$\begin{aligned} P(X \in I_1, Y \in I_2) &= \int\int_{I_1 \times I_2} f_{XY}(x,y)\, dx\, dy = \int_{I_1} f_X(x)\, dx \int_{I_2} f_Y(y)\, dy \\ &= P(X \in I_1)\, P(Y \in I_2)\,. \end{aligned}$$

   □

3. Se X ed Y sono indipendenti, allora le variabili aleatorie $Z = g(X)$ e $W = h(Y)$ sono indipendenti.

   Prova. Si ha:

   $$F_{ZW}(z,w) = P(Z \le z, W \le w) = P(X \in R_z, Y \in R_w)\,,$$

   dove $R_z \triangleq \{x \in \mathbb{R} \text{ tali che } g(x) \le z\}$ e $R_w \triangleq \{y \in \mathbb{R} \text{ tali che } h(y) \le w\}$. Per l'indipendenza di X ed Y, si ha (in base alla proprietà 2 precedentemente dimostrata):

   $$F_{ZW}(z,w) = P(X \in R_z) P(Y \in R_w) = P(Z \le z) P(W \le w) = F_Z(z)\, F_W(w)$$

   per cui resta provato l'asserto. □

Osserviamo che se le variabili aleatorie sono costruite su uno spazio di probabilità prodotto $\Omega_1 \times \Omega_2$, e in maniera tale che:

$$\begin{aligned} X[(\omega_1,\omega_2)] &= X(\omega_1)\,, \\ Y[(\omega_1,\omega_2)] &= Y(\omega_2)\,, \end{aligned}$$

allora, se gli esperimenti Ω_1 ed Ω_2 sono indipendenti, anche le variabili aleatorie X ed Y sono indipendenti.

► *Esempio* 6.3. Sia Ω_1 lo spazio campione associato all'esperimento "lancio di una moneta", e sia Ω_2 lo spazio campione associato all'esperimento "lancio di un dado". Qualsiasi variabile aleatoria X costruita su $\Omega_1 \times \Omega_2$ che dipende solo da Ω_1 e qualunque variabile aleatoria Y costruita su $\Omega_1 \times \Omega_2$ che dipende solo da Ω_2 sono indipendenti. ◄

► *Esempio* 6.4. Spesso l'indipendenza "statistica" tra due variabili aleatorie si può assumere sulla base dell'indipendenza "fisica". Ad esempio, appare chiaro che se X rappresenta l'altezza ed Y il peso di un individuo, X ed Y non sono indipendenti (le persone più alte mediamente pesano di più). Viceversa, se X rappresenta il peso e Y rappresenta il numero di fratelli e sorelle di una persona, pare ragionevole ritenere queste due variabili indipendenti. ◄

► *Esempio* 6.5. Abbiamo osservato (cfr. esempio 6.2) che due variabili aleatorie X ed Y marginalmente gaussiane non sono necessariamente anche congiuntamente gaussiane. Questo risultato però non vale se le variabili aleatorie sono marginalmente gaussiane *e* indipendenti: infatti, se $X \sim \mathrm{N}(\mu_X, \sigma_X)$ e $Y \sim \mathrm{N}(\mu_Y, \sigma_Y)$, indipendenti, la loro pdf congiunta si ottiene come:

$$f_{XY}(x,y) = f_X(x)\, f_Y(y) = \frac{1}{\sigma_X\sqrt{2\pi}}\, e^{-\frac{1}{2\sigma_X^2}(x-\mu_X)^2} \frac{1}{\sigma_Y\sqrt{2\pi}}\, e^{-\frac{1}{2\sigma_Y^2}(y-\mu_Y)^2},$$

per cui è facile verificare che essa è una pdf del tipo gaussiano bidimensionale (6.5), con $\rho = 0$. Viceversa, se si hanno due variabili aleatorie X, Y congiuntamente gaussiane e con $\rho = 0$, si vede che esse sono indipendenti, in quanto la loro pdf congiunta $f_{XY}(x,y)$ si fattorizza nel prodotto di due pdf gaussiane monodimensionali. Quindi, sebbene non ne abbiamo ancora dato una interpretazione rigorosa, intuiamo che ρ misura il grado di dipendenza tra due variabili aleatorie congiuntamente gaussiane: quando $\rho = 0$ le variabili aleatorie sono indipendenti; quando $|\rho| = 1$ le due variabili aleatorie sono massimamente dipendenti. ◄

6.7 Trasformazioni di coppie di variabili aleatorie

Vogliamo adesso estendere il nostro studio sulle trasformazioni di variabili aleatorie, condotto nel capitolo 4 per il caso di trasformazioni di una sola variabile aleatoria, al caso di coppie di variabili aleatorie. Qui però la situazione si presenta più articolata, in quanto possiamo avere *una* trasformazione $Z = g(X,Y)$, ovvero la trasformazione di una coppia di variabili aleatorie in una singola variabile aleatoria (trasformazione $2 \to 1$), oppure una *coppia* di trasformazioni $Z = g(X,Y)$ e $W = h(X,Y)$, ovvero la trasformazione di una coppia di variabili aleatorie in una coppia di nuove variabili aleatorie (trasformazione $2 \to 2$). Studiamo separatamente i due casi.

6.7.1 Trasformazione 2→1

In questo caso, abbiamo una coppia (X,Y) di variabili aleatorie, caratterizzate dalle loro CDF congiunta $F_{XY}(x,y)$ e pdf congiunta $f_{XY}(x,y)$, e a partire da esse costruiamo una nuova variabile aleatoria $Z = g(X,Y)$, dove $g(x,y)$ è una funzione di due variabili. Vogliamo caratterizzare Z, in particolare calcolandone la CDF $F_Z(z)$ e la pdf $f_Z(z)$. Il problema formalmente si risolve in maniera semplice, in quanto si ha:

$$\begin{aligned} F_Z(z) &= P(Z \le z) = P(g(X,Y) \le z) \\ &= P((X,Y) \in D_z) = \iint_{D_z} f_{XY}(x,y)\, dx\, dy\,, \end{aligned} \tag{6.8}$$

dove $D_z = \{(x,y) \in \mathbb{R}^2 \text{ tali che } g(x,y) \le z\}$ è un dominio di $\mathbb{R}^2$. Per determinare poi la pdf di Z, possiamo o derivare la CDF, o direttamente ricavarla come:

$$\begin{aligned} f_Z(z)\, dz &= P(z < Z \le z + dz) \\ &= P(z < g(X,Y) \le z + dz) \\ &= P((X,Y) \in \Delta D_z) = \iint_{\Delta D_z} f_{XY}(x,y)\, dx\, dy\,, \end{aligned} \tag{6.9}$$

dove $\Delta D_z = \{(x,y) \in \mathbb{R}^2 \text{ tali che } z < g(x,y) \le z + dz\}$ è un dominio di $\mathbb{R}^2$.

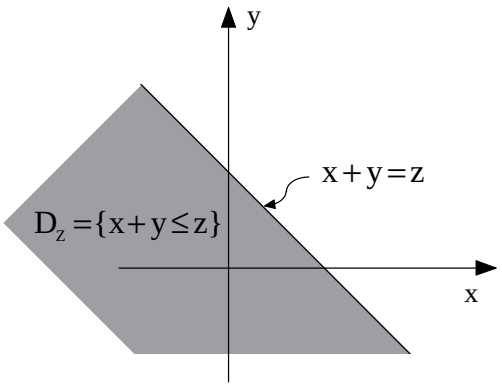


Fig. 6.10. Il dominio $D_z = \{(x,y) \in \mathbb{R}^2 \text{ tali che } x+y \le z\}$ è raffigurato in grigio.

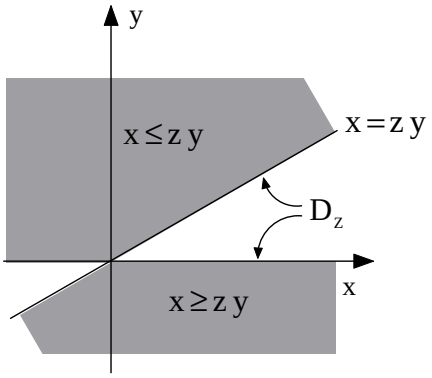


Fig. 6.11. Il dominio $D_z = \{(x,y) \in \mathbb{R}^2 \text{ tali che } \frac{x}{y} \le z\}$ è raffigurato in grigio (per $z > 0$).

► *Esempio* 6.6. Consideriamo la trasformazione $Z = X + Y$. Si ha:

$$F_Z(z) \triangleq P(Z \le z) = P(X + Y \le z)\,.$$

Per un fissato z, il dominio $D_z = \{(x,y) \in \mathbb{R}^2 \text{ tali che } x + y \le z\}$ è quello raffigurato in grigio in Fig. 6.10. Tale dominio si può riguardare ad esempio come normale rispetto all'asse x, ed in tal caso si descrive come $D_z = \{x \in \mathbb{R}, y \le z - x\}$. Pertanto applicando la (6.8) si trova:

$$F_Z(z) = \iint_{D_z} f_{XY}(x,y)\,dx\,dy = \int_{-\infty}^{\infty} dx \int_{-\infty}^{z-x} f_{XY}(x,y)\,dy\,.$$

Possiamo ottenere la pdf derivando la CDF precedente rispetto a z. Si ha, applicando il teorema fondamentale del calcolo integrale,

$$f_Z(z) = \frac{d}{dz}\int_{-\infty}^{\infty} dx \int_{-\infty}^{z-x} f_{XY}(x,y)\,dy = \int_{-\infty}^{\infty} f_{XY}(x, z-x)\,dx = \int_{-\infty}^{\infty} f_{XY}(z-x, x)\,dx\,,$$

dove l'ultimo integrale si ottiene con un semplice cambio di variabile. Osserviamo che, se X ed Y sono indipendenti, allora $f_{XY}(x,y) = f_X(x)\,f_Y(y)$, e quindi la pdf di $Z = X + Y$ diventa:

$$f_Z(z) = \int_{-\infty}^{\infty} f_X(x)\,f_Y(z-x)\,dx\,,$$

ovvero è data dal *prodotto di convoluzione* o semplicemente dalla *convoluzione* tra le funzioni $f_X(x)$ ed $f_Y(y)$, che si denota sinteticamente con $f_X * f_Y$. Pertanto, la pdf della somma di due variabili aleatorie *indipendenti* si ottiene effettuando la *convoluzione* delle rispettive pdf. ◄

► *Esempio* 6.7. Consideriamo la trasformazione $Z = X/Y$. Si ha:

$$F_Z(z) \triangleq P(Z \le z) = P\left(\frac{X}{Y} \le z\right)\,.$$

Osserviamo che, per un fissato z, la disuguaglianza $\frac{x}{y} \le z$ si scrive come $x \le z\,y$, se $y > 0$, oppure come $x \ge z\,y$, se $y < 0$. Pertanto, il dominio $D_z = \{(x,y) \in \mathbb{R}^2 \text{ tali che } \frac{x}{y} \le z\}$ è quello raffigurato in grigio in Fig. 6.11. Tale dominio è normale rispetto all'asse y, e si descrive come $D_z = \{y \in \mathbb{R}, x \le z\,y, \text{se } y > 0; x \ge z\,y, \text{se } y < 0\}$. Pertanto, applicando la (6.8) si trova:

$$F_Z(z) = \int_0^{\infty} dy \int_{-\infty}^{zy} f_{XY}(x,y)\,dx + \int_{-\infty}^{0} dy \int_{zy}^{\infty} f_{XY}(x,y)\,dx\,.$$

Possiamo ottenere la pdf derivando la CDF precedente rispetto a z: si ha[4]

$$f_Z(z) = \frac{d}{dz}F_Z(z) = \int_0^\infty y\, f_{XY}(zy,y)\, dy - \int_{-\infty}^0 y\, f_{XY}(zy,y)\, dy = \int_{-\infty}^\infty |y|\, f_{XY}(zy,y)\, dy\,.$$

Ad esempio, se $(X,Y) \sim \mathrm{N}(0,0,1,1,0)$, sono cioè gaussiane standard indipendenti, applicando la precedente relazione si trova:

$$\begin{aligned} f_Z(z) &= \int_{-\infty}^\infty |y|\, \frac{1}{2\pi}\, e^{-\frac{1}{2}(z^2y^2+y^2)}\, dy = \frac{1}{\pi}\int_0^\infty y\, e^{-\frac{1}{2}[y^2(z^2+1)]}\, dy = \\ &= \frac{1}{\pi}\int_0^\infty \frac{1}{z^2+1} y(z^2+1)\, e^{-\frac{1}{2}[y^2(z^2+1)]}\, dy = \\ &= \frac{1}{\pi}\frac{1}{z^2+1}\int_0^\infty \frac{d}{dy}\left[-e^{-\frac{1}{2}[y^2(z^2+1)]}\right] dy = \\ &= \frac{1}{\pi}\frac{1}{z^2+1}\left[-e^{-\frac{1}{2}[y^2(z^2+1)]}\right]_{y=0}^{y=\infty} \\ &= \frac{1/\pi}{z^2+1}\,, \end{aligned}$$

cioè risulta $Z \sim$ Cauchy(1). Pertanto il rapporto X/Y tra due variabili aleatorie gaussiane standard e indipendenti è una variabile aleatoria di Cauchy. ◀

6.7.2 Trasformazione 2→2

In questo caso abbiamo una coppia (X,Y) di variabili aleatorie, caratterizzate dalle loro CDF congiunta $F_{XY}(x,y)$ e pdf congiunta $f_{XY}(x,y)$, e a partire da esse costruiamo una nuova *coppia* di variabili aleatorie (Z,W), con $Z = g(X,Y)$ e $W = h(X,Y)$, dove $g(x,y)$ e $h(x,y)$ sono funzioni di due variabili. Il problema che si pone in tal caso è quello di determinare la CDF $F_{ZW}(z,w)$ o la pdf congiunta $f_{ZW}(z,w)$ di Z e W. È possibile applicare il seguente *teorema fondamentale sulle trasformazioni di coppie di variabili aleatorie*, che generalizza al caso di coppie di variabili aleatorie il teorema 4.1, valido per trasformazioni di una singola variabile aleatoria, e che enunciamo senza dimostrazione:

Teorema 6.1 (teorema fondamentale sulle trasformazioni di coppie di variabili aleatorie). Sia (X,Y) una coppia di variabili aleatorie con pdf $f_{XY}(x,y)$, e siano $Z = g(X,Y)$ e $W = h(X,Y)$ due nuove variabili aleatorie ottenute per trasformazione da (X,Y). Si consideri il sistema di equazioni:

$$\begin{cases} z &= g(x,y) \\ w &= h(x,y) \end{cases} \tag{6.10}$$

La pdf congiunta di (Z,W) è data da:

$$f_{ZW}(z,w) = \begin{cases} 0, & \text{se il sistema (6.10) non ha soluzioni;} \\ \displaystyle\sum_i \frac{f_{XY}(x_i,y_i)}{|\det[\boldsymbol{J}(x_i,y_i))]|}, & \text{dove } (x_i,y_i) \text{ è una soluzione del sistema (6.10);} \end{cases}$$

in cui $\det(\cdot)$ denota il determinante, e

$$\boldsymbol{J}(x,y) = \frac{\partial(z,w)}{\partial(x,y)} = \begin{pmatrix} \frac{\partial z}{\partial x} & \frac{\partial z}{\partial y} \\ \frac{\partial w}{\partial x} & \frac{\partial w}{\partial y} \end{pmatrix}$$

è la matrice jacobiana della trasformazione.

[4] Per la derivazione, si applichi la formula di Leibnitz, riportata in Appendice F.

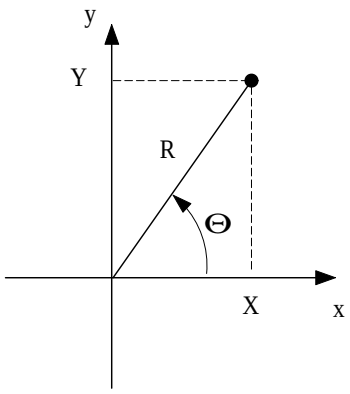


Fig. 6.12. Trasformazione da coordinate cartesiane a coordinate polari.

Si noti che per l'applicazione del teorema si richiede che il sistema (6.10) abbia al più una infinità numerabile di soluzioni. Nella pratica, risulta spesso utile la seguente osservazione: se il sistema è invertibile e denotiamo con $\boldsymbol{J}(z,w)$ la matrice jacobiana del sistema inverso, risulta:

$$\boldsymbol{J}(z,w) = \boldsymbol{J}(x,y)^{-1}\,, \quad (x,y) \text{ soluzione del sistema,}$$

e quindi:

$$\det[\boldsymbol{J}(z,w)] = \frac{1}{\det[\boldsymbol{J}(x,y)]}\,, \quad (x,y) \text{ soluzione del sistema.}$$

► *Esempio* 6.8 *(trasformazione da coordinate cartesiane a polari).* Consideriamo la coppia di variabili aleatorie (X,Y), che possiamo interpretare come le coordinate cartesiane di un punto di $\mathbb{R}^2$. Ha senso allora considerare la trasformazione che fornisce le coordinate *polari* (R,Θ) dello stesso punto, con $R \geq 0$ e $\Theta \in [0,2\pi[$. Tale trasformazione si può esprimere come:

$$\begin{cases} R &= \sqrt{X^2+Y^2} \\ \Theta &= \tan^{-1}(Y/X) \end{cases},$$

dove la funzione $\tan^{-1}(Y/X)$ (da non confondere con la funzione $\arctan(\cdot)$), determina *univocamente*, per ogni valore della coppia (X,Y), l'angolo $\Theta \in [0,2\pi[$ formato dal segmento di estremi $(0,0)$ ed (X,Y) con il semiasse positivo delle x, misurato in senso antiorario (Fig. 6.12). Se vogliamo ricavare la pdf di (R,Θ), applicando il teorema fondamentale 6.1 consideriamo il seguente sistema di equazioni, nelle incognite (x,y):

$$\begin{cases} r = \sqrt{x^2+y^2} \\ \theta = \tan^{-1}(y/x) \end{cases},$$

che ha se $r \geq 0$ e $\theta \in [0,2\pi[$ una sola soluzione, data da

$$\begin{cases} x = r\cos\theta \\ y = r\sin\theta \end{cases},$$

mentre non ha nessuna soluzione se $r < 0$. Il calcolo della matrice jacobiana, inoltre, fornisce:

$$J(r,\theta) = \frac{\partial(x,y)}{\partial(r,\theta)} = \begin{pmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{pmatrix},$$

e quindi

$$|J(r,\theta)| = |r| = r \geq 0\,.$$

Si ha allora:

$$f_{R\Theta}(r,\theta) = \begin{cases} 0, & \text{se } r < 0; \\ r f_{XY}(r\cos\theta, r\sin\theta), & \text{se } r \geq 0; \end{cases}$$

ovvero:

$$f_{R\Theta}(r,\theta) = r f_{XY}(r\cos\theta, r\sin\theta)\, u(r)\,, \tag{6.11}$$

dove $u(r)$ è la funzione gradino. Se poi vogliamo ottenere le distribuzioni *marginali* di R e Θ, basta integrare rispetto alla variabile che non interessa: si ha, cioè:

$$f_R(r) = u(r) \int_0^{2\pi} r f_{XY}(r\cos\theta, r\sin\theta)\, d\theta$$
$$f_\Theta(\theta) = \int_0^{\infty} r f_{XY}(r\cos\theta, r\sin\theta)\, dr$$

Si noti la scelta degli intervalli di integrazione: $[0,\infty[$ per l'integrale in dr, $[0,2\pi[$ per l'integrale in $d\theta$, corrispondenti ai valori assunti da R e da Θ, rispettivamente. ◄

► *Esempio* 6.9. Applichiamo i risultati della trasformazione da coordinate cartesiane a coordinate polari al caso in cui le variabili aleatorie X ed Y siano congiuntamente gaussiane, ed in particolare siano indipendenti ($\rho = 0$), a media nulla ($\mu_X = \mu_Y = 0$) e con la stessa deviazione standard ($\sigma_X = \sigma_Y = \sigma$), il che sinteticamente si denota come $(X,Y) \sim \mathrm{N}(0,0,\sigma,\sigma,0)$.

Per l'ipotesi di indipendenza, la pdf congiunta $f_{XY}(x,y)$ si scrive semplicemente come prodotto di due pdf gaussiane marginali a media nulla e con la stessa deviazione standard:

$$f_{XY}(x,y) = \frac{1}{2\pi\sigma^2}\, e^{-\frac{1}{2\sigma^2}(x^2+y^2)}\,.$$

Applicando la (6.11), si ha:

$$\begin{aligned} f_{R\Theta}(r,\theta) &= r f_{XY}(r\cos\theta, r\sin\theta)\, u(r) \\ &= r\frac{1}{2\pi\sigma^2}\, e^{-\frac{1}{2\sigma^2} r^2(\cos^2(\theta)+\sin^2(\theta))}\, u(r) \\ &= \frac{r}{2\pi\sigma^2}\, e^{-\frac{r^2}{2\sigma^2}}\, u(r)\,. \end{aligned}$$

Se ricaviamo le due pdf marginali, otteniamo per R:

$$\begin{aligned} f_R(r) &= \int_0^{2\pi} f_{R\Theta}(r,\theta)\, d\theta = \int_0^{2\pi} \frac{r}{2\pi\sigma^2}\, e^{-\frac{r^2}{2\sigma^2}}\, u(r)\, d\theta = \\ &= \frac{r}{\sigma^2}\, e^{-\frac{r^2}{2\sigma^2}}\, u(r)\,, \end{aligned}$$

mentre per Θ si ha:

$$\begin{aligned} f_\Theta(\theta) &= \int_0^{\infty} f_{R\Theta}(r,\theta)\, dr = \int_0^{\infty} \frac{r}{2\pi\sigma^2}\, e^{-\frac{r^2}{2\sigma^2}}\, dr = \\ &= \frac{1}{2\pi}\int_0^{\infty} \left[-\frac{d}{dr} e^{-\frac{r^2}{2\sigma^2}}\right] dr = \frac{1}{2\pi}\left[-e^{-\frac{r^2}{2\sigma^2}}\right]_{r=0}^{r=\infty} = \frac{1}{2\pi}\,, \end{aligned}$$

per cui si osserva che $f_{R\Theta}(r,\theta) = f_R(r)\, f_\Theta(\theta)$, per cui R e Θ sono indipendenti, ed inoltre R ha una distribuzione di tipo Rayleigh con parametro $b = 2\sigma^2$, ovvero $R \sim \text{Rayleigh}(2\sigma^2)$, mentre $\Theta \sim \mathrm{U}(0,2\pi)$, cioè è uniforme in $(0,2\pi)$. ◄

► *Esempio* 6.10 *(generazione di variabili aleatorie gaussiane).* Il precedente esempio suggerisce un metodo, alternativo a quello della CDF inversa o della trasformazione percentile (cfr. § 4.3.1) per generare variabili aleatorie gaussiane. Infatti, l'ostacolo principale all'applicazione della tecnica della trasformazione percentile al caso gaussiano risiede nel fatto che la CDF gaussiana non ammette un'espressione analitica in forma

chiusa, e quindi la sua inversione si ottiene solo attraverso tecniche numeriche. Viceversa, non ci sono problemi a generare con la tecnica della trasformazione percentile due variabili aleatorie R e Θ, rispettivamente di tipo Rayleigh e uniforme in $(0, 2\pi)$ (cfr. esercizio 4.13), in quanto le rispettive CDF sono facilmente invertibili. Pertanto, la generazione di variabili aleatorie gaussiane si può effettuare con un algoritmo in due passi:

1. utilizzando due generatori di variabili aleatorie $\mathrm{U}(0,1)$ indipendenti,[5] e adoperando la tecnica della trasformazione percentile, si generano due variabili aleatorie R e Θ rispettivamente di tipo Rayleigh ed uniforme in $(0, 2\pi)$; tali variabili aleatorie, essendo ottenute per trasformazione da variabili aleatorie indipendenti, saranno ancora indipendenti;
2. si costruiscono le due variabili aleatorie $X = R\cos(\Theta)$ ed $Y = R\sin(\Theta)$; esse risulteranno variabili aleatorie gaussiane indipendenti, a media nulla e con la stessa varianza.

Per completezza, osserviamo che se si desidera generare variabili aleatorie gaussiane non indipendenti, è sufficiente partire da variabili aleatorie Z_1 e Z_2 indipendenti e standard (a media nulla e varianza unitaria) generate con l'algoritmo precedentemente esposto, e successivamente applicare la seguente trasformazione $2 \to 2$:

$$\begin{cases} X = \rho\,\sigma_X\,Z_1 + \sigma_X\sqrt{1-\rho^2}\,Z_2 + \mu_X\,, \\ Y = \sigma_Y\,Z_1 + \mu_Y\,. \end{cases}$$

Infatti, applicando il teorema fondamentale sulle trasformazioni, si verifica facilmente che le variabili aleatorie sono congiuntamente gaussiane, vale a dire $(X, Y) \sim \mathrm{N}(\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho)$. ◀

6.7.3 Metodo della variabile ausiliaria

Il teorema fondamentale 6.1 per le trasformazioni del tipo $2 \to 2$ può servire anche per il caso visto nel § 6.7.1, nel quale ho una sola trasformazione $Z = g(X, Y)$. Per applicarlo, è sufficiente rendere la trasformazione $2 \to 1$ una trasformazione $2 \to 2$ ("quadrare" la trasformazione), ricorrendo all'artificio di introdurre una trasformazione fittizia o una *variabile ausiliaria* $W = h(X, Y)$ (tipicamente si sceglie $W = X$ oppure $W = Y$ per semplicità). Una volta ottenuta la $f_{ZW}(z, w)$ dall'applicazione del teorema fondamentale, è poi possibile *eliminare* la variabile ausiliaria, ricavando la pdf marginale $f_Z(z)$ per integrazione di $f_{ZW}(z, w)$ rispetto alla variabile w.

► *Esempio* 6.11. Consideriamo ad esempio la trasformazione $Z = XY$. In questo caso, scegliamo come variabile ausiliaria $W = X$, ottenendo così la seguente trasformazione $2 \to 2$:

$$\begin{cases} Z &= XY\,, \\ W &= X\,. \end{cases}$$

Il sistema di equazioni corrispondenti è il seguente:

$$\begin{cases} z &= xy\,, \\ w &= x\,; \end{cases}$$

e ammette, per ogni (z, w), con $w \neq 0$, una sola soluzione (x, y) data da

$$\begin{cases} x &= w\,, \\ y &= \frac{z}{w}\,; \end{cases}$$

mentre per $w = 0, z \neq 0$ il sistema non ha soluzioni, e infine per $w = 0, z = 0$, si ha $x = 0$ e y qualsiasi. Il calcolo della matrice jacobiana, inoltre, fornisce:

$$\boldsymbol{J}(x, y) = \frac{\partial(z, w)}{\partial(x, y)} = \begin{pmatrix} y & x \\ 1 & 0 \end{pmatrix}$$

[5] In pratica, è possibile utilizzare un unico generatore di numeri pseudo-casuali inizializzato con due semi differenti.

e quindi il determinante in valore assoluto vale

$$|\det[\boldsymbol{J}(x,y)]| = |x|\,,$$

per cui l'applicazione del teorema fondamentale fornisce per $w \neq 0$ la pdf congiunta di (Z, W):

$$f_{ZW}(z,w) = \frac{1}{|w|}\, f_{XY}\left(w, \frac{z}{w}\right)$$

da cui ricaviamo quella di $Z = XY$ integrando rispetto a w:

$$f_Z(z) = \int_{-\infty}^{\infty} \frac{1}{|w|}\, f_{XY}\left(w, \frac{z}{w}\right)\, dw\,.$$

Se, ad esempio, $X \sim \mathrm{U}(0,1)$ e $Y \sim \mathrm{U}(0,1)$, con X ed Y indipendenti e $Z = X\,Y$, la precedente si scrive:

$$f_Z(z) = \int_0^1 \frac{1}{|w|}\, f_X(w)\, f_Y\left(\frac{z}{w}\right)\, dw$$

ma $f_X(x) = 1$ per $x \in [0,1]$, e $f_Y(y) = 1$ per $y \in [0,1]$. Allora $f_Y(z/w) = 1$ se $z/w \in [0,1]$, ovvero se $0 \leq z \leq w$. Pertanto, l'intervallo di integrazione per w va da z a 1, e quindi l'integrale si scrive:

$$f_Z(z) = \int_z^1 \frac{1}{w}\, dw = [\ln w]_{w=z}^{w=1} = -\ln z\,.$$

La pdf di Z è pertanto:

$$f_Z(z) = \begin{cases} -\ln z\,, & \text{se } z \in [0,1]; \\ 0\,, & \text{altrove.} \end{cases}$$

◄

Notiamo che la scelta $W = X$ oppure $W = Y$ è opportuna in molti casi, ma non sempre, come illustrato chiaramente dall'esempio che segue.

► *Esempio* 6.12. Si consideri la trasformazione $Z = \sqrt{X^2 + Y^2}$. In questo caso, poiché possiamo interpretare Z come il modulo di una trasformazione di coordinate cartesiane a polari, conviene considerare, come variabile ausiliaria, una variabile W che rappresenti la fase della stessa trasformazione. Pertanto, sulla base anche dell'esempio 6.8, possiamo considerare la variabile ausiliaria W definita da

$$W = \tan^{-1}(Y/X)\,.$$

A questo punto, possiamo sfruttare i risultati già ottenuti nell'esempio 6.8, per scrivere direttamente la pdf congiunta di (Z, W):

$$f_{ZW}(z,w) = z\, f_{XY}(z\cos w, z\sin w)\, u(z)\,,$$

e successivamente eliminare la variabile ausiliaria W ricavando la pdf marginale di Z per integrazione:

$$f_Z(z) = u(z) \int_0^{2\pi} z\, f_{XY}(z\cos w, z\sin w)\, dw$$

Ad esempio, se $(X,Y) \sim \mathrm{N}(0,0,\sigma,\sigma,0)$, applicando anche i risultati dell'esempio 6.9, si trova banalmente che $Z \sim \text{Rayleigh}(2\sigma^2)$. ◄

6.8 Variabili aleatorie complesse★

È utile in taluni casi estendere la definizione di variabile aleatoria al caso complesso:

Definizione (variabile aleatoria complessa). una variabile aleatoria complessa Z è definita come

$$Z = X + jY\,,$$

con X, Y variabili aleatorie reali, e $j \triangleq \sqrt{-1}$.

Si osservi esplicitamente che nel caso complesso non ha senso indagare se $\{Z \leq z\}$ sia un evento, perchè il campo dei numeri complessi non è dotato di *ordinamento*. Pertanto, una variabile aleatoria complessa rappresenta solo un modo sintetico di denotare un coppia di variabili aleatorie reali. Poichè poi il piano complesso si identifica con $\mathbb{R}^2$, se D è una regione del piano complesso è possibile calcolare probabilità del tipo $P(Z \in D)$ semplicemente utilizzando la pdf congiunta della coppia di variabili aleatorie (X, Y), come

$$P(Z \in D) = \int\int_D f_{XY}(x,y)\, dx\, dy\,.$$

Possiamo definire formalmente la media di una variabile aleatoria complessa, applicando la proprietà di linearità. Infatti, se $Z = X + jY$ è una variabile aleatoria complessa, si ha:

$$E(Z) \triangleq E(X) + jE(Y)\,.$$

Come si vede, la media di Z si esprime attraverso la media di X (reale) e quella di Y (reale).

Il discorso si complica quando passiamo a definire il valore quadratico medio. In linea di principio, si potrebbe pensare che una definizione appropriata sia $E(Z^2)$, tuttavia se Z è complesso la quantità $E(Z^2)$ non è né reale né positiva. Pertanto una definizione appropriata di valor quadratico medio di una variabile aleatoria complessa Z è la seguente:

$$E(|Z|^2) = E(X^2 + Y^2) = E(X^2) + E(Y^2)\,.$$

Anche qui, il valor quadratico medio si ottiene combinando (sommando) i valori quadratici medi di X ed Y.

Infine, per la varianza una definizione appropriata è, in accordo a quella per il valor quadratico medio, la seguente:

$$\mathrm{Var}(Z) = E[|Z - E(Z)|^2] = E[(X - \mu_X)^2] + E[(Y - \mu_Y)^2] = \mathrm{Var}(X) + \mathrm{Var}(Y)\,,$$

e coincide con la somma delle varianze di X ed Y. Notiamo che vale anche in questo caso la relazione fondamentale tra varianza, valor quadratico medio e media, che si scrive:

$$\mathrm{Var}(X) = E(|Z|^2) - |E(Z)|^2$$

Naturalmente, è possibile definire un qualunque momento di Z, semplicemente sviluppando l'espressione risultante in termini di X ed Y. Ad esempio, per $E(Z^2)$ si ha:

$$E(Z^2) = E[(X + jY)^2] = E(X^2) - E(Y^2) + 2jE(XY)\,.$$

Come si vede, oltre ai valori quadratici medi di X ed Y, compare anche un *momento congiunto* $E(XY)$ (correlazione) di X ed Y, che introdurremo nel prossimo capitolo.

► *Esempio* 6.13. Data la variabile aleatoria $\Theta \sim \mathrm{U}(0, 2\pi)$, consideriamo come esempio di variabile aleatoria complessa la seguente:

$$Z = e^{j\Theta}\,.$$

Per l'identità di Eulero, si ha anche:

$$Z = \cos(\Theta) + j\sin(\Theta)\,,$$

per cui possiamo identificare la coppia (X, Y) come $X = \cos(\Theta)$ e $Y = \sin(\Theta)$. Il calcolo della media e della varianza di Z si conduce semplicemente applicando il teorema fondamentale della media. Infatti, si ha:

$$E(Z) = \int_0^{2\pi} e^{j\theta}\, \frac{1}{2\pi}\, d\theta = \frac{1}{2\pi j}\left[e^{j\theta}\right]_0^{2\pi} = 0\,,$$

per la periodicità (di periodo 2π) della funzione $e^{j\theta}$; inoltre, banalmente, si ha $E(|Z|^2) = E(1) = 1$, per cui $\mathrm{Var}(Z) = 1$. Si noti, invece, che sempre l'applicazione del teorema fondamentale della media ci consente di riconoscere che $E(Z^2) = 0$. Infatti:

$$E(Z^2) = \int_0^{2\pi} e^{j2\theta} \frac{1}{2\pi}\, d\theta = \frac{1}{4\pi j} \left[e^{j2\theta}\right]_0^{2\pi} = 0\,,$$

stavolta per la periodicità (di periodo π) della funzione $e^{j2\theta}$. ◄

6.9 Esercizi proposti

Esercizio 6.1. Si consideri l'esperimento del lancio di due dadi bilanciati, e si costruiscano due variabili aleatorie X ed Y nel seguente modo:

$$X \triangleq \text{somma dei risultati}$$
$$Y \triangleq \text{valore assoluto della differenza dei risultati}$$

Dopo aver individuato i possibili valori assunti da X ed Y, determinare la loro DF congiunta.

Esercizio 6.2. Una coppia di variabili aleatorie ha la seguente CDF:

$$F_{XY}(x,y) = \begin{cases} 0, & \text{se } x < 0 \text{ oppure } y < 0; \\ xy, & 0 \leq x \leq 1 \text{ e } 0 \leq y \leq 1; \\ x, & 0 \leq x \leq 1 \text{ e } y > 1; \\ y, & x > 1 \text{ e } 0 \leq y \leq 1; \\ 1, & \text{se } x > 1 \text{ e } y > 1. \end{cases}$$

Calcolare in termini della CDF congiunta le seguenti probabilità:

a) $P(X \leq 0.5, Y \leq 0.5)$;

b) $P(0.2 \leq X \leq 0.5, Y \leq 0.2)$;

c) $P(-0.5 \leq X \leq 0.5, -0.5 \leq Y \leq 0.5)$;

d) $P(X \geq 0.2, Y \geq 0.3)$;

e) $P(X \leq 0.2, Y \geq 0.4)$.

[*Risposta:* $\frac{1}{4}$; $\frac{3}{50}$; $\frac{1}{4}$; $\frac{14}{25}$; $\frac{3}{25}$.]

Esercizio 6.3. La pdf di una coppia di variabili aleatorie è definita da:

$$f_{XY}(x,y) = \begin{cases} 6xy^2, & \text{se } 0 < x < 1 \text{ e } 0 < y < 1; \\ 0, & \text{altrove.} \end{cases}$$

a) Verificare la condizione di normalizzazione;

b) calcolare $P(X + Y \geq 1)$;

c) calcolare $P(1/2 < X < 3/4)$.

Esercizio 6.4. Le variabili aleatorie (X, Y) sono uniformemente distribuite nel quadrato avente vertici nei punti $(1,1)$, $(1,-1)$, $(-1,1)$, $(-1,-1)$. Determinare la probabilità dei seguenti eventi:

a) $X^2 + Y^2 < 1$;

b) $2X - Y > 0$;

c) $|X + Y| < 2$.

Esercizio 6.5. La pdf di una coppia di variabili aleatorie è definita da:

$$f_{XY}(x,y) = \begin{cases} k(x+2y), & \text{se } 0 < x < 2 \text{ e } 0 < y < 1, \\ 0, & \text{altrove.} \end{cases}$$

a) Determinare il valore di k;

b) determinare le pdf marginali di X ed Y;

c) verificare se X ed Y sono indipendenti.

Esercizio 6.6. Si supponga che le variabili aleatorie X ed Y abbiano la seguente pdf:

$$f_{XY}(x,y) = \begin{cases} k, & \text{se } x^2 + y^2 \leq 1, \\ 0, & \text{altrimenti.} \end{cases}$$

a) Determinare il valore di k;

b) determinare le pdf marginali di X ed Y e stabilire se esse sono indipendenti.

Esercizio 6.7. Determinare $P(X > \sqrt{Y})$ se la pdf congiunta di X ed Y è $f_{XY}(x,y) = x + y$, $0 \leq x \leq 1$, $0 \leq y \leq 1$.

Esercizio 6.8. Determinare $P(X^2 < Y < X)$ se la pdf congiunta di X ed Y è $f_{XY}(x,y) = 2x$, $0 \leq x \leq 1$, $0 \leq y \leq 1$.

Esercizio 6.9. Date due variabili aleatorie con pdf congiunta $f_{XY}(x,y)$, ricavare la pdf di $Z = X + Y$, $Z = X - Y$, $Z = X/Y$, $Z = XY$.

Esercizio 6.10. Date due variabili aleatorie con pdf congiunta $f_{XY}(x,y)$, ricavare la pdf di $Z = \max(X,Y)$ e $Z = \min(X,Y)$. Particolarizzare il risultato ottenuto al caso in cui X ed Y sono indipendenti.

Esercizio 6.11. Determinare la pdf di $Z = X/Y$ dove X e Y sono variabili aleatorie indipendenti, ciascuna delle quali $N(0,\sigma)$.

Esercizio 6.12. Siano X ed Y due variabili aleatorie indipendenti, con distribuzione uniforme nell'intervallo $(0,1)$. Determinare la pdf della variabile aleatoria $Z = |X - Y|$.

Esercizio 6.13. Siano X ed Y due variabili aleatorie congiuntamente gaussiane, di parametri $\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho$. Provare che la somma $Z = X + Y$ è ancora una variabile aleatoria gaussiana, con media $\mu_X + \mu_Y$ e varianza $\sigma_X^2 + \sigma_Y^2 + 2\rho\,\sigma_X\,\sigma_Y$.

Esercizio 6.14. Sia $X \sim \text{Poiss}(\lambda)$ e $Y \sim \text{Poiss}(\mu)$, indipendenti. Provare che $Z = X + Y \sim \text{Poiss}(\lambda + \mu)$.

Esercizio 6.15. Siano U e V due variabili aleatorie gaussiane standard $N(0,1)$ ed indipendenti. Si consideri la trasformazione lineare:

$$\begin{cases} X = \rho\,\sigma_X\,U + \sigma_X\,\sqrt{1-\rho^2}\,V + \mu_X \\ Y = \sigma_Y\,U + \mu_Y \end{cases}$$

Verificare che $X, Y \sim N(\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho)$, ovvero X ed Y sono congiuntamente gaussiane con i parametri indicati.
Questo esercizio suggerisce un modo per generare coppie di variabili aleatorie congiuntamente gaussiane a partire da variabili aleatorie gaussiane standard indipendenti.

Esercizio 6.16. Siano X ed Y due variabili aleatorie indipendenti, con X avente CDF $F_X(x)$ e $Y \sim \text{U}(0,1)$. Mostrare che la pdf di $Z = X + Y$ è $f_Z(z) = F_X(z) - F_X(z-1)$.

Esercizio 6.17. Siano X ed Y due variabili aleatorie con pdf $f_{XY}(x,y)$. Determinare la pdf delle variabili aleatorie centrate $Z = X - \mu_X$ e $W = Y - \mu_Y$, e delle variabili aleatorie standardizzate $Z = (X - \mu_X)/\sigma_X$ e $W = (Y - \mu_Y)/\sigma_Y$.

★ *Esercizio* 6.18. Siano X ed Y due variabili aleatorie e sia $Z = \max(X,Y)$ e $W = \min(X,Y)$. Esprimere la CDF congiunta di Z e W in termini di quella di X ed Y.

Esercizio 6.19. Sia $X \sim \text{Exp}(\lambda)$ e $Y \sim \text{Exp}(\mu)$, con X ed Y indipendenti. Determinare le pdf delle seguenti variabili aleatorie:

a) $Z = 2X + Y$;

b) $Z = X - Y$;

c) $Z = X/Y$;

d) $Z = \max(X,Y)$;

e) $Z = \min(X,Y)$.

★ *Esercizio* 6.20. Siano $X \sim N(\mu_X, \sigma)$ e $Y \sim N(\mu_Y, \sigma)$, indipendenti, e si consideri la trasformazione di variabili aleatorie

$$\begin{cases} R = \sqrt{X^2 + Y^2} \\ \Theta = \tan^{-1} \frac{Y}{X} \end{cases}$$

Determinare la pdf di R.

[*Suggerimento: si faccia uso della funzione* $I_0(x) \triangleq \frac{1}{2\pi} \int_0^{2\pi} \exp(x \cos \alpha) d\alpha$*, funzione di Bessel modificata di prima specie ed ordine 0.*]

Capitolo 7

Caratterizzazione sintetica di una coppia di variabili aleatorie

Anche per una coppia di variabili aleatorie è possibile fornire alcuni parametri numerici (momenti) che ne forniscono una caratterizzazione sintetica. In questo capitolo, in particolare, dopo aver generalizzato il teorema fondamentale della media, si introducono i momenti congiunti di una coppia di variabili aleatorie, con particolare riferimento alla correlazione, alla covarianza, e al coefficiente di correlazione: tali quantità forniscono una misura della dipendenza lineare esistente tra due variabili aleatorie. All'interpretazione probabilistica di tali grandezze viene affiancata una interpretazione "geometrica", consistente nel riguardare le variabili aleatorie come vettori appartenenti ad un opportuno spazio vettoriale. Il problema della stima lineare, introdotto al termine del capitolo e risolto mediante l'applicazione del principio di ortogonalità, mostra i benefici derivanti dall'interpretazione geometrica.

7.1 Introduzione

Nel capitolo 5, abbiamo introdotto i *momenti* di una singola variabile aleatoria X, tra i quali la media, la varianza ed il valor quadratico medio sono sicuramente i più utilizzati. Abbiamo visto che attraverso tali momenti è possibile fornire una caratterizzazione *sintetica* della variabile aleatoria X, che non si basa cioè sulla conoscenza della sua CDF e pdf. In questo capitolo, vogliamo estendere la definizione dei momenti al caso di coppie (X, Y) di variabili aleatorie, così da poter fornire una caratterizzazione sintetica anche in questo caso: i momenti associati ad una coppia di variabili aleatorie prendono il nome di *momenti congiunti*. Osserviamo peraltro che nel caso di coppie di variabili aleatorie la caratterizzazione sintetica appare ancora più interessante rispetto al caso di una singola variabile aleatoria, vista la difficoltà di manipolare, e talvolta di interpretare, le funzioni (di due variabili) che forniscono la caratterizzazione completa, quali la CDF, la pdf e la DF congiunta.

7.2 Teorema fondamentale della media per una coppia di variabili aleatorie

Il primo e fondamentale passo da seguire per definire i momenti congiunti per coppie di variabili aleatorie è quello di estendere al caso di una coppia di variabili aleatorie il teorema fondamentale della media (teorema 5.1), che abbiamo introdotto nel capitolo 5 per una singola variabile aleatoria.

Iniziamo col considerare la trasformazione (cosiddetta $2 \rightarrow 1$, cfr. § 6.7.1) mediante la quale a partire da una coppia (X, Y) di variabili aleatorie si ottiene una nuova variabile aleatoria $Z = g(X, Y)$. Nel precedente capitolo, abbiamo studiato vari metodi per determinare CDF e pdf di Z, conoscendo la CDF o la pdf congiunta della coppia (X, Y). Una volta determinata la pdf di Z, in particolare, siamo in grado di calcolare la media di Z, utilizzando la definizione di media per una singola variabile aleatoria:

$$E(Z) = \int_{-\infty}^{\infty} z\, f_Z(z)\, dz\,.$$

Tuttavia, non è necessario conoscere la pdf di Z per calcolarne la media, poiché è sufficiente la conoscenza della pdf congiunta di (X, Y), come affermato dal seguente teorema, il quale estende il teorema fondamentale della media al caso di coppie di variabili aleatorie, e che enunciamo senza dimostrazione:

Teorema 7.1 (teorema fondamentale della media per coppie di variabili aleatorie). Sia $Z = g(X, Y)$ una trasformazione della coppia di variabili aleatorie (X, Y) aventi pdf congiunta $f_{XY}(x, y)$; si ha:

$$E(Z) = E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y)\, f_{XY}(x, y)\, dx\, dy\,,$$

se tale integrale esiste finito.

Nel caso in cui (X, Y) siano variabili aleatorie discrete, osserviamo che anche $Z = g(X, Y)$ è una variabile aleatoria discreta, ed il teorema precedente si esprime in termini della *DF congiunta* $p_{XY}(x, y)$ di (X, Y) come:

$$E(Z) = E[g(X, Y)] = \sum_{x \in \mathcal{X}, y \in \mathcal{Y}} g(x, y)\, p_{XY}(x, y)\,.$$

In questo caso, è anche semplice delineare la dimostrazione del teorema fondamentale (si veda [4] oppure [5]).

È immediato verificare che l'operatore di media gode sempre della proprietà di *linearità*, anche se in questo caso il risultato è più generale, visto che coinvolge coppie di variabili aleatorie. Infatti, siano g_1 e g_2 arbitrarie funzioni di due variabili, e siano a_1 e a_2 costanti reali; si ha:

$$E[a_1\, g_1(X, Y) + a_2\, g_2(X, Y)] = a_1\, E[g_1(X, Y)] + a_2\, E[g_2(X, Y)]\,.$$

La dimostrazione di questo risultato è banale, basandosi direttamente sul teorema fondamentale della media precedentemente enunciato, e si lascia al lettore per esercizio. Se, in particolare, si sceglie $g_1(X, Y) = X$ e $g_2(X, Y) = Y$, si ha:

$$E(a_1\, X + a_2\, Y) = a_1\, E(X) + a_2\, E(Y)\,,$$

e pertanto la media della combinazione lineare di due variabili aleatorie coincide con la combinazione lineare delle medie. Notiamo esplicitamente che tale proprietà di linearità vale in generale, sia che X ed Y siano *indipendenti*, sia che non lo siano.

7.3 Momenti congiunti di una coppia di variabili aleatorie

Sulla base del teorema fondamentale della media, possiamo definire i *momenti congiunti* della coppia di variabili aleatorie (X, Y):

Definizione (momento congiunto di ordine $n = k + r$). Il momento congiunto (di ordine $n = k + r$) di una coppia di variabili aleatorie (X, Y) è:

$$\mu_{kr} \triangleq E(X^k\, Y^r) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x^k y^r f_{XY}(x,y)\, dx\, dy\,,$$

se l'integrale esiste finito.

Definizione (momento congiunto centrale di ordine $n = k + r$). Il momento congiunto centrale (di ordine $n = k + r$) di una coppia di variabili aleatorie (X, Y), con medie $\mu_X = E(X)$ e $\mu_Y = E(Y)$, è:

$$\sigma_{kr} \triangleq E[(X - \mu_X)^k (Y - \mu_Y)^r] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_X)^k (y - \mu_Y)^r\, f_{XY}(x,y)\, dx\, dy\,,$$

se l'integrale esiste finito.

Osserviamo che, ponendo $k = 0$ oppure $r = 0$, e sfruttando la relazione tra statistiche congiunte e marginali, è possibile ritrovare i momenti e i momenti centrali delle singole variabili aleatorie X ed Y. Ad esempio, per $k = 1$ e $r = 0$, si verifica facilmente che il momento congiunto μ_{10} coincide con la media di X, in quanto:

$$\begin{aligned}\mu_{10} &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{XY}(x,y)\, dx\, dy = \int_{-\infty}^{\infty} x \left[\int_{-\infty}^{\infty} f_{XY}(x,y)\, dy \right] dx = \\ &= \int_{-\infty}^{\infty} x\, f_X(x)\, dx = E(X)\end{aligned}$$

poiché $\int_{-\infty}^{\infty} f_{XY}(x,y)\, dy = f_X(x)$. Similmente si trova $\mu_{20} = E(X^2)$ (valor quadratico medio) e $\sigma_{20} = E[(X - \mu_X)^2)]$ (varianza), e analogamente per i corrispondenti momenti di Y. Ponendo poi $k = r = 0$ nella definizione di momenti, si ottiene la proprietà di normalizzazione della pdf congiunta, ovvero $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x,y)\, dx\, dy = 1$.

Come abbiamo accennato, la conoscenza di un sottoinsieme dei momenti congiunti va sotto il nome di *caratterizzazione sintetica* della coppia di variabili aleatorie (X, Y). La caratterizzazione completa consiste invece nella conoscenza della CDF, pdf o DF congiunta. Ovviamente, se si dispone della caratterizzazione completa, è possibile ricavare un qualunque momento congiunto; viceversa, se si conosce un sottoinsieme dei momenti congiunti, non è possibile in generale ricavare CDF, pdf o DF congiunta.[1]

[1] Anche qui, come accade per il caso di una singola variabile aleatoria, il discorso è diverso se si suppone di conoscere *tutti* i momenti congiunti; in tal caso, sotto opportune ipotesi, è possibile risalire alla CDF, pdf o DF congiunta attraverso l'uso della *funzione caratteristica congiunta* (si veda ad esempio [3, § 7-2])

7.4 Misure di correlazione di una coppia di variabili aleatorie

Tra i momenti congiunti di una coppia di variabili aleatorie (X, Y), quelli più utilizzati sono i momenti del secondo ordine ($n = 2$), che vanno sotto il nome di *correlazione* e *covarianza*.

7.4.1 Correlazione

Definizione (correlazione). La correlazione di una coppia di variabili aleatorie (X, Y) è il momento congiunto μ_{11} di ordine $n = 2$, ovvero:

$$\mathrm{Corr}(X, Y) \triangleq \mu_{11} = E(X\,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\,y\,f_{XY}(x,y)\,dx\,dy\,,$$

se l'integrale esiste finito.

Per fornire una prima interpretazione della correlazione, osserviamo che essa compare naturalmente se calcoliamo il valore quadratico medio della somma di due variabili aleatorie X ed Y:

$$E[(X+Y)^2] = E(X^2) + E(Y^2) + 2\,E(X\,Y)\,.$$

Poiché la correlazione può essere positiva, negativa o nulla, allora il valor quadratico medio della somma di due variabili aleatorie può essere maggiore, minore o uguale alla somma dei valori quadratici medi.

Una seconda interpretazione della correlazione è di tipo geometrico, e precisamente come *prodotto scalare* tra X ed Y; pertanto, essa richiede l'introduzione del fondamentale concetto di *spazio vettoriale* di variabili aleatorie, che sarà sviluppato nella successiva sezione.

7.4.2 Spazio vettoriale di variabili aleatorie

L'idea è quella di interpretare le variabili aleatorie come vettori appartenenti ad un opportuno spazio vettoriale. Osserviamo preliminarmente che, affinché si possa parlare legittimamente di vettori, è necessario che siano definite ed abbiano senso l'operazione di *somma* di due vettori e l'operazione di *prodotto* di un vettore per uno scalare. Ma tali operazioni corrispondono alla somma $X + Y$ di due variabili aleatorie ed al prodotto $a\,X$ di una variabile aleatoria per una costante reale, per cui sono perfettamente definite.

Una volta assimilate le variabili aleatorie a vettori, è possibile introdurre una serie di concetti *geometrici* di grande importanza. In particolare, sui vettori appartenenti a questo spazio vettoriale, è possibile definire, con diretta interpretazione geometrica,

- una *norma* $\|X\| \triangleq \sqrt{E(X^2)}$;
- una *distanza* $d(X, Y) \triangleq \|X - Y\| = \sqrt{E[(X-Y)^2]}$;
- un *prodotto scalare* $\langle X, Y\rangle \triangleq E(XY)$.

Tali definizioni non sono arbitrarie; in effetti si può far vedere che la norma, la distanza e il prodotto scalare così definiti soddisfano le proprietà caratteristiche di tali operazioni. Osserviamo, in particolare, che la norma coincide con il valore efficace (e quindi la norma al quadrato coincide con il valore quadratico medio $E(X^2)$), mentre il prodotto scalare coincide proprio con la *correlazione* tra le variabili aleatorie X ed Y.

Di particolare importanza, in uno spazio vettoriale dotato di prodotto scalare, risulta la seguente *disuguaglianza di Schwartz*:

Teorema 7.2 (disuguaglianza di Schwartz). In uno spazio vettoriale di variabili aleatorie dotato di prodotto scalare, vale la seguente disuguaglianza:

$$|E(XY)| \leq \sqrt{E(X^2)}\sqrt{E(Y^2)}\,,$$

con uguaglianza se e solo se $Y = aX$ (in media quadratica).

Prova. È sufficiente considerare l'espressione quadratica in a, non negativa,

$$E[(aX - Y)^2] \geq 0\,,$$

che si sviluppa facilmente, utilizzando la linearità della media:

$$a^2E(X^2) - 2\,a\,E(XY) + E(Y^2) \geq 0\,.$$

Poiché tale forma quadratica è non negativa, allora il suo discriminante Δ è minore o uguale a 0. Il calcolo del discriminante fornisce:

$$\Delta = 4\,E(XY)^2 - 4\,E(X^2)\,E(Y^2) \leq 0\,,$$

per cui si ha $E(XY)^2 \leq E(X^2)\,E(Y^2)$ e quindi, prendendo la radice quadrata, l'asserto. Osserviamo che, se il discriminante si annulla, allora esiste un valore di a, sia esso a^*, tale che

$$E[(a^*X - Y)^2] = 0\,.$$

Questa condizione è equivalente a dire che $Y = a^*\,X$ in media quadratica. Ovviamente se l'uguaglianza $Y = a^*\,X$ vale puntualmente (ovvero si ha $Y(\omega) = a^*\,X(\omega), \forall\omega \in \Omega$), essa vale a maggior ragione anche in media quadratica, ma il viceversa non è vero. □

La disuguaglianza di Schwartz afferma che, in valore assoluto, la correlazione non può eccedere il prodotto dei valori efficaci delle due variabili aleatorie X ed Y. Inoltra, essa consente anche di riesprimere il prodotto scalare, e quindi la correlazione, come:

$$\langle X, Y\rangle = \|X\|\|Y\|\cos(\theta) \tag{7.1}$$

dove θ è l'angolo (compreso tra 0 e 2π) formato dai due vettori.[2] Si trova allora che tale prodotto scalare è massimo (in modulo) quando $\theta = 0$ (vettori allineati e nello stesso verso) oppure quando $\theta = \pi$ (vettori allineati ma di verso opposto). Viceversa, il prodotto scalare è nullo, e quindi minimo in modulo, quando $\cos(\theta) = 0$, ovvero per $\theta = \pi/2$ o $3\pi/2$; in questo caso i vettori X ed Y sono *ortogonali*. Possiamo allora fornire la seguente definizione di *ortogonalità* tra due variabili aleatorie X ed Y:

Definizione (ortogonalità). Due variabili aleatorie X ed Y si dicono ortogonali ($X \perp Y$) se e solo se:

$$E(XY) = 0\,,$$

ovvero se la loro correlazione è nulla.

[2] In realtà la (7.1) consente di definire l'angolo θ tra due vettori sulla base del prodotto scalare, anche nei casi in cui l'interpretazione geometrica non è direttamente applicabile.

7.4.3 Covarianza

Definizione (covarianza). La covarianza di una coppia di variabili aleatorie (X, Y) è il momento congiunto centrale σ_{11} di ordine $n = 2$, ovvero:

$$\begin{aligned}\mathrm{Cov}(X,Y) \triangleq \sigma_{11} &= E[(X-\mu_X)(Y-\mu_Y)] = \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}(x-\mu_X)\,(y-\mu_Y)\,f_{XY}(x,y)\,dx\,dy\,,\end{aligned}$$

se l'integrale esiste finito.

Esiste una relazione semplice tra correlazione e covarianza: sviluppando infatti la media che compare nella definizione di covarianza, si trova banalmente:

$$\boxed{\mathrm{Cov}(X,Y) = E(X\,Y) - E(X)\,E(Y) = \mathrm{Corr}(X,Y) - \mu_X\,\mu_Y\,.} \tag{7.2}$$

Una prima interpretazione della covarianza è che essa compare naturalmente se proviamo a calcolare la varianza della *somma* di due variabili aleatorie X ed Y. Infatti si ha, con semplici passaggi

$$\boxed{\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\,\mathrm{Cov}(X,Y)\,.} \tag{7.3}$$

Osserviamo che, poichè la covarianza può essere positiva, negativa o nulla, allora la varianza della somma di due variabili aleatorie può essere maggiore, minore o uguale alla somma delle varianze.

Una seconda interpretazione della covarianza è di tipo geometrico, e precisamente in termini di prodotto scalare; infatti essa rappresenta il prodotto scalare tra le variabili aleatorie *centrate* $X - \mu_X$ e $Y - \mu_Y$. Essendo la covarianza un prodotto scalare, la disuguaglianza di Schwartz si può applicare anche ad essa, ed assume la seguente forma:

$$|\mathrm{Cov}(X,Y)| \leq \sqrt{E[(X-\mu_X)^2]}\sqrt{E[(Y-\mu_Y)^2]} = \sigma_X\,\sigma_Y \tag{7.4}$$

con uguaglianza se e solo se $Y - \mu_Y = a(X - \mu_X)$ (in media quadratica o quasi ovunque), e quindi X ed Y sono legati da una dipendenza lineare *esatta*, del tipo $Y = a\,X + b$, con $b = \mu_Y - a\,\mu_X$. La (7.4) afferma che, in valore assoluto, la covarianza non può eccedere il prodotto delle deviazioni standard delle due variabili aleatorie X ed Y. Inoltre, essa consente di affermare che la covarianza misura la dipendenza *lineare* tra due variabili aleatorie, in quanto è massima in modulo se le due variabili aleatorie sono legate da una relazione lineare.

Osserviamo che, se a scostamenti (rispetto alla media) $X - \mu_X$ positivi corrispondono in media scostamenti $Y - \mu_Y$ positivi, e analogamente per scostamenti negativi, la covarianza sarà positiva, e le variabili si diranno *positivamente correlate*; ciò accade se ad esempio si considera altezza e peso di una persona. In altri termini, ad un incremento di X (l'altezza) corrisponde un incremento di Y (il peso). Viceversa, se a scostamenti positivi di una variabile corrispondono scostamenti negativi dell'altra, la covarianza sarà negativa, e le variabili si diranno *negativamente correlate*; ciò accade ad esempio se si considerano il numero di sigarette fumate giornalmente e la speranza di vita di una persona. In questo caso, ad un incremento di X (il numero di sigarette) corrisponde un decremento di Y (la speranza di vita).

7.4.4 Coefficiente di correlazione

La covarianza è una misura *assoluta* di dipendenza lineare: per avere una misura relativa, è sufficiente normalizzarla al suo valore massimo (in modulo) $\sigma_X \sigma_Y$, ottenendo così il *coefficiente di correlazione*:

Definizione (coefficiente di correlazione). Il coefficiente di correlazione ρ_{XY} di una coppia di variabili aleatorie (X, Y) è:

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

Sulla base dell'interpretazione come covarianza normalizzata, ed in particolare richiamando la (7.4), si osserva che il coefficiente di correlazione gode della seguente proprietà:

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

con uguaglianza se e solo se X ed Y presentano una dipendenza di tipo lineare esatta, ovvero $Y = a\,X + b$ (in media quadratica o quasi ovunque).

► *Esempio* 7.1. Mostriamo che il parametro ρ che compare nella pdf congiunta di due variabili aleatorie congiuntamente gaussiane X ed Y è proprio il coefficiente di correlazione ρ_{XY}. A tal scopo, consideriamo prima il calcolo della covarianza $\mathrm{Cov}(X, Y)$: si ha:

$$\mathrm{Cov}(X,Y) = E[(X-\mu_X)(Y-\mu_Y)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} (x-\mu_X)(y-\mu_Y)\, f_{XY}(x,y)\, dx\, dy .$$

Per risolvere tale integrale, conviene decomporre la $f_{XY}(x, y)$ come già fatto nell'esempio 6.2, e precisamente come

$$f_{XY}(x,y) = \left[\frac{1}{\sigma_X\sqrt{2\pi}}\, e^{-\frac{1}{2\sigma_X^2}(x-\mu_X)^2}\right]\left[\frac{1}{\sigma_Y\sqrt{1-\rho^2}\sqrt{2\pi}}\, e^{-\frac{1}{2\sigma_Y^2(1-\rho^2)}\left[y-\mu_Y-\rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)\right]^2}\right] .$$

Sostituendo nell'espressione della covarianza, si ottiene:

$$\begin{aligned}\mathrm{Cov}(X,Y) = &\int_{-\infty}^{\infty}(x-\mu_X)\,\frac{1}{\sigma_X\sqrt{2\pi}}\, e^{-\frac{1}{2\sigma_X^2}(x-\mu_X)^2} \\ &\times\left[\int_{-\infty}^{\infty}(y-\mu_Y)\,\frac{1}{\sigma_Y\sqrt{1-\rho^2}\sqrt{2\pi}}\, e^{-\frac{1}{2\sigma_Y^2(1-\rho^2)}\left[y-\mu_Y-\rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)\right]^2} dy\right] dx\end{aligned} \tag{7.5}$$

Concentriamo l'attenzione sull'integrale in dy: poiché la pdf che vi compare è ancora gaussiana, a media $\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)$, aggiungendo e sottraendo $\rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)$ nel termine $(y-\mu_Y)$ avremo che l'integrale si può decomporre come:

$$\begin{aligned}&\int_{-\infty}^{\infty}\left[y-\mu_Y-\rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)\right]\,\frac{1}{\sigma_Y\sqrt{1-\rho^2}\sqrt{2\pi}}\, e^{-\frac{1}{2\sigma_Y^2(1-\rho^2)}\left[y-\mu_Y-\rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)\right]^2} dy + \\ &+\rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)\int_{-\infty}^{\infty}\frac{1}{\sigma_Y\sqrt{1-\rho^2}\sqrt{2\pi}}\, e^{-\frac{1}{2\sigma_Y^2(1-\rho^2)}\left[y-\mu_Y-\rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)\right]^2} dy .\end{aligned}$$

Dei due integrali risultanti, il primo è nullo per la definizione di media, mentre il secondo è unitario, per la condizione di normalizzazione della pdf. In definitiva, il risultato del calcolo è semplicemente $\rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)$ che, sostituito nella (7.5), fornisce:

$$\begin{aligned}\mathrm{Cov}(X,Y) &= \int_{-\infty}^{\infty}\rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)^2\,\frac{1}{\sigma_X\sqrt{2\pi}}\, e^{-\frac{1}{2\sigma_X^2}(x-\mu_X)^2} = \\ &= \rho\frac{\sigma_Y}{\sigma_X}\int_{-\infty}^{\infty}(x-\mu_X)^2\,\frac{1}{\sigma_X\sqrt{2\pi}}\, e^{-\frac{1}{2\sigma_X^2}(x-\mu_X)^2} = \\ &= \rho\frac{\sigma_Y}{\sigma_X}\,\sigma_X^2 = \rho\,\sigma_X\sigma_Y ,\end{aligned}$$

per cui si ricava, dividendo la covarianza per $\sigma_X\,\sigma_Y$, che $\rho = \rho_{XY}$, come si voleva provare. ◄

7.4.5 Incorrelazione tra due variabili aleatorie

Abbiamo visto che, come la covarianza, anche il coefficiente di correlazione misura la dipendenza *lineare* esistente tra le due variabili aleatorie. Il vantaggio è che esso, essendo normalizzato, è più facilmente interpretabile: tale relazione lineare è tanto più forte quanto più ρ_{XY} in modulo è prossimo ad uno. La completa assenza di dipendenza lineare, viceversa, si ha se $\rho_{XY} = 0$, il che ovviamente equivale anche a $\mathrm{Cov}(X, Y) = 0$; tale condizione va sotto il nome di *incorrelazione*:

Definizione (incorrelazione). Due variabili aleatorie X ed Y si dicono incorrelate se $\mathrm{Cov}(X, Y) = 0$ o, equivalentemente, se $\rho_{XY} = 0$.

Notiamo che, per la relazione (7.2) esistente tra covarianza e correlazione, la condizione di incorrelazione si può esprimere equivalentemente come:

$$E(X\,Y) = E(X)\,E(Y)\,,$$

che si interpreta come una proprietà di *fattorizzazione* della correlazione (la media del prodotto $X\,Y$ è uguale al prodotto delle medie di X ed Y). Questa proprietà va messa in relazione con quella di indipendenza, che rappresenta invece una proprietà di fattorizzazione per la pdf congiunta: è facile verificare che vale il seguente teorema:

Teorema 7.3 (relazione tra incorrelazione e indipendenza). Se X ed Y sono due variabili aleatorie indipendenti, allora esse sono anche incorrelate.

Prova. Se scriviamo $E(X\,Y)$ esplicitamente, si ha:

$$E(X\,Y) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x\,y\,f_{XY}(x,y)\,dx\,dy\,,$$

per cui, se le variabili aleatorie sono indipendenti, risulta $f_{XY}(x,y) = f_X(x)\,f_Y(y)$, e quindi:

$$E(X\,Y) = \left[\int_{-\infty}^{\infty} x\,f_X(x)\,dx\right]\left[\int_{-\infty}^{\infty} y\,f_X(y)\,dy\right] = E(X)\,E(Y)\,,$$

pertanto resta dimostrato l'asserto. □

È altrettanto ovvio che, viceversa, l'incorrelazione *non* implica l'indipendenza: infatti se si fattorizzano le medie (gli integrali), non è detto che si fattorizzino le pdf (le funzioni integrande). Una eccezione degna di nota è il caso delle variabili aleatorie congiuntamente gaussiane, come mostrato dall'esempio che segue.

► *Esempio* 7.2. Siamo $(X, Y) \sim \mathrm{N}(\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho)$ due variabili aleatorie congiuntamente gaussiane. Supponiamo che (X, Y) siano incorrelate, il che equivale alla condizione $\rho = 0$, poichè tale parametro è il coefficiente di correlazione. Se allora si considera l'espressione della pdf bidimensionale gaussiana (6.5) per $\rho = 0$, si vede che essa si fattorizza nel prodotto delle pdf marginali di due variabili aleatorie $X \sim \mathrm{N}(\mu_X, \sigma_X)$ e $Y \sim \mathrm{N}(\mu_Y, \sigma_Y)$, per cui X ed Y sono indipendenti. ◄

Notiamo che l'incorrelazione tra X ed Y garantisce solo la fattorizzazione della media del prodotto $X\,Y$; viceversa, l'indipendenza tra X ed Y, essendo più forte dell'incorrelazione, garantisce

la fattorizzazione della media di un qualunque prodotto del tipo $g(X)\,h(Y)$; per dimostrarlo formalmente, osserviamo che se X ed Y sono indipendenti, anche le variabili aleatorie $Z = g(X)$ e $W = h(Y)$ sono indipendenti (cfr. § 6.6.1 proprietà 2) e quindi incorrelate, e si ha allora

$$E[g(X)\,h(Y)] = E[g(X)]\,E[h(Y)]\,.$$

Una conseguenza importante di tale proprietà è il fatto che, per variabili aleatorie indipendenti, qualunque momento congiunto (centrale oppure no) si fattorizza nel prodotto dei corrispondenti momenti marginali; ad esempio, si ha:

$$\mu_{kr} = E(X^k\,Y^r) = E(X^k)\,E(Y^r) = \mu_k\,\mu_r\,.$$

Un'altra proprietà interessante delle variabili aleatorie incorrelate è che, per esse, risulta

$$\boxed{\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)\,,}$$

ovvero la varianza della somma è pari alla somma delle varianze. Tale risultato si ottiene banalmente dalla (7.3) ponendovi $\mathrm{Cov}(X,Y) = 0$, e ovviamente vale a maggior ragione se X ed Y sono indipendenti.

Notiamo infine che, nonostante una terminologia poco felice, non bisogna confondere i concetti di ortogonalità ed incorrelazione: l'ortogonalità corrisponde all'annullarsi della *correlazione*, l'incorrelazione corrisponde all'annullarsi della *covarianza* o del *coefficiente di correlazione*. Stante la relazione (7.2), ortogonalità e incorrelazione coincidono se almeno una delle due variabili aleatorie è a media nulla. Inoltre, dire che X ed Y sono incorrelate equivale a dire che le variabili centrate $X - \mu_X$ e $Y - \mu_Y$ sono ortogonali.

7.5 Stima lineare a minimo errore quadratico medio★

Un'applicazione interessante dei concetti introdotti nelle precedenti sezioni, basata in particolare sull'interpretazione delle variabili aleatorie come vettori, è il problema della *stima*. Possiamo formalizzarlo come segue: abbiamo a disposizione una variabile aleatoria X, e a partire da un valore x assunto da X, vogliamo ottenere la stima di una seconda variabile aleatoria Y, collegata ad X da qualche relazione.

► *Esempio* 7.3. Sia Y l'altezza ed X il peso di una persona; sappiamo che il peso è $X = x$ e vogliamo stimare l'altezza Y (tale problema è significativo per esempio in ambito pediatrico, per controllare la crescita dei neonati). ◄

Chiameremo *stimatore* di Y, a partire dai dati X, una funzione $\widehat{Y} = g(X)$, dove $\widehat{Y}$ rappresenta la stima di Y. Notiamo che uno stimatore non è altro che una trasformazione della variabile aleatoria X: sebbene la forma funzionale di g possa essere arbitraria, ci limiteremo a considerare *stimatori lineari*, per i quali g è una funzione lineare dei dati X, e si ha quindi:

$$\widehat{Y} = a\,X + b\,, \tag{7.6}$$

con a, b parametri reali.

Per determinare l'espressione esplicita di uno stimatore (vale a dire la forma della funzione g, o per uno stimatore lineare i coefficienti a e b) occorre introdurre una *misura di qualità* dello

stimatore stesso, per determinare quanto "buona" sia la stima $\widehat{Y}$. Una misura molto utilizzata, per la sua semplicità matematica, è l'errore quadratico medio (*mean square error*, MSE) di stima:

$$\mathrm{MSE}(Y, \widehat{Y}) \triangleq E[(Y - \widehat{Y})^2] \,. \tag{7.7}$$

Il criterio di stima *a minimo errore quadratico medio* (*minimum mean square error*, MMSE) consiste nel determinare lo stimatore g che minimizza l'errore quadratico medio; nel caso di stima lineare, si tratta semplicemente di determinare i parametri a e b che minimizzano l'errore quadratico medio. Si ha:

$$a = \rho_{XY} \frac{\sigma_Y}{\sigma_X} \,, \tag{7.8}$$

$$b = \mu_Y - \mu_X \rho_{XY} \frac{\sigma_Y}{\sigma_X} \,. \tag{7.9}$$

Prova. Si consideri l'errore quadratico medio (7.7): sostituendo l'espressione dello stimatore data dalla (7.6) nella (7.7) si trova:

$$\mathrm{MSE} = E[(Y - aX - b)^2] \,.$$

Per determinare i valori di a e b che rendono minimo l'MSE, si calcolano le derivate parziali dell'MSE rispetto ad a e b e si eguagliano a zero (derivando sotto il segno di media):

$$\frac{\partial}{\partial a} \mathrm{MSE} = 2E[(Y - aX - b)X] = 0 \,,$$
$$\frac{\partial}{\partial b} \mathrm{MSE} = 2\,E[(Y - aX - b)] = 0 \,,$$

Sviluppando le medie, si ottiene un sistema di due equazioni nelle incognite a e b:

$$\begin{cases} a\,E(X^2) + b\,E(X) = E(XY) \\ a\,E(X) + b = E(Y) \end{cases}$$

che risolto rispetto ad a e b fornisce i risultati (7.8) e (7.9). □

Osserviamo che la media dello stimatore ottimo vale

$$E(\widehat{Y}) = a\,E(X) + b = \rho_{XY} \frac{\sigma_Y}{\sigma_X} \mu_X + \mu_Y - \mu_X \rho_{XY} \frac{\sigma_Y}{\sigma_X} = \mu_Y$$

cioè è uguale alla media della variabile aleatoria Y da stimare. Uno stimatore che soddisfa una tale proprietà non commette un errore sistematico di stima, e si dice quindi *non polarizzato* (in inglese, "unbiased").

Passiamo ora a sostituire i valori di a e b appena determinati nella (7.7), per trovare il valore dell'errore quadratico medio minimo; con facili passaggi algebrici, si ottiene

$$\mathrm{MSE}_{\min} = \sigma_Y^2 [1 - \rho_{XY}^2] \,,$$

dove ρ_{XY} è il coefficiente di correlazione tra le variabili aleatorie X ed Y. Poiché $|\rho_{XY}| \leq 1$, notiamo che l'errore minimo risulta maggiore o uguale a zero, come è naturale; inoltre esso è minore o uguale alla varianza di Y; in particolare, possiamo considerare i due casi limite:

1. se $\rho_{XY} = 0$, ovvero se le variabili aleatorie X ed Y sono *incorrelate*, risulta $a = 0$ e $b = \mu_Y$ nelle (7.8) ed (7.9), per cui lo stimatore diventa $\widehat{Y} = \mu_Y$ e l'errore minimo è σ_Y^2. In questo caso, la migliore stima lineare di Y è indipendente dai dati X e coincide con la sua media μ_Y, mentre l'errore quadratico medio coincide con la varianza di Y. È chiaro che in questo caso X non fornisce alcuna indicazione utile per determinare Y;

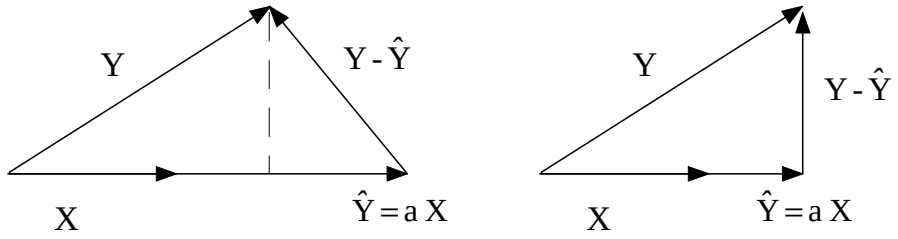


Fig. 7.1. Principio di ortogonalità: l'errore quadratico medio (MSE) rappresenta la norma del vettore $Y - \widehat{Y}$ (a sinistra); al variare di a, il valore minimo dell'MSE si ottiene quando il vettore $Y - \widehat{Y}$ è ortogonale ad X (a destra).

2. se $\rho_{XY} = \pm 1$, ovvero se le variabili aleatorie X ed Y sono legate da una dipendenza lineare esatta, allora l'errore quadratico medio minimo è pari a zero. In questo caso, uno stimatore lineare si adatta perfettamente alla dipendenza lineare posseduta dalle variabili aleatorie, per cui fornisce risultati assai soddisfacenti.

7.5.1 Principio di ortogonalità

Una formulazione geometrica interessante del problema della stima lineare MMSE si ottiene supponendo che X ed Y siano a media nulla, ovvero $\mu_X = \mu_Y = 0$. In tal caso, per avere uno stimatore non polarizzato, è necessario porre $b = 0$ (si noti che tale risultato discende anche dalla (7.9) per $\mu_X = \mu_Y = 0$) e scegliere quindi uno stimatore lineare *omogeneo*:

$$\widehat{Y} = a\,X\,,$$

in quanto solo in tal caso risulta $E(\widehat{Y}) = a\,E(X) = 0$, qualunque sia il parametro a. Quest'ultimo va determinato in modo da minimizzare l'errore quadratico medio (7.7).

In effetti, tale problema è un caso particolare del precedente, con $\mu_X = \mu_Y = 0$, per cui il valore di a è dato ancora dalla (7.8),

$$a = \rho_{XY}\,\frac{\sigma_Y}{\sigma_X}\,.$$

Vogliamo tuttavia reinterpretare tale problema da un punto di vista differente, ed in particolare vogliamo fornirne una interessante interpretazione geometrica. Osserviamo che, se X ed Y sono due vettori, lo stimatore $\widehat{Y} = a\,X$, al variare di a, sarà un vettore proporzionale (Fig. 7.1) ad X, cioè allineato ad esso. L'errore quadratico medio $E[(Y - \widehat{Y})^2]$ rappresenta allora la norma al quadrato del vettore *differenza* $Y - \widehat{Y}$. Con semplici considerazioni geometriche, è facile convincersi che, al variare di a, tale norma è *minima* quando il vettore $Y - \widehat{Y}$ è *ortogonale* ad X. D'altra parte, se consideriamo l'espressione esplicita di tale norma (ovvero dell'errore quadratico medio):

$$\mathrm{MSE}(Y, \widehat{Y}) = E[(Y - \widehat{Y})^2] = E[(Y - aX)^2]$$

e deriviamo rispetto ad a, annullando tale derivata si ottiene:

$$E[(Y - aX)X] = E[(Y - \widehat{Y})X] = 0\,,$$

che rappresenta proprio la condizione di annullamento del prodotto scalare tra i vettori $Y - \widehat{Y}$ e X, ovvero la condizione di ortogonalità precedentemente menzionata. Tale risultato, in sintesi,

afferma che la stima lineare MMSE è quella che rende l'errore di stima $Y - \widehat{Y}$ ortogonale ai dati X, e va sotto il nome di *principio di ortogonalità*.

Per quanto riguarda il valore minimo dell'errore quadratico medio, sfruttando ancora l'ortogonalità tra $Y - \widehat{Y}$ e X, si trova:

$$\begin{aligned} \mathrm{MSE}_{\min} &= E[(Y-\widehat{Y})^2] = E[(Y-\widehat{Y})\,(Y-\widehat{Y})] = \\ &= E[(Y-\widehat{Y})Y] - E[(Y-\widehat{Y})\widehat{Y}] = \\ &= E[(Y-\widehat{Y})Y] - a\,\underbrace{E[(Y-\widehat{Y})X]}_{=0} = \\ &= E(Y^2) - E(\widehat{Y}\,Y) = \sigma_Y^2[1-\rho_{XY}^2]\,, \end{aligned} \tag{7.10}$$

che è lo stesso valore trovato per il caso $b \neq 0$, e per il quale valgono le stesse considerazioni, adesso suscettibili di una chiara interpretazione geometrica. Infatti, se osserviamo la Fig. 7.1 (a destra), in condizioni di ortogonalità risulta, per il teorema di Pitagora,

$$E(Y^2) = E(\widehat{Y}^2) + \mathrm{MSE}_{\min}\,,$$

e inoltre si ha:

$$E(\widehat{Y}^2) = a^2\,E(X^2) = \rho_{XY}^2\sigma_Y^2\,,$$

dove abbiamo sostituito l'espressione di a data dalla (7.8); in definitiva, si trova proprio la (7.10).

In conclusione, va osservato che, tra tutti i tipi di stima possibile, la stima lineare è raramente ottima, in quanto la dipendenza tra X ed Y non è generalmente di tipo lineare, ma spesso è di tipo non lineare. Tuttavia, la stima lineare è ampiamente considerata nelle applicazioni, principalmente per la semplicità matematica della sua formulazione e per l'interpretazione geometrica. Degno di nota è ancora il caso delle variabili aleatorie congiuntamente gaussiane, per le quali si può dimostrare l'importante risultato che la stima lineare è ottima (in senso MMSE) tra tutti i possibili tipi di stima; ciò significa intuitivamente che le variabili aleatorie congiuntamente gaussiane presentano tra loro esclusivamente dipendenze di tipo lineare.

7.6 Esercizi proposti

Esercizio 7.1. Un esperimento aleatorio consiste nello scegliere a caso ed in modo indipendente due punti X ed Y nell'intervallo $(0,1)$. Calcolare il valor medio della distanza tra i due punti. [*Risposta:* $1/3$]

Esercizio 7.2. Un rettangolo ha i due lati X ed Y che sono modellati come variabili aleatorie aventi pdf $f_{XY}(x,y) = x+y$, $0<x<1$, $0<y<1$. Calcolare il valor medio dell'area del rettangolo. [*Risposta:* $1/3$]

Esercizio 7.3. In un sistema di riferimento cartesiano, si sceglie a caso ed in modo indipendente una lunghezza R nell'intervallo $(0,1)$ ed un angolo Θ nell'intervallo $(0,2\pi)$, e si costruisce un vettore centrato nell'origine di lunghezza R e che forma con l'asse x un angolo Θ (valutato in senso antiorario). Calcolare la lunghezza media delle proiezioni X ed Y del vettore sui due assi cartesiani. [*Risposta:* $1/\pi$]

Esercizio 7.4. L'energia cinetica $\mathcal{E}$ di un corpo è pari a $\mathcal{E} = \frac{1}{2}MV^2$, dove M rappresenta la massa (in kg) e V la velocità (scalare) del corpo (in m/s). Se la pdf congiunta di M e V è $f_{MV}(x,y) = x+y$, per $0<x<1$ e $0<y<1$, determinare l'energia cinetica media posseduta dal corpo. [*Risposta:* 0.12 *Joule*]

Esercizio 7.5. Una particella di massa $m = 10^{-7}$ kg si muove su un sottile strato superficiale, assimilabile ad un piano. Le componenti lungo x ed y della sua velocità (in m/s) sono modellate come variabili aleatorie a media nulla e varianza unitaria. Calcolare l'energia cinetica media posseduta dalla particella. [*Risposta:* 10^{-7} *Joule*]

Esercizio 7.6. Due aste X ed Y hanno lunghezze modellabili come variabili aleatorie indipendenti ed uniformi in $(0,1)$.

a) Determinare la lunghezza media della più lunga tra la due.

b) Determinare la lunghezza media della più corta tra le due.

[*Risposta: a)* $2/3$*; b)* $1/3$]

Esercizio 7.7. Siano X, Y due variabili aleatorie con pdf congiunta $f_{XY}(x,y) = 1/24$, $0<x<6$, $0<y<4$. Calcolare il momento congiunto $E(X^2\,Y^2)$.

Esercizio 7.8. Siano X, Y due variabili aleatorie indipendenti con medie $\mu_X = 2$, $\mu_Y = 4$ e valori quadratici medi $E(X^2) = 8$ ed $E(Y^2) = 25$. Calcolare media, valor quadratico medio e varianza di $Z = 3X - Y$.

Esercizio 7.9. Siano X, Y due variabili aleatorie indipendenti, con medie μ_X, μ_Y e varianze σ_X^2, σ_Y^2, rispettivamente. Esprimere la correlazione tra $Z = XY$ ed Y in funzione dei precedenti parametri.

Esercizio 7.10. Sia X una variabile aleatoria con media $\mu_X = 3$ e varianza $\sigma_X^2 = 2$, e sia $Y = -6X + 22$.

a) Calcolare correlazione, covarianza e coefficiente di correlazione tra X ed Y;

b) stabilire se X ed Y sono ortogonali, incorrelate, indipendenti.

Esercizio 7.11. Siano X, Y due variabili aleatorie con la seguente pdf congiunta:

$$f_{XY}(x,y) = \begin{cases} \frac{1}{40}(x+y)^2\,, & -1<x<1,\ -3<y<3\,; \\ 0\,, & \text{altrimenti}\,. \end{cases}$$

Determinare il coefficiente di correlazione tra X ed Y.

Esercizio 7.12. Siano X ed Y due variabili aleatorie con pdf congiunta $f_{XY}(x,y) = x+y$, $0 \le x \le 1$, $0 \le y \le 1$. Calcolare correlazione, covarianza e coefficiente di correlazione tra X ed Y.

Esercizio 7.13. Siano X ed Y due variabili aleatorie con pdf congiunta $f_{XY}(x,y) = 2x$, $0 \le x \le 1$, $0 \le y \le 1$. Calcolare correlazione, covarianza e coefficiente di correlazione tra X ed Y.

Esercizio 7.14. Siano X ed Y due variabili aleatorie con pdf congiunta $f_{XY}(x,y) = 1, 0 \leq x \leq 1, x \leq y \leq x+1$. Calcolare il coefficiente di correlazione tra X ed Y. [*Risposta:* $\rho_{XY} = 1/\sqrt{2}$]

Esercizio 7.15. Siano X ed Y due variabili aleatorie con pdf congiunta $f_{XY}(x,y) = 10, 0 \leq x \leq 1, x \leq y \leq x + 1/10$. Calcolare il coefficiente di correlazione tra X ed Y. [*Risposta:* $\rho_{XY} = \sqrt{100/101}$]

Esercizio 7.16. Sia $X \sim \mathrm{U}(-1,1)$, e sia $Y = X^2$. Mostrare che X ed Y sono *incorrelate* anche se sono chiaramente *dipendenti*.

Esercizio 7.17. Mostrare che ogni variabile aleatoria X è incorrelata con una costante.

★ *Esercizio* 7.18. Mostrare che se $u(a - X)$ e $u(b - Y)$ sono incorrelate per ogni a e b, allora X e Y sono indipendenti.

Esercizio 7.19. Siano U, V due variabili aleatorie ottenute da X, Y mediante la seguente trasformazione:

$$\begin{cases} U &= X + a\,Y \\ V &= X - a\,Y \end{cases}$$

Determinare, in funzione dei momenti di X ed Y, i valori di a per i quali le variabili aleatorie U e V sono (i) ortogonali; (ii) incorrelate.

Esercizio 7.20. Siano X, Y due variabili aleatorie a media nulla, varianze $\sigma_X^2 = 4$, $\sigma_Y^2 = 16$, e coefficiente di correlazione $\rho_{XY} = -0.5$; a partire da esse si costruisca $W = a\,X + 3\,Y$.

a) Determinare il valore di a che rende minimo il valore quadratico medio di W;

b) determinare il valore quadratico medio minimo.

Esercizio 7.21. Siano X, Y due variabili aleatorie incorrelate. Trovare il coefficiente di correlazione tra $X + Y$ ed $X - Y$ in funzione di σ_X^2 e σ_Y^2. [*Risposta:* $\rho_{XY} = (\sigma_X^2 - \sigma_Y^2)/(\sigma_X^2 + \sigma_Y^2)$]

Capitolo 8

Vettori di variabili aleatorie

In questo capitolo si generalizzano al caso di n variabili aleatorie molti dei concetti già esposti per il caso di una coppia di variabili aleatorie; la generalizzazione è quasi sempre naturale, portando alla definizione di CDF, pdf e DF congiunte, che risultano in tal caso funzioni di n variabili e consentono la caratterizzazione statistica completa delle n variabili aleatorie. Successivamente vengono introdotte le trasformazioni di n variabili aleatorie, generalizzando il teorema fondamentale sulle trasformazioni già studiato per $n = 1, 2$. La definizione di indipendenza viene anch'essa generalizzata, e si introducono i concetti di indipendenza a coppie e a gruppi. Per quanto riguarda la caratterizzazione sintetica, l'attenzione viene rivolta principalmente alle matrici di correlazione e di covarianza, la cui introduzione consente l'importante generalizzazione al caso n-dimensionale delle variabili aleatorie congiuntamente gaussiane, discusso nell'esempio 8.5. Il capitolo si conclude con una breve introduzione ai teoremi limite (per $n \to \infty$), nella quale si espongono la legge dei grandi numeri (nella versione debole e forte) ed il teorema limite fondamentale.

8.1 Introduzione

Abbiamo visto nel capitolo 6 come descrivere probabilisticamente una coppia di variabili aleatorie X ed Y. Tuttavia, è evidente che esistono casi in cui si presenta la necessità di descrivere congiuntamente *più di due* variabili aleatorie. Ad esempio, un insieme di misure di tensione effettuate su un circuito elettrico può essere rappresentato da una n-pla di variabili aleatorie, in cui X_1 rappresenta la tensione nel punto 1, X_2 rappresenta la tensione nel punto 2, e così via. Allo stesso modo, un'analisi medica volta a individuare una malattia potrebbe essere modellata come una n-pla di variabili aleatorie, in cui X_1 rappresenta il livello di glucosio del sangue, X_2 il livello di azoto, e così via. È necessario allora introdurre gli strumenti matematici per caratterizzare statisticamente n variabili aleatorie, con $n > 2$. Fortunatamente, vedremo che la maggior parte

dei concetti necessari si ottengono generalizzando in maniera semplice definizioni e risultati già ottenuti per il caso di coppie di variabili aleatorie.

Infine, in alcuni casi interessa studiare il comportamento *limite* o *asintotico* di n variabili aleatorie quando si faccia tendere n all'infinito, ottenendo così una *sequenza* di variabili aleatorie. I principali risultati sono raccolti nei cosiddetti *teoremi limite* (legge dei grandi numeri e teorema limite fondamentale o *central limit theorem*, CLT), sulla base dei quali è tra l'altro possibile approfondire il legame esistente tra la teoria assiomatica della probabilità e l'interpretazione frequentista.

8.2 Caratterizzazione statistica di n variabili aleatorie

Sia $(\Omega, \mathcal{S}, P)$ uno spazio di probabilità, e siano $X_1, X_2, \ldots, X_n$ n variabili aleatorie costruite su tale spazio. Per adoperare una notazione sintetica, possiamo organizzare le n variabili aleatorie in un vettore *colonna*[1]

$$\boldsymbol{X} = [X_1, X_2, \ldots, X_n]^T ,$$

dove con l'apice T abbiamo denotato l'operazione di trasposizione. Abbiamo costruito in questo modo un *vettore di variabili aleatorie*, e adopereremo indifferentemente la terminologia "n variabili aleatorie", "n-pla di variabili aleatorie", oppure "vettore di n variabili aleatorie".

8.2.1 Funzione di distribuzione cumulativa (CDF)

Per caratterizzare statisticamente le n variabili aleatorie, dobbiamo generalizzare il concetto di CDF congiunta per una coppia di variabili aleatorie al caso di n variabili aleatorie:

Definizione (CDF congiunta di n variabili aleatorie). Date n variabili aleatorie $X_1, X_2, \ldots, X_n$ costruite su uno stesso spazio di probabilità $(\Omega, \mathcal{S}, P)$, la loro CDF congiunta è:

$$F_{X_1 X_2 \cdots X_n}(x_1, x_2, \ldots, x_n) \triangleq P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n) ,$$

dove $(x_1, x_2, \ldots, x_n) \in \overline{\mathbb{R}}^n$.

La CDF congiunta è una funzione reale di n variabili reali, e spesso viene denominata *CDF di ordine n*.

Per adoperare una notazione sintetica, possiamo utilizzare una notazione vettoriale anche per i valori $x_1, x_2, \ldots, x_n$, ponendo $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]$, e denotare la CDF congiunta con $F_{\boldsymbol{X}}(\boldsymbol{x})$. Va notato che, poiché la rappresentazione grafica di una funzione di n variabili avviene in uno spazio $(n+1)$-dimensionale, per $n > 2$ tale rappresentazione è praticamente impossibile.

8.2.2 Funzione densità di probabilità (pdf)

Analogamente al caso di coppie di variabili aleatorie, a partire dalla CDF congiunta si ottiene la *pdf congiunta* per derivazione mista:

[1] In questo capitolo, faremo sovente uso di nozioni elementari di algebra lineare, quali vettori, matrici, prodotti tra matrici / vettori, etc; si assume pertanto che il lettore abbia familiarità con tali concetti; per agevolare la lettura, le definizioni e proprietà di uso più frequente sono brevemente richiamate nell'Appendice E.

Definizione (pdf congiunta di *n* variabili aleatorie). Date n variabili aleatorie $X_1, X_2, \ldots, X_n$ con CDF congiunta $F_{X_1 X_2 \ldots X_n}(x_1, x_2, \ldots, x_n)$, la loro pdf congiunta è:

$$f_{X_1 X_2 \cdots X_n}(x_1, x_2, \ldots, x_n) \triangleq \frac{\partial^n}{\partial x_1 \partial x_2 \cdots \partial x_n} F_{X_1 X_2 \cdots X_n}(x_1, x_2, \ldots, x_n) .$$

Anche in questo caso assumeremo che valga la condizione di Schwartz, in modo che la derivata mista non dipenda dall'ordine di integrazione. Inoltre, denoteremo sinteticamente la pdf congiunta con $f_{\boldsymbol{X}}(\boldsymbol{x})$.

8.2.3 Funzione di distribuzione di probabilità (DF)

Infine, per vettori di variabili aleatorie discrete, possiamo definire la DF congiunta:

Definizione (DF congiunta di *n* variabili aleatorie). Date n variabili aleatorie $X_1, X_2, \ldots, X_n$ discrete costruite su uno stesso spazio di probabilità $(\Omega, \mathcal{S}, P)$, a valori in $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_n$, rispettivamente, la loro DF congiunta è:

$$p_{X_1 X_2 \cdots X_n}(x_1, x_2, \ldots, x_n) \triangleq P(X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n) ,$$

dove $(x_1, x_2, \ldots, x_n) \in \mathcal{X}_1 \times \mathcal{X}_2 \cdots \times \mathcal{X}_n$.

Anche per la DF congiunta utilizzeremo spesso la notazione sintetica $p_{\boldsymbol{X}}(\boldsymbol{x})$.

8.2.4 Proprietà delle distribuzioni congiunte di *n* variabili aleatorie

Le CDF, pdf e DF di n variabili aleatorie godono di proprietà che sono la naturale generalizzazione delle corrispondenti proprietà valide per il caso $n = 2$. Senza elencarle tutte in maniera sistematica, limitiamoci a considerare quelle più importanti nelle applicazioni.

Ad esempio, notiamo che a partire dalla pdf congiunta è possibile ricavare la CDF congiunta per integrazione, come:

$$F_{X_1 X_2 \cdots X_n}(x_1, x_2, \ldots, x_n) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \cdots \int_{-\infty}^{x_n} f_{X_1 X_2 \cdots X_n}(u_1, u_2, \ldots, u_n)\, du_1\, du_2 \cdots du_n ,$$

o equivalentemente, utilizzando la notazione sintetica, come

$$F_{\boldsymbol{X}}(\boldsymbol{x}) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \cdots \int_{-\infty}^{x_n} f_{\boldsymbol{X}}(\boldsymbol{u})\, d\boldsymbol{u} .$$

Poichè, poi, $F_{X_1 X_2 \cdots X_n}(\infty, \infty, \ldots, \infty) = 1$, dalla precedente ricaviamo la condizione di *normalizzazione della pdf*:

$$\int_{\mathbb{R}^n} f_{\boldsymbol{X}}(\boldsymbol{x})\, d\boldsymbol{x} = 1 ,$$

dove abbiamo utilizzato una notazione sintetica per l'integrale n-plo. Pertanto la pdf ha "volume" unitario nello spazio n-dimensionale.

L'interpretazione della pdf congiunta come *densità di probabilità* scaturisce dalla seguente relazione:

$$f_{\boldsymbol{X}}(\boldsymbol{x})\, d\boldsymbol{x} = P(x_1 < X_1 \leq x_1 + dx_1, x_2 < X_2 \leq x_2 + dx_2, \ldots, x_n < X_n \leq x_n + dx_n)\,,$$

per cui la pdf congiunta nel punto $\boldsymbol{x}$ rappresenta la probabilità che la n-pla di variabili aleatorie $X_1, X_2, \ldots, X_n$ appartengano ad un rettangolino n-dimensionale di lati infinitesimi centrato su $\boldsymbol{x}$, divisa per il "volume" $dx_1\, dx_2 \cdots dx_n$ del rettangolino. Anche in questo caso, tale risultato prova implicitamente che $f_{\boldsymbol{X}}(\boldsymbol{x}) \geq 0$; inoltre, se D è un dominio qualsiasi di $\mathbb{R}^n$, si ha:

$$P(\boldsymbol{X} \in D) = \int_D f_{\boldsymbol{X}}(\boldsymbol{x})\, d\boldsymbol{x}\,.$$

Ponendo pari a $+\infty$ uno o più argomenti della CDF, è poi possibile ottenere tutte le statistiche di ordine $k < n$ a partire da quelle di ordine n. Consideriamo, ad esempio, il caso di tre variabili aleatorie X_1, X_2, X_3, caratterizzate dalla loro CDF congiunta $F_{X_1X_2X_3}(x_1, x_2, x_3)$. È facile verificare che si ha, ad esempio:

$$F_{X_1X_2}(x_1, x_2) = F_{X_1X_2X_3}(x_1, x_2, +\infty)\,,$$
$$F_{X_1}(x_1) = F_{X_1X_2X_3}(x_1, +\infty, +\infty)\,.$$

È possibile procedere analogamente con le pdf, semplicemente *integrando* rispetto alle variabili che non interessano. Le relazioni precedenti si scrivono, in termini di pdf, come:

$$f_{X_1X_2}(x_1, x_2) = \int_{-\infty}^{\infty} f_{X_1X_2X_3}(x_1, x_2, x_3)\, dx_3$$
$$f_{X_1}(x_1) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{X_1X_2X_3}(x_1, x_2, x_3)\, dx_2\, dx_3$$

Analoghe relazioni valgono per le DF, per le quali, anziché integrare rispetto alle variabili che non interessano, si effettua la somma.

In definitiva, osserviamo che assegnare le CDF, pdf o DF congiunte di ordine n equivale ad assegnare implicitamente tutte le CDF, pdf e DF congiunte di ordine $k < n$.

8.3 Trasformazioni di n variabili aleatorie

Generalizziamo adesso il concetto di trasformazione già presentato per il caso di una e due variabili aleatorie, considerando trasformazioni di n variabili aleatorie. Il caso più generale che possiamo considerare è quello in cui, a partire da un vettore di n variabili aleatorie $\boldsymbol{X} = [X_1, X_2, \ldots, X_n]^T$, si ottiene un vettore di k variabili aleatorie $\boldsymbol{Y} = [Y_1, Y_2, \ldots, Y_k]^T$, utilizzando k funzioni $g_1, g_2, \ldots, g_k$ di n variabili. Tale trasformazione di variabili aleatorie si esprime esplicitamente come:

$$\begin{cases} Y_1 &= g_1(X_1, X_2, \ldots, X_n) \\ Y_2 &= g_2(X_1, X_2, \ldots, X_n) \\ \vdots & \vdots \qquad\qquad \vdots \\ Y_k &= g_k(X_1, X_2, \ldots, X_n) \end{cases}.$$

L'obiettivo è quello di determinare la pdf congiunta di $\boldsymbol{Y}$ a partire dalla pdf congiunta di $\boldsymbol{X}$. Possiamo distinguere tre casi: (i) $k < n$ (sistema "sottodeterminato"); (ii) $k > n$ (sistema "sovradeterminato"); (iii) $k = n$ (sistema "quadrato"). In particolare, nel caso $k = n$ è possibile fornire una interessante generalizzazione del teorema fondamentale sulle trasformazioni di variabili aleatorie, che abbiamo già introdotto per $n = 1$ (cfr. § 4.2.3) ed $n = 2$ (cfr. § 6.7.2):

Teorema 8.1 (teorema fondamentale sulle trasformazioni di n variabili aleatorie). Sia $\boldsymbol{X} = [X_1, X_2, \ldots, X_n]^T$ un vettore di variabili aleatorie con pdf $f_{X_1 X_2 \cdots X_n}(x_1, x_2, \ldots, x_n)$, e sia $\boldsymbol{Y} = [Y_1, Y_2, \ldots, Y_n]^T$ un vettore di variabili aleatorie ottenuto per trasformazione da $\boldsymbol{X}$, come:

$$\begin{cases} Y_1 &= g_1(X_1, X_2, \ldots, X_n) \\ Y_2 &= g_2(X_1, X_2, \ldots, X_n) \\ \vdots & \vdots \\ Y_n &= g_n(X_1, X_2, \ldots, X_n) \end{cases} .$$

Si consideri il sistema di equazioni:

$$\begin{cases} y_1 &= g_1(x_1, x_2, \ldots, x_n) \\ y_2 &= g_2(x_1, x_2, \ldots, x_n) \\ \vdots & \vdots \\ y_n &= g_n(x_1, x_2, \ldots, x_n) \end{cases} . \tag{8.1}$$

La pdf congiunta di $\boldsymbol{Y}$ è data da:

$$f_{Y_1 Y_2 \cdots Y_n}(y_1, y_2, \ldots, y_n) = \begin{cases} 0, & \text{se il sistema (8.1) non ha soluzioni;} \\ \displaystyle\sum_i \frac{f_{X_1 X_2 \cdots X_n}(x_1^i, x_2^i, \ldots, x_n^i)}{|\det[\boldsymbol{J}(x_1^i, x_2^i, \ldots, x_n^i))]|}, & \text{dove } (x_1^i, x_2^i, \ldots, x_n^i) \text{ è una soluzione del sistema (8.1);} \end{cases}$$

in cui $\det(\cdot)$ denota il determinante, e

$$\boldsymbol{J}(x_1, x_2, \ldots, x_n) = \frac{\partial(y_1, y_2, \ldots, y_n)}{\partial(x_1, x_2, \ldots, x_n)} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial y_n}{\partial x_1} & \frac{\partial y_n}{\partial x_2} & \cdots & \frac{\partial y_n}{\partial x_n} \end{pmatrix}$$

è la matrice jacobiana della trasformazione.

Come nel caso $n = 1$ ed $n = 2$, l'applicazione del teorema richiede che il sistema (8.1) abbia al più una infinità numerabile di soluzioni.

► *Esempio* 8.1 *(trasformazione lineare).* Consideriamo il caso particolarmente semplice di una trasformazione *lineare* di variabili aleatorie:

$$\begin{cases} Y_1 &= a_{11} X_1 + a_{12} X_2 + \ldots a_{1n} X_n \\ Y_2 &= a_{21} X_1 + a_{22} X_2 + \ldots a_{2n} X_n \\ \vdots & \vdots \\ Y_n &= a_{n1} X_1 + a_{n2} X_2 + \ldots a_{nn} X_n \end{cases}$$

Tale trasformazione si può esprimere in forma assai compatta utilizzando la notazione vettoriale:

$$\boldsymbol{Y} = \boldsymbol{A}\,\boldsymbol{X}\,,$$

dove $\boldsymbol{Y} = [Y_1, Y_2, \ldots, Y_n]^T$ ed $\boldsymbol{X} = [X_1, X_2, \ldots, X_n]^T$ sono vettori colonna n-dimensionali, e la matrice $\boldsymbol{A}$ è definita come:

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} & \ldots & a_{1n} \\ a_{21} & a_{22} & \ldots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \ldots & a_{nn} \end{pmatrix} .$$

Se assumiamo che $\det(\boldsymbol{A}) \neq 0$, il sistema numerico $\boldsymbol{y} = \boldsymbol{A}\,\boldsymbol{x}$ ammetterà, per ogni $\boldsymbol{y} \in \mathbb{R}^n$, una ed una sola soluzione nell'incognita $\boldsymbol{x}$, data da:

$$\boldsymbol{x} = \boldsymbol{A}^{-1}\,\boldsymbol{y}\,,$$

dove $\boldsymbol{A}^{-1}$ denota la matrice *inversa* di $\boldsymbol{A}$. È facile, inoltre, verificare che la matrice jacobiana $\boldsymbol{J}$ della trasformazione è pari proprio ad $\boldsymbol{A}$, per cui $|\det(\boldsymbol{J})| = |\det(\boldsymbol{A})|$. Per il teorema fondamentale, allora, la pdf del vettore $\boldsymbol{Y}$ si può esprimere, con sintetica notazione vettoriale, come:

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{|\det(\boldsymbol{A})|}\, f_{\boldsymbol{X}}(\boldsymbol{A}^{-1}\,\boldsymbol{y})\,,$$

dove $\boldsymbol{y} = [y_1, y_2, \ldots, y_n] \in \mathbb{R}^n$. ◀

Gli altri due casi menzionati (sistema "sottodeterminato" e "sovradeterminato") si possono ricondurre al caso di sistema "quadrato". Infatti, nel caso $k < n$, possiamo introdurre $n - k$ variabili aleatorie ausiliarie, ad esempio $Y_{k+1} = X_{k+1}, Y_{k+2} = X_{k+2}, \ldots, Y_n = X_n$ per ricondurci al caso $k = n$ ("quadrando", per così dire, il sistema). Una volta determinata la pdf congiunta di $\boldsymbol{Y}' = [Y_1, Y_2, \ldots, Y_k, Y_{k+1}, \ldots, Y_n]$, è possibile ottenere quella di $\boldsymbol{Y} = [Y_1, Y_2, \ldots, Y_k]$ semplicemente integrando la pdf di $\boldsymbol{Y}'$ rispetto alle variabili $y_{k+1}, y_{k+2}, \ldots, y_n$, corrispondenti alle variabili aleatorie ausiliarie $Y_{k+1}, Y_{k+2}, \ldots, Y_n$. che non interessano.

Nel caso $k > n$, un teorema di analisi (teorema di Dini) assicura che $k - n$ variabili aleatorie appartenenti a $\boldsymbol{Y}$, ad esempio $Y_{n+1}, Y_{n+2}, \ldots, Y_k$ possono essere espresse in funzione delle n rimanenti, siano esse $Y_1, Y_2, \ldots, Y_n$. In questo caso, si può dimostrare che la pdf congiunta del vettore k-dimensionale $\boldsymbol{Y}$ è *singolare*, ovvero è definita su un sottospazio n-dimensionale di $\mathbb{R}^k$, e può essere espressa in termini della pdf delle sole $Y_1, Y_2, \ldots, Y_n$. Tale pdf si può determinare considerando il sottosistema quadrato composto dalle prime n equazioni, e quindi riconducendosi ancora al caso $k = n$.

8.4 Variabili aleatorie indipendenti

Vogliamo ora estendere il concetto di indipendenza a vettori di n variabili aleatorie:

Definizione (variabili aleatorie indipendenti). Le variabili aleatorie X_1, X_2, ..., X_n si dicono indipendenti se

$$F_{X_1 X_2 \cdots X_n}(x_1, x_2, \ldots, x_n) = F_{X_1}(x_1)\, F_{X_2}(x_2) \cdots F_{X_n}(x_n)\,, \tag{8.2}$$

per ogni $(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$.

Come si vede, l'indipendenza equivale anche in questo caso alla fattorizzazione della CDF congiunta (si confronti con la definizione di indipendenza per coppie di variabili aleatorie, data nel § 6.6). È chiaro d'altronde che la fattorizzazione della CDF congiunta è equivalente a quella della pdf congiunta, per cui si ha anche:

$$f_{X_1 X_2 \cdots X_n}(x_1, x_2, \ldots, x_n) = f_{X_1}(x_1)\, f_{X_2}(x_2) \cdots f_{X_n}(x_n)\,,$$

per ogni $(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$.

Si può osservare, data la definizione della CDF congiunta, che l'indipendenza delle variabili aleatorie $X_1, X_2, \ldots, X_n$ equivale all'indipendenza degli eventi $\{X_1 \leq x_1\}$, $\{X_2 \leq x_2\}$, ..., $\{X_n \leq x_n\}$. Tuttavia, nel § 2.3.1, abbiamo visto che per specificare l'indipendenza di tre o più

eventi, oltre alla fattorizzazione della probabilità congiunta di tutti gli eventi, bisogna richiedere anche la fattorizzazione della probabilità congiunta di un qualunque sottoinsieme di tutti gli eventi. In questo caso, tale fattorizzazione di ordine inferiore, che sarebbe equivalente a richiedere che anche le CDF di ordine $k < n$ si fattorizzino nel prodotto delle CDF marginali, non è necessaria, in quanto discende necessariamente dalla fattorizzazione della CDF di ordine n: infatti, le statistiche di ordine $k < n$ sono univocamente determinate dalle statistiche di ordine n. Ad esempio, consideriamo il caso $n = 3$, per il quale la relazione di indipendenza si scrive esplicitamente come

$$F_{X_1X_2X_3}(x_1, x_2, x_3) = F_{X_1}(x_1)\, F_{X_2}(x_2)\, F_{X_3}(x_3)\,.$$

Ponendo ad esempio $x_3 = +\infty$, si ha al primo membro $F_{X_1X_2X_3}(x_1, x_2, +\infty) = F_{X_1X_2}(x_1, x_2)$ ed al secondo $F_{X_3}(+\infty) = 1$, per cui si trova:

$$F_{X_1X_2}(x_1, x_2) = F_{X_1}(x_1)\, F_{X_2}(x_2)\,,$$

ovvero la fattorizzazione della CDF di X_1 ed X_2. Con analogo ragionamento si può ricavare la fattorizzazione della CDF di X_1 ed X_3 e tra X_2 ed X_3. La solo apparente discrepanza tra tale definizione di indipendenza e quella fornita nel § 2.3.1 sta nel fatto che in realtà la fattorizzazione (8.2), poichè deve valere $\forall (x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$, è in realtà una condizione assai più forte di quella fornita nel § 2.3.1.

Con riferimento alle trasformazioni di variabili aleatorie, analogamente al caso di due variabili, è possibile provare che se si considera la trasformazione "diagonale":

$$\begin{cases} Y_1 &= g_1(X_1) \\ Y_2 &= g_2(X_2) \\ \vdots & \vdots \quad\quad \vdots \\ Y_n &= g_n(X_n) \end{cases}$$

e le variabili aleatorie $X_1, X_2, \ldots, X_n$ sono indipendenti, allora sono indipendenti anche le variabili aleatorie $Y_1, Y_2, \ldots, Y_n$. La prova è semplice e ricalca quella per il caso di due variabili (cfr. § 6.6).

Anche per le variabili aleatorie, così come per gli eventi (cfr. § 2.3.1), si può definire il concetto di *indipendenza a coppie*:

Definizione (variabili aleatorie indipendenti a coppie). Le variabili aleatorie X_1, X_2, ..., X_n si dicono indipendenti a coppie se

$$F_{X_iX_j}(x_i, x_j) = F_{X_i}(x_i)\, F_{X_j}(x_j)\,, \quad \forall i \neq j \quad \text{e} \quad \forall (x_i, x_j) \in \mathbb{R}^2\,.$$

È evidente che l'indipendenza implica sempre l'indipendenza a coppie, mentre il viceversa non è vero.[2]

È possibile anche definire l'indipendenza tra *gruppi* di variabili aleatorie appartenenti ad un vettore $\boldsymbol{X}$:

[2]Salvo per il caso delle variabili aleatorie gaussiane, per le quali, peraltro, è sufficiente una condizione ancora più debole dell'indipendenza a coppie, ovvero l'incorrelazione (cfr. § 8.5.4).

Definizione (variabili aleatorie indipendenti a gruppi). Le variabili aleatorie $X_1, X_2, \ldots, X_k$ si dicono indipendenti dalle variabili aleatorie $X_{k+1}, X_{k+2}, \ldots, X_n$ se:

$$F_{X_1 X_2 \cdots X_n}(x_1, x_2, \ldots, x_n) = F_{X_1 X_2 \cdots X_k}(x_1, x_2, \ldots, x_k)\, F_{X_{k+1} X_{k+2} \cdots X_n}(x_{k+1}, x_{k+2}, \ldots, x_n)\,,$$

per ogni $(x_1, x_2, \ldots, x_n) \in \mathbb{R}^n$.

Infine, un concetto che spesso ricorre insieme a quello di variabili aleatorie indipendenti, ma che non ha niente a che vedere con l'indipendenza, è quello di variabili aleatorie *identicamente distribuite* .

Definizione (variabili aleatorie identicamente distribuite). Le variabili aleatorie $X_1, X_2, \ldots, X_n$ si dicono identicamente distribuite se

$$F_{X_i}(x) = F(x)\,, \quad \forall i \in \{1, 2, \ldots, n\}\,.$$

In altri termini, variabili aleatorie identicamente distribuite sono caratterizzate dall'avere la stessa CDF del primo ordine (ad esempio, sono tutte gaussiane con la stessa media e la stessa varianza). Spesso si considerano n variabili aleatorie che sono *sia* indipendenti *sia* identicamente distribuite; in tal caso si parla di variabili aleatorie *indipendenti ed identicamente distribuite (iid)*. Si noti che per caratterizzare completamente n variabili iid è sufficiente assegnare la CDF del primo ordine $F(x)$, che è la stessa per tutte le variabili. Infatti, data l'indipendenza, qualsiasi CDF di ordine $k > 1$ si ottiene moltiplicando tra loro k CDF del primo ordine.

8.5 Momenti di n variabili aleatorie

Il punto di partenza per definire i momenti di n variabili aleatorie è introdurre la generalizzazione del teorema fondamentale della media:

Teorema 8.2 (teorema fondamentale della media per n variabili aleatorie). Sia $Z = g(X_1, X_2, \ldots, X_n)$ una trasformazione delle variabili aleatorie X_1, X_2, ..., X_n aventi pdf congiunta $f_{X_1 X_2 \cdots X_n}(x_1, x_2, \ldots, x_n)$; si ha:

$$\begin{aligned} E(Z) &= E[g(X_1, X_2, \ldots, X_n)] = \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, x_2, \ldots, x_n)\, f_{X_1 X_2 \cdots X_n}(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \cdots dx_n\,, \end{aligned} \quad (8.3)$$

se tale integrale esiste finito.

Notiamo che con notazione vettoriale la (8.3) si scrive molto più concisamente come:

$$E(Z) = E[g(\boldsymbol{X})] = \int_{\mathbb{R}^n} g(\boldsymbol{x})\, f_{\boldsymbol{X}}(\boldsymbol{x})\, d\boldsymbol{x}\,.$$

Ovviamente anche per vettori di n variabili aleatorie la media gode della proprietà di linearità. Infatti, siano g_k n arbitrarie funzioni di n variabili, e siano a_k n costanti reali, con $k = 1, 2, \ldots, n$. Si ha:

$$E\left[\sum_{k=1}^{n} a_k\, g_k(\boldsymbol{X})\right] = \sum_{k=1}^{n} a_k E[g_k(\boldsymbol{X})]\,,$$

e scegliendo $g_1(\boldsymbol{X}) = X_1, g_2(\boldsymbol{X}) = X_2, \ldots, g_n(\boldsymbol{X}) = X_n$,

$$\boxed{E\left(\sum_{k=1}^{n} a_k\, X_k\right) = \sum_{k=1}^{n} a_k E(X_k)\,,}$$

ovvero la media della combinazione lineare di n variabili aleatorie coincide con la combinazione lineare delle medie.

► *Esempio* 8.2 *(media di una variabile aleatoria binomiale).* Nel § 5.2 abbiamo dimostrato, utilizzando le proprietà del coefficiente binomiale, che la media di una variabile aleatoria $X \sim \mathrm{B}(n, p)$ è pari a $E(X) = n\,p$. Una dimostrazione più semplice si basa sull'osservazione che una variabile aleatoria binomiale si può interpretare come la somma di n variabili aleatorie bernoulliane, di parametro p, indipendenti tra loro, cioè:

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

con $X_i \sim \mathrm{Bern}(p)$. Infatti, per contare il numero di successi in n prove, è sufficiente sommare i valori ottenuti associando ad un successo il valore 1 e ad un insuccesso il valore 0. Poichè allora $E(X_i) = p$, applicando la linearità della media si ha:

$$E(X) = \sum_{i=1}^{n} E(X_i) = n\,p\,,$$

che è lo stesso risultato ottenuto nel § 5.2. ◄

8.5.1 Vettore delle medie

Dato un vettore di variabili aleatorie $\boldsymbol{X} = [X_1, X_2, \ldots, X_n]^T$, possiamo calcolare le medie delle sue componenti, date da

$$\mu_{X_i} = E(X_i) \triangleq \int_{-\infty}^{\infty} x_i\, f_{X_i}(x_i)\, dx_i\,,$$

per $i = 1, 2, \ldots, n$, e raccoglierle in un vettore colonna $\boldsymbol{\mu_X} \triangleq [\mu_{X_1}, \mu_{X_2}, \ldots, \mu_{X_n}]^T$. Tale vettore prende il nome di *vettore delle medie*, e formalmente possiamo scrivere $\boldsymbol{\mu_X} = E(\boldsymbol{X})$, dove per media di un vettore intendiamo l'operatore che calcola la media di ciascuna componente del vettore, restituendo un vettore di uguale dimensione. Notiamo che, per calcolare il vettore delle medie, non è necessario conoscere la pdf di ordine n, ma è sufficiente conoscere la pdf del primo ordine di ciascuna componente del vettore. Questo è in accordo con il fatto che la media è un momento del *primo* ordine.

8.5.2 Matrice di correlazione

Dato un vettore di variabili aleatorie $\boldsymbol{X} = [X_1, X_2, \ldots, X_n]^T$, possiamo calcolare le correlazioni tra le sue componenti prese due a due, date da

$$\mathrm{Corr}(X_i, X_j) = E(X_i\, X_j) \triangleq \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x_i\, x_j\, f_{X_i X_j}(x_i, x_j)\, dx_i\, dx_j\,,$$

per $i, j = 1, 2, \ldots, n$, e raccoglierle in una matrice $\boldsymbol{R_X}$ quadrata, di dimensioni $n \times n$, che prende il nome di *matrice di correlazione* del vettore $\boldsymbol{X}$ o delle n variabili aleatorie $X_1, X_2, \ldots, X_n$. Notiamo che per calcolare tale matrice non è necessario conoscere la pdf di ordine n, ma è sufficiente conoscere la pdf del secondo ordine di qualsiasi coppia di componenti del vettore, coerentemente con il fatto che la correlazione è un momento del *secondo* ordine.

Applicando semplici risultati di algebra lineare (cfr. Appendice E), è facile verificare che la matrice di correlazione si può scrivere nella seguente forma, particolarmente compatta:

$$\boxed{\boldsymbol{R} = E(\boldsymbol{X}\,\boldsymbol{X}^T)\,,} \tag{8.4}$$

dove $\boldsymbol{X}\,\boldsymbol{X}^T$ è il prodotto[3] di un vettore colonna $n \times 1$ per un vettore riga $1 \times n$, il cui risultato è una matrice $n \times n$ i cui elementi sono (è facile verificarlo) proprio $X_i\,X_j$, e per media di una matrice intendiamo l'operatore che calcola la media di ciascun elemento della matrice, restituendo una matrice di uguali dimensioni.

Osserviamo che, per $i = j$, risulta $\mathrm{Corr}(X_i, X_i) = E(X_i^2)$, cioè la correlazione si riduce al valore quadratico medio; pertanto, sulla diagonale principale di $\boldsymbol{R_X}$ sono presenti i valori quadratici medi $E(X_1^2), E(X_2^2), \ldots, E(X_n^2)$. Notiamo poi che la matrice $\boldsymbol{R_X}$ è simmetrica, cioè l'elemento di posto (i,j) è uguale all'elemento di posto (j,i), come si ricava facilmente dalla simmetria della correlazione:

$$\mathrm{Corr}(X_i, X_j) = E(X_i\,X_j) = E(X_j\,X_i) = \mathrm{Corr}(X_j, X_i)\,.$$

Una formulazione equivalente della proprietà di simmetria di $\boldsymbol{R_X}$ è che tale matrice coincide con la sua trasposta, cioè si ha $\boldsymbol{R_X} = \boldsymbol{R_X}^T$; per verificare direttamente tale relazione, basta considerare la definizione (8.4) e applicare semplici relazioni di algebra lineare:

$$\boldsymbol{R_X}^T = E[(\boldsymbol{X}\,\boldsymbol{X}^T)^T] = E[(\boldsymbol{X}^T)^T\boldsymbol{X}^T] = E(\boldsymbol{X}\,\boldsymbol{X}^T) = \boldsymbol{R_X}\,.$$

Un'importante e non banale proprietà della matrice di correlazione si trova considerando la combinazione lineare Y, con coefficienti arbitrari, delle n variabili aleatorie:

$$Y = \sum_{k=1}^{n} a_k\,X_k = \boldsymbol{a}^T\,\boldsymbol{X}\,, \tag{8.5}$$

che abbiamo interpretato, introducendo il vettore colonna $\boldsymbol{a} = [a_1, a_2, \ldots, a_n]^T \in \mathbb{R}^n$, come prodotto scalare tra $\boldsymbol{a}$ ed $\boldsymbol{X}$. Calcoliamo il valor quadratico medio di Y adoperando semplici relazioni di algebra lineare:

$$E(Y^2) = E[(\boldsymbol{a}^T\,\boldsymbol{X})^2] = E[\boldsymbol{a}^T\,\boldsymbol{X}(\boldsymbol{a}^T\boldsymbol{X})^T] = E(\boldsymbol{a}^T\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{a}) = \boldsymbol{a}^T E(\boldsymbol{X}\boldsymbol{X}^T)\,\boldsymbol{a} = \boldsymbol{a}^T\boldsymbol{R_X}\,\boldsymbol{a}\,.$$

Poichè evidentemente $E(Y^2) \geq 0$, si ha che

$$\boxed{\boldsymbol{a}^T\,\boldsymbol{R_X}\,\boldsymbol{a} \geq 0, \quad \forall \boldsymbol{a} \in \mathbb{R}^n\,,}$$

che si esprime dicendo che la matrice $\boldsymbol{R_X}$ è *semidefinita positiva* (cfr. Appendice E). Se vale la disuguaglianza stretta, ovvero se $\boldsymbol{a}^T\,\boldsymbol{R_X}\,\boldsymbol{a} > 0$, $\forall \boldsymbol{a} \in \mathbb{R}^n - \{\boldsymbol{0}\}$, allora la matrice $\boldsymbol{R_X}$ è *definita positiva*. Si noti che la differenza tra i due casi è la seguente: se la matrice è solo semidefinita positiva, allora esiste un valore di $\boldsymbol{a} \neq \boldsymbol{0}$ tale che la *forma quadratica* $\boldsymbol{a}^T\,\boldsymbol{R_X}\,\boldsymbol{a} = 0$. Poichè tale forma quadratica coincide con il valor quadratico medio della combinazione lineare (8.5), allora una condizione *sufficiente* affinchè ciò accada è che le variabili aleatorie siano *linearmente dipendenti*, cioè che esista un vettore $\boldsymbol{a}$ di coefficienti non tutti nulli tali che:

$$Y = a_1\,X_1 + a_2\,X_2 + \ldots + a_n\,X_n = 0\,.$$

[3] Il prodotto $\boldsymbol{x}\,\boldsymbol{y}^T$, il cui risultato è una matrice, viene chiamato talvolta *prodotto esterno* tra i vettori (colonna) $\boldsymbol{x}$ e $\boldsymbol{y}$, e non va confuso con il *prodotto scalare* o *prodotto interno* $\boldsymbol{x}^T\boldsymbol{y}$, il cui risultato è uno scalare.

In questo caso, allora, almeno una variabile aleatoria può essere espressa come combinazione lineare delle rimanenti variabili aleatorie. Viceversa, se risulta $E[(\sum_{k=1}^{n} a_k X_k)^2] > 0$ per ogni $\boldsymbol{a} \neq \boldsymbol{0}$, le variabili si diranno *linearmente indipendenti*, e la loro matrice di correlazione sarà definita positiva. Si può dimostrare che una matrice definita positiva è senz'altro *non singolare*, e quindi è *invertibile*, mentre una matrice che è solo semidefinita positiva non ha tale proprietà.

8.5.3 Matrice di covarianza

Così come la matrice di correlazione raccoglie le correlazioni tra tutte le possibili coppie di variabili aleatorie, è possibile definire una *matrice di covarianza* $\boldsymbol{C_X}$ quadrata, di dimensioni $n \times n$, il cui elemento di posto (i, j) rappresenta la covarianza tra X_i ed X_j:

$$\mathrm{Cov}(X_i, X_j) = E[(X_i - \mu_{X_i})(X_j - \mu_{X_j})] \triangleq = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x_i - \mu_{X_i})(x_j - \mu_{X_j})\, f_{X_i X_j}(x_i, x_j)\, dx_i\, dx_j\,,$$

per $i, j = 1, 2, \ldots, n$. Notiamo che la matrice $\boldsymbol{C_X}$ si può scrivere con notazione compatta come:

$$\boxed{\boldsymbol{C_X} = E[(\boldsymbol{X} - \boldsymbol{\mu_X})(\boldsymbol{X} - \boldsymbol{\mu_X})^T]\,.} \tag{8.6}$$

Poichè, per $i = j$, la covarianza si riduce alla varianza, sulla diagonale principale di $\boldsymbol{C_X}$ sono presenti le varianze $\sigma_1^2, \sigma_2^2, \ldots, \sigma_n^2$. Inoltre, poichè la covarianza, come la correlazione, è simmetrica, allora la matrice $\boldsymbol{C_X}$ è anch'essa simmetrica, ovvero $\boldsymbol{C_X^T} = \boldsymbol{C_X}$.

In effetti, si può notare che la matrice di correlazione e quella di covarianza condividono numerose proprietà; ciò consegue dal fatto che la matrice di covarianza di $\boldsymbol{X}$ è *anche* una matrice di correlazione; in particolare, essa si può riguardare come la matrice di correlazione del *vettore centrato* $\boldsymbol{Y} = \boldsymbol{X} - \boldsymbol{\mu_X}$, in quanto si ha:

$$\boldsymbol{R_Y} = E(\boldsymbol{Y}\boldsymbol{Y}^T) = E[(\boldsymbol{X} - \boldsymbol{\mu_X})(\boldsymbol{X} - \boldsymbol{\mu_X})^T] = \boldsymbol{C_X}\,.$$

Per questo motivo, la matrice di covarianza possiede tutte le proprietà delle matrice di correlazione, ed in particolare è *semidefinita positiva*, ovvero

$$\boxed{\boldsymbol{a}^T\, \boldsymbol{C_X}\, \boldsymbol{a} \geq 0, \quad \forall \boldsymbol{a} \in \mathbb{R}^n\,.}$$

Ragionando analogamente a quanto fatto per la matrice di correlazione, una condizione sufficiente affinché $\boldsymbol{a}^T\, \boldsymbol{C_x}\, \boldsymbol{a} = 0$ è che gli scarti $X_1 - \mu_{X_1}, X_2 - \mu_{X_2}, \ldots, X_n - \mu_{X_n}$ siano linearmente dipendenti, nel senso che esiste un vettore $\boldsymbol{a}$ di coefficienti non tutti nulli tali che:

$$Y = a_1\,(X_1 - \mu_{X_1}) + a_2\,(X_2 - \mu_{X_2}) + \ldots + a_n\,(X_n - \mu_{X_n}) = 0\,.$$

In questo caso, almeno una variabile aleatoria può essere espressa come combinazione lineare delle rimanenti variabili aleatorie *a meno di una quantità costante*. Se ciò non accade, la matrice $\boldsymbol{C_X}$ è *definita positiva* e quindi invertibile (cfr. Appendice E).

Infine, così come vale la seguente relazione tra la covarianza e la correlazione di una coppia di variabili aleatorie (cfr. capitolo 7)

$$\mathrm{Cov}(X_i, X_j) = \mathrm{Corr}(X_i, X_j) - \mu_{X_i}\, \mu_{X_j}\,,$$

allora sussiste la seguente relazione tra le matrici di covarianza e di correlazione:

$$\boxed{\boldsymbol{C_X} = \boldsymbol{R_X} - \boldsymbol{\mu_X}\, \boldsymbol{\mu_X^T}\,.}$$

Tale relazione si può anche ricavare sviluppando la definizione (8.6) ed applicando semplici risultati di algebra lineare.

► *Esempio* 8.3 *(matrice di covarianza di una coppia di variabili aleatorie).* Per $n = 2$, possiamo porre $\boldsymbol{X} = [X, Y]^T$, $\boldsymbol{\mu_X} = [\mu_X, \mu_Y]^T$, per cui la matrice di covarianza è una matrice 2×2, data da

$$\boldsymbol{C_X} = \begin{pmatrix} \mathrm{Cov}(X,X) & \mathrm{Cov}(X,Y) \\ \mathrm{Cov}(Y,X) & \mathrm{Cov}(Y,Y) \end{pmatrix} = \begin{pmatrix} \sigma_X^2 & \rho\,\sigma_X\,\sigma_Y \\ \rho\,\sigma_X\,\sigma_Y & \sigma_Y^2 \end{pmatrix} .$$

Il calcolo del determinante di tale matrice fornisce:

$$\det(\boldsymbol{C_X}) = \sigma_X^2\,\sigma_Y^2(1-\rho_{XY}^2) \geq 0$$

in quanto $|\rho_{XY}| \leq 1$. Si verifica allora facilmente che il determinante è diverso da zero, e quindi la matrice è *definita positiva*, se e solo se $\rho_{XY} \neq \pm 1$; viceversa, esso si annulla, e quindi la matrice è solo *semidefinita positiva*, se e solo se $\rho_{XY} = \pm 1$, il che accade se $Y = a\,X + b$. Ritroviamo allora le stesse condizioni espresse nel § 7.4.4 e dimostrate attraverso la disuguaglianza di Schwartz. ◄

8.5.4 Incorrelazione

Concludiamo questa sezione estendendo il concetto di *incorrelazione* ad un vettore di variabili aleatorie:

Definizione (incorrelazione). Le variabili aleatorie $X_1, X_2, \ldots, X_n$ si dicono incorrelate se $\mathrm{Cov}(X_i, X_j) = 0, \forall i \neq j$.

Notiamo che la condizione di incorrelazione è equivalente al fatto che la matrice di covarianza $\boldsymbol{C_X}$ è *diagonale*. Notiamo altresì che l'indipendenza tra le variabili aleatorie $X_1, X_2, \ldots, X_n$ implica l'incorrelazione; in realtà, poichè la correlazione è un momento del *secondo* ordine, è sufficiente, in luogo dell'indipendenza, l'indipendenza *a coppie*, basta cioè che si fattorizzi la pdf congiunta (del secondo ordine) di qualunque coppia di variabili aleatorie. Viceversa, l'incorrelazione non implica l'indipendenza, e neppure l'indipendenza a coppie, salvo nel caso di vettori di variabili aleatorie congiuntamente gaussiane, come vedremo nel seguito.

Un'altra importante conseguenza della proprietà di incorrelazione è che, per variabili aleatorie $X_1, X_2, \ldots, X_n$ incorrelate, risulta:

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

Prova. Sviluppando la definizione di varianza, si ha:

$$\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) = E\left\{\left[\sum_{i=1}^{n}(X_i - \mu_{X_i})\right]^2\right\} = \sum_{i=1}^{n}\sum_{j=1}^{n} E[(X_i - \mu_{X_i})(X_j - \mu_{X_j})] = \sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{Cov}(X_i, X_j)\,.$$

Se le variabili aleatorie sono incorrelate, allora $\mathrm{Cov}(X_i, X_j) = 0$ per $i \neq j$, mentre $\mathrm{Cov}(X_i, X_i) = \mathrm{Var}(X_i)$; in tal caso, la doppia sommatoria si riduce ad una singola sommatoria, per cui si ha l'asserto. □

► *Esempio* 8.4 *(varianza di una variabile aleatoria binomiale).* Come applicazione del precedente risultato, osserviamo che la varianza di una variabile aleatoria $X \sim \mathrm{B}(n, p)$ è pari a $\mathrm{Var}(X) = n\,p\,q$, dove $q = 1 - p$. Infatti, abbiamo già osservato (cfr. esempio 8.2) che una variabile aleatoria binomiale si può esprimere come somma di n variabili aleatorie bernoulliane X_i indipendenti: poichè l'indipendenza implica l'incorrelazione,

tali variabili aleatorie bernoulliane saranno anche incorrelate, e quindi, poiché la varianza di una variabile aleatoria $X_i \sim \text{Bern}(p)$ è pari a $\text{Var}(X_i) = p\,q$, si ha:

$$\text{Var}(X) = \text{Var}\left(\sum_{i=1}^{n} X_i\right) = \sum_{i=1}^{n} \text{Var}(X_i) = n\,p\,q\,.$$

come annunciato. ◄

► *Esempio* 8.5 *(n variabili aleatorie congiuntamente gaussiane).* Un esempio particolarmente importante di n variabili aleatorie è la generalizzazione del concetto di coppie di variabili aleatorie congiuntamente gaussiane al caso n-dimensionale.

Le variabili aleatorie $\boldsymbol{X} = [X_1, X_2, \ldots, X_n]^T$ si dicono *congiuntamente gaussiane* se la loro pdf congiunta ammette la seguente espressione:

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}\,[\det(\boldsymbol{C_X})]^{1/2}}\ \exp\left[-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu_X})^T \boldsymbol{C_X}^{-1}(\boldsymbol{x}-\boldsymbol{\mu_X})\right]\,, \tag{8.7}$$

dove $\boldsymbol{C_X}$ è la matrice di covarianza di $\boldsymbol{X}$ e $\boldsymbol{\mu_X}$ è il vettore delle medie di $\boldsymbol{X}$. Notiamo che, affinché la (8.7) abbia significato, occorre che: (i) il determinante della matrice di covarianza $\boldsymbol{C_X}$, del quale si calcola la radice, sia non negativo; ciò è garantito dalla natura *semidefinita positiva* della matrice di covarianza (cfr. Appendice E); (ii) l'inversa della matrice di covarianza esista; ciò è garantito se la matrice di covarianza è *definita positiva*, il che non è vero sempre, ma solo se gli scarti $X_1 - \mu_{X_1}, X_2 - \mu_{X_2}, \ldots, X_n - \mu_{X_n}$ sono linearmente indipendenti, ipotesi che assumeremo senz'altro vera nel seguito. Notiamo che tale ipotesi assicura anche che $\det(\boldsymbol{C_X}) > 0$.

La pdf (8.7) di un vettore $\boldsymbol{X}$ di n variabili aleatorie congiuntamente gaussiane dipende solo dal vettore delle medie $\boldsymbol{\mu_X}$ e dalla matrice di covarianza $\boldsymbol{C_X}$, per cui si denota sinteticamente $\boldsymbol{X} \sim \text{N}(\boldsymbol{\mu_X}, \boldsymbol{C_X})$.

Possiamo osservare che la definizione (8.7) è consistente con i risultati già noti per il caso $n = 1$ ed $n = 2$. Infatti, per $n = 1$, possiamo porre $\boldsymbol{X} = X$, $\boldsymbol{\mu_X} = \mu_X$, $\boldsymbol{C_X} = E[(X-\mu_X)^2] = \sigma_X^2$, e quindi $\boldsymbol{C_X}^{-1} = 1/\sigma_X^2$ e $\det(\boldsymbol{C_X}) = \sigma_X^2$, per cui la (8.7) si riduce a:

$$f_X(x) = \frac{1}{\sigma_X\sqrt{2\pi}}\, e^{-\frac{1}{2\sigma_X^2}(x-\mu_X)^2}\,.$$

Per $n = 2$, possiamo porre $\boldsymbol{X} = [X, Y]^T$, $\boldsymbol{\mu_X} = [\mu_X, \mu_Y]^T$, mentre la matrice di covarianza è una matrice 2×2, data da (cfr. esempio 8.3)

$$\boldsymbol{C_X} = \begin{pmatrix} \text{Cov}(X,X) & \text{Cov}(X,Y) \\ \text{Cov}(Y,X) & \text{Cov}(Y,Y) \end{pmatrix} = \begin{pmatrix} \sigma_X^2 & \rho\,\sigma_X\,\sigma_Y \\ \rho\,\sigma_X\,\sigma_Y & \sigma_Y^2 \end{pmatrix}\,,$$

per cui la sua inversa si calcola facilmente (vedi Appendice E) come:

$$\boldsymbol{C_X}^{-1} = \frac{1}{\det(\boldsymbol{C_X})} \begin{pmatrix} \sigma_Y^2 & -\rho\,\sigma_X\,\sigma_Y \\ -\rho\,\sigma_X\,\sigma_Y & \sigma_X^2 \end{pmatrix}\,,$$

dove $\det(\boldsymbol{C_X}) = \sigma_X^2\sigma_Y^2(1-\rho^2)$. Sostituendo l'espressione di $\boldsymbol{C_X}$, dell'inversa e del determinante nella (8.7), si ottiene una pdf bidimensionale di tipo gaussiano, come espressa dalla (6.5).

Le principali proprietà delle variabili aleatorie congiuntamente gaussiane sono le seguenti:

1. *Se n variabili aleatorie sono congiuntamente gaussiane, allora qualsiasi sottoinsieme composto da $k < n$ tra queste variabili aleatorie sono ancora congiuntamente gaussiane. In particolare, le n variabili aleatorie sono anche marginalmente gaussiane.*

   Per provare questo risultato, bisognerebbe dimostrare che integrando la pdf gaussiana rispetto a $n-k$ variabili arbitrarie si ottiene ancora una pdf gaussiana, un calcolo laborioso in generale; ricordiamo solo che un calcolo simile è stato effettuato nell'esempio 6.2 per dimostrare che una coppia di variabili aleatorie congiuntamente gaussiane sono anche marginalmente gaussiane. Ciò risulta vero ovviamente anche nel caso di n variabili aleatorie: in particolare, X_1 è *marginalmente gaussiana*, di parametri μ_{X_1} e σ_{X_1}, ed analogamente per $X_2, X_3, \ldots, X_n$. Notiamo che il viceversa non è vero: n variabili aleatorie marginalmente gaussiane non sono necessariamente anche congiuntamente gaussiane, *salvo nel caso in cui siano indipendenti* (vedi proprietà 2).

2. *Se n variabili aleatorie marginalmente gaussiane sono anche indipendenti, allora esse sono anche congiuntamente gaussiane.*

   Prova. Poiché $X_i \sim \mathrm{N}(\mu_{X_i}, \sigma_{X_i})$, $i = 1, 2, \ldots, n$, e le X_i sono indipendenti, la pdf congiunta di $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)^T$ sarà:

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \prod_{i=1}^{n} f_{X_i}(x_i) = \prod_{i=1}^{n} \frac{1}{\sigma_{X_i}\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma_{X_i}^2}(x_i - \mu_{X_i})^2\right]$$
$$= \frac{1}{(2\pi)^{n/2}\prod_{i=1}^{n}\sigma_{X_i}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\frac{(x_i - \mu_{X_i})^2}{\sigma_{X_i}^2}\right]$$

   Confrontando con la (8.7), si vede che questa è l'espressione della pdf di n variabili aleatorie congiuntamente gaussiane, a patto che risulti

$$\det(\boldsymbol{C}_{\boldsymbol{X}}) = \sigma_{X_1}^2\,\sigma_{X_2}^2\,\cdots\sigma_{X_n}^2\,,$$
$$\boldsymbol{C}_{\boldsymbol{X}}^{-1} = \mathrm{diag}(1/\sigma_{X_1}^2, 1/\sigma_{X_2}^2, \ldots, 1/\sigma_{X_n}^2)\,.$$

   D'altra parte, l'indipendenza delle variabili aleatorie $X_1, X_2, \ldots, X_n$ implica che esse sono anche incorrelate, e quindi la loro matrice di covarianza è effettivamente diagonale, ovvero $\boldsymbol{C}_{\boldsymbol{X}} = \mathrm{diag}(\sigma_{X_1}^2, \sigma_{X_2}^2, \ldots, \sigma_{X_n}^2)$, per cui, tenendo conto delle proprietà delle matrici diagonali (cfr. Appendice E), le precedenti relazioni sono vere e la proprietà è dimostrata. □

3. *Se n variabili aleatorie congiuntamente gaussiane sono incorrelate, allora esse sono indipendenti.*

   Prova. La dimostrazione ricalca, con un ordine logico differente, quella della precedente proprietà. Infatti, se le variabili aleatorie $X_1, X_2, \ldots, X_n$ sono incorrelate, la loro matrice di covarianza risulta diagonale, ovvero $\boldsymbol{C}_{\boldsymbol{X}} = \mathrm{diag}(\sigma_{X_1}^2, \sigma_{X_2}^2, \ldots, \sigma_{X_n}^2)$. In tal caso, l'inversa è ancora diagonale:

$$\boldsymbol{C}_{\boldsymbol{X}}^{-1} = \mathrm{diag}(1/\sigma_{X_1}^2, 1/\sigma_{X_2}^2, \ldots, 1/\sigma_{X_n}^2)\,,$$

   ed il determinante è il prodotto dei valori della diagonale:

$$\det(\boldsymbol{C}_{\boldsymbol{X}}) = \sigma_{X_1}^2\,\sigma_{X_2}^2\,\cdots\sigma_{X_n}^2$$

   per cui la (8.7) si semplifica, riducendosi a:

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \frac{1}{(2\pi)^{n/2}\prod_{i=1}^{n}\sigma_{X_i}} \exp\left[-\frac{1}{2}\sum_{i=1}^{n}\frac{(x - \mu_{X_i})^2}{\sigma_{X_i}^2}\right] = \prod_{i=1}^{n}\frac{1}{\sigma_{X_i}\sqrt{2\pi}} \exp\left[-\frac{1}{2\sigma_{X_i}^2}(x - \mu_{X_i})^2\right],$$

   cioè al prodotto delle pdf marginali, per cui le $X_1, X_2, \ldots, X_n$ sono indipendenti. □

   Il risultato ottenuto, in sintesi, afferma che per le variabili aleatorie gaussiane l'incorrelazione implica l'indipendenza; poichè in generale l'indipendenza implica l'incorrelazione, possiamo concludere che, per le variabili aleatorie gaussiane, l'incorrelazione è del tutto *equivalente* all'indipendenza.[4]

4. *Una trasformazione lineare trasforma vettori gaussiani in vettori gaussiani*

   Questa è probabilmente la proprietà più importante delle variabili aleatorie congiuntamente gaussiane, e viene anche denominata *proprietà di chiusura* delle variabili aleatorie rispetto alle trasformazioni lineari.

[4] Una conseguenza errata che si potrebbe trarre, collegando impropriamente le proprietà 2 e 3, è la seguente: *se n variabili aleatorie sono marginalmente gaussiane ed incorrelate, allora esse sono indipendenti e quindi anche congiuntamente gaussiane.* Si invita il lettore ad individuare il punto debole del precedente ragionamento.

Prova. Per semplicità, dimostreremo la proprietà solo nel caso di trasformazioni lineari "quadrate" e non singolari. Sia $\boldsymbol{X}$ un vettore gaussiano, e consideriamo la trasformazione lineare (non omogenea)

$$\boldsymbol{Y} = \boldsymbol{A}\boldsymbol{X} + \boldsymbol{b}\,,$$

dove $\boldsymbol{A}$ è una matrice quadrata $n \times n$, non singolare (e quindi invertibile), e $\boldsymbol{b}$ è un vettore colonna n-dimensionale, cosicché il vettore $\boldsymbol{Y}$ è n-dimensionale.

Il calcolo della pdf di $\boldsymbol{Y}$ si ottiene facilmente applicando il teorema fondamentale sulle trasformazioni (si noti che questa è una generalizzazione dell'esempio 8.1). La soluzione del sistema numerico $\boldsymbol{y} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$ è unica, ed è data da:

$$\boldsymbol{x} = \boldsymbol{A}^{-1}(\boldsymbol{y} - \boldsymbol{b})\,,$$

mentre la matrice jacobiana $\boldsymbol{J}$ della trasformazione è pari ad $\boldsymbol{A}$, per cui la pdf del vettore $\boldsymbol{Y}$ si scrive come:

$$f_{\boldsymbol{Y}}(\boldsymbol{y}) = \frac{1}{|\det(\boldsymbol{A})|}\, f_{\boldsymbol{X}}[\boldsymbol{A}^{-1}(\boldsymbol{y} - \boldsymbol{b})]\,.$$

Sostituendo l'espressione di $f_{\boldsymbol{X}}(\boldsymbol{x})$ data dalla (8.7), e tenendo conto che

$$\boldsymbol{x} - \boldsymbol{\mu}_{\boldsymbol{X}} = \boldsymbol{A}^{-1}(\boldsymbol{y} - \boldsymbol{b}) - \boldsymbol{\mu}_{\boldsymbol{X}} = \boldsymbol{A}^{-1}(\boldsymbol{y} - \boldsymbol{b} - \boldsymbol{A}\,\boldsymbol{\mu}_{\boldsymbol{X}})\,,$$

si trova:

$$f_{\boldsymbol{Y}}(\boldsymbol{Y}) = \frac{1}{(2\pi)^{n/2}\det(\boldsymbol{C}_{\boldsymbol{X}})^{1/2}|\det(\boldsymbol{A})|}\, e^{-\frac{1}{2}(\boldsymbol{y}-\boldsymbol{b}-\boldsymbol{A}\,\boldsymbol{\mu}_{\boldsymbol{X}})^T(\boldsymbol{A}^{-1})^T\,\boldsymbol{C}_{\boldsymbol{X}}^{-1}\boldsymbol{A}^{-1}(\boldsymbol{y}-\boldsymbol{b}-\boldsymbol{A}\,\boldsymbol{\mu}_{\boldsymbol{X}})}\,,$$

che per ispezione si riconosce essere la pdf di un vettore di variabili aleatorie gaussiane di media $\boldsymbol{\mu}_{\boldsymbol{Y}} = \boldsymbol{A}\,\boldsymbol{\mu}_{\boldsymbol{X}} + \boldsymbol{b}$ e con matrice di covarianza $\boldsymbol{C}_{\boldsymbol{Y}} = \boldsymbol{A}\,\boldsymbol{C}_{\boldsymbol{X}}\,\boldsymbol{A}^T$. Infatti, si osservi che $\boldsymbol{C}_{\boldsymbol{Y}}^{-1} = (\boldsymbol{A}^{-1})^T\,\boldsymbol{C}_{\boldsymbol{X}}^{-1}\,\boldsymbol{A}^{-1}$ e $\det(\boldsymbol{C}_{\boldsymbol{Y}}) = \det(\boldsymbol{C}_{\boldsymbol{X}})\det(\boldsymbol{A})^2$. □

L'uso appropriato delle proprietà delle variabili aleatorie gaussiane consente di semplificare i calcoli in molti casi. La proprietà di chiusura, in particolare, consente di semplificare notevolmente la determinazione di talune probabilità che coinvolgono più variabili aleatorie congiuntamente gaussiane.

Ad esempio, si consideri il seguente problema: siano (X, Y, Z) tre variabili aleatorie marginalmente gaussiane, indipendenti, con medie nulle e deviazioni standard $2\,\sigma_X = \sigma_Y = \sigma_Z = 1$, e si vuole calcolare $P(X > Y + Z)$. In generale, bisognerebbe calcolare l'integrale triplo della pdf $f_{XYZ}(x, y, z)$ (fattorizzabile, per l'indipendenza) sul dominio $D = \{(x, y, z) \in \mathbb{R}^3$ tali che $x > y + z\}$. Invece, osservando che

$$P(X > Y + Z) = P(X - Y - Z > 0)$$

e tenendo conto delle proprietà 2 e 4, notiamo che la variabile aleatoria $W = X - Y - Z$, essendo ottenuta per combinazione lineare di variabili aleatorie congiuntamente gaussiane, è essa stessa gaussiana, con media $E(W) = E(X) - E(Y) - E(Z) = 0$ e varianza $\mathrm{Var}(X) = \mathrm{Var}(X) + \mathrm{Var}(Y) + \mathrm{Var}(Z) = 0.5 + 1 + 1 = 2.5$, per cui la probabilità richiesta è quella che la variabile aleatoria W, a media nulla e con pdf pari, assuma valori positivi, che vale $1/2$ per banali considerazioni di simmetria. In casi più complicati (ad esempio, se le variabili aleatorie non hanno tutte media nulla) è possibile sempre ricondursi al calcolo di una probabilità che coinvolge solo la variabile aleatoria W, probabilità che può comunque essere espressa in termini della funzione $\mathbb{G}(x)$. ◄

8.6 Teoremi limite e convergenza di una sequenza di variabili aleatorie★

Concludiamo il capitolo introducendo due fondamentali teoremi, denominati *teoremi limite* perchè descrivono il comportamento al limite (per $n \to \infty$) di una sequenza di variabili aleatorie $X_1, X_2, \ldots, X_n$. I due teoremi sono la *legge dei grandi numeri* (nella forma *debole* e *forte*) ed il *teorema limite fondamentale*. Vedremo che tali teoremi definiscono varie forme di *convergenza* associate alla sequenza di variabili aleatorie $X_1, X_2, \ldots, X_n$.

8.6.1 Legge dei grandi numeri

Teorema 8.3 (legge debole dei grandi numeri). Siano $X_1, X_2, \ldots, X_n$ variabili aleatorie indipendenti, aventi la stessa media $E(X_k) = \mu$ e la stessa varianza $\mathrm{Var}(X_k) = \sigma^2 < \infty$, e si consideri la variabile aleatoria

$$\widehat{\mu}_n \triangleq \frac{1}{n} \sum_{k=1}^{n} X_k \,.$$

Si ha:

$$\lim_{n\to\infty} P(|\widehat{\mu}_n - \mu| < \varepsilon) = 1 \,, \qquad \forall \varepsilon > 0 \,. \tag{8.8}$$

Prova. Iniziamo col calcolare media e varianza di $\widehat{\mu}_n$. Per la media, utilizzando la linearità, si ha:

$$E(\widehat{\mu}_n) = \frac{1}{n} \sum_{k=1}^{n} E(X_k) = \frac{1}{n} n\,\mu = \mu \,;$$

inoltre, sfruttando le proprietà della varianza e l'ipotesi di indipendenza (che implica l'incorrelazione), si ha:

$$\mathrm{Var}(\widehat{\mu}_n) = \frac{1}{n^2} \sum_{k=1}^{n} \mathrm{Var}(X_k) = \frac{1}{n^2} n\,\sigma^2 = \frac{\sigma^2}{n} \,.$$

A questo punto il teorema è una conseguenza diretta della disuguaglianza di Chebishev:

$$P(|\widehat{\mu}_n - \mu| < \varepsilon) \geq 1 - \frac{\mathrm{Var}(\widehat{\mu}_n)}{\varepsilon^2} = 1 - \frac{\sigma^2}{n\,\varepsilon^2}$$

Al tendere di $n \to \infty$ si ha $P(|\widehat{\mu}_n - \mu| < \varepsilon) \geq 1$ e quindi necessariamente $P(|\widehat{\mu}_n - \mu| < \varepsilon) = 1$. □

Dal punto di vista matematico, il teorema[5] afferma in sostanza che la probabilità di avere $\widehat{\mu}_n$ prossimo a piacere a μ tende ad 1 al tendere di $n \to \infty$, e pertanto che $\widehat{\mu}_n$ *converge* a μ per $n \to \infty$. La convergenza definita dalla (8.8) viene chiamata *convergenza in probabilità*; la legge dei grandi numeri si dice *debole* perchè la convergenza in probabilità è una forma *debole* di convergenza, vale a dire che non richiede condizioni molto restrittive. Notiamo inoltre che l'assunzione di indipendenza tra le variabili aleatorie $X_1, X_2, \ldots, X_n$ non è in realtà richiesta per la dimostrazione del teorema, in quanto è sufficiente l'incorrelazione.

Proviamo ora dare una interpretazione "intuitiva" della legge dei grandi numeri. La quantità $\widehat{\mu}_n$ rappresenta la media aritmetica delle variabili aleatorie $X_1, X_2, \ldots, X_n$; se interpretiamo $X_1, X_2, \ldots, X_n$, anziché come variabili aleatorie, come *valori osservati*, allora la $\widehat{\mu}_n$ rappresenta la media *empirica* delle osservazioni. Nell'ipotesi che le osservazioni siano ripetute nelle medesime condizioni, si osserva sperimentalmente che la media empirica, al divergere del numero delle osservazioni, presenta fluttuazioni sempre meno marcate, convergendo ad un valore costante. La legge dei grandi numeri afferma, in sostanza, che tale valore costante è la media delle variabili aleatorie $X_1, X_2, \ldots, X_n$ associate alle osservazioni.

Un altra interpretazione della legge dei grandi numeri consiste nel riguardare $\widehat{\mu}_n$ come uno *stimatore* della media μ delle variabili aleatorie. Il fatto che $E(\widehat{\mu}_n) = \mu$ si esprime dicendo che lo stimatore è *non polarizzato*, ovvero che non si commette un errore sistematico di stima, almeno in media. È desiderabile che uno stimatore sia non polarizzato, ma evidentemente questa condizione da sola non caratterizza la bontà dello stimatore: infatti, un buon stimatore dovrà presentare una varianza piccola, e tendente a zero al divergere di n; tale proprietà è chiamata *consistenza*. Notiamo che la legge debole dei grandi numeri esprime proprio il fatto che lo stimatore $\widehat{\mu}_n$ è

[5] Il teorema è stato enunciato e dimostrato per la prima volta dal matematico svizzero J. Bernoulli (1654–1705) nel trattato "Ars Conjectandi".

consistente, in quanto la sua varianza è inversamente proporzionale ad n. Possiamo pensare ad uno stimatore consistente e non polarizzato come ad una quantità aleatoria che però, per $n \to \infty$, presenta una pdf sempre più stretta centrata intorno alla media, per cui tende a diventare una quantità deterministica.

Notiamo infine che la condizione di consistenza si può esprimere esplicitamente come segue:

$$\lim_{n\to\infty} \mathrm{Var}(\widehat{\mu}_n) = \lim_{n\to\infty} E[(\widehat{\mu}_n - \mu)^2] = 0\,,$$

per cui equivale alla *convergenza in media quadratica* di $\widehat{\mu}_n$ a μ. Utilizzando la disuguaglianza di Chebishev, come si è fatto nella dimostrazione della legge debole dei grandi numeri, si prova facilmente che la convergenza in media quadratica *implica* quella in probabilità, il che giustifica anche il motivo per cui la convergenza in probabilità è ritenuta una forma debole di convergenza.

È possibile dimostrare che $\widehat{\mu}_n$ converge a μ in un senso più *forte* di quello espresso dalla legge debole dei grandi numeri. Tale risultato è stato dimostrato dal matematico francese E. Borel (1871-1956) e prende il nome di *legge forte dei grandi numeri*:

Teorema 8.4 (legge forte dei grandi numeri). Siano $X_1, X_2, \ldots, X_n$ variabili aleatorie indipendenti, aventi la stessa media $E(X_k) = \mu$ e la stessa varianza $\mathrm{Var}(X_k) = \sigma^2 < \infty$, e si consideri la variabile aleatoria

$$\widehat{\mu}_n \triangleq \frac{1}{n}\sum_{k=1}^{n} X_k\,.$$

Si ha:

$$P\left(\lim_{n\to\infty} |\widehat{\mu}_n - \mu| < \varepsilon\right) = 1\,, \qquad \forall \varepsilon > 0\,. \tag{8.9}$$

La tesi (8.9) della legge forte dei grandi numeri (per una dimostrazione si veda ad esempio [1]) sembra quasi identica a quella della legge debole, ma fondamentale è lo scambio tra l'operazione di limite e la probabilità (si confrontino la (8.8) e la (8.9)) La convergenza definita dalla (8.9) è un tipo di convergenza più forte della convergenza in media quadratica o in probabilità, e prende il nome di *convergenza con probabilità* 1 o *convergenza quasi certa* o *quasi ovunque*. È altresì interessante notare che la legge forte dei grandi numeri vale nelle stesse ipotesi della legge debole.

Le leggi dei grandi numeri (sia la versione forte che quella debole) descrivono il comportamento della media aritmetica di n variabili aleatorie al divergere di n. È tuttavia semplice verificare che tali leggi possono applicarsi anche per dimostrare che la frequenza di successo di un evento A in n prove indipendenti e ripetute sotto identiche condizioni *tende* alla probabilità $p = p(A)$ dell'evento al divergere di n (si noti che questo problema è esattamente quello delle *prove ripetute*, affrontato nel § 3.5.2). Per mostrare ciò, è sufficiente associare, ad ogni ripetizione dell'esperimento, la variabile aleatoria indicatrice dell'evento A, definita come:

$$X_k = \begin{cases} 1, & \text{se } A \text{ si verifica nella } k\text{-esima ripetizione;} \\ 0, & \text{altrimenti.} \end{cases}$$

È chiaro che le variabili aleatorie così definite sono $X_k \sim \mathrm{Bern}(p)$, e inoltre sono iid. Si ha, per le proprietà delle variabili aleatorie bernoulliane:

$$E(X_k) = p\,,$$
$$\mathrm{Var}(X_k) = p\,q\,.$$

Osserviamo che in questo caso la media aritmetica delle variabili aleatorie $X_1, X_2, \ldots, X_n$, ovvero

$$\widehat{p}_n = \frac{1}{n}\sum_{k=1}^{n} X_k$$

rappresenta proprio la *frequenza di successo* dell'evento A nelle n prove. Notiamo, poi, che:

$$E(\widehat{p}_n) = p\,, \tag{8.10}$$

$$\mathrm{Var}(\widehat{p}_n) = \frac{p\,q}{n}\,; \tag{8.11}$$

per cui possiamo affermare che $\widehat{p}_n$ è uno stimatore non polarizzato e consistente della probabilità p.

Poichè ci troviamo esattamente nelle ipotesi delle leggi dei grandi numeri, possiamo affermare che $\widehat{p}_n$ tende a p, al divergere di n, sia in probabilità (legge debole) che con probabilità 1 (legge forte). Questo risultato è di grande importanza, in quanto costituisce il legame tra la teoria assiomatica della probabilità e l'interpretazione frequentista.

► *Esempio* 8.6. Un'applicazione estremamente importante della legge dei grandi numeri è la seguente. Supponiamo di voler stimare la probabilità p di un evento A con una certa affidabilità: sappiamo che per $n \to \infty$ l'affidabilità può essere migliorata a piacere, ma vogliamo avere un'indicazione su quale dev'essere il valore *effettivo* di n per avere un determinato livello di affidabilità.

Come misura di affidabilità potremmo prendere la varianza (8.11) dello stimatore $\widehat{p}_n$; tuttavia notiamo che una misura *assoluta* di affidabilità non ha molto senso: infatti un errore di 0.01 su una probabilità di 0.3 potrebbe essere considerato trascurabile, lo stesso errore su una probabilità di 0.03 è inaccettabile! È allora più significativo considerare una misura *relativa*, ottenuta normalizzando la varianza al valore da stimare. Più precisamente, poiché la varianza è un momento quadratico, è opportuno normalizzare la sua radice (la deviazione standard) alla media dello stimatore, in modo da avere due quantità dimensionalmente omogenee. Si ottiene così la seguente misura di qualità, denominata *coefficiente di variazione* di $\widehat{p}_n$:

$$\chi = \frac{\sqrt{\mathrm{Var}(\widehat{p}_n)}}{E(\widehat{p}_n)}\,.$$

Sostituendo i valori dati dalle (8.10) e (8.11), si trova:

$$\chi = \frac{1}{p}\sqrt{\frac{p\,q}{n}} = \sqrt{\frac{q}{n\,p}} = \sqrt{\frac{1-p}{n\,p}}\,.$$

Un caso che spesso si presenta è quello in cui l'evento A è poco probabile, ovvero $p \ll 1$, per cui possiamo approssimare nella precedente relazione $1 - p \approx 1$ e scrivere:

$$\chi \approx \frac{1}{\sqrt{np}}\,.$$

Se allora imponiamo che il coefficiente di variazione sia 0.1, corrispondente ad un errore relativo del 10% (non eccezionalmente piccolo, ma sufficiente in molte applicazioni), troviamo:

$$n = \frac{100}{p}\,, \tag{8.12}$$

cioè il numero di prove deve eccedere di due ordini di grandezza l'inverso della probabilità da stimare. Ad esempio, se $p = 10^{-2}$, allora $n = 10^4$, e così via. La (8.12) è una regola pratica molto utilizzata per determinare il numero di prove da effettuare negli esperimenti di simulazione.

Un problema che può sorgere in pratica è il seguente: poichè non conosciamo in anticipo p, come facciamo a determinare a priori il numero di prove da effettuare? Osserviamo che se effettuiamo n prove, e l'evento A si verifica k volte, allora $\widehat{p}_n = \frac{k}{n}$. Sostituendo $\widehat{p}_n$ in luogo di p nella (8.12), troviamo $k = 100$. Questo significa che per avere l'affidabilità desiderata l'evento A si deve verificare almeno 100 volte. Pertanto, sebbene non sappiamo calcolare a priori il numero di prove da effettuare, abbiamo una condizione di "arresto" del nostro algoritmo: *ripetere l'esperimento finché l'evento A non si è verificato* 100 *volte.* Se l'evento A è poco probabile, questo può significare che dobbiamo effettuare un numero molto elevato di prove. ◄

8.6.2 Teorema limite fondamentale

Le legge dei grandi numeri, sia nella versione forte che in quella debole, afferma che la media aritmetica $\widehat{\mu}_n$ converge a quella statistica μ al crescere di n. Abbiamo visto, inoltre, che essa può essere applicata per dimostrare che la frequenza di successo $\widehat{p}_n$ converge alla probabilità p al crescere di n. L'importanza, teorica ed applicativa, di tale risultato è enorme, come abbiamo discusso nel precedente paragrafo; inoltre, conoscendo la varianza di $\widehat{\mu}_n$, ed applicando la disuguaglianza di Chebishev, possiamo maggiorare la probabilità che $\widehat{\mu}_n$ si discosti arbitrariamente da μ, in quanto si ha:

$$P(|\widehat{\mu}_n - \mu| \geq \varepsilon) \leq \frac{\mathrm{Var}(\widehat{\mu}_n)}{\varepsilon^2} = \frac{\sigma^2}{n\,\varepsilon^2}\,. \tag{8.13}$$

Tuttavia nel § 5.5 abbiamo verificato che la disuguaglianza di Chebishev fornisce risultati anche assai lontani dal vero, cioè non è una disuguaglianza stretta. Per una valutazione più accurata della probabilità (8.13), allora, bisogna necessariamente conoscere la CDF di $\widehat{\mu}_n$, eventualmente per valori elevati di n (CDF "asintotica"). La determinazione di tale CDF è l'oggetto proprio dal *teorema limite fondamentale*,[6] che fornisce un risultato per certi versi sorprendente: la CDF di $\widehat{\mu}_n$, per $n \to \infty$, tende a diventare *gaussiana*, indipendentemente dalle CDF delle variabili aleatorie $X_1, X_2, \ldots, X_n$; ciò giustifica l'enfasi che abbiamo dato alle variabili aleatorie gaussiane durante tutta la nostra trattazione.

Teorema 8.5 (teorema limite fondamentale). Siano $X_1, X_2, \ldots, X_n$ variabili aleatorie indipendenti, aventi la stessa media $E(X_k) = \mu$ e la stessa varianza $\mathrm{Var}(X_k) = \sigma^2 < \infty$, e si consideri la variabile aleatoria

$$\widehat{\mu}_n \triangleq \frac{1}{n}\sum_{k=1}^{n} X_k$$

e la sua versione normalizzata (a media nulla e varianza unitaria)

$$Z_n = \frac{\widehat{\mu}_n - E(\widehat{\mu}_n)}{\sqrt{\mathrm{Var}(\widehat{\mu}_n)}}\,.$$

Detta $F_n(x)$ la CDF di Z_n, si ha:

$$\lim_{n\to\infty} F_n(x) = \mathbb{G}(x) = \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{u^2}{2}}\,du\,.$$

ovvero, per $n \to \infty$, Z_n ha la CDF di una variabile aleatoria $Z \sim \mathrm{N}(0,1)$ (normale standard).

Dal punto di vista matematico, notiamo che anche il teorema limite fondamentale esprime una forma di convergenza della sequenza di variabili aleatorie $Z_1, Z_2, \ldots, Z_n$ alla variabile aleatoria $Z \sim \mathrm{N}(0,1)$; poichè tale convergenza è in realtà una convergenza delle rispettive CDF, viene chiamata *convergenza in distribuzione*, e si può provare che è un tipo di convergenza *debole*.[7]

Dal punto di vista pratico, sebbene sia stato enunciato con riferimento alla media aritmetica di n variabili aleatorie, il teorema limite fondamentale stabilisce in pratica che la somma di un

[6] In inglese, tale teorema viene denominato "Central Limit Theorem" (CLT), che spesso viene tradotto come "teorema del limite centrale". Tale terminologia è spesso impropriamente adottata in taluni testi italiani di probabilità e statistica; la traduzione corretta è, invece, quella di "teorema limite fondamentale", in quanto esso rappresenta un risultato "centrale" (nel senso, appunto, di fondamentale) dell'intera teoria della probabilità.

[7] Notiamo anche che la formulazione precedente del teorema è una formulazione *integrale*, perchè riguarda la CDF che si può esprimere come un integrale; vedremo successivamente che, sotto ipotesi più restrittive, è possibile darne anche una formulazione *puntuale* o *locale*, con riferimento cioè alla pdf.

gran numero di variabili aleatorie *indipendenti* tende ad assumere la distribuzione gaussiana. Si noti che il teorema si può generalizzare anche al caso in cui le variabili aleatorie non abbiano tutte la stessa media e la stessa varianza, mantenendo sempre l'assunzione di indipendenza. Ad esempio, se le variabili aleatorie indipendenti $X_1, X_2, \ldots, X_n$ hanno medie $E(X_k) = \mu_k$ e varianze $\mathrm{Var}(X_k) = \sigma_k^2 < \infty$, e consideriamo la somma $S_n = \sum_{k=1}^n X_k$, che ha media $E(S_n) = \sum_{k=1}^n \mu_k$ e varianza $\mathrm{Var}(S_n) = \sum_{k=1}^n \sigma_k^2$, la versione normalizzata di S_n, sia essa

$$Z_n = \frac{S_n - E(S_n)}{\sqrt{\mathrm{Var}(S_n)}},$$

tende ad assumere una distribuzione gaussiana standard, cioè $Z_n \to Z \sim N(0,1)$. Per n sufficientemente grande, notiamo che questo equivale a dire che S_n ha approssimativamente una distribuzione gaussiana con media $E(S_n)$ e varianza $\mathrm{Var}(S_n)$, e questo risultato ha una grossa rilevanza applicativa nei calcoli di probabilità riguardanti S_n, in quanto consente di sostituire alla *vera* CDF di S_n (complicata da calcolare, in generale) una CDF gaussiana con media e varianza pari a quelle di S_n.

► *Esempio* 8.7. Il teorema limite fondamentale si può applicare per ricavare la distribuzione limite della somma S_n di n variabili aleatorie iid bernoulliane, cioè $X_k \sim \mathrm{Bern}(p)$. Notiamo peraltro che è possibile calcolare esattamente tale distribuzione per ogni valore di n, in quanto risulta $S_n \sim \mathrm{B}(n,p)$, cioè tale distribuzione è quella di una variabile aleatoria binomiale. Poichè $E(X_k) = p$ e $\mathrm{Var}(X_k) = pq$, allora $E(S_n) = np$ e $\mathrm{Var}(S_n) = npq$, per cui la variabile aleatoria normalizzata si scrive:

$$Z_n = \frac{S_n - np}{\sqrt{npq}}.$$

Se allora vogliamo calcolare la probabilità che $k_1 \leq S_n \leq k_2$, per n sufficientemente grande, possiamo scrivere:

$$P(k_1 \leq S_n \leq k_2) = P\left(\frac{k_1 - np}{\sqrt{npq}} \leq \frac{S_n - np}{\sqrt{npq}} \leq \frac{k_2 - np}{\sqrt{npq}}\right) = \\ = \mathbb{G}\left(\frac{k_2 - np}{\sqrt{npq}}\right) - \mathbb{G}\left(\frac{k_1 - np}{\sqrt{npq}}\right),$$

cioè ritroviamo il teorema di de Moivre-Laplace (cfr. § 3.5.12) nella forma *integrale*, che adesso possiamo riguardare come una semplice applicazione del teorema limite fondamentale alla somma di n variabili aleatorie bernoulliane. ◄

Come accennato precedentemente, è possibile anche fornire una formulazione *puntuale* o *locale* del teorema limite fondamentale. Nelle stesse ipotesi già enunciate per la formulazione integrale, con in più l'assunzione che le variabili aleatorie $X_1, X_2, \ldots, X_n$ siano *continue*, si può mostrare che la successione di variabili aleatorie $Z_1, Z_2, \ldots, Z_n$ ha, per $n \to \infty$, la pdf di una variabile aleatoria $Z \sim \mathrm{N}(0,1)$ (normale standard), ovvero:

$$\lim_{n\to\infty} f_n(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

dove $f_n(x)$ è la pdf di Z_n.

Se le $Z_1, Z_2, \ldots, Z_n$ sono variabili aleatorie discrete, le loro pdf sono costituite da sovrapposizioni di impulsi di Dirac, per cui evidentemente $f_n(x)$ non può convergere ad una pdf ordinaria, quale quella gaussiana. Tuttavia, per variabili aleatorie discrete di tipo *reticolare*, che assumono cioè i valori $a + bk$, con $a, b \in \mathbb{R}$ e $k \in K \subseteq \mathbb{Z}$, vale un risultato molto interessante. Infatti, osserviamo che se $X_1, X_2, \ldots, X_n$ sono variabili aleatorie di tipo reticolare, anche la loro somma

$S_n = \sum_{k=1}^n X_k$ è di tipo reticolare, in quanto può assumere i valori $n\,a + b\,k$. Nell'ipotesi che le $X_1, X_2, \ldots, X_n$ siano indipendenti, con medie $\mu_k = E(X_k)$ e varianze $\mathrm{Var}(X_k) = \sigma_k^2 < \infty$, posto $\mu \triangleq E(S_n) = \sum_{k=1}^n \mu_k$ e $\sigma^2 \triangleq \mathrm{Var}(S_n) = \sum_{k=1}^n \sigma_k^2$, si ha:

$$\lim_{n\to\infty} P(S_n = a + b\,k) = \frac{1}{\sigma\sqrt{2\pi}}\, e^{-\frac{1}{2\sigma^2}(a+b\,k-\mu)^2} \tag{8.14}$$

per cui in pratica la DF della somma S_n, nei punti $x = a + b\,k$, può essere approssimata dai campioni di una pdf gaussiana, avente media e varianza uguali a quelle di S_n.

► *Esempio* 8.8. Il risultato precedente può essere applicato al caso della somma S_n di n variabili aleatorie iid bernoulliane $X_k \sim \mathrm{Bern}(p)$. Tale somma ha una distribuzione binomiale, che è di tipo reticolare, in quanto assume i valori $\{0, 1, \ldots, n\}$. Poichè si ha $E(S_n) = n\,p$ e $\mathrm{Var}(S_n) = n\,p\,q$, la (8.14) si scrive:

$$\lim_{n\to\infty} P(S_n = k) = \frac{1}{\sqrt{2\pi\, n\, p\, q}}\, e^{-\frac{(k-np)^2}{2npq}}$$

che esprime proprio il teorema di de Moivre-Laplace nella forma locale (cfr. equazione (3.12)). ◄

In conclusione, vale la pena fare qualche considerazione critica dell'utilità *pratica* del teorema limite fondamentale. In primo luogo, notiamo che la bontà dell'approssimazione gaussiana, per n finito, non è quantificabile a priori, e dipende criticamente dalla distribuzione delle variabili aleatorie $X_1, X_2, \ldots, X_n$. Ciò nonostante, nel passato anche recente il teorema limite fondamentale era ampiamente utilizzato nelle applicazioni, in quanto il calcolo della pdf di un gran numero di variabili aleatorie risultava un problema matematicamente poco trattabile. Al giorno d'oggi, la disponibilità di calcolatori sempre più veloci ha reso tale problema relativamente semplice da affrontare con tecniche numeriche, per cui l'utilità pratica del teorema limite fondamentale è diminuita. Tuttavia, per motivi teorici, esso resta uno dei risultati più importanti e noti dell'intera teoria della probabilità, e tale da giustificare l'uso della distribuzione gaussiana in tanti problemi della fisica, della statistica, e dell'ingegneria.

8.7 Esercizi proposti

Esercizio 8.1. In un ufficio postale, esistono tre sportelli ed una fila unica per tutti e tre gli sportelli. Quando il signor Rossi arriva all'ufficio, è il primo della fila, ma ciascuno degli sportelli è occupato da un cliente. Se i tempi residui di servizio T_1, T_2 e T_3 per i clienti agli sportelli sono modellabili come variabili aleatorie esponenziali indipendenti, di media 20 minuti, 10 minuti e 5 minuti, rispettivamente, calcolare:

a) la probabilità che il signor Rossi debba aspettare più di 10 minuti prima che uno degli sportelli si liberi;

b) il tempo medio di attesa del signor Rossi.

Esercizio 8.2. Siano $X_1, X_2, \ldots, X_n$ n variabili aleatorie iid, aventi ciascuna CDF $F(x)$ e pdf $f(x)$.

a) Determinare la CDF e la pdf di $Z = \max(X_1, X_2, \ldots, X_n)$;

b) Determinare la CDF e la pdf di $W = \min(X_1, X_2, \ldots, X_n)$.

[*Risposta: a)* $F_Z(z) = [F(z)]^n$, $f_Z(z) = n\,[F(z)]^{n-1}\,f(z)$; *b)* $F_W(w) = 1 - [1 - F(w)]^n$, $f_W(w) = n\,[1 - F(w)]^{n-1}\,f(w)$.]

★ *Esercizio* 8.3. Siano $X_1, X_2, \ldots, X_n$ n variabili aleatorie iid, aventi ciascuna CDF $F(x)$ e pdf $f(x)$. Determinare la pdf *congiunta* di $Z = \max(X_1, X_2, \ldots, X_n)$ e $W = \min(X_1, X_2, \ldots, X_n)$.
[*Risposta:* $F_{ZW}(z, w) = n(n-1)[F(z) - F(w)]^{n-2} f(z)\,f(w)$, *per* $z \geq w$.]

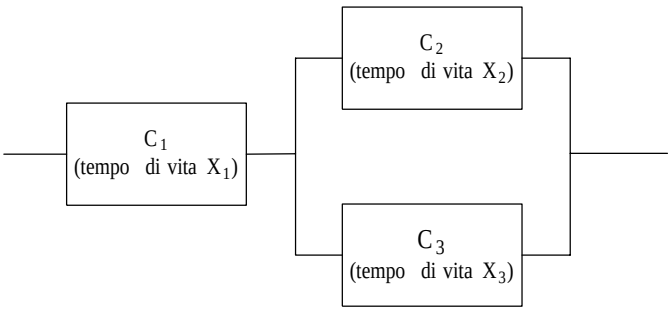
Esercizio 8.4. Il tempo di vita di una lampadina è modellabile come una variabile aleatoria $X \sim \text{Exp}(\lambda)$, con media $E(X) = 100$ (in ore). Se 10 lampadine vengono installate contemporaneamente, assumendo indipendenti i tempi di vita delle diverse lampadine, determinare la distribuzione del tempo di vita della lampadina che si esaurisce per prima e la sua durata media.

★ *Esercizio* 8.5. Il sistema indicato in figura funziona finché il componente C_1 ed almeno uno tra C_2 e C_3 funzionano. Il tempo di vita del componente C_i è modellabile come una variabile aleatoria $X_i \sim \text{Exp}(\lambda)$, con media $E(X_i) = 1$ (in anni); i tempi di vita X_1, X_2 ed X_3 sono indipendenti.

a) Determinare la distribuzione del tempo di vita Z del sistema complessivo;

b) Determinare $E(Z)$ (tempo medio di vita) e $\text{Var}(Z)$.

[*Suggerimento: esprimere il tempo di vita* Z *in funzione di* X_1, X_2 *ed* X_3 *utilizzando le funzioni* max *e* min.]



★ *Esercizio* 8.6. Si supponga che le variabili aleatorie X, Y, Z abbiano la seguente pdf:

$$f_{XYZ}(x, y, z) = \begin{cases} k, & \text{se } x^2 + y^2 + y^2 \leq 1, \\ 0, & \text{altrimenti.} \end{cases}$$

a) Determinare il valore di k;

b) determinare le pdf $f_{XY}(x, y)$ e $f_X(x)$;

c) stabilire se le variabili aleatorie X, Y, Z sono indipendenti.

[*Risposta: a)* $k = 3/(4\pi)$*; b)* $f_{XY}(x,y) = \frac{3}{2\pi}\sqrt{1-(x^2+y^2)}$*, per* $x^2+y^2 \leq 1$*;* $f_X(x) = \frac{3}{4}(1-x^2)$*, per* $|x| \leq 1$.]
[*Suggerimento: Nei calcoli per la* $f_X(x)$*, si sfrutti il seguente integrale indefinito (valido per* $|y| < a$*):* $\int \sqrt{a^2-y^2}\,dy = \frac{y\sqrt{a^2-y^2}}{2} + \frac{a^2}{2}\arcsin\left(\frac{y}{a}\right)$]

Esercizio 8.7. Siano $X_1, X_2, \ldots, X_n$ n variabili aleatorie indipendenti. Adoperando il teorema fondamentale sulle trasformazioni di variabili aleatorie, determinare la pdf di $Z = \sum_{i=1}^{n} X_i$.

Esercizio 8.8. Siano $X_1, X_2, \ldots, X_n$ n variabili aleatorie indipendenti, con $X_i \sim \mathrm{N}(\mu_{X_i}, \sigma_{X_i})$. *Senza* adoperare il teorema fondamentale sulle trasformazioni di variabili aleatorie, determinare la pdf di $Z = \sum_{i=1}^{n} X_i$.

Esercizio 8.9. Siano $X_1, X_2, \ldots, X_n$ n variabili aleatorie indipendenti ed identicamente distribuite, con pdf del primo ordine di tipo *Pareto*:

$$f_X(x) = \frac{1}{x^2}\,u(x-1)\,,$$

e sia $Y = \min(X_1, X_2, \ldots, X_n)$.

a) Determinare $E(X_i)$ (se esiste).

b) Determinare $E(Y)$ (se esiste).

★ *Esercizio* 8.10. Siano X_1, X_2, X_3, X_4 variabili aleatorie con pdf congiunta

$$f_{\boldsymbol{X}}(x_1,x_2,x_3,x_4) = e^{-x_1-x_2-x_3-x_4}\,, \quad x_1 \geq 0, x_2 \geq 0, x_3 \geq 0, x_4 \geq 0\,.$$

Si consideri la seguente trasformazione di variabili aleatorie:

$$\begin{cases} Y_1 &= X_1 \\ Y_2 &= X_2 - X_1 \\ Y_3 &= X_3 - X_2 \\ Y_4 &= X_4 - X_3 \end{cases}$$

a) Calcolare la pdf congiunta di Y_1, Y_2, Y_3, Y_4;

b) a partire dalla pdf congiunta calcolata al punto 1, calcolare successivamente la pdf di Y_1, Y_2, Y_3, quella di Y_1, Y_2, ed infine quella di Y_1.

★ *Esercizio* 8.11. Siano X_1, X_2, X_3 variabili aleatorie iid, con $X_i \sim \mathrm{Exp}(\lambda)$. Determinare la pdf congiunta delle variabili aleatorie $Y = X_2 - X_1$ e $Z = X_3 - X_1$.

Esercizio 8.12. Siano X_1, X_2, X_3 variabili aleatorie incorrelate con la stessa varianza σ^2. Determinare il coefficiente di correlazione tra $X_1 + X_2$ e $X_2 + X_3$.

Esercizio 8.13. Siano X_1, X_2, X_3 variabili aleatorie indipendenti con la stessa media μ e la stessa varianza σ^2. Determinare il coefficiente di correlazione tra $X_2 - X_1$ e $X_3 + X_1$.

Esercizio 8.14. Sia $\boldsymbol{X}$ un vettore di n variabili aleatorie con vettore delle medie $\boldsymbol{\mu_X}$, matrice di correlazione $\boldsymbol{R_X}$ e matrice di covarianza $\boldsymbol{C_X}$. Calcolare le corrispondenti grandezze per il vettore $\boldsymbol{Y} = \boldsymbol{A}\,\boldsymbol{X}$, dove $\boldsymbol{A}$ è una matrice $n \times n$.

★ *Esercizio* 8.15. Un vettore $\boldsymbol{X} = [X_1, X_2, X_3]^T$ di tre variabili aleatorie congiuntamente gaussiane, a media nulla e con matrice di covarianza:

$$\boldsymbol{C_X} = \begin{pmatrix} 4 & 2.05 & 1.05 \\ 2.05 & 4 & 2.05 \\ 1.05 & 2.05 & 4 \end{pmatrix}$$

è sottoposto alla seguente trasformazione:

$$\begin{cases} Y_1 &= 5\,X_1 + 2\,X_2 - X_3 \\ Y_2 &= -X_1 + 3\,X_2 + X_3 \\ Y_3 &= 2\,X_1 - X_2 + 2\,X_3 \end{cases}$$

Calcolare la pdf congiunta del vettore $\boldsymbol{Y} = [Y_1, Y_2, Y_3]^T$.
[*Suggerimento: la risoluzione di questo esercizio è agevolata dall'uso del calcolatore (Matlab).*]

Esercizio 8.16. Siano X_1, X_2, X_3 tre variabili aleatorie indipendenti con medie $\mu_{X_1} = 3$, $\mu_{X_2} = 6$ e $\mu_{X_3} = -2$. Calcolare la media delle seguenti variabili aleatorie:

a) $Z = X_1 + 3\,X_2 + 4\,X_3$;

b) $Z = X_1\,X_2\,X_3$;

c) $Z = -2\,X_1\,X_2 - 3\,X_1\,X_3 + 4\,X_2\,X_3$;

d) $Z = X_1 + X_2 + X_3$.

Esercizio 8.17. Tre variabili aleatorie incorrelate X_1, X_2, X_3 hanno medie $\mu_{X_1} = 1$, $\mu_{X_2} = -3$ e $\mu_{X_3} = 1.5$, e valori quadratici medi $E(X_1^2) = 2.5$, $E(X_2^2) = 11$ e $E(X_3^2) = 3.5$. Sia $Z = X_1 - 2\,X_2 + 3\,X_3$ una nuova variabile aleatoria. Determinare media e varianza di Z.

Esercizio 8.18. Si scelgono a caso ed indipendentemente l'uno dall'altro n numeri nell'intervallo $[0, 1]$.

a) Se $n = 10$, determinare la probabilità che esattamente 5 numeri scelti siano minori di 0.5.

b) Se $n = 10$, determinare in media quanti numeri sono minori di 0.5.

c) Se $n = 100$, determinare la probabilità che la media aritmetica dei numeri scelti sia compresa tra 0.49 e 0.51.

[*Suggerimento: per la risposta c), applicare il teorema limite fondamentale*]

Esercizio 8.19. Si lancia $n = 10\,000$ volte una moneta ben bilanciata. Calcolare la probabilità di ottenere un numero di teste compreso tra 4950 e 5050.
[*Risposta:* 0.683]
[*Suggerimento: applicare il teorema limite fondamentale*]

★ *Esercizio* 8.20. Si collegano in serie n spezzoni di tubo, le cui lunghezze (in metri) sono modellate come variabili aleatorie $X_1, X_2, \ldots, X_n$ discrete, indipendenti e identicamente distribuite, con $X_i \sim \text{Geom}(1/2)$. Sia L la lunghezza totale del tubo.

a) Se $n = 400$, determinare la probabilità che la lunghezza totale del tubo sia superiore a 820 metri.

b) Se $n = 400$, determinare il valore di lunghezza che viene superato con probabilità 0.841 (circa).

c) Determinare il valore di n in modo che la probabilità che la lunghezza L sia almeno pari a 200 metri sia 0.841 (circa).

[*Risposta: a)* 0.221*; b)* 772*; c)* 108.]
[*Suggerimento: applicare il teorema limite fondamentale*]

Capitolo 9

Distribuzioni e medie condizionali

In questo capitolo si riprende il concetto di probabilità condizionale, applicandolo alle variabili aleatorie per costruire le cosiddette distribuzioni (CDF, pdf o DF) condizionali. Il problema viene affrontato prima per una singola variabile aleatoria, poi per una coppia di variabili aleatorie, ed infine esteso al caso generale di n variabili aleatorie. Successivamente si introducono i momenti condizionali, tra i quali la media condizionale è il più semplice e ricorre frequentemente nelle applicazioni. Infine viene esposto il teorema della media condizionale, che rappresenta un utile strumento di calcolo per la risoluzione di numerosi problemi applicativi.

9.1 Introduzione

La funzione di distribuzione cumulativa (CDF) di una singola variabile aleatoria, di una coppia di variabili aleatorie, o più in generale di un vettore di variabili aleatorie, rappresenta in ultima analisi la probabilità di un evento, semplice o composto. Poiché la probabilità condizionale definita nel capitolo 2 è una valida legge di probabilità, ci chiediamo se sia possibile definire valide CDF anche in termini di probabilità condizionali. La risposta è naturalmente affermativa, e conduce all'introduzione delle cosiddette *distribuzioni condizionali* (CDF, pdf e DF). Tali distribuzioni condizionali[1] consentono di approfondire le relazioni esistenti tra le variabili aleatorie e gli eventi dello spazio campione su cui esse sono definite, nonché le relazioni esistenti tra le variabili aleatorie stesse. A tali distribuzioni condizionali sono associati i corrispondenti momenti, denominati *momenti condizionali*, la cui definizione si basa sul concetto fondamentale di *media condizionale*.

9.2 Distribuzioni condizionali per una variabile aleatoria

In questo paragrafo inizieremo col considerare le distribuzioni condizionali per il caso di *una* singola variabile aleatoria X.

[1] Si usa indifferentemente la terminologia "distribuzioni condizionali" o "condizionate".

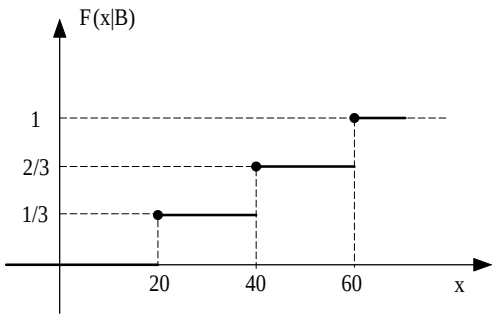


Fig. 9.1. La CDF condizionale $F(x|B)$ della variabile aleatoria dell'esempio 9.1.

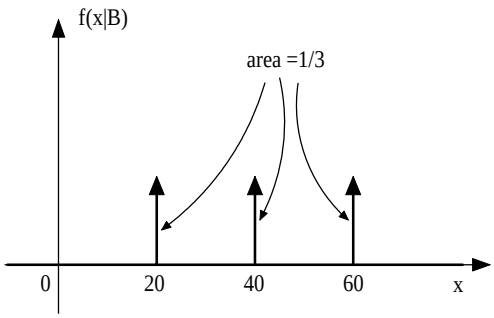


Fig. 9.2. La pdf condizionale $f(x|B)$ della variabile aleatoria dell'esempio 9.1.

9.2.1 Funzione di distribuzione cumulativa (CDF) condizionale

Ricordiamo che dati due eventi A e B, con $P(B) \neq 0$, la probabilità condizionale di A dato B (cfr. § 2.2) si definisce come:

$$P(A|B) = \frac{P(AB)}{P(B)}\,.$$

Scegliendo allora l'evento $A = \{X \leq x\}$ e B qualsiasi, con $P(B) \neq 0$, possiamo definire la *CDF condizionale* di X dato un evento B:

Definizione (CDF condizionale di una variabile aleatoria). Sia X una variabile aleatoria definita su uno spazio di probabilità $(\Omega, \mathcal{S}, P)$, e sia B un evento di Ω, con $P(B) \neq 0$. La *CDF condizionale* di X dato l'evento B è:

$$F(x|B) \triangleq P(X \leq x|B) = \frac{P(X \leq x, B)}{P(B)}\,.$$

Osserviamo subito che, poiché la probabilità condizionale è una valida legge di probabilità, ne consegue che la CDF condizionale è una valida CDF, e pertanto gode di tutte le proprietà caratteristiche della CDF di una variabile aleatoria (cfr. § 3.2.1). In particolare, essa è una funzione continua da destra, e si ha:

1. $F(+\infty|B) = 1$, $F(-\infty|B) = 0$;
2. $P(x_1 < X \leq x_2|B) = F(x_2|B) - F(x_1|B) = \dfrac{P(x_1 < X \leq x_2, B)}{P(B)}$.

Per denotare che la CDF condizionale di una variabile aleatoria, dato B, è $F(x|B)$, si scrive talvolta $X|B \sim F(x|B)$.

► *Esempio* 9.1. Consideriamo lo spazio $\Omega = \{\omega_1, \omega_2, \ldots, \omega_6\}$ (lancio di un dado) e la variabile aleatoria

$$X(\omega_i) = 10\,i\,,$$

che abbiamo già considerato nell'esempio 3.7. Sia $B = \{\text{pari}\} = \{\omega_2, \omega_4, \omega_6\}$ e calcoliamo la $F(x|B)$. Si ha:

$$x < 20 \Rightarrow \{X \leq x\} \cap B = \varnothing \Rightarrow F(x|B) = 0\,;$$

$$20 \leq x < 40 \Rightarrow \{X \leq x\} \cap B = \{\omega_2\} \Rightarrow F(x|B) = \frac{1/6}{1/2} = 1/3\,;$$

$$40 \leq x < 60 \Rightarrow \{X \leq x\} \cap B = \{\omega_2, \omega_4\} \Rightarrow F(x|B) = \frac{1/3}{1/2} = 2/3\,;$$

$$x \geq 60 \Rightarrow \{X \leq x\} \cap B = \{\omega_2, \omega_4, \omega_6\} \Rightarrow F(x|B) = \frac{1/2}{1/2} = 1\,;$$

per cui la CDF condizionale $F(x|B)$ è costante a tratti (è la CDF di una variabile aleatoria discreta) ed è mostrata in Fig. 9.1. ◄

9.2.2 Funzione densità di probabilità (pdf) condizionale

In maniera naturale, passiamo ora a definire la *pdf condizionale* di una variabile aleatoria X dato un evento B:

Definizione (pdf condizionale di una variabile aleatoria). Sia X una variabile aleatoria definita su uno spazio di probabilità $(\Omega, \mathcal{S}, P)$, e sia B un evento di Ω, con $P(B) \neq 0$. La *pdf condizionale* di X dato l'evento B è la derivata (in senso generalizzato) di $F(x|B)$ rispetto a x:

$$f(x|B) \triangleq \frac{d}{dx} F(x|B) \,.$$

Valgono per la pdf condizionale considerazioni analoghe a quelle per la CDF condizionale: poiché essa è a tutti gli effetti una pdf, gode di tutte le proprietà della pdf (cfr. § 3.3.1). In particolare, vale la proprietà di normalizzazione, cioè si ha

$$\int_{-\infty}^{\infty} f(x|B)\, dx = 1 \,.$$

► *Esempio* 9.2. Consideriamo la CDF condizionale dell'esempio 9.1. Poiché la CDF ha un andamento costante a tratti, la pdf condizionale sarà puramente impulsiva, ed è data da:

$$f(x|B) = \frac{1}{3}\,\delta(x-20) + \frac{1}{3}\,\delta(x-40) + \frac{1}{3}\,\delta(x-60) \,,$$

che è rappresentata in Fig. 9.2. ◄

9.2.3 Funzione distribuzione di probabilità (DF) condizionale

Infine, per variabili aleatorie discrete è utile definire la *DF condizionale*:

Definizione (DF condizionale di una variabile aleatoria). Sia X una variabile aleatoria discreta definita su uno spazio di probabilità $(\Omega, \mathcal{S}, P)$ e a valori in $\mathcal{X}$, e sia B un evento di Ω, con $P(B) \neq 0$. La *DF condizionale* di X dato l'evento B è

$$p(x|B) \triangleq P(X = x|B) \,,$$

con $x \in \mathcal{X}$.

Anche la DF condizionale, essendo una valida DF, gode delle proprietà caratteristiche della DF (cfr. § 3.4).

► *Esempio* 9.3. Riprendiamo l'esempio 9.1, in cui B è l'evento "pari"; poiché X è una variabile aleatoria discreta, risulta più immediato calcolare, in luogo della CDF condizionale, la DF condizionale:

$$p(x|B) = P(X = x|B) = \frac{P(X = x, B)}{P(B)} = \frac{P(X = x, B)}{1/2} \,.$$

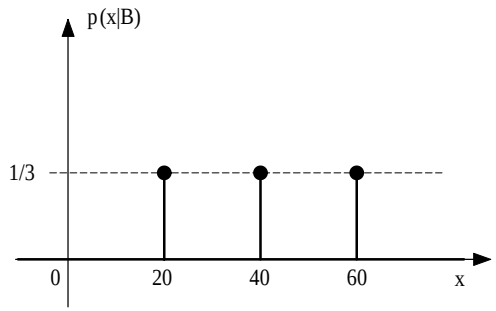


Fig. 9.3. La DF condizionale $p(x|B)$ della variabile aleatoria dell'esempio 9.1.

La variabile aleatoria X assume valori in $\mathcal{X} = \{10, 20, 30, 40, 50, 60\}$; si verifica immediatamente che per $x = 10, 30, 50$ (valori corrispondenti a risultati *dispari* dell'esperimento) risulta $P(X = x, B) = 0$, mentre per $x = 20, 40, 60$ (valori corrispondenti a risultati *pari* dell'esperimento) si ha:

$$P(X = x, B) = P(X = x) = \frac{1}{6}\,,$$

e quindi in definitiva la DF cercata è:

$$p(x|B) = \begin{cases} \frac{1}{3}\,, & x = 20, 40, 60; \\ 0\,, & \text{altrimenti,} \end{cases}$$

che è rappresentata in Fig. 9.3. ◄

Osserviamo che, per determinare la CDF, la pdf o la DF condizionale, dobbiamo in genere conoscere in dettaglio l'esperimento su cui è costruita la variabile aleatoria. Tuttavia, in molti casi si assegna una variabile aleatoria X attraverso la sua CDF, pdf o DF, senza descrivere esplicitamente l'esperimento probabilistico sul quale tale variabile aleatoria è definita. Diventa allora particolarmente interessante il caso in cui l'evento B possa essere espresso esso stesso in termini della variabile aleatoria X. Ciò accade se, ad esempio, $B = \{X \leq a\}$ oppure $B = \{X > a\}$, con $a \in \overline{\mathbb{R}}$. In casi del genere, è sufficiente la conoscenza della sola CDF $F(x)$ (o della sola pdf o DF) di X per calcolare le corrispondenti distribuzioni condizionali, e non occorre quindi conoscere l'esperimento probabilistico. Approfondiamo meglio questo concetto negli esempi che seguono.

► *Esempio* 9.4. Sia X una variabile aleatoria con CDF $F(x)$ (supposta continua) e sia $B = \{X \leq a\}$. Si ha:

$$F(x|B) = P(X \leq x | X \leq a) = \frac{P(X \leq x, X \leq a)}{P(X \leq a)}\,.$$

Se $x \geq a$, allora $\{X \leq x, X \leq a\} = \{X \leq a\}$ e quindi

$$F(x|B) = \frac{P(X \leq a)}{P(X \leq a)} = 1\,.$$

Viceversa, se $x < a$, allora $\{X \leq x, X \leq a\} = \{X \leq x\}$, e quindi

$$F(x|B) = \frac{P(X \leq x)}{P(X \leq a)} = \frac{F(x)}{F(a)}\,.$$

In definitiva, allora

$$F(x|X \leq a) = \begin{cases} \dfrac{F(x)}{F(a)}, & x < a\,; \\ 1, & x \geq a\,. \end{cases}$$

Derivando, si ottiene la pdf

$$f(x|X \leq a) = \begin{cases} \dfrac{f(x)}{F(a)}, & x < a\,; \\ 0, & x \geq a\,. \end{cases}$$

Notiamo che nella derivazione il punto $x = a$ potrebbe essere punto di discontinuità per la $F(x|B)$, e quindi potrebbe comparire un impulso di Dirac in $x = a$. Tuttavia, calcolando i limiti da sinistra e da destra della $F(x|B)$ nel punto $x = a$, si ha:

$$F(a^-|B) = \frac{1}{F(a)} F(a^-) = 1 = F(a^+|B)$$

per l'ipotesi di continuità di $F(x)$. Pertanto, $F(x|B)$ è continua e quindi la pdf condizionale non contiene impulsi. Più in generale, bisogna applicare qualche cautela nella derivazione, per non ignorare possibili discontinuità della CDF. ◄

► *Esempio* 9.5. Sia X una variabile aleatoria con CDF $F(x)$ (supposta continua), e sia $B = \{a < X \leq b\}$. Si ha:

$$F(x|B) = P(X \leq x | a < X \leq b) = \frac{P(X \leq x, a < X \leq b)}{P(a < X \leq b)}\,.$$

Se $x \geq b$, allora $\{X \leq x, a < X \leq b\} = \{a < X \leq b\}$ e quindi

$$F(x|B) = \frac{P(a < X \leq b)}{P(a < X \leq b)} = 1\,.$$

Se $a < x < b$, allora $\{X \leq x, a < X \leq b\} = \{a < X \leq x\}$, e quindi

$$F(x|B) = \frac{P(a < X \leq x)}{P(a < X \leq b)} = \frac{F(x) - F(a)}{F(b) - F(a)}\,.$$

Infine, se $x \leq a$, allora $\{X \leq x, a < X \leq b\} = \varnothing$, e quindi

$$F(x|B) = 0\,.$$

In definitiva, allora:

$$F(x|a < X \leq b) = \begin{cases} 0, & x \leq a\,; \\ \frac{F(x)-F(a)}{F(b)-F(a)}, & a < x < b\,; \\ 1, & x \geq b\,. \end{cases}$$

Derivando, si ottiene la pdf:

$$f(x|a < X \leq b) = \begin{cases} 0, & x \leq a\,; \\ \frac{f(x)}{F(b)-F(a)}, & a < x < b\,; \\ 0, & x \geq b\,. \end{cases}$$

Anche qui, i punti $x = a$ e $x = b$ potrebbero essere di discontinuità per la CDF condizionale. Tuttavia, nell'ipotesi che $F(x)$ sia continua, è facile verificare che anche $F(x|B)$ lo è, e quindi nella pdf condizionale non compaiono impulsi di Dirac.[2] ◄

9.2.4 Teorema della probabilità totale per CDF, pdf, DF

Poichè le CDF, pdf e DF condizionali si definiscono a partire dalla probabilità condizionale, tutta una serie di relazioni e teoremi (probabilità totale, Bayes, probabilità a posteriori) visti per le probabilità condizionali si possono estendere anche alle CDF, pdf e DF di variabili aleatorie. Ad esempio, ricordiamo che per il teorema della probabilità totale (vedi § 2.2.4) si ha:

$$P(B) = \sum_{i=1}^{n} P(B|A_i)\, P(A_i)\,.$$

[2] In realtà, calcolando i limiti da destra, si può facilmente verificare che la CDF condizionale è continua in $x = a$ anche se la variabile aleatoria X non è continua; viceversa, l'ipotesi che X sia una variabile aleatoria continua è indispensabile per garantire la continuità della CDF condizionale in $x = b$.

dove gli A_i sono eventi mutuamente esclusivi tali che $B \subseteq \cup_{i=1}^{n} A_i$. Scegliendo allora $B = \{X \leq x\}$, si ha $P(B) = P(X \leq x) = F(x)$ e $P(B|A_i) = P(X \leq x|A_i) = F(x|A_i)$, per cui si ottiene il *teorema della probabilità totale per la CDF*:

$$\boxed{F(x) = \sum_{i=1}^{n} F(x|A_i)\, P(A_i)\,,}$$

e, derivando, si ottiene il *teorema della probabilità totale per la pdf*:

$$\boxed{f(x) = \sum_{i=1}^{n} f(x|A_i)\, P(A_i)\,.}$$

Se X è una variabile aleatoria discreta, è possibile considerare direttamente l'evento $B = \{X = x\}$, ottenendo quindi il *teorema della probabilità totale per la DF*:

$$\boxed{p(x) = \sum_{i=1}^{n} p(x|A_i)\, P(A_i)\,.}$$

► *Esempio* 9.6 *(variabili aleatorie di tipo mixture).* Consideriamo il seguente problema: abbiamo una collezione di dispositivi, che possono essere suddivisi in due insiemi: l'insieme A rappresenta dispositivi a bassa affidabilità, mentre l'insieme $\overline{A}$ rappresenta dispositivi ad alta affidabilità. Supponiamo che il tempo di vita dei dispositivi appartenenti al primo insieme sia modellabile come una variabile aleatoria $X|A \sim \text{Exp}(\lambda_1)$, mentre il tempo di vita dei dispositivi appartenenti al secondo insieme sia modellabile come una variabile aleatoria $X|\overline{A} \sim \text{Exp}(\lambda_2)$. Poichè la media di una generica variabile aleatoria esponenziale $X \sim \text{Exp}(\lambda)$ è pari a $1/\lambda$, allora deve risultare $1/\lambda_1 < 1/\lambda_2$, e quindi $\lambda_1 > \lambda_2$, perché abbiamo supposto che l'insieme A abbia affidabilità più bassa.

La variabile aleatoria X che descrive il tempo di vita (l'affidabilità) di un qualunque dispositivo scelto a caso tra quelli appartenenti ai due insiemi ha una pdf di tipo "mixture", che si può calcolare applicando il teorema della probabilità totale. Infatti, se denotiamo con $p = P(A)$ la probabilità che un dispositivo appartenga al primo insieme, e con $q = 1 - p$ la probabilità che un dispositivo appartenga al secondo insieme, si ha (probabilità totale):

$$F(x) = F(x|A)\, P(A) + F(x|\overline{A})\, P(\overline{A})\,.$$

Poiché

$$F(x|A) = (1 - e^{-\lambda_1 x})\, u(x)\,;$$
$$F(x|\overline{A}) = (1 - e^{-\lambda_2 x})\, u(x)\,;$$

allora si ha:

$$F(x) = [(1 - e^{-\lambda_1 x})p + (1 - e^{-\lambda_2 x})(1 - p)]\, u(x)\,,$$

e derivando

$$f(x) = [\lambda_1\, e^{-\lambda_1 x} p + \lambda_2\, e^{-\lambda_2 x}(1 - p)]\, u(x)\,.$$

Abbiamo ottenuto in questo modo una variabile aleatoria di tipo *mixture* (vedi § 3.5.11) con $\gamma = p$.

La particolare interpretazione della variabile aleatoria di tipo mixture fornita da questo esempio suggerisce anche una pratica strategia per la sua *generazione*. Infatti, per generare una variabile aleatoria del tipo precedentemente visto, è sufficiente avere due generatori (Fig. 9.4), uno per la variabile aleatoria $X|A \sim \text{Exp}(\lambda_1)$ ed un altro per la variabile aleatoria $X|\overline{A} \sim \text{Exp}(\lambda_2)$, e scegliere l'uscita di un generatore oppure di un altro in accordo con i valori di una terza variabile aleatoria binaria W (riconducibile ad una bernoulliana), che assume il valore 1 con probabilità p ed il valore 2 con probabilità $q = 1 - p$. ◄

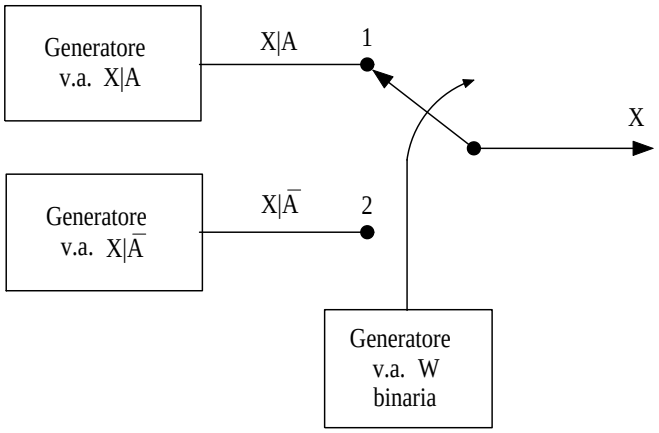


Fig. 9.4. Schema per la generazione di una variabile aleatoria di tipo mixture (esempio 9.6): l'interruttore è comandato dalla variabile aleatoria W, ed è chiuso su 1 con probabilità p e su 2 con probabilità $q = 1 - p$.

9.2.5 Probabilità a posteriori di un evento★

Un'altra relazione utile è quella che calcola la *probabilità a posteriori* di un evento in termini di CDF condizionale. Partiamo dall'identità

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}\,,$$

valida per $P(B) \neq 0$. Scegliendo $B = \{X \leq x\}$, possiamo scrivere:

$$P(A|X \leq x) = \frac{P(X \leq x|A)\,P(A)}{P(X \leq x)} = \frac{F(x|A)}{F(x)}\,P(A)\,.$$

Questa relazione si interpreta come la *probabilità a posteriori* dell'evento A, sapendo che il valore della variabile aleatoria X è minore o uguale ad un certo numero x. La denominazione di "probabilità a posteriori" per $P(A|X \leq x)$ è utilizzata per contrasto con la probabilità $P(A)$, denominata "probabilità a priori". In altri termini, "a priori" sappiamo che la probabilità di A è pari a $P(A)$, poi veniamo a conoscenza del fatto che $X \leq x$, per cui "a posteriori" calcoliamo la probabilità $P(A|X \leq x)$. Si osservi che tale probabilità a posteriori è maggiore, uguale o minore alla probabilità a priori, in dipendenza del fatto che il rapporto tra la CDF condizionale $F(x|A)$ e la CDF $F(x)$ sia maggiore, uguale, o minore dell'unità.

In maniera analoga, se scegliamo $B = \{x_1 < X \leq x_2\}$, si ha:

$$P(A|x_1 < X \leq x_2) = \frac{P(x_1 < X \leq x_2|A)\,P(A)}{P(x_1 < X \leq x_2)} = \frac{F(x_2|A) - F(x_1|A)}{F(x_2) - F(x_1)}P(A)\,, \tag{9.1}$$

che si interpreta come la *probabilità a posteriori* dell'evento A sapendo che il valore della variabile aleatoria X è compreso nell'intervallo (aperto a sinistra) $]x_1, x_2]$. Gli esempi precedenti si possono estendere facilmente al caso in cui B è un qualunque evento (con $P(B) \neq 0$) che può essere espresso in termini della variabile aleatoria X.

9.2.6 Probabilità a posteriori dato $X = x$★

Notiamo che se la variabile aleatoria X è discreta possiamo senza difficoltà estendere il calcolo della probabilità a posteriori effettuato nel precedente paragrafo al caso in cui $B = \{X = x\}$, in

quanto $P(B) \neq 0$. Si ha in tal caso:

$$P(A|X=x) = \frac{P(X=x|A)\,P(A)}{P(X=x)} = \frac{p(x|A)}{p(x)}\,P(A)\,, \tag{9.2}$$

ovvero tale probabilità a posteriori si esprime in termini del rapporto tra la DF condizionale $p(x|A)$ e la DF $p(x)$. Se invece X è una variabile aleatoria continua, sappiamo che $P(X = x) = 0$, e quindi non possiamo procedere direttamente come nel caso precedente, in quanto il denominatore della (9.2) si annulla. D'altra parte, si ha anche:

$$P(X=x|A) = \frac{P(\{X=x\}\cap A)}{P(A)} = 0\,,$$

perché $\{X=x\}\cap A \subseteq \{X=x\}$ e quindi $P(\{X=x\}\cap A) \leq P(\{X=x\}) = 0$. In sostanza, nella (9.2) sia il numeratore che il denominatore sono nulli, per cui $P(A|X=x)$ si presenta in forma indeterminata, e può darsi che il risultato esista lo stesso finito al *limite*. Supponiamo allora che X sia una variabile aleatoria continua, e procediamo ponendo[3]

$$P(A|X=x) \triangleq \lim_{\varepsilon\to 0} P(A|x-\varepsilon < X \leq x)\,.$$

con $\varepsilon \geq 0$. Possiamo calcolare facilmente la probabilità al secondo membro, in quanto essa è del tipo (9.1) con $x_1 = x-\varepsilon$ e $x_2 = x$. Si ha:

$$\begin{aligned} P(A|x-\varepsilon < X \leq x) &= \frac{P(x-\varepsilon < X \leq x|A)\,P(A)}{P(x-\varepsilon < X \leq x)} = \frac{F(x|A)-F(x-\varepsilon|A)}{F(x)-F(x-\varepsilon)}\,P(A) \\ &= \frac{[F(x|A)-F(x-\varepsilon|A)]/\varepsilon}{[F(x)-F(x-\varepsilon)]/\varepsilon}\,P(A)\,, \end{aligned}$$

da cui, al limite per $\varepsilon \to 0$, e ricordando la definizione di pdf e di pdf condizionale dato A, si ottiene la relazione cercata:

$$\boxed{P(A|X=x) = \frac{f(x|A)}{f(x)}\,P(A)\,,} \tag{9.3}$$

che costituisce una generalizzazione della (9.2) al caso di variabili aleatorie continue. Si noti che si è supposto che le pdf che compaiono nella (9.3) esistano e che $f(x) \neq 0$.

► *Esempio* 9.7. Consideriamo una popolazione di individui, che possiamo suddividere nell'insieme $A = \{\text{maschi}\}$ e nell'insieme $\overline{A} = \{\text{femmine}\}$. Sia X una variabile aleatoria che rappresenta l'altezza di un individuo appartenente alla popolazione in esame: è chiaro che tale variabile aleatoria è la mixture delle altezze $X|A$ (altezza di un maschio) ed $X|\overline{A}$ (altezza di una femmina), per cui la pdf di X è

$$f(x) = f(x|A)\,P(A) + f(x|\overline{A})\,P(\overline{A})$$

da cui risulta

$$P(A|X=x) = \frac{f(x|A)}{f(x|A)P(A) + f(x|\overline{A})P(\overline{A})}\,P(A)\,,$$

Intuitivamente, tale probabilità *a posteriori* rappresenta la probabilità che un individuo sia maschio, *sapendo che* la sua altezza è pari ad x; essa può essere confrontata con la probabilità *a priori* $P(A)$ che un individuo sia maschio senza sapere nulla sulla sua altezza. Poichè mediamente i maschi sono più alti delle femmine, ci

[3] Per un maggior rigore formale, bisognerebbe effettuare il limite considerando una successione *discreta* di eventi ($\varepsilon = 1/n$) ed utilizzando la proprietà di continuità della probabilità; tuttavia si giungerebbe allo stesso risultato.

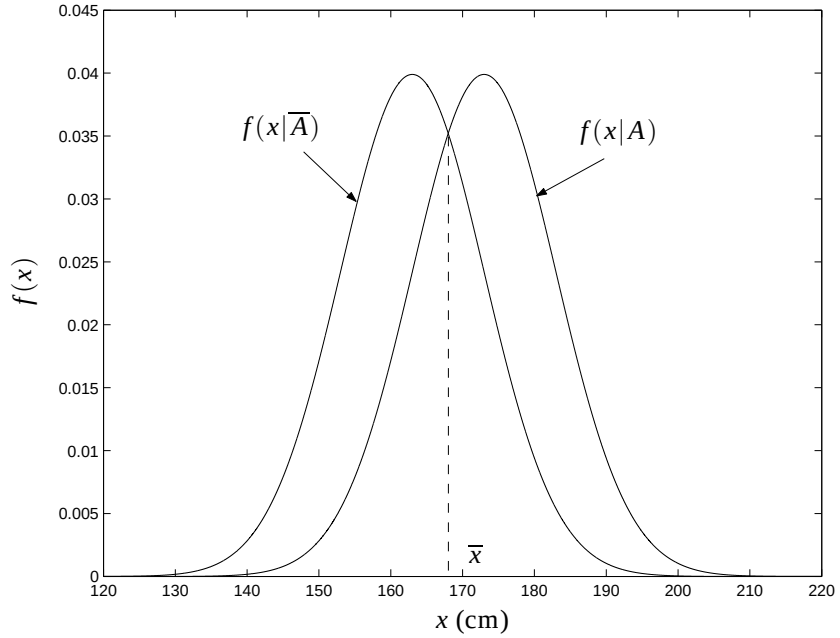


Fig. 9.5. Le due pdf rappresentano (vedi esempio 9.7) la pdf dell'altezza della popolazione femminile (a sinistra) e quella della popolazione maschile (a destra); il valore $x = \overline{x}$ è quella particolare altezza che non fornisce informazioni sull'appartenenza ad un sesso .

aspettiamo che se x è grande, risulterà $P(A|X = x) > P(A)$, viceversa se x è piccolo. Ci sarà un particolare valore di x per il quale $P(A|X = x) = P(A)$, che rappresenta la condizione per cui gli eventi A ed $\{X = x\}$ sono indipendenti, e quindi la conoscenza dell'altezza non fornisce informazione sull'appartenenza ad un sesso. Imponendo la condizione $P(A|X = x) = P(A)$ si trova

$$\frac{f(x|A)}{f(x|A)P(A) + f(x|\overline{A})P(\overline{A})} = 1\,,$$

sviluppando la quale si ha la condizione $f(x|A) = f(x|\overline{A})$. Il punto $x = \overline{x}$ si può allora determinare (Fig. 9.5) riportando su uno stesso diagramma le due pdf condizionali e trovando la loro intersezione (si noti che tale intersezione potrebbe non essere unica, in dipendenza dalla forma della pdf). ◄

► *Esempio* 9.8 *(test di ipotesi).* Riprendiamo l'esempio 9.6, e supponiamo di voler risolvere il seguente problema, tipico del controllo industriale di qualità: si prende a caso un dispositivo, e si misura il suo tempo di vita, ottenendo un valore x; si vuole determinare se il dispositivo sia a bassa oppure ad alta affidabilità. Matematicamente, ciò equivale al seguente problema: si vuole valutare, osservato un valore x della variabile aleatoria mixture X, se sia più probabile che si sia verificato l'evento A (dispositivo a bassa affidabilità) oppure $\overline{A}$ (dispositivo ad alta affidabilità). Questo equivale a valutare quale delle due quantità $P(A|X = x)$ e $P(\overline{A}|X = x)$ sia la più grande, il che si può formalizzare come un *test di ipotesi*:

$$P(A|X = x) \underset{\overline{A}}{\overset{A}{\gtrless}} P(\overline{A}|X = x)$$

che si interpreta nel modo seguente: se il primo membro è maggiore del secondo, allora diremo che l'evento A è più probabile, per cui sceglieremo l'ipotesi "il componente appartiene alla classe A", viceversa se il primo membro è minore del secondo. Poiché tale test sceglie l'ipotesi che ha probabilità a posteriori maggiore, esso prende il nome di *test a massima probabilità a posteriori* (*maximum a posteriori probability*, MAP). Applicando la (9.3), tale test si può riscrivere come segue:

$$\frac{f(x|A)}{f(x)}\,P(A) \underset{\overline{A}}{\overset{A}{\gtrless}} \frac{f(x|\overline{A})}{f(x)}\,P(\overline{A})\,,$$

e semplificando un termine $f(x) > 0$ in ambo i membri, si ha:

$$f(x|A)\,P(A) \underset{\overline{A}}{\overset{A}{\gtrless}} f(x|\overline{A})\,P(\overline{A})\,.$$

Se si suppone poi che gli eventi A e $\overline{A}$ siano equiprobabili, il test MAP si semplifica ulteriormente:

$$f(x|A) \underset{\overline{A}}{\overset{A}{\gtrless}} f(x|\overline{A})\,.$$

Tale test prende il nome di *test a massima verosimiglianza* (*maximum likelihood*, ML), e la funzione $f(x|A)$ prende il nome di *funzione di verosimiglianza* dell'evento A. Pertanto il test ML si ottiene come particolarizzazione del test MAP nel caso di ipotesi equiprobabili.

Sostituendo le pdf condizionali di tipo esponenziale (cfr. esempio 9.6), il test si può scrivere nella forma esplicita

$$\lambda_1\, e^{-\lambda_1 x} \underset{\overline{A}}{\overset{A}{\gtrless}} \lambda_2\, e^{-\lambda_2 x}\,,$$

che, nel caso $\lambda_1 > \lambda_2$ (corrispondente al caso in cui A sia la classe a più bassa affidabilità rispetto a $\overline{A}$), può essere posto nella forma:

$$\frac{1}{\lambda_1 - \lambda_2} \ln \frac{\lambda_1}{\lambda_2} \underset{\overline{A}}{\overset{\overline{A}}{\gtrless}} x\,.$$

Notiamo che nelle ipotesi fatte il primo membro è positivo. Osserviamo che il test si riduce a confrontare il tempo di vita osservato x con una soglia positiva $\gamma = \frac{1}{\lambda_1-\lambda_2} \ln \frac{\lambda_1}{\lambda_2}$; se il tempo di vita è inferiore a tale soglia, si dichiara che il dispositivo appartiene alla classe A a più bassa affidabilità; viceversa, se il tempo di vita è superiore a tale soglia, si dichiara che il dispositivo appartiene alla classe $\overline{A}$ a più alta affidabilità. Il risultato pare intuitivamente accettabile, meno intuitiva è l'espressione della soglia che abbiamo ricavato, e che dipende dalla particolare distribuzione esponenziale scelta per il tempo di vita. Se il tempo di vita è *esattamente* uguale alla soglia (il che peraltro accade, essendo X una variabile aleatoria continua, con probabilità zero), scegliere un'ipotesi oppure un'altra è *indifferente* (le due ipotesi hanno la stessa probabilità a posteriori). ◀

9.2.7 Teorema della probabilità totale (versione continua)★

Sulla base della (9.3), possiamo a questo punto ottenere una generalizzazione del teorema della probabilità totale visto al § 2.2.4. Si riscriva infatti la (9.3) nella forma:

$$f(x|A)\,P(A) = P(A|X = x)\,f(x)\,. \tag{9.4}$$

Poichè $f(x|A)$ è una valida pdf, allora avrà area unitaria:

$$\int_{-\infty}^{\infty} f(x|A)\,dx = 1\,,$$

per cui, integrando membro a membro la (9.4), si ha:

$$\boxed{P(A) = \int_{-\infty}^{\infty} P(A|X = x) f(x)\,dx\,.} \tag{9.5}$$

Questa relazione rappresenta una versione *continua* del teorema della probabilità totale $P(A) = \sum_i P(A|B_i)\,P(B_i)$, nella quale gli eventi condizionanti sono del tipo $\{X = x\}$ e costituiscono una infinità continua (e non finita o numerabile).

9.2.8 Teorema di Bayes per le pdf★

Come ultima relazione utile, introduciamo una relazione per le pdf affine a quella di Bayes. Dalla (9.3), si ottiene:

$$f(x|A) = \frac{P(A|X=x)}{P(A)} f(x)\,,$$

per cui, sostituendo a $P(A)$ il valore dato dalla (9.5) si ha:

$$\boxed{f(x|A) = \frac{P(A|X=x)f(x)}{\int_{-\infty}^{\infty} P(A|X=x)f(x)dx}\,,}$$

che rappresenta una sorta di *teorema di Bayes per le pdf*.

9.3 Distribuzioni condizionali per coppie di variabili aleatorie

Analogamente a quanto fatto nel paragrafo precedente per il caso di una variabile aleatoria, è possibile definire distribuzioni condizionali (CDF, pdf e DF) anche per una coppia di variabili aleatorie. Ad esempio, date due variabili aleatorie (X,Y) ed un evento B con $P(B) \neq 0$, possiamo definire la *CDF condizionale di* (X,Y) *dato* B:

Definizione (CDF condizionale di una coppia di variabili aleatorie). Siano (X,Y) una coppia di variabili aleatorie definite su uno spazio di probabilità $(\Omega, \mathcal{S}, P)$, e sia $B \in \mathcal{S}$ un evento di Ω, con $P(B) \neq 0$. La *CDF condizionale* di (X,Y) dato l'evento B è:

$$F_{XY}(x,y|B) \triangleq P(X \leq x, Y \leq y|B) = \frac{P(X \leq x, Y \leq y, B)}{P(B)}\,.$$

La corrispondente *pdf condizionale* si ricava per derivazione dalla CDF, ed è:

$$\boxed{f_{XY}(x,y|B) \triangleq \frac{\partial^2}{\partial x\,\partial y} F_{XY}(x,y|B)\,.}$$

Se le variabili aleatorie X ed Y sono discrete, a valori in $\mathcal{X}$ e $\mathcal{Y}$, rispettivamente, è utile definire la *DF condizionale* di (X,Y) dato B:

$$\boxed{p_{XY}(x,y|B) \triangleq P(X=x, Y=y|B) = \frac{P(X=x, Y=y, B)}{P(B)}\,,}$$

con $(x,y) \in \mathcal{X} \times \mathcal{Y}$.

► *Esempio* 9.9. Come per il caso di una singola variabile aleatoria, il calcolo della CDF condizionale è particolarmente semplice se l'evento B si esprime in funzione delle variabili aleatorie (X,Y) o di una di esse. Supponiamo, ad esempio, che $B = \{X \leq a\}$, con a numero reale (cfr. esempio 9.4 per un calcolo simile per il caso di una singola variabile aleatoria). Si ha:

$$F_{XY}(x,y|B) \triangleq P(X \leq x, Y \leq y|B) = \frac{P(X \leq x, Y \leq y, X \leq a)}{P(X \leq a)}\,.$$

Se $x < a$, si ha che $\{X \leq x, Y \leq y, X \leq a\} = \{X \leq x, Y \leq y\}$, per cui:

$$F_{XY}(x, y|B) \triangleq \frac{F_{XY}(x, y)}{F_X(a)} ;$$

mentre se $x \geq a$ si ha che $\{X \leq x, Y \leq y, X \leq a\} = \{X \leq a, Y \leq y\}$, per cui:

$$F_{XY}(x, y|B) \triangleq \frac{F_{XY}(a, y)}{F_X(a)} .$$

Calcolando la derivata mista rispetto ad x ed y (escludiamo la presenza di impulsi) si ottiene la pdf condizionale:

$$f_{XY}(x, y|B) = \begin{cases} \dfrac{f_{XY}(x, y)}{F_X(a)}, & x < a ; \\ 0 , & x \geq a . \end{cases}$$

Verifichiamo che la precedente è una valida pdf, osservando se è soddisfatta la condizione di normalizzazione. Si ha:

$$\begin{aligned} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y|B)\, dx\, dy &= \frac{1}{F_X(a)} \int_{-\infty}^{\infty} \int_{-\infty}^{a} f_{XY}(x, y)\, dx\, dy \\ &= \frac{1}{F_X(a)} F_{XY}(a, +\infty) = 1 , \end{aligned}$$

dove abbiamo sfruttato la (6.2) e la relazione tra CDF congiunte e marginali, secondo la quale $F_{XY}(a, +\infty) = F_X(a)$. ◀

► *Esempio* 9.10. Consideriamo adesso il caso in cui $B = \{a < X \leq b\}$ (cfr. esempio 9.5 per un calcolo simile per il caso di una singola variabile aleatoria). Si ha:

$$F_{XY}(x, y|B) = P(X \leq x, Y \leq y | a < X \leq b) = \frac{P(X \leq x, Y \leq y, a < X \leq b)}{P(a < X \leq b)} .$$

Se $x \geq b$, allora $\{X \leq x, Y \leq y, a < X \leq b\} = \{a < X \leq b, Y \leq y\}$, e quindi

$$F_{XY}(x, y|B) = \frac{P(a < X \leq b, Y \leq y)}{P(a < X \leq b)} = \frac{F_{XY}(b, y) - F_{XY}(a, y)}{F_X(b) - F_X(a)} .$$

Se $a < x < b$, allora $\{X \leq x, Y \leq y, a < X \leq b\} = \{a < X \leq x, Y \leq y\}$, e quindi

$$F_{XY}(x, y|B) = \frac{P(a < X \leq x, Y \leq y)}{P(a < X \leq b)} = \frac{F_{XY}(x, y) - F_{XY}(a, y)}{F_X(b) - F_X(a)} .$$

Infine, se $x \leq a$, allora $\{X \leq x, Y \leq y, a < X \leq b\} = \varnothing$, e quindi

$$F_{XY}(x, y|B) = 0 .$$

In definitiva, allora:

$$F_{XY}(x, y|a < X \leq b) = \begin{cases} 0, & x \leq a ; \\ \frac{F_{XY}(x,y) - F_{XY}(a,y)}{F_X(b) - F_X(a)}, & a < x < b ; \\ \frac{F_{XY}(b,y) - F_{XY}(a,y)}{F_X(b) - F_X(a)}, & x \geq b . \end{cases}$$

Calcolando la derivata mista rispetto ad x e ad y, si ottiene la corrispondente pdf:

$$f_{XY}(x, y|a < X \leq b) = \begin{cases} 0, & x \leq a ; \\ \frac{f_{XY}(x,y)}{F_X(b) - F_X(a)}, & a < x < b ; \\ 0, & x \geq b . \end{cases}$$

Anche in questo caso, verifichiamo che la condizione di normalizzazione delle pdf sia soddisfatta. Si ha:

$$\begin{aligned}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} f_{XY}(x,y|a<X\leq b)\,dx\,dy &= \int_a^b dx \int_{-\infty}^{\infty} \frac{f_{XY}(x,y)}{F_X(b)-F_X(a)}\,dy \\ &= \frac{1}{F_X(b)-F_X(a)}\int_a^b dx\int_{-\infty}^{\infty} f_{XY}(x,y)\,dy \\ &= \frac{1}{F_X(b)-F_X(a)}\int_a^b f_X(x)\,dx = 1\,,\end{aligned}$$

dove abbiamo sfruttato la relazione tra pdf congiunte e marginali e le proprietà della pdf. ◄

9.3.1 Distribuzioni condizionali dato $X = x$ ed $Y = y$

Consideriamo ora il caso particolarmente interessante in cui l'evento condizionante è $B = \{X = x\}$, che non può rientrare come caso particolare delle precedenti definizioni, in quanto, se X è una variabile aleatoria continua, allora risulta $P(B) = 0$. L'obiettivo è calcolare le distribuzioni condizionali di Y dato $B = \{X = x\}$, per capire come si modifica la distribuzione marginale di Y se fissiamo un valore $X = x$ di un'altra variabile aleatoria; ad esempio, come si modifica la distribuzione del peso Y di una persona se conosciamo la sua altezza X. Per aggirare il problema insito nell'annullamento di $P(B)$, procediamo al limite, ponendo per definizione:

$$F_Y(y|X=x) \triangleq \lim_{\varepsilon\to 0} P(Y\leq y|x-\varepsilon<X\leq x)\,,$$

con $\varepsilon \geq 0$. Si trova:

$$\boxed{F_Y(y|X=x) = \frac{1}{f_X(x)}\,\frac{\partial}{\partial x}F_{XY}(x,y)\,.} \tag{9.6}$$

Prova. Si ha:

$$P(Y\leq y|x-\varepsilon<X\leq x) = \frac{P(Y\leq y, x-\varepsilon<X\leq x)}{P(x-\varepsilon<X\leq x)} = \frac{F_{XY}(x,y)-F_{XY}(x-\varepsilon,y)}{F_X(x)-F_X(x-\varepsilon)}\,,$$

da cui, dividendo numeratore e denominatore per ε e passando al limite per $\varepsilon \to 0$, si ha l'asserto (supponendo l'esistenza della derivata parziale rispetto ad x di $F_{XY}(x,y)$). □

Scambiando i ruoli di X ed Y, si ottiene la relazione simmetrica:

$$\boxed{F_X(x|Y=y) = \frac{1}{f_Y(y)}\,\frac{\partial}{\partial y}F_{XY}(x,y)\,.} \tag{9.7}$$

Particolarmente interessante è l'espressione delle *pdf condizionali*, che si ottengono derivando la (9.6) rispetto a y e la (9.7) rispetto ad x. Si ha, infatti:

$$f_Y(y|X=x) = \frac{\partial}{\partial y}F_Y(y|X=x) = \frac{1}{f_X(x)}\,\frac{\partial^2}{\partial y\,\partial x}F_{XY}(x,y) = \frac{f_{XY}(x,y)}{f_X(x)}\,,$$

e similmente:

$$f_X(x|Y=y) = \frac{f_{XY}(x,y)}{f_Y(y)}\,.$$

Spesso le relazioni precedenti si esprimono, in forma più sintetica, come:

$$\boxed{\begin{aligned} f_X(x|y) &= \frac{f_{XY}(x,y)}{f_Y(y)}\,; \\ f_Y(y|x) &= \frac{f_{XY}(x,y)}{f_X(x)}\,. \end{aligned}} \tag{9.8}$$

Si noti l'affinità formale tra tali relazioni e la definizione (2.1) di probabilità condizionale. Ricordiamo, inoltre, che $f_X(x|y)$ è una pdf (monodimensionale) vista come funzione di x, ma non di y, per cui risulta verificata la condizione di normalizzazione in x:

$$\int_{-\infty}^{\infty} f_X(x|y)\, dx = 1\,,$$

ma il corrispondente integrale in dy non è unitario. Analogo discorso, scambiando i ruoli di x ed y, vale per $f_Y(y|x)$. Notiamo poi che, per denotare che $f_Y(y|x)$ è la pdf condizionale di Y dato $\{X = x\}$, si usa la notazione sintetica $Y|x \sim f_Y(y|x)$.

Osserviamo infine che se X ed Y sono indipendenti, la fattorizzazione $f_{XY}(x,y) = f_X(x)\, f_Y(y)$ della pdf congiunta implica che

$$\begin{aligned} f_X(x|y) &= f_X(x)\,, \\ f_Y(y|x) &= f_Y(y)\,, \end{aligned}$$

ovvero la pdf condizionale è uguale a quella marginale (l'evento $\{X = x\}$ non modifica la pdf di Y, e simmetricamente l'evento $\{Y = y\}$ non modifica la pdf di X).

Data la somiglianza formale tra l'espressione delle pdf condizionali e la definizione di probabilità condizionale, non sorprende che alcuni teoremi tipici della probabilità condizionale abbiano una loro controparte per le pdf condizionali. Ad esempio, sulla base delle definizioni (9.8), la pdf congiunta ammette due distinte fattorizzazioni in termini di pdf condizionali:

$$\boxed{f_{XY}(x,y) = f_X(x|y)\, f_Y(y) = f_Y(y|x)\, f_X(x)\,,} \tag{9.9}$$

che è una relazione simile alla *legge della probabilità composta* (2.2). Utilizzando la relazione tra statistiche congiunte e marginali, si ha poi:

$$\boxed{f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x,y)\, dx = \int_{-\infty}^{\infty} f_Y(y|x)\, f_X(x)\, dx\,,} \tag{9.10}$$

che è una relazione analoga al *teorema della probabilità totale* (cfr. teorema 2.1, capitolo 2) e generalizza il teorema della probabilità totale per le pdf (cfr. § 9.2) al caso in cui gli eventi condizionanti siano una infinità continua. Per questo motivo, tale relazione costituisce una versione continua del teorema della probabilità totale per le pdf. Infine, combinando la legge della probabilità composta (9.9) e quella della probabilità totale (9.10), possiamo ottenere la relazione

$$\boxed{f_X(x|y) = \frac{f_Y(y|x)\, f_X(x)}{f_Y(y)} = \frac{f_Y(y|x)\, f_X(x)}{\int_{-\infty}^{\infty} f_Y(y|x)\, f_X(x)\, dx}} \tag{9.11}$$

che consente di esprimere una pdf condizionale in funzione dell'altra, ed è una relazione analoga al *teorema di Bayes* (cfr. teorema 2.2, capitolo 2).

► *Esempio* 9.11. Consideriamo il caso di una coppia di variabili aleatorie congiuntamente gaussiane $(X, Y) \sim \mathrm{N}(\mu_X, \mu_Y, \sigma_X, \sigma_Y, \rho)$, e calcoliamo le pdf condizionali $f_X(x|y)$ e $f_Y(y|x)$. Il risultato si ottiene semplicemente

se ricordiamo la fattorizzazione della pdf congiunta ricavata nell'esempio 6.2, che si riporta di seguito per comodità del lettore:

$$f_{XY}(x,y) = \left[\frac{1}{\sigma_X\sqrt{2\pi}}\, e^{-\frac{1}{2\sigma_X^2}(x-\mu_X)^2}\right]\left[\frac{1}{\sigma_Y\sqrt{1-\rho^2}\sqrt{2\pi}}\, e^{-\frac{1}{2\sigma_Y^2(1-\rho^2)}\left[y-\mu_Y-\rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)\right]^2}\right].$$

Infatti, poiché nell'esempio 6.2 abbiamo dimostrato che il primo fattore rappresenta la pdf marginale $f_X(x)$, allora il secondo fattore per la (9.9) deve necessariamente rappresentare la pdf condizionale $f_Y(y|x)$, ovvero si ha:

$$f_Y(y|x) = \frac{f_{XY}(x,y)}{f_X(x)} = \frac{1}{\sigma_Y\sqrt{1-\rho^2}\sqrt{2\pi}}\, e^{-\frac{1}{2\sigma_Y^2(1-\rho^2)}\left[y-\mu_Y-\rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)\right]^2}.$$

Tale pdf (come funzione di y) ha ancora la forma gaussiana monodimensionale, con parametri media e varianza che si individuano facilmente per ispezione, per cui $Y|x \sim \mathrm{N}(\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X), \sigma_Y\sqrt{1-\rho^2})$, cioè Y dato $X = x$ è *condizionalmente gaussiana*, con i parametri indicati (dipendenti dal valore di y). In maniera simmetrica, si trova:

$$f_X(x|y) = \frac{f_{XY}(x,y)}{f_Y(y)} = \frac{1}{\sigma_X\sqrt{1-\rho^2}\sqrt{2\pi}}\, e^{-\frac{1}{2\sigma_X^2(1-\rho^2)}\left[x-\mu_X-\rho\frac{\sigma_X}{\sigma_Y}(y-\mu_Y)\right]^2},$$

per cui $X|y \sim \mathrm{N}(\mu_X + \rho\frac{\sigma_X}{\sigma_Y}(y-\mu_Y), \sigma_X\sqrt{1-\rho^2})$, per cui anche X, dato $Y = y$, è *condizionalmente gaussiana*, con i parametri indicati. In conclusione, possiamo affermare che *variabili aleatorie* (X,Y) *congiuntamente gaussiane sono non solo marginalmente gaussiane, ma anche condizionalmente gaussiane.*

È interessante interpretare intuitivamente i parametri caratteristici della distribuzione condizionale di X dato $Y = y$, con riferimento ad un esempio concreto. Si consideri ad esempio il caso in cui X rappresenti l'altezza ed Y il peso di una persona; supponendo di modellare tali quantità approssimativamente (perché?) come variabili aleatorie congiuntamente gaussiane, esse risulteranno sicuramente *positivamente correlate*, perché ad un incremento dell'una corrisponde in media un'incremento dell'altra.

Se infatti fissiamo $Y = y$, con $y > \mu_Y$ (un peso al di sopra della media), la media della distribuzione condizionale di $X|Y = y$ si sposta verso destra (cioè verso altezze superiori alla media μ_X. Se invece fissiamo $y < \mu_Y$ (un peso al di sotto della media) la media della distribuzione condizionale di $X|Y = y$ si sposta verso sinistra (cioè verso altezze inferiori alla media μ_X); si sarebbe verificato il contrario se X ed Y fossero state negativamente correlate.

Per quanto riguarda la varianza della distribuzione condizionale di X dato $Y = y$, notiamo che essa *non* dipende dal valore y che fissiamo di volta in volta, ma solo da σ_X e da ρ; in particolare, tale varianza assume il valore massimo σ_X^2 per $\rho = 0$, e diminuisce al crescere di ρ (in modulo). Tale comportamento si interpreta come segue: se fissiamo un valore del peso, si riduce l'incertezza che abbiamo sull'altezza, e quindi la varianza condizionale dell'altezza *dato il peso* è più piccola; questa riduzione della varianza è tanto più grande quanto più il coefficiente di correlazione è prossimo (in modulo) ad uno, cioè quanto più le variabili aleatorie X ed Y sono correlate. ◀

9.4 Distribuzioni condizionali per vettori di variabili aleatorie

È possibile generalizzare la definizione di distribuzioni condizionali introdotte per due variabili aleatorie al caso di vettori di variabili aleatorie. Ad esempio, la definizione di CDF congiunta delle variabili aleatorie $X_1, X_2, \ldots, X_n$ dato un evento B si generalizza come segue:

Definizione (CDF condizionale di un vettore di variabili aleatorie). Siano $X_1, X_2, \ldots, X_n$ n variabili aleatorie definite su uno spazio di probabilità $(\Omega, \mathcal{S}, P)$, e sia $B \in \mathcal{S}$ un evento di Ω, con $P(B) \neq 0$. La *CDF condizionale* di $X_1, X_2, \ldots, X_n$ dato l'evento B è:

$$F_{X_1X_2\cdots X_n}(x_1, x_2, \ldots, x_n|B) \triangleq P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n|B).$$

La corrispondente pdf condizionale si ricava per derivazione della CDF condizionale.

Particolarmente interessante è calcolare la pdf condizionale di $X_1, X_2, \ldots, X_k$ dati i valori assunti da $X_{k+1}, X_{k+2}, \ldots, X_n$, siano essi $x_{k+1}, x_{k+2}, \ldots, x_n$, che si ottiene generalizzando le (9.8):

$$f_{X_1 X_2 \cdots X_k}(x_1, x_2, \ldots, x_k | x_{k+1}, x_{k+2}, \ldots, x_n) \triangleq \frac{f_{X_1 X_2 \cdots X_n}(x_1, x_2, \ldots, x_n)}{f_{X_{k+1} X_{k+2} \cdots X_n}(x_{k+1} x_{k+2}, \ldots, x_n)} .$$

Ovviamente tale funzione è una pdf vista come funzione di $x_1, x_2, \ldots, x_k$, ed in particolare soddisfa la condizione di normalizzazione, per ogni $(x_{k+1}, x_{k+2}, \ldots, x_n)$:

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_{X_1 X_2 \cdots X_k}(x_1, x_2, \ldots, x_k | x_{k+1}, x_{k+2}, \ldots, x_n)\, dx_1\, dx_2 \cdots dx_k .$$

Il modo per costruire una qualunque pdf condizionale dovrebbe essere a questo punto chiaro al lettore: al numeratore va la pdf congiunta di tutte le variabili aleatorie in gioco, al denominatore quella delle sole variabili aleatorie condizionanti.

► *Esempio* 9.12. Consideriamo il caso di quattro variabili aleatorie X_1, X_2, X_3, X_4, e calcoliamo esplicitamente alcune distribuzioni condizionali:

$$f_{X_2}(x_2 | x_1, x_3, x_4) = \frac{f_{X_1 X_2 X_3 X_4}(x_1, x_2, x_3, x_4)}{f_{X_1 X_3 X_4}(x_1, x_3, x_4)} ;$$
$$f_{X_1 X_2}(x_1, x_2 | x_3, x_4) = \frac{f_{X_1 X_2 X_3 X_4}(x_1, x_2, x_3, x_4)}{f_{X_3 X_4}(x_3, x_4)} ;$$
$$f_{X_1 X_2 X_4}(x_1, x_2, x_4 | x_1) = \frac{f_{X_1 X_2 X_3 X_4}(x_1, x_2, x_3, x_4)}{f_{X_1}(x_1)} .$$

◄

Le corrispondenti CDF condizionali si possono ottenere per integrazione, ad esempio si ha:

$$F_{X_1 X_2 \cdots X_k}(x_1, x_2, \ldots, x_k | x_{k+1}, x_{k+2}, \ldots, x_n) = \\ = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \cdots \int_{-\infty}^{x_k} f_{X_1 X_2 \cdots X_k}(u_1, u_2, \ldots, u_k | x_{k+1}, x_{k+2}, \ldots, x_n)\, du_1\, du_2 \cdots du_k . \quad (9.12)$$

Per variabili aleatorie discrete, è possibile estendere in maniera analoga anche la definizione di DF condizionali.

9.4.1 Indipendenza condizionale e regola della catena per le pdf

Consideriamo il caso in cui le variabili aleatorie $X_1, X_2, \ldots X_k$ siano indipendenti dalle variabili aleatorie $X_{k+1}, X_{k+2}, \ldots, X_n$, evidentemente si ha:

$$f_{X_1 X_2 \cdots X_k}(x_1, x_2, \ldots, x_k | x_{k+1}, x_{k+2}, \ldots, x_n) = f_{X_1 X_2 \cdots X_k}(x_1, x_2, \ldots, x_k)$$

cioè il condizionamento non opera. Avendo introdotto le CDF e pdf condizionali, possiamo anche introdurre il concetto di *indipendenza condizionale* tra le variabili aleatorie componenti il vettore $\boldsymbol{X}$. Ad esempio, X_1 ed X_2 si diranno condizionalmente indipendenti, data una terza variabile X_3, se vale la seguente fattorizzazione della pdf condizionale:

$$f_{X_1 X_2}(x_1, x_2 | x_3) = f_{X_1}(x_1 | x_3)\, f_{X_2}(x_2 | x_3) ,$$

che con semplici manipolazioni algebriche si può mostrare che implica le:

$$f_{X_1}(x_1 | x_2, x_3) = f_{X_1}(x_1 | x_3) ;$$
$$f_{X_2}(x_2 | x_1, x_3) = f_{X_2}(x_2 | x_3) ;$$

che si interpretano nel seguente modo: dato $X_3 = x_3$, il condizionamento $X_2 = x_2$ o $X_1 = x_1$ non opera. Il concetto di indipendenza condizionale si può estendere banalmente anche a gruppi di variabili aleatorie.

Una relazione interessante che scaturisce dalla definizione di pdf condizionale è la cosiddetta *regola della catena* per le pdf. Infatti, notiamo che, con successivi condizionamenti, la pdf congiunta di $\boldsymbol{X}$ si può fattorizzare nel prodotto di n pdf condizionali monodimensionali, come:

$$\begin{aligned} f_{X_1 X_2 \cdots X_n}(x_1, x_2, \ldots, x_n) &= f_{X_1}(x_1)\, f_{X_2 X_3 \cdots X_n}(x_2, x_3, \ldots, x_n | x_1) \\ &= f_{X_1}(x_1)\, f_{X_2}(x_2|x_1)\, f_{X_3 \cdots X_n}(x_3, \ldots, x_n | x_1, x_2) \\ &= \ldots \\ &= f_{X_1}(x_1)\, f_{X_2}(x_2|x_1)\, f_{X_3}(x_3|x_1, x_2) \cdots f_{X_n}(x_n | x_1, x_2, \ldots, x_{n-1})\,. \end{aligned}$$

9.5 Media condizionale e momenti condizionali

La definizione di *media condizionale* di una variabile aleatoria X, dato un evento B, è una semplice estensione dalla definizione di media di una variabile aleatoria, ottenuta sostituendo alla pdf $f(x)$ la pdf condizionale $f(x|B)$:

Definizione (media condizionale di una variabile aleatoria). La media condizionale $E(X|B)$ di una variabile aleatoria X con pdf condizionale $f(x|B)$ è:

$$E(X|B) \triangleq \int_{-\infty}^{\infty} x\, f(x|B)\, dx\,,$$

se tale integrale esiste finito.

► *Esempio* 9.13. Abbiamo già visto (cfr. esempio 9.4) che se $B = \{X \leq a\}$, allora:

$$f(x|X \leq a) = \begin{cases} \frac{f(x)}{F(a)}, & x < a\,; \\ 0, & x \geq a\,. \end{cases}$$

Pertanto, si ha:

$$E(X|B) = \int_{-\infty}^{a} x\, \frac{f(x)}{F(a)}\, dx = \frac{1}{F(a)} \int_{-\infty}^{a} x\, f(x)\, dx = \frac{\int_{-\infty}^{a} x\, f(x)\, dx}{\int_{-\infty}^{a} f(x)\, dx}\,.$$

◄

Osserviamo che la media condizionale gode di tutte le proprietà della media (cfr. § 5.2): in particolare ad essa si applica il teorema fondamentale della media. Infatti, se si vuole calcolare la media condizionale di $g(X)$ dato un evento B, si ha:

$$E[g(X)|B] = \int_{-\infty}^{\infty} g(x)\, f_X(x|B)\, dx\,,$$

mentre per variabili aleatorie discrete il teorema si può particolarizzare come:

$$E[g(X)|B] = \sum_{x \in \mathcal{X}} g(x)\, P(X = x|B) = \sum_{x \in \mathcal{X}} g(x)\, p_X(x|B)\,,$$

ovvero si esprime in termini della DF condizionale $p(x|B)$ della variabile aleatoria X. A partire dal teorema fondamentale della media, poi, è possibile definire qualunque *momento condizionale*: ad esempio, il valor quadratico medio condizionale è dato da:

$$E(X^2|B) \triangleq \int_{-\infty}^{\infty} x^2\, f(x|B)\, dx\,,$$

mentre la varianza condizionale si può esprimere facilmente in termini del valor quadratico medio condizionale e della media condizionale, come:

$$\mathrm{Var}(X|B) = E(X^2|B) - E^2(X|B)\,. \tag{9.13}$$

Si noti che *non* è possibile scrivere $\mathrm{Var}(X|B) = E[(X-\mu_X)^2|B]$ in quanto μ_X *non* è la media condizionale; viceversa, si verifica facilmente che la definizione corretta è:

$$\mathrm{Var}(X|B) \triangleq E[(X - E(X|B))^2|B]\,;$$

infatti, sviluppando la precedente relazione, si ottiene la (9.13).

La definizione di media condizionale dato un evento B si estende naturalmente al caso di coppie di variabili aleatorie e, più in generale, al caso di vettori di variabili aleatorie; basta sostituire alla pdf, nell'integrale che definisce la media, la pdf condizionale dato B. Il teorema fondamentale della media si estende anche al caso in cui desideriamo calcolare la media condizionale di $g(X,Y)$ dato un evento B, e conosciamo la pdf condizionale $f_{XY}(x,y|B)$. Si ha:

$$E[g(X,Y)|B] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)\, f_{XY}(x,y|B)\, dx\, dy\,. \tag{9.14}$$

Tale teorema consente di definire i momenti congiunti condizionali di una coppia di variabili aleatorie, come la correlazione condizionale e la covarianza condizionale.

Infine, più in generale, nel caso in cui abbiamo un vettore $\boldsymbol{X} = [X_1, X_2, \ldots, X_n]^T$ di n variabili aleatorie, e vogliamo calcolare la media condizionale di $g(\boldsymbol{X})$ dato un evento B, conoscendo la pdf condizionale $f_{\boldsymbol{X}}(\boldsymbol{x}|B)$, si ha:

$$E[g(\boldsymbol{X})|B] = \int_{\mathbb{R}^n} g(\boldsymbol{x})\, f_{\boldsymbol{X}}(\boldsymbol{x}|B)\, d\boldsymbol{x}\,.$$

Anche in questo caso il teorema fondamentale della media consente di definire un qualunque momento condizionale associato al vettore $\boldsymbol{X}$.

9.5.1 Teorema della media condizionale

La media condizionale può essere calcolata anche per le pdf condizionali $f_X(x|y)$ e $f_Y(y|x)$ viste nel § 9.3. Infatti, abbiamo visto che la pdf condizionale $f_Y(y|x)$ rappresenta la pdf di Y, per un fissato valore x della variabile aleatoria X. A tale pdf è associato un valor medio, che prende il nome di *media condizionale* di Y dato $X = x$:

Definizione (media condizionale di Y dato $X = x$). La media condizionale di Y dato $X = x$ è:

$$E(Y|x) \triangleq \int_{-\infty}^{\infty} y\, f_Y(y|x)\, dy\,,$$

se tale integrale esiste finito.

La definizione precedente si può estendere anche al caso in cui consideriamo una trasformazione $g(Y)$ di Y, e vogliamo calcolarne la media. Vale infatti anche in questo caso il teorema fondamentale della media, per cui:

$$E[g(Y)|x] = \int_{-\infty}^{\infty} g(y)\, f_Y(y|x)\, dy\,.$$

Osserviamo che, per ogni fissato x, la media condizionale $E[g(Y)|x]$ è un numero; se allora facciamo variare x, la media condizionale $E[g(Y)|x]$ definisce una funzione $\psi(x)$ di x. Possiamo allora costruire una variabile aleatoria $Z = \psi(X) = E[g(Y)|X]$ semplicemente associando ad ogni valore $X = x$ il valore $z = E[g(Y)|x]$. Il calcolo della media di Z rappresenta l'oggetto del seguente *teorema della media condizionale*:

Teorema 9.1 (media condizionale). Sia $E[g(Y)|x] = \psi(x)$ la media condizionale di $g(Y)$ dato $X = x$, e costruiamo la variabile aleatoria $Z = \psi(X) = E[g(Y)|X]$. Si ha:

$$E[E(g(Y)|X)] = E[g(Y)]\,,$$

se tale media esiste finita.

Prova. Con facili passaggi, si ha:

$$\begin{aligned} E[E(g(Y)|X)] &= \int_{-\infty}^{\infty} E[g(Y)|x]\, f_X(x)dx = \int_{-\infty}^{\infty} \left[\int_{-\infty}^{\infty} g(y)\, f_Y(y|x)\, dy\right] f_X(x)dx = \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(y)\, f_Y(y|x)\, f_X(x)\, dx\, dy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(y)\, f_{XY}(x,y)\, dx\, dy = \\ &= \int_{-\infty}^{\infty} g(y) \left[\int_{-\infty}^{\infty} f_{XY}(x,y)\, dx\right] dy = \int_{-\infty}^{\infty} g(y)\, f_Y(y)\, dy = E[g(Y)]\,, \end{aligned}$$

dove abbiamo applicato la legge della probabilità composta per le pdf, le relazioni tra statistiche congiunte e marginali, ed il teorema fondamentale della media. □

Il teorema si applica anche al caso in cui $g(Y) = Y$, ed in questo caso assume una forma particolarmente semplice:

$$\boxed{E[E(Y|X)] = E(Y)\,;}$$

questa relazione suggerisce una procedura in due passi per il calcolo della media di una variabile aleatoria Y che dipende da un'altra variabile aleatoria X; (i) si fissa prima un valore di $X = x$, e si calcola la media condizionale $E(Y|x)$; (ii) successivamente si media tale risultato rispetto a tutti i possibili valori di X, ottenendo la media $E(Y)$ cercata.

► *Esempio* 9.14. Consideriamo il caso di due variabili aleatorie congiuntamente gaussiane. Abbiamo osservato che $Y|x \sim \mathrm{N}(\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X), \sigma_Y\sqrt{1-\rho^2})$, per cui:

$$E(Y|x) = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x-\mu_X)\,.$$

Si ha, pertanto, mediando su X:

$$E[E(Y|X)] = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}\underbrace{E(X-\mu_X)}_{=0} = \mu_Y = E(Y)\,,$$

come previsto dal teorema della media condizionale. Questo esempio serve più per illustrare l'applicazione del teorema che per evidenziarne l'utilità pratica; si veda l'esempio 9.16 per un'applicazione più significativa. ◄

9.5.2 Generalizzazione al caso di coppie di variabili aleatorie★

Il teorema della media condizionale si può generalizzare al caso di coppie di variabili aleatorie, e precisamente si può applicare per calcolare la media di $g(X,Y)$. Supponiamo infatti di fissare

$X = x$ e di voler calcolare la media condizionale $E[g(X,Y)|X = x]$, che possiamo denotare sinteticamente come $E[g(X,Y)|x]$, essendo essa una funzione di x. Poiché l'evento condizionante è $B = \{X = x\}$, possiamo applicare il teorema fondamentale della media (9.14)

$$E[g(X,Y)|x] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(u,v)\, f_{XY}(u,v|x)\, du\, dv\,,$$

dove $f_{XY}(u,v|x)$ denota sinteticamente $f_{XY}(u,v|X = x)$. Il calcolo di tale pdf pone, tuttavia, qualche problema, se X è una variabile aleatoria continua; infatti, potremmo pensare di ottenere $f_{XY}(u,v|x)$ sulla base del risultato dell'esempio 9.10 che fornisce $f_{XY}(u,v|a < x \leq b)$, ponendo $a = x - \varepsilon$ e $b = x$, e facendo tendere ε a zero. Si avrebbe (cfr. esempio 9.10):

$$f_{XY}(u,v|x-\varepsilon < X \leq x) = \begin{cases} 0, & u \leq x-\varepsilon\,; \\ \frac{f_{XY}(u,v)}{F_X(x)-F_X(x-\varepsilon)}, & x-\varepsilon < u \leq x\,; \\ 0, & u > x\,. \end{cases}$$

Il problema è che passando poi al limite per $\varepsilon \to 0$ tale espressione diverge, in quanto $F_X(x) - F_X(x-\varepsilon) \to 0$, per cui la pdf $f_{XY}(u,v|x-\varepsilon < X \leq x)$ è singolare. Possiamo aggirare tale difficoltà calcolando direttamente la $E[g(X,Y)|x]$ con procedura al limite, ponendo cioè:

$$E[g(X,Y)|x] = E[g(X,Y)|X = x] = \lim_{\varepsilon\to 0} E[g(X,Y)|x-\varepsilon < X \leq x]\,.$$

Si trova:

$$E[g(X,Y)|x] = \int_{-\infty}^{\infty} g(x,y)\, f_Y(y|x)\, dy\,.$$

Prova. Si ha:

$$\begin{aligned} E[g(X,Y)|x-\varepsilon < X \leq x] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(u,v)\, f_{XY}(u,v|x-\varepsilon < X \leq x)\, du\, dv = \\ &= \int_{-\infty}^{\infty} dv \int_{x-\varepsilon}^{x} g(u,v)\frac{f_{XY}(u,v)}{F_X(x)-F_X(x-\varepsilon)}\, du = \\ &\approx \int_{-\infty}^{\infty} g(x,v)\,\frac{f_{XY}(x,v)}{F_X(x)-F_X(x-\varepsilon)}\,\varepsilon\, dv\,. \end{aligned}$$

Facendo tendere $\varepsilon \to 0$, si ha che $\frac{F_X(x)-F_X(x-\varepsilon)}{\varepsilon} \to f_X(x)$ (supposta esistente), per cui:

$$E[g(X,Y)|X = x] = \int_{-\infty}^{\infty} g(x,v)\,\frac{f_{XY}(x,v)}{f_X(x)} dv = \int_{-\infty}^{\infty} g(x,v)\, f_Y(v|x)\, dv\,,$$

cioè l'asserto, cambiando nome alla variabile di integrazione v. □

Siamo in grado adesso di formulare l'annunciata generalizzazione del teorema della media condizionale. Osserviamo che $E[g(X,Y)|x]$ rappresenta, anche in questo caso, al variare di x, una funzione $\psi(x)$; definiamo allora una variabile aleatoria $Z = \psi(X) = E[g(X,Y)|X]$, della quale calcoliamo la media. Si trova:

$$E[E[g(X,Y)|X]] = E[g(X,Y)]\,.$$

Prova. La prova è analoga a quella del teorema della media condizionale. Si ha:

$$\begin{aligned} E[E(g(X,Y)|X)] &= \int_{-\infty}^{\infty} E[g(X,Y)|x]\, f_X(x)dx = \int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty} g(x,y)\, f_Y(y|x)\, dy\right] f_X(x)dx = \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)\, f_Y(y|x)\, f_X(x)\, dx\, dy = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} g(x,y)\, f_{XY}(x,y)\, dx\, dy = \\ &= E[g(X,Y)]\,, \end{aligned}$$

cioè l'asserto. □

► *Esempio* 9.15. Applichiamo il risultato precedente per calcolare la correlazione tra due variabili aleatorie gaussiane. In questo caso $g(X, Y) = X\,Y$, e si ha:

$$E(X\,Y) = E[E(X\,Y|X)]\,.$$

Inoltre, risulta:

$$E(X\,Y|x) = E(x\,Y|x) = x\,E(Y|x)\,,$$

poiché x è fissato; poiché poi (cfr. esempio 9.14)

$$E(Y|x) = \mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x - \mu_X)\,,$$

allora si ha:

$$E(X\,Y|x) = x\,\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}(x^2 - x\,\mu_X)\,.$$

Mediando il risultato precedente rispetto ad X troviamo il risultato cercato:

$$E(X\,Y) = \mu_X\,\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}[E(X^2) - \mu_X^2] = \mu_X\,\mu_Y + \rho\frac{\sigma_Y}{\sigma_X}\,\sigma_X^2 = \mu_X\,\mu_Y + \rho\,\sigma_X\,\sigma_Y\,,$$

da cui si ha anche

$$\mathrm{Cov}(X, Y) = E(X\,Y) - \mu_X\,\mu_Y = \rho\,\sigma_X\,\sigma_Y\,,$$

per cui ritroviamo anche che $\rho_{XY} = \rho$, cioè il parametro ρ coincide con il coefficiente di correlazione. ◄

I concetti precedenti si estendono al caso di n variabili aleatorie in maniera naturale. Ad esempio, possiamo calcolare la media condizionale di X_1 per fissati valori $x_2, x_3, \ldots, x_n$ delle variabili aleatorie $X_2, X_3, \ldots, X_n$:

$$E(X_1|x_2, x_3, \ldots, x_n) \triangleq \int_{-\infty}^{\infty} x_1\, f_{X_1}(x_1|x_2, x_3, \ldots, x_n)\, dx_1\,. \tag{9.15}$$

La relazione precedente definisce una funzione $\psi(x_2, x_3, \ldots, x_n)$; se allora consideriamo la variabile aleatoria $Z = \psi(X_2, X_3, \ldots, X_n) \triangleq E(X_1|X_2, X_3, \ldots, X_n)$ e ne calcoliamo la media, si trova:

$$E[E(X_1|X_2, X_3, \ldots, X_n)] = E(X_1)\,.$$

che rappresenta la generalizzazione del teorema della media condizionale.

Prova. Applicando il teorema fondamentale della media, si ha:

$$E[E(X_1|X_2,X_3,\ldots,X_n)] = E[\psi(X_2,X_3,\ldots,X_n)] = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\psi(x_2,x_3,\ldots,x_n)\, f_{X_2X_3\cdots X_n}(x_2,x_3,\ldots,x_n)dx_2\,dx_3\cdots dx_n$$

Sostituendo la (9.15), si ha:

$$\begin{aligned}
&E[E(X_1|X_2,X_3,\ldots,X_n)] = \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}\left[\int_{-\infty}^{\infty}x_1\, f_{X_1}(x_1|x_2,x_3,\ldots,x_n)\,dx_1\right] f_{X_2X_3\cdots X_n}(x_2,x_3,\ldots,x_n)\,dx_2\,dx_3\cdots dx_n \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}x_1\,\underbrace{f_{X_1}(x_1|x_2,x_3,\ldots,x_n)\, f_{X_2X_3\cdots X_n}(x_2,x_3,\ldots,x_n)}_{=\, f_{X_1X_2\cdots X_n}(x_1,x_2,\ldots,x_n)}\,dx_1\,dx_2\cdots dx_n \\
&= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}x_1\, f_{X_1X_2\cdots X_n}(x_1,x_2,\ldots,x_n)\,dx_1\,dx_2\cdots dx_n \\
&= \int_{-\infty}^{\infty}x_1\,\underbrace{\left[\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty}f_{X_1X_2\cdots X_n}(x_1,x_2,\ldots,x_n)\,dx_2\,dx_3\cdots dx_n\right]}_{=\, f_{X_1}(x_1)} \\
&= \int_{-\infty}^{\infty}x_1\, f_{X_1}(x_1)\,dx_1 = E(X_1)\,,
\end{aligned}$$

cioè l'asserto. □

► *Esempio* 9.16 *(somma di un numero aleatorio di variabili aleatorie).* Si considerino le variabili aleatorie iid $X_1, X_2, \ldots, X_n$, con media μ e varianza σ^2, ed una variabile aleatoria N discreta, indipendente dalle precedenti, a valori in $\{1,2,\ldots,n\}$. Costruiamo la variabile aleatoria S come:

$$S = \sum_{k=1}^{N} X_k\,,$$

dove l'estremo superiore della somma è aleatorio. Calcolare media, valor quadratico medio e varianza di S.

Il problema si risolve semplicemente adoperando il teorema della media condizionata, ed in particolare condizionando ai possibili valori assunti da N. Infatti, per quanto riguarda il calcolo della media di S, si ha:

$$E(S) = E[E(S|N)]\,,$$

e, per un fissato valore $N = n$, risulta:

$$E(S|n) = E\left(\sum_{k=1}^{N}X_k\,\middle|\,N=n\right) = E\left(\sum_{k=1}^{n}X_k\right) = \sum_{k=1}^{n}E(X_k) = n\,\mu\,,$$

dove abbiamo sfruttato l'indipendenza tra N e le $X_1, X_2, \ldots, X_n$, per cui:

$$E(S) = E(N\,\mu) = E(N)\,\mu\,.$$

In maniera simile si può calcolare il valore quadratico medio, ovvero:

$$E(S^2) = E[E(S^2|N)]\,,$$

e si ha:

$$E(S^2|n) = E\left(\sum_{k=1}^{N}\sum_{h=1}^{N}X_k\,X_h\,\middle|\,N=n\right) = \sum_{k=1}^{n}\sum_{h=1}^{n}E(X_k\,X_h) = \sum_{k=1}^{n}\sum_{h=1}^{n}\Big[\mathrm{Cov}(X_k,X_h)+\mu^2\Big]\,.$$

Poichè le variabili aleatorie $X_1, X_2, \ldots, X_n$ sono indipendenti, allora esse sono anche incorrelate, per cui $\mathrm{Cov}(X_k, X_h) = \sigma^2\,\delta_{kh}$, e si ha quindi:

$$E(S^2|n) = n\,\sigma^2 + n^2\mu^2\,,$$

per cui

$$E(S^2) = E(N\,\sigma^2 + N^2\mu^2) = E(N)\,\sigma^2 + E(N^2)\,\mu^2\,.$$

La varianza si ottiene infine come:

$$\mathrm{Var}(S) = E(S^2) - E^2(S) = E(N)\,\sigma^2 + \mu^2[E(N^2) - E^2(N)] = E(N)\,\sigma^2 + \mu^2\,\mathrm{Var}(N)\,.$$

Se assumiamo N deterministico ($N = n$ con probabilità 1), ritroviamo $E(S) = n\,\mu$ e $\mathrm{Var}(S) = n\,\sigma^2$, come è naturale. ◄

9.6 Esercizi proposti

Esercizio 9.1. Il tempo di vita X di un dispositivo è modellato come una variabile aleatoria $X \sim \text{Exp}(\lambda)$. Sapendo che il dispositivo è vissuto fino al tempo $a > 0$, calcolare CDF e pdf del *tempo residuo di vita* $Y = X - a$. [*Risposta:* $Y \sim \text{Exp}(\lambda)$]

Esercizio 9.2. Il tempo di vita (misurato in settimane) di un componente elettronico è modellato come una variabile aleatoria $X \sim \text{Rayleigh}(b)$, con $b = 30$. Se per qualche motivo è noto che il dispositivo non durerà più di 20 settimane, determinare la CDF e la pdf del nuovo tempo di vita X.

Esercizio 9.3. Il numero di prove che intercorrono tra due successi consecutivi in un esperimento di prove ripetute è modellato come una variabile aleatoria $X \sim \text{Geom}(p)$. Sapendo che sono già trascorse $\overline{k} > 0$ prove senza alcun successo, calcolare la DF del *numero residuo di prove* $Y = X - \overline{k}$.

Esercizio 9.4. Sia $X \sim \text{U}(0, 2\pi)$. Determinare la CDF e la pdf della variabile aleatoria X condizionata all'evento $B = \{\cos(X) \geq 0\}$.

Esercizio 9.5. Siano X ed Y due variabili aleatorie con pdf congiunta

$$f_{XY}(x,y) = \begin{cases} 2, & \text{se } 0 \leq x \leq 1 \text{ e } 0 \leq y \leq x, \\ 0, & \text{altrimenti.} \end{cases}$$

a) Determinare le pdf condizionali $f_X(x|y)$ e $f_Y(y|x)$;

b) verificare che le pdf condizionali determinate al punto 1 soddisfino la condizione di normalizzazione per le pdf.

Esercizio 9.6. Siano X ed Y due variabili aleatorie con pdf congiunta

$$f_{XY}(x,y) = u(x)\, u(y)\, x\, e^{-x(y+1)}, \qquad (x,y) \in \mathbb{R}^2$$

a) Determinare le pdf condizionali $f_X(x|y)$ e $f_Y(y|x)$;

b) verificare che le pdf condizionali determinate al punto 1 soddisfino la condizione di normalizzazione per le pdf;

c) utilizzando le pdf condizionali precedentemente calcolate, determinare il valore di $P(Y \leq 2|X = 1)$.

Esercizio 9.7. Si supponga che le variabili aleatorie X ed Y abbiano la seguente pdf:

$$f_{XY}(x,y) = \begin{cases} k, & \text{se } x^2 + y^2 \leq 1, \\ 0, & \text{altrimenti.} \end{cases}$$

a) Determinare il valore di k;

b) determinare le pdf condizionali $f_X(x|y)$ e $f_Y(y|x)$.

Esercizio 9.8. Siano $X \sim \text{Geom}(p$ ed $Y \sim \text{Geom}(p)$ due variabili aleatorie indipendenti, aventi entrambe distribuzione geometrica. Calcolare $P(X = Y)$.
[*Risposta:* $p^2/(1-q^2)$]

Esercizio 9.9. Si generalizzi il concetto di variabile aleatoria binomiale nel seguente modo: la probabilità p di un successo non è più una costante, ma una variabile aleatoria $P \sim \text{U}(0,1)$, per cui il numero di successi in n prove ha la distribuzione condizionale $X|p \sim \text{B}(n,p)$. Calcolare la DF di X.
[*Risposta:* $p_X(k) = 1/(n+1)$ *per* $0 \leq k \leq n$ *(uniforme).*]
[*Suggerimento: sfruttare l'integrale notevole* $\int_0^1 x^k (1-x)^{n-k}\, dx = \frac{k!(n-k)!}{(n+1)!}$]

Esercizio 9.10. Sia assegnata la variabile aleatoria $X = G\,Y$, con $Y \sim \text{N}(0,\sigma)$ e G variabile aleatoria discreta con pdf $f_G(x) = 0.5[\delta(x-1) + \delta(x+1)]$, indipendente da Y. Valutare la pdf di X.
[*Suggerimento: condizionare ai possibili valori assunti da G.*]

Esercizio 9.11. Utilizzando i concetti relativi a CDF e pdf condizionali, provare che se X ed Y sono due variabili aleatorie indipendenti, si ha:

$$P(Y \leq X) = \int_{-\infty}^{\infty} F_Y(x)\, f_X(x)\, dx\,.$$

[*Suggerimento: condizionare ai possibili valori assunti da x.*]

Esercizio 9.12. Siano X ed Y due variabili aleatorie indipendenti ed uniformi in $(0,1)$, e si consideri la seguente trasformazione di variabili aleatorie:

$$\begin{cases} Z &= X - Y \\ V &= X + Y \end{cases}$$

Determinare la pdf condizionata $f_Z(z|V \leq 1)$.

Esercizio 9.13. Siano X ed Y due variabili aleatorie indipendenti ed esponenziali di parametro λ. Mostrare che la pdf di X dato $X + Y = v$ ($v \geq 0$) è uniforme in $(0, v)$.
[*Suggerimento: Porre $Z = X + Y$ e $W = X$ e calcolare la pdf congiunta di Z e W.*]

Esercizio 9.14. Disponendo di un sottoprogramma che genera variabili aleatorie uniformi $\mathrm{U}(0,1)$ e di uno che genera variabili aleatorie gaussiane standard, delineare una procedura per generare osservazioni di una variabile aleatoria X "mixture" di più gaussiane, avente, cioè, la seguente pdf

$$f_X(x) = \sum_{i=1}^{N} \varepsilon_i \frac{1}{\sigma_i\sqrt{2\pi}} \exp\left\{-\frac{(x-\mu_i)^2}{2\sigma_i^2}\right\}; \qquad \varepsilon_i \geq 0, \qquad \sum_{i=1}^{N} \varepsilon_i = 1$$

Esercizio 9.15. La pdf congiunta di quattro variabili aleatorie X_1, X_2, X_3, X_4 è:

$$f_{X_1X_2X_3X_4}(x_1, x_2, x_3, x_4) = \prod_{i=1}^{4} \exp(-2|x_i|)$$

Determinare le seguenti pdf condizionali:

a) $f_{X_1X_2X_3}(x_1, x_2, x_3|x_4)$;

b) $f_{X_1,X_2}(x_1, x_2|x_3, x_4)$;

c) $f_{X_1}(x_1|x_2, x_3, x_4)$.

Esercizio 9.16. Si consideri una variabile aleatoria $X \sim \mathrm{N}(0,\sigma)$. Calcolare $E(X|X \geq 0)$.

Esercizio 9.17. Siano X ed Y due variabili aleatorie con pdf congiunta $f_{XY}(x,y) = 8\,x\,y$, per $0 < y < x < 1$. Determinare $E(X|y)$ e $E(Y|x)$.
[*Risposta:* $E(X|y) = \frac{2}{3}\left(\frac{1-y^3}{1-y^2}\right)$; $E(Y|x) = \frac{2}{3}\,x$.]

Esercizio 9.18. Siano X ed Y due variabili aleatorie con pdf congiunta $f_{XY}(x,y) = 8\,x\,y$, per $0 < x < y < 1$.

a) Determinare $E(Y|x)$;

b) determinare $E(XY|x)$;

c) determinare $\mathrm{Var}(Y|x)$.

Esercizio 9.19. Siano X ed Y due variabili aleatorie dipendenti, con $Y \sim \mathrm{U}(0,5)$. Calcolare $E(X)$, sapendo che:

$$f_X(x|Y=y) = \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}(x-y)^2\right]\,.$$

[*Suggerimento: applicare il teorema della media condizionata.*]

Esercizio 9.20. Si consideri la variabile aleatoria $Y = X^{2B}$, con $X \sim \mathrm{N}(0,\sigma)$ e $B \sim \mathrm{Bern}(p)$, indipendenti tra loro. Calcolare $E(Y)$ e $\mathrm{Var}(Y)$.
[*Suggerimento: applicare il teorema della media condizionata.*]

Capitolo 10

Elementi di teoria dell'informazione

Molti dei risultati della teoria della probabilità trovano applicazione in questo capitolo nell'ambito della teoria dell'informazione, ed in particolare della codifica di sorgente. Dopo aver introdotto la definizione di autoinformazione ed entropia, si riprende la definizione di sorgente di informazione già introdotta nel cap. 2 e si mostra come la quantità di informazione da essa prodotta possa essere misurata mediante l'entropia della sorgente ed il tasso di informazione della sorgente. Successivamente si considera il problema della codifica di sorgente: si definiscono i codici a lunghezza fissa e variabile, i codici univocamente decifrabili e a prefisso, e si mostra come per lo studio di questi ultimi possa essere utilmente introdotto il concetto di albero di codice. Infine si studiano le fondamentali relazioni esistenti tra la lunghezza media di un codice e l'entropia di sorgente (primo teorema di Shannon), e si introducono i codici di Shannon e di Huffmann; questi ultimi, in particolare, risultano ottimi tra tutti i codici che operano su blocchi di sorgente di lunghezza prefissata.

10.1 Introduzione

Lo scopo della *teoria dell'informazione* è quello di individuare le basi teoriche per lo studio dei problemi riguardanti la trasmissione, la ricezione, l'elaborazione e la memorizzazione dell'informazione. Tale disciplina è relativamente recente, essendo nata solo negli anni '40 e principalmente per effetto di un singolo e decisivo contributo del ricercatore statunitense della Bell Claude E. Shannon, che pubblicò nel 1948 il fondamentale "A mathematical theory of communication", nel quale si sviluppano i principali concetti della teoria.[1]

Il punto di partenza della teoria dell'informazione è ovviamente definire il concetto stesso di *informazione*, che ricorre in varie discipline e assume significati e sfumature differenti a seconda dei contesti nei quali viene utilizzato. Noi ci riferiremo al caso di un sistema di comunicazione (lo

[1] Si veda l'URL `http://cm.bell-labs.com/cm/ms/what/shannonday/paper.html` per una versione Postscript o pdf del lavoro di Shannon.

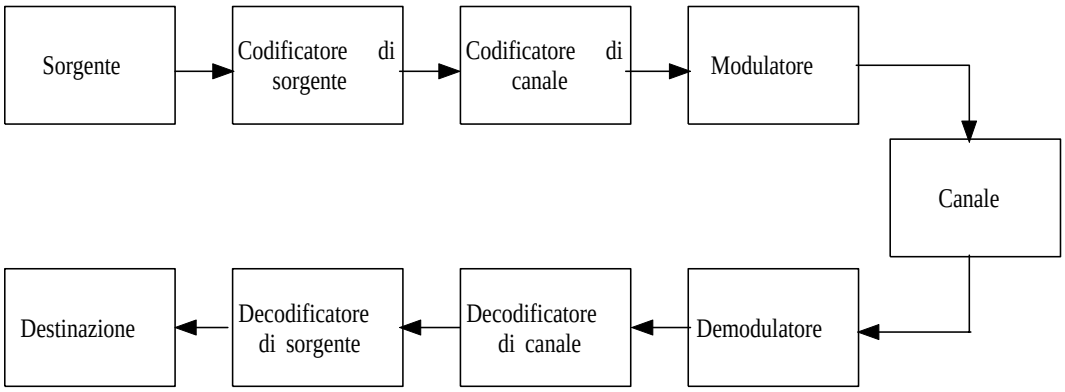


Fig. 10.1. Schema di Shannon di un sistema di comunicazione per la trasmissione di informazione da una sorgente ad una destinazione.

schema in Fig. 10.1 è dovuto allo stesso Shannon), nel quale l'informazione emessa da una *sorgente* viene trasportata fino ad una *destinazione*, mediante un *canale* di comunicazione; sorgente e destinazione possono essere due persone (es. comunicazione telefonica), due apparecchiature (es. comunicazione tra due calcolatori), o due parti di una stessa apparecchiatura (es. comunicazione tra microprocessore e memoria RAM di un calcolatore).

Spesso l'informazione è di natura *simbolica*, o può comunque essere espressa mediante un insieme di simboli (si pensi alle lettere dell'alfabeto); tale rappresentazione non è necessariamente efficiente, anzi contiene sovente un elevato grado di *ridondanza*. Poiché il trasporto e la memorizzazione di ridondanza comporta uno spreco di risorse, compito del *codificatore di sorgente* è quello di fornire una rappresentazione il più possibile compatta e sintetica dei simboli emessi dalla sorgente, eliminando se possibile ogni ridondanza (il decodificatore di sorgente opera la trasformazione inversa alla destinazione).

Poiché poi il canale di comunicazione è un canale *fisico* (ad esempio, un doppino telefonico o una fibra ottica nella comunicazione cablata, oppure lo spazio libero nella comunicazione radio), si richiede per la trasmissione che i simboli vengano rappresentati con segnali fisici (elettrici, ottici o di altra natura): questo compito è svolto dal *modulatore*, mentre il demodulatore opera la trasformazione inversa alla destinazione (il blocco modulatore / demodulatore è spesso comunemente denominato *modem*).

Osserviamo infine che qualunque canale di comunicazione è *rumoroso*, cioè è affetto da rumore, il quale a sua volta introduce *errori* nella trasmissione; per contrastare l'effetto di tali errori, e consentire comunque una comunicazione affidabile, prima della modulazione si effettua generalmente una *codifica di canale*, con lo scopo di introdurre una quantità controllata di *ridondanza* per "irrobustire" la trasmissione dell'informazione attraverso il canale (un semplice esempio di codifica di canale è costituito dal *bit di parità* che viene concatenato ad una stringa di bit prima della trasmissione). Tale ridondanza viene rimossa alla destinazione dal decodificatore di canale.

► *Esempio* 10.1. Per fornire un esempio tratto dall'esperienza quotidiana, supponiamo di voler invitare il nostro amico Mario Rossi, che vive all'estero, alla nostra laurea, e di volerlo fare per telegramma. La *codifica di sorgente* in questo caso consiste nel trasformare il nostro invito in una formula telegrafica, del tipo "GRADITA TUA PARTECIPAZIONE MIA LAUREA PROSSIMO 15 GIUGNO ORE 9:00 FACOLTA INGEGNERIA NAPOLI", nella quale abbiamo fornito le informazioni essenziali, eliminando un certo livello di ridondanza tipica della lingua parlata. A questo punto, telefoniamo al servizio dettatura telegrammi, e

per essere sicuri che l'impiegato (un po' duro d'orecchi) capisca bene tutte le parole del telegramma, le ripetiamo scandendole più volte; in particolare per fargli capire che il cognome è Rossi usiamo frasi del tipo "R come Roma, O come Orvieto, S come Sassari, etc.": in questo caso stiamo introducendo ridondanza controllata, ovvero stiamo effettuando una *codifica di canale*. ◄

► *Esempio* 10.2. Un altro esempio significativo è quello dell'invio di un documento di testo come allegato (attachment) ad un messaggio di posta elettronica. Poiché un file generato da un editor di testo evoluto (ad esempio, Microsoft Word) contiene un elevato grado di ridondanza, come testimoniato dalle considerevoli dimensioni dei file `.doc` anche per documenti consistenti in poche righe di testo, al fine di utilizzare efficientemente le risorse di comunicazione a disposizione (e quindi spendere di meno in un collegamento via linea telefonica) è opportuno effettuare una *codifica di sorgente*, ad esempio mediante il programma Winzip, che riduce considerevolmente le dimensioni del file. Prima tuttavia di inviare il messaggio, attraverso il modulatore, sul canale (in questo caso, un doppino telefonico oppure un cavo Ethernet) è necessario proteggerlo introducendo una *codifica di canale* (ad esempio dei bit di parità) in modo da poter rivelare, ed eventualmente correggere, eventuali errori inevitabilmente causati dal rumore presente sul canale. ◄

Le principali aree di studio della teoria dell'informazione corrispondono all'incirca ai blocchi funzionali dello schema di Shannon (Fig. 10.1); in particolare, essa si occupa dei seguenti problemi:

1. la rappresentazione dell'informazione nella forma più efficiente possibile, eliminando ogni possibile ridondanza, in modo da ridurre il numero di simboli necessari per la sua descrizione (*codifica di sorgente*);

2. la determinazione della massima quantità di informazione che è possibile trasmettere con degradazione piccola a piacere su un canale di trasmissione (*capacità di canale*);

3. l'introduzione di ridondanza controllata in trasmissione, così da limitare le degradazioni introdotte dal canale (*codifica di canale*).

In particolare, il successo di una particolare soluzione per la trasmissione dell'informazione risiede tutto in un accorto bilanciamento tra la *riduzione della ridondanza* (effettuata dal codificatore di sorgente) e *l'aumento della ridondanza* (effettuata dal codificatore di canale).

La teoria dell'informazione è una disciplina dal forte contenuto matematico, e noi ci limiteremo semplicemente ad introdurre i concetti fondamentali, quali la misura dell'informazione, ed a fornire qualche elemento di codifica di sorgente, tralasciando del tutto la codifica di canale. Per non sottovalutare l'importanza applicativa di tale disciplina, tuttavia, menzioniamo solo che alcuni tra i più importanti standard tecnologici utilizzati al giorno d'oggi (JPEG, MPEG, MP3, ADSL tra essi) devono la loro nascita ai risultati della teoria dell'informazione.

10.2 Misura dell'informazione ed entropia

Come già accennato, "informazione" è un concetto necessariamente vago, che talvolta assume caratteristiche soggettive; tuttavia, per costruire una teoria matematica, dovremo definirla in maniera più rigorosa, tanto rigorosa da fornire degli strumenti per *misurarla.*

L'osservazione fondamentale per arrivare ad introdurre una misura dell'informazione è che il concetto di informazione è intrinsecamente associato a quello di *impredicibilità* o di *incertezza.* Ad esempio, se telefoniamo al servizio informazioni meteorologiche in pieno agosto, e ci viene detto "domani sarà una bella giornata", sicuramente attribuiremo a tale asserzione un minore contenuto informativo rispetto ad una previsione del tipo "domani si scatenerà un uragano tropicale", semplicemente perchè alle nostre latitudini e nel mese di agosto la prima eventualità è

sicuramente di gran lunga *più probabile* della seconda. Per questo motivo, a livello intuitivo accettiamo che l'informazione associata ad un evento sia *inversamente proporzionale* alla probabilità con la quale quel dato evento può verificarsi. Sulla base di questa osservazione, possiamo passare ad introdurre una definizione operativa di "misura" dell'informazione. Parlando di eventi e di probabilità, è naturale modellare l'oggetto del nostro studio come un esperimento aleatorio, dotato di struttura di spazio di probabilità *discreto*[2] $(\Omega, \mathcal{S}, P)$. Poichè intendiamo misurare l'informazione associata ad eventi di Ω, supponiamo (senza ledere la generalità) che ai possibili risultati dell'esperimento siano associati biunivocamente i valori $x \in \mathcal{X} = \{x_1, x_2, \ldots, x_n, \ldots\}$ assunti da una variabile aleatoria discreta X, avente DF $p_X(x) \triangleq P(X = x)$. Per comodità di notazione, porremo talvolta $p_k \triangleq p_X(x_k)$; supporremo poi per semplicità che la variabile aleatoria assuma un numero *finito* di valori $x_1, x_2, \ldots, x_K$, dove $K = \text{card}(\Omega)$.

10.2.1 Autoinformazione

Avendo osservato che ad una minore probabilità corrisponde una maggiore quantità di informazione, definiamo *l'autoinformazione* dell'evento $\{X = x\}$:

Definizione (autoinformazione). Sia X una variabile aleatoria discreta a valori $x \in \mathcal{X}$ e con DF $p_X(x)$: l'autoinformazione associata all'evento $\{X = x\}$ è data da:

$$I(x) \triangleq \log \frac{1}{p_X(x)} = -\log p_X(x)\,,$$

dove il logaritmo è in una base qualsiasi maggiore di 1.

La notazione $I(x)$ è leggermente ambigua, in quanto il valore dell'autoinformazione non dipende in effetti dal *valore* di x, ma solo dalla sua *probabilità* $p_X(x)$; essa può assumere solo valori maggiori o uguali a zero (in quanto $p_X(x)$, essendo una probabilità, è minore o uguale ad uno), e assume valori tanto maggiori quanto *meno* probabile è l'evento $\{X = x\}$: in particolare, se $p_X(x) \to 0$, l'autoinformazione $I(x)$ tende all'infinito, mentre se $p_X(x) \to 1$, l'autoinformazione $I(x)$ tende a zero. Tale comportamento soddisfa il ragionamento intuitivo effettuato in precedenza: il verificarsi di un evento poco probabile possiede un maggior contenuto informativo rispetto ad un evento molto probabile o addirittura certo. La presenza del logaritmo nella definizione di autoinformazione si può poi giustificare per la proprietà di tale funzione di trasformare prodotti in somme. Infatti, siano X ed Y due variabili aleatorie discrete con DF congiunta $p_{XY}(x,y)$: l'autoinformazione associata all'evento $\{X = x, Y = y\}$ è

$$I(x,y) = \log \frac{1}{p_{XY}(x,y)}\,.$$

Se gli eventi $\{X = x\}$ ed $\{Y = y\}$ sono indipendenti, la DF congiunta $p_{XY}(x,y)$ si fattorizza nel prodotto delle DF marginali, e quindi si ha:

$$I(x,y) = \log \frac{1}{p_X(x)\,p_Y(y)} = \log \frac{1}{p_X(x)} + \log \frac{1}{p_Y(y)} = I(x) + I(y)\,,$$

per cui l'autoinformazione associata ad eventi indipendenti è la *somma* delle autoinformazioni associate ai singoli eventi, come pare intuitivamente accettabile.

[2] La misura dell'informazione associata a spazi di probabilità continui è un problema matematicamente più complesso, che non prenderemo in cosiderazione.

Sebbene in teoria il logaritmo possa essere calcolato in una base qualsiasi maggiore di uno, le scelte di gran lunga più comuni sono il logaritmo in base *e* (logaritmo naturale o neperiano), che denoteremo con $\ln(\cdot)$, oppure il logaritmo in base 2, che denoteremo semplicemente con $\log(\cdot)$; nel primo caso, l'autoinformazione si misura in "nat", nel secondo si misura in "bit". [3] Poiché $\log x = \ln x / \ln 2$, per convertire l'informazione da nat a bit, e viceversa, basta applicare le seguenti relazioni:

$$\begin{array}{lll} [I(x)]_{\text{nat}} & = \ln 2\, [I(x)]_{\text{bit}} & = 0.693\, [I(x)]_{\text{bit}}\,; \\ [I(x)]_{\text{bit}} & = \frac{1}{\ln 2}\, [I(x)]_{\text{nat}} & = 1.443\, [I(x)]_{\text{nat}}\,. \end{array}$$

Nel seguito, misureremo l'informazione sempre in bit. Notiamo che non bisogna confondere il "bit" come unità di misura dell'informazione con il "bit" inteso come simbolo binario (0 oppure 1), come il seguente esempio dovrebbe chiarire.

► *Esempio* 10.3. Supponiamo che X assuma K valori equiprobabili, per cui $p_X(x) = \frac{1}{K}$. In tal caso, l'autoinformazione associata ad un qualunque valore di X è la stessa, e vale

$$I(x) = \log \frac{1}{p_X(x)} = \log K\,.$$

Ad esempio, supponiamo di avere una stringa composta da n simboli binari (bit); possiamo costruire $K = 2^n$ di tali stringhe, e se esse sono ugualmente probabili, l'autoinformazione $I(x)$ associata a ciascuna di tali stringhe sarà $I(x) = \log 2^n = n$ (in bit). Pare abbastanza naturale che l'informazione associata ad una stringa di n bit sia pari ad n bit! Osserviamo, tuttavia, che questo è vero solo nell'ipotesi che le K stringhe siano *equiprobabili*: se ciò non accade, l'autoinformazione di ogni stringa potrà essere maggiore o minore di n bit. Quindi la conclusione leggermente paradossale è: "una stringa di n bit non equivale sempre ad n bit di informazione"! ◄

10.2.2 Entropia

A questo punto, osserviamo che l'autoinformazione $I(x) = -\log p_X(x)$ precedentemente definita è una funzione che associa ad ogni x il numero reale e positivo $I(x)$. Pertanto, al variare di $x \in \mathcal{X}$, tale funzione definisce una variabile aleatoria $I(X) = -\log p_X(X)$, funzione della variabile aleatoria X. La media statistica di tale variabile aleatoria (facilmente calcolabile utilizzando il teorema fondamentale della media) rappresenta una misura *media* dell'autoinformazione associata alla variabile aleatoria X che, per affinità con la corrispondente grandezza termodinamica, prende il nome di *entropia* (informazionale):

Definizione (entropia). Data una variabile aleatoria X, l'entropia di X è la media dell'autoinformazione $I(x)$, ed è data da:

$$H(X) \triangleq E[-\log p_X(X)] = -\sum_{x\in\mathcal{X}} p_X(x)\, \log p_X(x) = \sum_{x\in\mathcal{X}} p_X(x)\, \log \frac{1}{p_X(x)}\,,$$

dove il logaritmo è in una base qualsiasi maggiore di 1.

Come già osservato per l'autoinformazione, anche l'entropia $H(X)$ non dipende in effetti dai *valori* assunti dalla variabile aleatoria, ma soltanto dalle *probabilità* $p_X(x)$ con cui vengono assunti tali valori. Quindi in sostanza l'entropia non dipende dal "significato" dell'informazione ma solo

[3] Il termine "bit", proposto da J.W. Tukey, è l'acronimo per *bi*nary digi*t*.

dalle probabilità con cui essa si può manifestare. Per enfatizzare tale dipendenza, se denotiamo tali probabilità (in numero finito) con $p_k = p_X(x_k)$, $k = 1, 2, \ldots, K$, e costruiamo il vettore $\boldsymbol{p} = [p_1, p_2, \ldots, p_K]$, possiamo parlare di entropia associata al vettore di probabilità $\boldsymbol{p}$, e scrivere anche $H(\boldsymbol{p})$ in luogo di $H(X)$.

► *Esempio* 10.4. . Supponiamo che lo spazio di probabilità contenga due soli eventi di interesse, ad esempio $A = \{\text{oggi piove}\}$ e $\overline{A} = \{\text{oggi non piove}\}$. È chiaro che possiamo descrivere numericamente tale esperimento mediante una variabile aleatoria bernoulliana $X \sim \text{Bern}(p)$, a valori 0 ed 1, dove possiamo convenzionalmente associare il valore 1 ad A ed il valore 0 a $\overline{A}$. In ogni caso, l'entropia associata ad X non dipende dai valori della variabile aleatoria, ma solo dalle probabilità con cui tali valori sono assunti; essa si calcola immediatamente dalla definizione, e vale

$$\boxed{H(X) = H(p) = -p \log p - (1-p)\log(1-p)\,.} \tag{10.1}$$

Tale entropia si denota anche con $H(p)$, poiché dipende solo dal valore di p, e prende il nome di *entropia binaria*. Il suo andamento è diagrammato in Fig. 10.2, dalla quale si vede che essa vale 0 per $p = 0$ oppure $p = 1$, mentre è massima (vale 1 bit) per $p = 1/2$ (osserviamo che poniamo $0 \log 0 = \lim_{p\to 0} p \log p = 0$). Anche in questo caso, allora, per specificare una tra due alternative *equiprobabili* ($p = 1/2$) occorre un bit

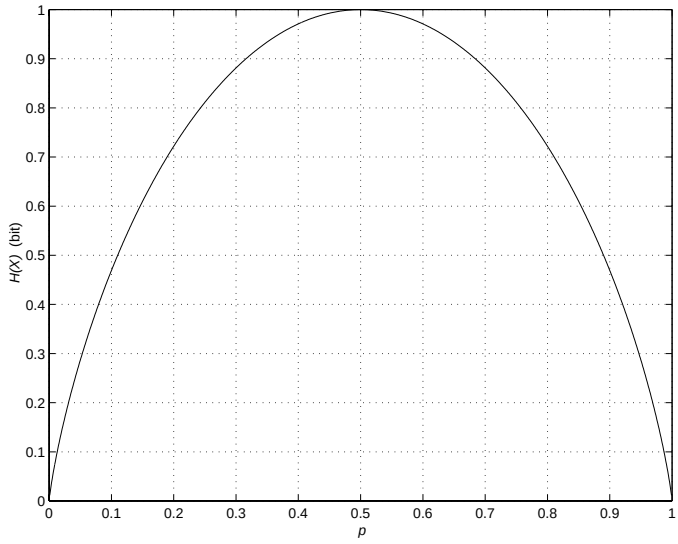


Fig. 10.2. Entropia binaria $H(X)$ (espressa in bit) in funzione della probabilità p.

di informazione, mentre per specificare una tra due alternative *non equiprobabili* è sufficiente una quantità di informazione inferiore ad 1 bit. Pertanto, l'equiprobabilità, essendo la situazione di massima incertezza, equivale anche alla massima informazione, il che pare intuitivamente accettabile. ◄

10.2.3 Proprietà dell'entropia

L'entropia gode delle seguenti proprietà fondamentali, alcune delle quali di immediata verifica ed interpretazione:

1. $H(X)$ è simmetrica rispetto al vettore di probabilità $\boldsymbol{p}$ da cui dipende, nel senso che se si effettua una permutazione degli elementi del vettore $\boldsymbol{p}$ l'entropia non cambia.

2. $H(X)$ è nulla se e solo se la distribuzione di probabilità è del tipo:

$$p_X(x) = \begin{cases} 1, & \text{per } x = \overline{x}; \\ 0, & \text{altrimenti.} \end{cases}$$

3. $H(X)$ è *massima*, e vale $H(X) = \log K$, se e solo se $p_X(x) = \frac{1}{K}$ (alternative equiprobabili).

Di queste proprietà, la prima riafferma che l'entropia non dipende dall'ordine in cui si considerano i possibili eventi; la seconda afferma che, se esiste un risultato certo (per cui gli altri hanno necessariamente probabilità nulla), il contenuto informativo medio è nullo; infine, la terza proprietà afferma che il contenuto informativo, a parità di alternative, è massimo se tali alternative sono equiprobabili. Notiamo che, a differenza dell'autoinformazione che può assumere il valore infinito, l'entropia, che ne rappresenta il valor medio, vale al più $\log K$, dove K è il numero delle possibili alternative; inoltre, al crescere di K, tale valore massimo dell'entropia aumenta, il che significa ovviamente che ad un maggior numero di alternative è associata potenzialmente una maggiore quantità di informazione.

10.2.4 Entropia congiunta

Nelle precedenti sezioni, abbiamo definito l'entropia di una singola variabile aleatoria X. È immediato estendere tale definizione al caso di due o più variabili aleatorie $X_1, X_2, \ldots, X_n$:

Definizione (entropia congiunta). Date n variabili aleatorie $\boldsymbol{X} = [X_1, X_2, \ldots, X_n]^T$, a valori $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]$ in $\mathcal{X} = \mathcal{X}_1 \times \mathcal{X}_2 \cdots \times \mathcal{X}_n$, con DF congiunta $p_{\boldsymbol{X}}(\boldsymbol{x})$, l'entropia congiunta $H(\boldsymbol{X})$ di $\boldsymbol{X}$ è data da:

$$H(\boldsymbol{X}) \triangleq E[-\log p_{\boldsymbol{X}}(\boldsymbol{X})] = -\sum_{\boldsymbol{x}\in\mathcal{X}} p_{\boldsymbol{X}}(\boldsymbol{x}) \log p_{\boldsymbol{X}}(\boldsymbol{x}) \,,$$

dove il logaritmo è in una base qualsiasi maggiore di 1.

Ad esempio, nel caso $n = 2$, ponendo $X_1 = X$ ed $X_2 = Y$, si ha esplicitamente:

$$H(X,Y) = -\sum_{x\in\mathcal{X}, y\in\mathcal{Y}} p_{XY}(x,y) \log p_{XY}(x,y) \,.$$

Se le variabili aleatorie X ed Y sono indipendenti, la DF congiunta si fattorizza, e per la proprietà del logaritmo di trasformare prodotti in somme, si ha:

$$\begin{aligned} H(X,Y) &= -\sum_{x\in\mathcal{X}, y\in\mathcal{Y}} p_{XY}(x,y) \log[p_X(x)\, p_Y(y)] = \\ &= -\sum_{x\in\mathcal{X}, y\in\mathcal{Y}} p_{XY}(x,y) \log p_X(x) - \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} p_{XY}(x,y) \log p_Y(x) \\ &= H(X) + H(Y) \,, \end{aligned}$$

dove abbiamo sfruttato la relazione tra DF congiunte e marginali, e la definizione di entropia. Pertanto l'entropia congiunta associata a due variabili aleatorie indipendenti è la somma delle entropie delle singole variabili aleatorie. Il risultato si generalizza ovviamente al caso di un vettore $\boldsymbol{X} = [X_1, X_2, \ldots, X_n]$ di n variabili aleatorie indipendenti, per il quale si ha:

$$H(\boldsymbol{X}) = \sum_{i=1}^{n} H(X_i) \,;$$

se poi le variabili aleatorie che compongono $\boldsymbol{X}$, oltre ad essere indipendenti, sono anche *identicamente distribuite*, si ha $H(X_i) = H(X_j) = H(X_1)$, per cui si ottiene semplicemente:

$$H(\boldsymbol{X}) = n\, H(X_1)\,.$$

Nel caso in cui le variabili aleatorie che compongono il vettore $\boldsymbol{X}$ non sono indipendenti, vale comunque la disuguaglianza seguente:

$$H(\boldsymbol{X}) \leq \sum_{i=1}^{n} H(X_i)\,;$$

per cui l'entropia congiunta risulta essere massima se e solo se le variabili aleatorie $X_1, X_2, \ldots, X_n$ sono indipendenti.

Osserviamo, infine, che sostituendo alle DF congiunte le *DF condizionali*, è possibile definire anche le *entropie condizionali*, la cui trattazione esula comunque dalla natura introduttiva di questo capitolo.

► *Esempio* 10.5. Consideriamo ancora il caso della stringa di n bit, che possiamo riguardare come n variabili aleatorie iid $X_1, X_2, \ldots, X_n$, con $X_i \sim \text{Bern}(p)$. In tal caso, si ha:

$$H(\boldsymbol{X}) = n\, H(X_1) = n\, H(p)\,,$$

dove $H(p)$ è l'entropia binaria data dalla (10.1). Pertanto l'andamento dell'entropia $H(\boldsymbol{X})$ al variare di p è, a meno di un fattore di scala n, lo stesso di quello di Fig. 10.2; la conclusione è che il valore massimo di $H(\boldsymbol{X})$ al variare di p vale n, e si ottiene quando $p = 1/2$. In tutti gli altri casi, l'entropia di una stringa di n bit assume un valore *inferiore* ad n bit. ◄

10.3 Sorgenti di informazione

Con riferimento allo schema di Shannon (Fig. 10.1), il punto di partenza per affrontare un qualunque problema di teoria dell'informazione è definire con esattezza cosa intendiamo per *sorgente di informazione* e misurare la quantità di informazione da essa prodotta: senza dare una definizione formale, possiamo assimilare una sorgente di informazione *discreta* ad un dispositivo (fisico, elettronico, astratto etc.) che emette "simboli" appartenenti ad un insieme discreto con una determinata regolarità statistica. Alcuni esempi di sorgenti discrete di informazione sono i seguenti: un utente che scrive ad una tastiera alfanumerica di un calcolatore (i simboli sono in tal caso le lettere dell'alfabeto più i caratteri speciali); un termometro che registra i valori della temperatura esterna a passi di 1^{o}C (i simboli sono in tal caso un sottoinsieme dei numeri interi relativi); la successione dei valori dell'indice di borsa italiana (Mibtel) nei diversi giorni della settimana (i simboli sono in tal caso numeri interi). Non tutte le sorgenti di informazione sono, ovviamente, discrete (sia nei valori prodotti, che nel tempo); molte sorgenti emettono simboli appartenenti ad un insieme continuo e con continuità nel tempo (ad esempio, un voltmetro analogico per la misura della tensione continua in un dispositivo elettronico può emettere in ogni istante *reale* un valore *reale* di tensione nell'intervallo $[-V, V]$, dove V è il valore di fondoscala). Comunque, nel seguito, coerentemente con la scelta di introdurre la misura dell'informazione solo negli spazi di probabilità discreti, ci limiteremo a considerare esclusivamente il caso di sorgenti discrete.[4]

[4] In molti casi, i risultati ottenuti sono applicabili anche al caso delle sorgenti continue, purché queste siano appropriatamente *discretizzate*, ad esempio con una procedura di *campionamento* (per la discretizzazione dei tempi) e *quantizzazione* (per la discretizzazione dei valori).

10.3.1 Entropia di sorgente

Consideriamo una sorgente discreta che emette simboli in istanti discreti di tempo, che denotiamo convenzionalmente con $n = 1, 2, \ldots$, ovvero $n \in \mathbb{N}$. In un generico istante $n \in \mathbb{N}$, data l'incertezza sul suo valore, il simbolo emesso può essere modellato come una variabile aleatoria X_n che assume valori in un un *alfabeto* numerico[5] di cardinalità K finita, sia esso $\mathcal{X} = \{x_1, x_2, \ldots, x_K\}$; notiamo esplicitamente che l'alfabeto di sorgente non cambia al variare del tempo, in altri termini l'insieme $\mathcal{X}$ rimane sempre lo stesso al variare di n. Dal punto di vista matematico, quindi, potremo assimilare una sorgente $\mathcal{S}$ discreta ad una sequenza infinita $X_1, X_2, \ldots$ di variabili aleatorie discrete, dove l'indice della variabile aleatoria fa riferimento all'istante di tempo in cui è stato emesso il simbolo; pertanto, X_1 è il primo simbolo emesso dalla sorgente, X_2 il secondo, e così via.

Poniamoci ora il problema di misurare l'informazione associata all'*intera* sequenza $X_1, X_2, \ldots$ dei simboli emessi dalla sorgente; è chiaro che se vogliamo "catturare" eventuali proprietà di struttura della sequenza di simboli, ovvero le dipendenze statistiche tra simbolo e simbolo, non possiamo limitarci a considerare l'entropia del solo primo simbolo $H(X_1)$ (detta anche "entropia dell'alfabeto di sorgente"), ma dobbiamo calcolare quella associata a *blocchi* di due simboli consecutivi $H(X_1, X_2)$, a blocchi di tre simboli consecutivi $H(X_1, X_2, X_3)$, e così via, il caso generale essendo $H(X_1, X_2, \ldots, X_n)$. Se teniamo presente che al crescere della dimensione n del blocco aumenta il numero delle alternative possibili (esistono K^n differenti blocchi di lunghezza n), ci rendiamo conto che tale sequenza di entropie potrebbe aumentare indefinitamente. D'altra parte possiamo considerare l'informazione *media per simbolo di sorgente* semplicemente dividendo $H(X_1, X_2, \ldots, X_n)$ per n. Possiamo allora definire l'*entropia* $H(\mathcal{S})$ *di sorgente* (misurata in bit / simbolo) come il limite:

$$\boxed{H(\mathcal{S}) \triangleq \lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n)\,,}$$

ammesso che esso esista finito. Tale quantità rappresenta il contenuto medio di informazione associata ad un qualunque simbolo della sorgente, con riferimento alla trasmissione di una sequenza infinita di simboli.

È interessante notare che, poiché $H(X_1, X_2, \ldots, X_n) \leq \sum_{i=1}^{n} H(X_i) \leq n \log K$, si ha

$$H(\mathcal{S}) \leq \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} H(X_i) \leq \log K \tag{10.2}$$

per cui l'entropia di sorgente, se esiste, è limitata superiormente da $\log K$. Tale valore massimo si ottiene se e solo se i simboli sono equiprobabili e tra loro indipendenti, il che corrisponde anche intuitivamente alla situazione di massima incertezza e quindi di massima informazione. In generale, tuttavia, è lecito attendersi che i simboli emessi in successione da una sorgente altamente strutturata presentino qualche forma di dipendenza statistica; ad esempio, se la temperatura in una giornata vale 18^o, è probabile che il valore nel giorno successivo sia compreso tra 16^o e 20^o; se l'indice di borsa in una giornata vale x, è probabile che il valore nel giorno successivo non si discosti di $\pm 2\%$ da x, e così via. Pertanto sorgenti altamente strutturate presentano valori di

[5] L'assunzione di alfabeto numerico non è limitativa, in quanto se i simboli emessi dalla sorgente non sono numerici (ad esempio, l'alfabeto italiano), essi possono essere messi in corrispondenza biunivoca con un sottoinsieme dei numeri interi. Inoltre abbiamo osservato che l'entropia non dipende dai "valori" della variabile aleatoria, ma solo dalle probabilità con cui tali valori sono assunti.

entropia di sorgente molto *minori* di $\log K$; la differenza $r = \log K - H(\mathcal{S})$ rappresenta proprio la *ridondanza* associata alla sorgente, per cui sorgenti altamente strutturate hanno ridondanza molto elevata. Tale livello di dipendenza statistica o di ridondanza può essere sfruttato con vantaggio per rappresentare l'informazione emessa dalla sorgente in maniera efficiente, così come avviene nella *codifica di sorgente* (vedi § 10.3.4).

► *Esempio* 10.6 *(entropia della lingua italiana).* Un esempio di sorgente discreta di informazione è il linguaggio (scritto), che possiamo vedere come una successione di simboli appartenenti ad un certo alfabeto, con in aggiunta lo spazio ed i simboli di interpunzione. L'assunzione di indipendenza tra i simboli non è certamente appropriato per il linguaggio, in quanto si intuisce che qualsiasi lingua presenta un elevato grado di *struttura*, come provato anche dalla possibilità, spesso sfruttata nei giochi enigmistici, di ricostruire parole o anche frasi a partire da poche lettere. Consideriamo ad esempio la lingua italiana: date le lettere "a" e "c" in successione, è più probabile che la successiva lettera sia una "q" piuttosto che una "t". Per approfondire questo aspetto con riferimento alla lingua italiana, consideriamo un alfabeto semplificato composto dalle 21 lettere dell'alfabeto (a, b, c, d, e, f, g, h, i, l, m, n, o, p, q, r, s, t, u, v, z) più lo spazio, che indichiamo con –, e l'apostrofo ' (23 simboli in totale); non consideriamo per semplicità la punteggiatura e non facciamo distinzione tra lettere maiuscole e minuscole,

Se si dispone di un campione di testo sufficientemente lungo, e di un po' di pazienza (meglio ancora, di un buon programma al calcolatore) è possibile calcolare empiricamente la frequenza di occorrenza dei 23 simboli, i cui risultati indicativi sono riportati in Tab. 10.1. Osserviamo che la massima entropia che

lettera	probabilità	lettera	probabilità	lettera	probabilità
–	0.161	d	0.038	h	0.009
a	0.108	u	0.027	z	0.008
e	0.085	v	0.025	g	0.006
o	0.079	m	0.016	q	0.004
i	0.073	p	0.015	'	0.004
s	0.060	f	0.014		
n	0.055	b	0.010		
l	0.053				
t	0.051				
r	0.050				
c	0.049				

Tab. 10.1. Le lettere dell'alfabeto italiano con le relative probabilità di occorrenza (in ordine decrescente di probabilità).

si può ottenere con 23 lettere è pari a $\log 23 = 4.52$ bit, mentre quella effettiva delle lettere dell'alfabeto italiano è pari a $H(X_1) \approx 4$ bit, con una ridondanza di soli 0.5 bit. L'esempio non deve però indurre a conclusioni errate: in realtà, la ridondanza della lingua italiana è molto più elevata, ma richiede che si considerino gruppi di 2 lettere, di 3 lettere e così via, cioè richiede il calcolo dell'entropia media per lettera $\frac{1}{n} H(X_1, X_2, \ldots, X_n)$ e, al limite per $n \to \infty$, dell'entropia di sorgente.

Un calcolo di questo tipo è riportato in [12] per la lingua inglese, con un alfabeto di 27 simboli (26 lettere ed uno spazio), per il quale la massima entropia è pari a $\log 27 = 4.76$ bit. L'entropia di una lettera isolata nella lingua inglese è invece pari a $H(X_1) \approx 4$ bit, quindi praticamente coincidente con quella dell'alfabeto italiano, nonostante il maggior numero di simboli dell'alfabeto; se si considerano gruppi di più lettere, l'entropia per lettera diminuisce; ad esempio, per 4 lettere l'entropia media per lettera $\frac{1}{4} H(X_1, X_2, X_3, X_4)$ è pari a 2.8 bit. Esperimenti condotti dallo stesso Shannon e da altri ricercatori stimano l'entropia di sorgente $H(\mathcal{S})$ della lingua inglese pari a circa 1.3 bit per lettera, che confrontato con il valore massimo di 4.76 bit mostra l'elevato grado di ridondanza della lingua inglese. ◄

10.3.2 Tasso d'informazione di una sorgente

Se vogliamo portare esplicitamente in conto nella nostra trattazione il tempo che intercorre tra l'emissione di due simboli consecutivi, immaginiamo che la sorgente $\mathcal{S}$ emetta i suoi simboli con

una cadenza regolare, e sia T_s l'intervallo temporale che intercorre tra due simboli consecutivi. Ovviamente è raro che sorgenti non artificiali emettano simboli con siffatta regolarità: ad esempio, una persona che scrive alla tastiera di un computer batte sui tasti ad intervalli irregolari: in questo caso, per semplicità, potremmo pensare che T_s rappresenti l'intervallo *medio* tra la battitura di due tasti, ovvero tra due simboli emessi consecutivamente dalla sorgente. In ogni caso, definiamo il *tasso di informazione* R_s emesso dalla sorgente semplicemente come

$$\boxed{R_s \triangleq \frac{H(\mathcal{S})}{T_s}\,.}$$

Se l'entropia $H(\mathcal{S})$ è misurata in bit/simbolo, il tasso d'informazione R_s si misura in bit/s o multipli (kbit/s, Mbit/s). Si può notare che in base alla (10.2), si ha

$$R_s \leq \frac{\log K}{T_s} \triangleq R_b\,,$$

dove R_b (ritmo binario o bit-rate) rappresenta il numero di bit al secondo generati dalla sorgente, e si misura anch'esso in bit/s o multipli. Nonostante adottino la stessa unità di misura, le due quantità R_b ed R_s sono profondamente differenti: il bit-rate R_b è una semplice misura della velocità binaria di emissione della sorgente, ma non porta in conto assolutamente le proprietà *statistiche* della sorgente; viceversa, tali proprietà sono misurate dal tasso d'informazione R_s, che rappresenta la vera misura della quantità di informazione emessa dalla sorgente, e rappresenta il dato reale da tener presente quando si progetta un sistema di comunicazione capace di garantire il trasporto affidabile dell'informazione dalla sorgente alla destinazione. Per sorgenti altamente strutturate risulta $H(\mathcal{S}) \ll \log K$ e quindi $R_s \ll R_b$.

10.3.3 Sorgenti discrete senza memoria (DMS)

Nonostante le sorgenti con simboli statisticamente dipendenti siano praticamente la norma (si pensi ad esempio alle forti dipendenze statistiche del linguaggio scritto o parlato), la loro trattazione matematica risulta estremamente difficoltosa; spesso risulta utile considerare il caso particolarmente semplice, seppure ideale, di sorgente *discreta senza memoria (DMS) stazionaria*: in questo caso i successivi simboli $X_1, X_2, \ldots$ emessi dalla sorgente sono assunti indipendenti (sorgente senza memoria), ed identicamente distribuiti (sorgente stazionaria), con DF comune $p_X(x)$, $x \in \mathcal{X}$. In questo caso, si ha $H(X_1, X_2, \ldots, X_n) = n\,H(X_1)$ e quindi

$$H(\mathcal{S}) = \lim_{n\to\infty} \frac{1}{n}\, H(X_1, X_2, \ldots, X_n) = \lim_{n\to\infty} \frac{n\,H(X_1)}{n} = H(X_1) = -\sum_{x\in\mathcal{X}} p_X(x) \log p_X(x)\,,$$

cioè, per una sorgente DMS stazionaria, l'entropia di sorgente coincide con l'entropia di un simbolo emesso dalla sorgente in un qualunque istante di tempo (ad esempio, il simbolo X_1 emesso per $n = 1$). In altri termini, per una sorgente DMS stazionaria l'entropia di sorgente risulta coincidere con l'entropia dell'alfabeto di sorgente. Similmente, il tasso di informazione di una sorgente DMS stazionaria vale

$$R_s = \frac{H(X_1)}{T_s}\,.$$

Nel seguito, per brevità, riterremo che ogni sorgente DMS sia anche stazionaria, quindi parleremo sinteticamente di sorgenti DMS omettendo l'aggettivo "stazionarie".

10.3.4 Codifica di sorgente

Un'importante applicazione dei concetti di misura dell'informazione e di entropia è rappresentata dalla cosiddetta *codifica di sorgente*, che consiste nella rappresentazione efficiente dei simboli emessi da una sorgente di informazione. Più precisamente, sulla base della definizione di sorgente di informazione data nel paragrafo precedente, possiamo formalizzare il problema della codifica di sorgente come segue: data una sorgente di informazione $\mathcal{S}$, si desidera codificare le sequenze di simboli emessi dalla sorgente, che appartengono ad un *alfabeto di sorgente* $\mathcal{X} = \{x_1, x_2, \ldots, x_K\}$ di cardinalità K, in sequenze binarie, ovvero composte da soli due valori, 0 ed 1, che costituiscono l'*alfabeto codice*.[6] Un esempio tipico di codifica binaria è quello fornito dal *codice ASCII*, mediante il quale si codificano caratteri alfanumerici, più alcuni caratteri di controllo, in stringhe composte da 8 bit (1 byte).

L'obiettivo della codifica di sorgente è quello di ridurre al minimo (compattare) la lunghezza delle stringhe binarie necessarie a codificare le sequenze di simboli emessi dalla sorgente, eliminando, o riducendo al minimo, la ridondanza di informazione presente nella sorgente. Tale riduzione di ridondanza, effettuata da un dispositivo denominato *codificatore di sorgente* (vedi Fig. 10.1), può essere anche assai rilevante, a spese tuttavia della possibilità di ricostruire *esattamente* i simboli emessi dalla sorgente a partire dalle stringhe codificate: si parla in tal caso di *codifica di sorgente con perdite* (*lossy coding*) o di *compressione dati*. Tale perdita di informazione in molte applicazioni è accettabile, in quanto il destinatario ultimo dell'informazione (tipicamente un essere umano) ha una sensibilità finita; la codifica con perdite si applica infatti utilmente nella trasmissione telefonica, nella telefonia via Internet, nella trasmissione del segnale televisivo, nella codifica di file audio, ed in numerose altre applicazioni. Ad esempio, gli standard JPEG (per la compressione di immagini fisse), MPEG (per la compressione di immagini in movimento) ed MP3 (per la compressione di file audio) sono tutti esempi di codifica *con perdite*, quindi senza la possibilità di ricostruire *esattamente* l'informazione originaria a partire da quella codificata. In questi casi, tipicamente, modificando i parametri del codificatore è possibile aumentare la compressione dell'informazione a spese della qualità percepita alla destinazione, e viceversa.

Risultati più modesti, ma comunque rilevanti, si conseguono se si richiede la perfetta ricostruibilità dei simboli emessi dalla sorgente, il che nella comunicazione tra macchine (si pensi alla compressione di un file eseguibile di un programma) è un requisito imprescindibile: si parla in tal caso di *codifica di sorgente senza perdite* (*lossless coding*) o di *compattazione dati*. Esempi di codifica *senza perdite* sono quelle effettuate dai popolari programmi per la compattazione di file, quali Winzip (per sistemi operativi Windows) o il comando `compress` o `gzip` (per sistemi operativi Unix/Linux).

► *Esempio* 10.7. La codifica senza perdite consente di rappresentare l'informazione in maniera esatta, ma meno efficiente: per convincersene, basta considerare un esempio concreto, facilmente replicabile dal lettore al calcolatore: un file audio in formato WAV di circa 16 MB, corrispondente a circa 1 minuto e mezzo di musica stereo con qualità CD, viene convertito *senza perdite* dal programma Winzip, basato sull'algoritmo cosiddetto di *Lempel-Ziv*, in un file ZIP di circa 14 MB; viene invece convertito *con perdite* in un file MP3 a 128 kbps di circa 1.5 MB. In questo caso, la codifica con perdite risulta quasi 10 volte più efficiente della codifica senza perdite, senza un'apprezzabile degradazione della qualità percepita all'ascolto. ◄

Nonostante l'esempio e le considerazioni precedenti mostrino che i vantaggi più significativi

[6] La codifica binaria non è l'unico tipo di codifica esistente: il caso più generale prevede un alfabeto codice composto da due o più valori. Tuttavia la rilevanza della codifica binaria discende dal fatto che l'informazione binaria può più facilmente essere trasmessa, elaborata, e memorizzata.

si ottengano utilizzando la codifica con perdite, lo studio di tale argomento richiede una serie di strumenti matematici avanzati; inoltre, vale la pena osservare che non è pensabile affrontare lo studio della codifica con perdite senza possedere le conoscenze di base sulla codifica senza perdite. Per questi motivi, ci occuperemo nel seguito esclusivamente della compattazione dati, ovvero della *codifica senza perdite*.

10.4 Codici per la compattazione dati

Prima di introdurre le tecniche più semplici per la compattazione dati, forniamo alcune nozioni di base sui *codici*, iniziando dalla definizione formale di *codice binario*:

Definizione (codice binario). Sia $\mathcal{S}$ una sorgente discreta, un codice binario $\mathcal{C}$ per la sorgente $\mathcal{S}$ è una regola che trasforma sequenze di simboli emessi da $\mathcal{S}$ in sequenze di simboli binari, per esempio appartenenti all'alfabeto di codice $\{0, 1\}$.

Svilupperemo tale definizione individuando differenti tipologie di codice con differenti proprietà. In particolare, tra le possibili strategie di codifica, considereremo due famiglie di codici: i codici *a lunghezza fissa* ed i codici *a lunghezza variabile*.

10.4.1 Codici a lunghezza fissa

Nei codici a lunghezza fissa, le sequenze di simboli di sorgente da codificare sono suddivise o "segmentate" in blocchi di lunghezza fissa, pari ad n simboli; ciascun blocco viene poi trasformato in un blocco codificato (binario) anch'esso di lunghezza fissa, pari ad ℓ cifre binarie o bit. Comunemente i blocchi codificati si chiamano "parole codice", e l'insieme delle parole codice prende il nome di "dizionario" del codice.

► *Esempio* 10.8 *(codice ASCII).* Un esempio particolarmente semplice di codice a lunghezza fissa è rappresentato dal codice ASCII, nel quale tutti i caratteri alfanumerici ed i caratteri speciali sono codificati con parole di lunghezza fissa e pari a $\ell = 8$ bit. ◄

10.4.2 Codici a lunghezza variabile

Nei codici a lunghezza variabile, le sequenze di simboli di sorgente da codificare sono ancora segmentate in blocchi di lunghezza fissa pari ad n, mentre le parole codice non sono più vincolate ad avere tutte la stessa lunghezza. Il motivo per cui si introduce questo grado di libertà è intuitivamente comprensibile: si tende a codificare simboli (o blocchi di simboli) di sorgente *meno* probabili con parole codice *lunghe*, e viceversa simboli (o blocchi di simboli) *più* probabili con parole codice *corte*; in questo modo si riduce la *lunghezza media* della sequenza codificata, rispetto ad un codice a lunghezza fissa.

► *Esempio* 10.9 *(codice Morse).* Un classico esempio di codice a lunghezza variabile è il codice telegrafico Morse (ormai in disuso), nel quale l'alfabeto codice è costituito da punti ("dot") e linee ("dash"), e il codice è costruito in modo da tener conto della frequenza relativa delle lettere nella lingua inglese: ad esempio, alla frequente lettera "e" è associata la parola codice breve "." (punto), mentre alla poco frequente lettera "q" è associata la parola codice lunga ". . – –" (punto, punto, linea, linea). ◄

Nel seguito, considereremo prevalentemente il caso in cui la codifica sia effettuata su blocchi di sorgente di lunghezza $n = 1$, ovvero su singoli simboli della sorgente (codifica "simbolo a simbolo"); il caso $n > 1$ si può trattare come generalizzazione del precedente, considerando una "macro-sorgente" che emette blocchi anziché simboli.

10.4.3 Codici univocamente decifrabili

In una codifica senza perdite, una proprietà irrinunciabile di un codice è che esso sia *univocamente decifrabile*:

Definizione (codice univocamente decifrabile). Un codice $\mathcal{C}$ si dice *univocamente decifrabile* se è possibile ricostruire senza ambiguità le sequenze di simboli originali a partire dalle sequenze codificate.

► *Esempio* 10.10. Consideriamo una sorgente discreta $\mathcal{S}$ che emette simboli X appartenenti all'alfabeto $\mathcal{X} = \{x_1, x_2, x_3, x_4\}$. Quattro possibili codici binari simbolo a simbolo per tale sorgente sono riportati in Tab. 10.2: i codici $\mathcal{C}_1$ e $\mathcal{C}_2$ sono a lunghezza fissa, mentre i codici $\mathcal{C}_3$ e $\mathcal{C}_4$ sono a lunghezza variabile. Affinchè

X	$\mathcal{C}_1$	$\mathcal{C}_2$	$\mathcal{C}_3$	$\mathcal{C}_4$
x_1	00	00	0	0
x_2	01	01	11	10
x_3	10	01	00	110
x_4	11	11	01	1110

Tab. 10.2. Alcuni codici binari per una sorgente con $K = 4$ simboli.

un codice risulti univocamente decifrabile, in primo luogo le parole codice devono essere tutte differenti; codici che non soddisfano tale proprietà, come il codice $\mathcal{C}_2$, si dicono *singolari*, e non saranno considerati più nel seguito. A questo punto, è facile verificare che, se il codice è a lunghezza fissa, affinché esso risulti univocamente decifrabile è necessario e sufficiente che esso sia non singolare, come il codice $\mathcal{C}_1$. Più complesso è il problema di riconoscere l'univoca decifrabilità per codici a lunghezza variabile, in quanto il fatto che il codice sia non singolare non garantisce che esso sia anche univocamente decifrabile. Ad esempio, i codici $\mathcal{C}_3$ e $\mathcal{C}_4$ sono entrambi non singolari; tuttavia, se si considera il codice $\mathcal{C}_3$, è facile verificare che la stringa codificata 0011 può corrispondere alla sequenza di sorgente $x_1x_1x_2$ ma anche alla sequenza di sorgente x_3x_2, per cui tale codice non è univocamente decifrabile. Viceversa, si verifica facilmente che il codice $\mathcal{C}_4$ è univocamente decifrabile, in quanto le sue parole codice terminano tutte per 0 (che può essere considerato come una specie di simbolo di separazione tra due parole codice consecutive). ◄

10.4.4 Codici a prefisso

Tra i codici a lunghezza variabile univocamente decifrabili, un'importante classe di codici è rappresentata dalla classe dei cosiddetti *codici a prefisso*:

Definizione (codice a prefisso). Un codice $\mathcal{C}$ si dice a prefisso se nessuna parola codice è prefissa di un'altra parola codice.

Si intende che una parola codice è prefissa di un'altra parola codice se ne costituisce la sottostringa iniziale; ad esempio, la parola codice 01 è prefissa delle parole codice 011, 0110, e 01111. A questo punto, è chiaro che, se un codice è a prefisso, nella decodifica di una sequenza codificata non possono sorgere ambiguità, e quindi un tale codice è sicuramente univocamente decifrabile.

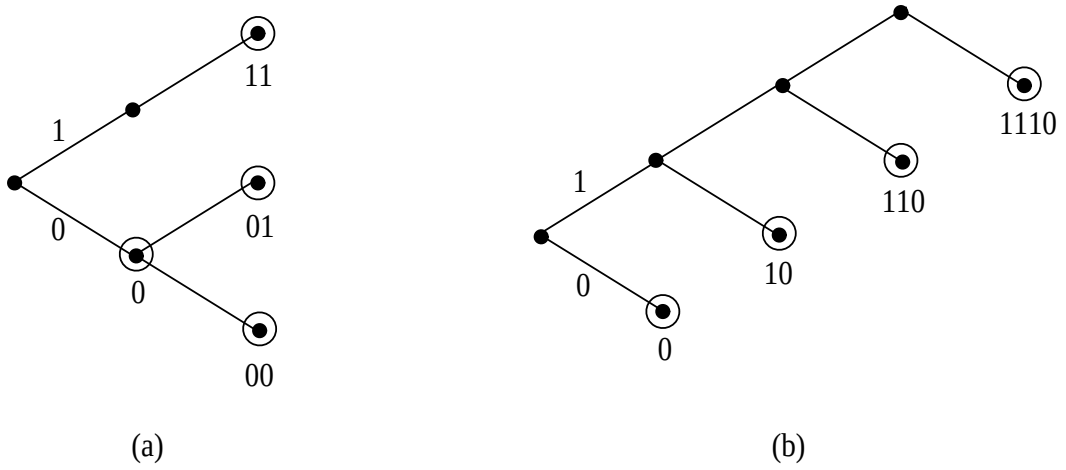


Fig. 10.3. Alberi di codice per il codice $\mathcal{C}_3$ (a) ed il codice $\mathcal{C}_4$ (b).

► *Esempio* 10.11. Consideriamo nuovamente i codici $\mathcal{C}_3$ e $\mathcal{C}_4$ dell'esempio 10.10. Per il primo, osserviamo che la parola codice 0 è prefissa delle parole codice 00 e 01, per cui tale codice *non* è un codice a prefisso (peraltro, abbiamo già verificato che esso non è univocamente decifrabile). Viceversa, se consideriamo il codice $\mathcal{C}_4$, osserviamo che nessuna parola codice è prefissa di un'altra parola codice, per cui tale codice è a prefisso, e quindi univocamente decifrabile. ◄

Per verificare se un codice è a prefisso oppure no, è assai utile la sua rappresentazione mediante un *albero di codice*, dove per albero intendiamo un grafo orientato (generalmente da sinistra a destra), composto da punti, detti "nodi", e da linee, dette "rami", con la condizione che da ogni nodo partano due rami (albero binario). Proseguendo nella similitudine "botanica", il nodo all'estrema sinistra dell'albero prende il nome di "radice", mentre i nodi all'estrema destra, da cui non partono rami, si dicono nodi "terminali" o "foglie"; i nodi che non sono nè radice nè terminali si dicono "interni". Se si contrassegnano i rami partenti da un nodo sempre allo stesso modo (ad esempio, il ramo superiore con 1 e quello inferiore con 0), ad ogni nodo viene assegnata un'etichetta univoca, ottenuta concatenando ordinatamente i contrassegni dei rami che si percorrono dalla radice fino al nodo in esame.[7] Un nodo i si dice predecessore di un nodo j se muovendosi dalla radice verso j si incontra prima i; equivalentemente, j si dirà successore di i.

È possibile allora costruire la rappresentazione ad albero di un codice semplicemente associando le parole codice (stringhe binarie) ai nodi corrispondenti dell'albero, scelto di lunghezza appropriata. A questo punto, è semplice verificare se un codice è a prefisso oppure no: infatti, se una parola codice c_i è prefissa di un'altra parola c_j, il nodo i è predecessore di j; pertanto, affinché il codice sia a prefisso, *tutte le parole codice devono corrispondere a nodi terminali dell'albero.*

► *Esempio* 10.12. Consideriamo la rappresentazione ad albero dei codici $\mathcal{C}_3$ e $\mathcal{C}_4$ dell'esempio 10.10, riportata in Fig. 10.3. Dall'esame degli alberi di codice, si nota chiaramente come il codice $\mathcal{C}_3$ (albero a sinistra) non sia a prefisso (la parola codice 0 è predecessore delle parole codice 01 e 00), mentre il codice $\mathcal{C}_4$ (albero a destra) è chiaramente a prefisso, in quanto tutte le sue parole codice corrispondono a nodi terminali. ◄

Osserviamo infine che un codice a prefisso è anche detto *istantaneo* perché, nella fase di decodifica, non appena percorrendo l'albero si riconosce una parola codice, è possibile decodificarla

[7] Notiamo che tale percorso sull'albero, dalla radice ad un nodo, è necessariamente unico.

istantaneamente, garantendo in questo modo un *ritardo di decodifica* nullo. In generale, un codice univocamente decifrabile (ma *non* a prefisso) non è detto che sia istantaneo, ma può presentare un ritardo di decodifica non nullo. Tale ritardo costituisce un problema nella trasmissione di informazione *in tempo reale*, in quanto può accadere, come caso limite, che per iniziare la decodifica del primo simbolo trasmesso si debba attendere la ricezione dell'*intera* sequenza di informazione.

10.4.5 Condizioni per l'univoca decifrabilità

Osserviamo che un codice univocamente decifrabile non è necessariamente a prefisso: in altri termini, la classe dei codici univocamente decifrabili comprende la classe dei codici a prefisso, ma non si limita ad essa. Pertanto, esistono codici univocamente decifrabili che non sono a prefisso, e quindi verificare mediante l'analisi dell'albero che il codice *non* è a prefisso non consente di affermare con sicurezza che esso *non* è univocamente decifrabile.

In effetti, esiste una procedura sistematica (metodo di Sardinas e Patterson [12]) per individuare se un dato codice (a prefisso oppure no) sia univocamente decifrabile, che tuttavia non discuteremo. Approfondiamo invece tale problema da un punto di vista leggermente diverso, che risulterà più proficuo per determinare i limiti ultimi dell'efficienza con cui è possibile compattare i simboli emessi da una sorgente. Sia $\mathcal{S}$ una sorgente che emette simboli appartenenti ad un alfabeto $\mathcal{X} = \{x_1, x_2, \ldots, x_K\}$ con K possibili simboli, e sia $\mathcal{C}$ un codice binario simbolo a simbolo, avente parole codice $c_1, c_2, \ldots, c_K$, di *lunghezze* $\ell_1, \ell_2, \ldots, \ell_K$. Se il codice è univocamente decifrabile, le lunghezze ℓ_k devono soddisfare al seguente teorema, che enunciamo senza dimostrazione:

Teorema 10.1 (disuguaglianza di Mc Millan). Se il codice binario $\mathcal{C}$ con parole codice aventi lunghezze $\ell_1, \ell_2, \ldots, \ell_K$ è univocamente decifrabile, risulta necessariamente

$$\sum_{k=1}^{K} 2^{-\ell_k} \leq 1 \,.$$

► *Esempio* 10.13. Applichiamo la disuguaglianza di Mc Millan ai codici $\mathcal{C}_3$ e $\mathcal{C}_4$ dell'esempio 10.10. Per il primo, che già sappiamo essere non univocamente decifrabile, si ha:

$$\sum_{k=1}^{K} 2^{-\ell_k} = 2^{-1} + 2^{-2} + 2^{-2} + 2^{-2} = \frac{5}{4} > 1$$

per cui, come ci attendiamo, la disuguaglianza di Mc Millan non è verificata. Per il secondo, che sappiamo essere a prefisso e quindi univocamente decifrabile, risulta invece:

$$\sum_{k=1}^{K} 2^{-\ell_k} = 2^{-1} + 2^{-2} + 2^{-3} + 2^{-3} = 1$$

per cui la disuguaglianza di Mc Millan è verificata (con il segno di uguaglianza). ◄

Notiamo che la disuguaglianza di Mc Millan fornisce solo una condizione *necessaria* per l'univoca decifrabilità, condizione che coinvolge tra l'altro soltanto le *lunghezze* delle parole codice; in altri termini, non è detto che un codice le cui lunghezze soddisfino la disuguaglianza di Mc Millan sia univocamente decifrabile: al limite, un tale codice potrebbe addirittura essere singolare, cioè avere due parole codice coincidenti! Tuttavia, la disuguaglianza di Mc Millan può essere utilizzata anche come condizione *sufficiente* per la costruzione di un codice *a prefisso* (e quindi univocamente decifrabile), come evidenziato dal seguente teorema, che enunciamo senza dimostrazione:

Teorema 10.2 (disuguaglianza di Kraft). Se i K numeri interi positivi $\ell_1, \ell_2, \ldots, \ell_K$ soddisfano la disuguaglianza

$$\sum_{k=1}^{K} 2^{-\ell_k} \leq 1 ,$$

allora è possibile costruire un codice binario $\mathcal{C}$ a prefisso (e quindi univocamente decifrabile) con K parole codice aventi lunghezze $\ell_1, \ell_2, \ldots, \ell_K$.

Notiamo che la disuguaglianza che compare nei due teoremi precedentemente enunciati è la stessa, e per questo motivo essi vengono spesso condensati in un unico teorema, che prende il nome di *disuguaglianza di Kraft-Mc Millan*. Una conseguenza notevole di tale disuguaglianza è che essa ci consente, senza ledere la generalità, di considerare, tra i codici univocamente decifrabili, solo quelli a prefisso. Infatti, se ho un codice univocamente decifrabile, le lunghezze delle parole codice soddisfano necessariamente il teorema 10.1; allora, in virtù del teorema 10.2, è possibile trovare un codice a prefisso avente lo stesso insieme di lunghezze (e quindi la stessa efficienza, in termini di lunghezza media delle parole codificate, cfr. §10.5).

10.5 Efficienza dei codici per la compattazione dati

Possiamo adesso affrontare il problema di misurare l'*efficienza* di una data strategia di codifica, facendo riferimento per il momento a strategie di codifica *simbolo a simbolo.* La domanda fondamentale a cui vogliamo dare risposta è la seguente: *per una data sorgente, qual è la lunghezza minima di un codice che rappresenti l'informazione emessa dalla sorgente in maniera non ambigua?*

Consideriamo una sorgente discreta di informazione $\mathcal{S}$, che emette simboli appartenenti ad un alfabeto $\mathcal{X} = \{x_1, x_2, \ldots, x_K\}$, di cardinalità K, con probabilità $p_1, p_2, \ldots, p_K$. Il contenuto di informazione associato *ad un singolo simbolo* X emesso dalla sorgente è misurato dall'entropia $H(X)$ dell'alfabeto:

$$H(X) = -\sum_{k=1}^{K} p_k \log p_k .$$

L'obiettivo della codifica di sorgente è senza perdite quello di costruire un codice univocamente decifrabile, in modo da ottenere sequenze codificate che risultino *mediamente* le più corte possibile. Non a caso abbiamo utilizzato la parola *mediamente*: infatti, mentre per i codici a lunghezza fissa la lunghezza ℓ è la stessa per tutte le parole codice, per i codici a lunghezza variabile la lunghezza di ogni parola codice è differente. Poiché la parola codice di lunghezza ℓ_k, essendo associata al simbolo di sorgente x_k, ricorre con probabilità p_k, la lunghezza delle parole codice è in effetti una variabile aleatoria discreta L, che assume i valori $\ell_1, \ell_2, \ldots, \ell_K$ con probabilità $p_1, p_2, \ldots, p_K$. Possiamo allora calcolare la *lunghezza media del codice* $\overline{\ell}$ come la media statistica della variabile aleatoria L:

$$\overline{\ell} \triangleq E[L] = \sum_{k=1}^{K} p_k \ell_k .$$

Nel caso di codici a lunghezza fissa, si ha ovviamente $\ell_k = \ell$ e $\overline{\ell} = \ell$.

Utilizzando la disuguaglianza di Kraft-Mc Millan, siamo allora in grado di dimostrare un fondamentale risultato, che mette in relazione la lunghezza media $\overline{\ell}$ di un codice simbolo a simbolo univocamente decifrabile con l'entropia $H(X)$ dei simboli emessi dalla sorgente:

Teorema 10.3. Per ogni codice $\mathcal{C}$ binario simbolo a simbolo univocamente decifrabile, la lunghezza media $\overline{\ell}$ del codice soddisfa la seguente disuguaglianza:

$$\overline{\ell} \geq H(X)\,,$$

dove $H(X)$ è l'entropia dell'alfabeto di sorgente (espressa in bit).

Prova. Proviamo che $H(X) - \overline{\ell} \leq 0$, scrivendo (si ricordi che i logaritmi sono in base 2):

$$H(X) - \overline{\ell} = -\sum_{k=1}^{K} p_k \log p_k - \sum_{k=1}^{K} p_k \ell_k = -\sum_{k=1}^{K} p_k \log p_k + \sum_{k=1}^{K} p_k \log 2^{-\ell_k} =$$
$$= \sum_{k=1}^{K} p_k \log \frac{2^{-\ell_k}}{p_k}\,.$$

Possiamo adoperare la seguente disuguaglianza, valida per i logaritmi naturali:

$$\ln x \leq x - 1\,,$$

che per i logaritmi in base 2 si scrive, adoperando le formula per il cambiamento di base, come

$$\log x \leq \log e\,(x-1)\,,$$

per cui si ha:

$$H(X) - \overline{\ell} = \sum_{k=1}^{K} p_k \log \frac{2^{-\ell_k}}{p_k} \leq \log e \sum_{k=1}^{K} p_k \left(\frac{2^{-\ell_k}}{p_k} - 1\right) = \log e \left(\sum_{k=1}^{K} 2^{-\ell_k} - \sum_{k=1}^{K} p_k\right)$$
$$= \log e \left(\sum_{k=1}^{K} 2^{-\ell_k} - 1\right)\,.$$

Poichè il codice $\mathcal{C}$ è univocamente decifrabile, allora esso soddisfa la disuguaglianza di Mc Millan (cfr. teorema 10.1) per cui $\sum_{k=1}^{K} 2^{-\ell_k} - 1 \leq 0$, ed essendo $\log e \geq 0$ si ha

$$H(X) - \overline{\ell} \leq 0\,,$$

cioè l'asserto. □

Il precedente teorema fornisce una interpretazione nuova ed estremamente importante dell'entropia dell'alfabeto dei simboli emessi da una sorgente; infatti, esso consente di interpretare tale entropia (in bit) come la *minima* lunghezza media di un codice binario simbolo a simbolo che rappresenti i simboli di sorgente in maniera non ambigua (vincolo di univoca decifrabilità). Di contro, il teorema fornisce anche un limite inferiore alla capacità di compattazione dati operata dalla codifica di sorgente: infatti la lunghezza media di un codice simbolo a simbolo univocamente decifrabile non potrà mai essere *inferiore* all'entropia dell'alfabeto di sorgente. Per confrontare tra loro differenti codici, definiamo allora l'*efficienza* di un codice, avente con lunghezza media $\overline{\ell}$, come:

$$\eta \triangleq \frac{H(X)}{\overline{\ell}}\,.$$

Tale efficienza assume ovviamente valori tra 0 ed 1, ed è sovente espressa in percentuale.

10.5.1 Codici di Shannon

Il teorema 10.3 stabilisce solo un limite *inferiore* per la lunghezza di un codice univocamente decifrabile: in pratica la lunghezza media $\overline{\ell}$ di un "cattivo" codice può anche essere molto maggiore dell'entropia dell'alfabeto $H(X)$, e quindi l'efficienza del codice può essere anche molto

minore dell'unità. È utile allora disporre di una procedura per costruire codici univocamente decifrabili la cui lunghezza media $\overline{\ell}$ sia, oltre che inferiormente, anche superiormente limitata. Tali codici vanno sotto il nome di *codici di Shannon*, e la procedura per costruirli è descritta nella dimostrazione del seguente teorema:

Teorema 10.4 (codice di Shannon). Data una sorgente discreta $\mathcal{S}$ di simboli appartenenti ad un alfabeto $\mathcal{X} = \{x_1, x_2, \ldots, x_k\}$ e con probabilità $p_1, p_2, \ldots, p_k$, è sempre possibile costruire un codice binario (codice di Shannon) simbolo a simbolo a prefisso (e quindi univocamente decifrabile) tale che la sua lunghezza media $\overline{\ell}$ soddisfi la seguente disuguaglianza:

$$\overline{\ell} < H(X) + 1\,,$$

dove $H(X)$ è l'entropia dell'alfabeto di sorgente (espressa in bit).

Prova. La dimostrazione si basa sulla costruzione di un codice che soddisfa la disuguaglianza. Partiamo fissando le lunghezze delle parole codice, secondo la:

$$\ell_k = \lceil -\log p_k \rceil\,, \tag{10.3}$$

dove il simbolo $\lceil x \rceil$ denota il più piccolo numero intero maggiore o uguale ad x. Notiamo che questa scelta equivale ad assegnare, come è ragionevole, parole codice più lunghe a simboli meno probabili, e viceversa. Risulta allora $\ell_k \geq -\log p_k$ e quindi $2^{-\ell_k} \leq p_k$. Sommando su tutti i valori di k, si ha:

$$\sum_{k=1}^{K} 2^{-\ell_k} \leq \sum_{k=1}^{K} p_k = 1\,,$$

per cui risulta verificata la disuguaglianza di Kraft (teorema 10.2), e pertanto esiste un codice a prefisso univocamente decifrabile con parole codice aventi lunghezze $\ell_1, \ell_2, \ldots, \ell_K$. Poiché poi, per la definizione (10.3) delle lunghezze ℓ_k, risulta anche

$$\ell_k < -\log p_k + 1\,,$$

allora si ha

$$\overline{\ell} = \sum_{k=1}^{K} p_k\,\ell_k < \sum_{k=1}^{K} p_k\,(-\log p_k + 1) = -\sum_{k=1}^{K} p_k \log p_k + \sum_{k=1}^{K} p_k = H(X) + 1\,,$$

per cui la disuguaglianza risulta provata. Un codice costruito secondo questa procedura prende il nome di *codice di Shannon*. □

Nella dimostrazione del teorema 10.4 si stabiliscono solo le *lunghezze* delle parole codice, ma la determinazione delle parole codice può essere fatta semplicemente, per un fissato insieme di lunghezze. Infatti, una volta determinate le lunghezze sulla base della (10.3), basta costruire un albero binario di lunghezza pari a $\ell_{\max} = \max_k \ell_k$ ed assegnare le parole codice ai nodi dell'albero, partendo dalle parole più corte ed eliminando via via dall'albero tutti i nodi discendenti dei nodi già assegnati, in modo da soddisfare la condizione di prefisso. Al termine di questa procedura, tipicamente si riconosce che alcuni rami che portano alle parole codice possono essere accorciati senza ledere la condizione di prefisso; a valle di tale "potatura" dell'albero, si ottiene allora un codice a prefisso con lunghezza media inferiore a quella del codice di Shannon originario.

Notiamo che un codice di Shannon, essendo univocamente decifrabile, deve soddisfare necessariamente anche la disuguaglianza stabilita dal teorema 10.3, per cui la sua lunghezza media risulta essere sia inferiormente che superiormente limitata:

$$H(X) \leq \overline{\ell} < H(X) + 1\,,$$

Sulla base di questa relazione notiamo che le prestazioni di un codice di Shannon (senza potatura) non sono necessariamente buone, in quanto la sua efficienza è compresa tra i seguenti limiti:

$$\frac{H(X)}{H(X)+1} < \eta \leq 1$$

per cui se $H(X) \ll 1$ l'efficienza può assumere valori estremamente bassi, come mostrato dal seguente esempio.

► *Esempio* 10.14. Sia $\mathcal{S}$ una sorgente che emette i simboli x_1 ed x_2 con probabilità $p_1 = 0.99$ e $p_2 = 0.01$. L'entropia dei simboli emessi da una tale sorgente è estremamente bassa:

$$H(X) = -0.99 \log 0.99 - 0.01 \log 0.01 = 8.08 \cdot 10^{-2} \text{ bit}\,.$$

Le parole codice del codice di Shannon avranno lunghezze date dalla (10.3), ovvero

$$\ell_1 = \lceil \log 0.99 \rceil = 1\,,$$
$$\ell_2 = \lceil \log 0.01 \rceil = 7\,,$$

per cui la lunghezza media del codice è:

$$\overline{\ell} = 1 \cdot 0.99 + 7 \cdot 0.01 = 1.06\,,$$

che risulta minore di $H(X)+1$, ma molto prossimo ad esso, per cui l'efficienza η è estremamente bassa, essendo pari a 0.076. D'altra parte, pare abbastanza stravagante utilizzare un codice a lunghezza variabile per codificare *due* simboli di sorgente, in quanto sarebbe sufficiente considerare un codice a lunghezza fissa, con parole codice 0 ed 1, la cui lunghezza media, esattamente pari ad 1, è tuttavia ancora molto distante dall'entropia dell'alfabeto $H(X)$. D'altra parte, questo rappresenta il meglio che possiamo fare se vincoliamo la codifica ad essere simbolo a simbolo. Una strategia più efficiente è quella della codifica a blocchi, discussa nel § 10.5.2. ◄

Va osservato che l'esempio precedente è un caso limite, in quanto spesso il codice di Shannon presenta valori di $\overline{\ell}$ non troppo lontani dall'entropia $H(X)$. In particolare, si può osservare che se le probabilità p_k sono del tipo $p_k = 2^{-\ell_k}$, con ℓ_k interi positivi, allora risulta per la (10.3) $\ell_k = -\log p_k$, ed inoltre

$$\overline{\ell} = \sum_{k=1}^{K} p_k\, \ell_k = -\sum_{k=1}^{K} p_k \log p_k = H(X)\,,$$

per cui si ottiene una lunghezza media esattamente pari all'entropia, e quindi il codice di Shannon è ottimo in questo caso; ovviamente è raro che la sorgente $\mathcal{S}$ presenti proprio probabilità esprimibili come $2^{-\ell_k}$.

10.5.2 Codifica a blocchi e primo teorema di Shannon

L'esempio 10.14 mostra che la codifica simbolo a simbolo non consente sempre di ottenere lunghezze media prossime all'entropia, in particolar modo per sorgenti con pochi simboli e con probabilità dei simboli fortemente diverse tra loro Per ovviare a ciò, dobbiamo rimuovere il vincolo di codifica simbolo a simbolo, passando a codificare blocchi di n simboli. Consideriamo allora un blocco di n simboli consecutivi emessi dalla sorgente negli sitanti $1, 2, \ldots, n$, siano essi $X_1, X_2, \ldots, X_n$. Per applicare i risultati della codifica simbolo a simbolo, è sufficiente interpretare la sorgente come una sorgente che emette "blocchi" anziché simboli, e sostituire all'entropia del simbolo $H(X)$ l'entropia del blocco $H(X_1, X_2, \ldots, X_n)$. Pertanto, detta $\overline{\ell}(n)$ la lunghezza media di un codice di Shannon per i blocchi di n simboli emessi dalla sorgente, risulta, per i teoremi 10.3 e 10.4,

$$H(X_1, X_2, \ldots, X_n) \leq \overline{\ell}(n) < H(X_1, X_2, \ldots, X_n) + 1\,. \tag{10.4}$$

Ovviamente, al crescere di n (dimensione del blocco), aumenterà anche il numero K^n dei differenti blocchi di sorgente, e crescerà l'entropia $H(X_1, X_2, \ldots, X_n)$; pertanto crescerà senza limiti anche la lunghezza media $\overline{\ell}(n)$ del codice di Shannon. Per avere un confronto equo per differenti valori di n, introduciamo la *lunghezza media per simbolo di sorgente* $\overline{\ell}_n = \overline{\ell}(n)/n$. Si ha allora, dividendo la (10.4) per n,

$$\frac{1}{n} H(X_1, X_2, \ldots, X_n) \leq \overline{\ell}_n < \frac{1}{n} H(X_1, X_2, \ldots, X_n) + \frac{1}{n}, \tag{10.5}$$

Osserviamo che tale strategia di codifica a blocchi consente di ottenere per n grandi valori arbitrariamente prossimi all'entropia media per simbolo $\frac{1}{n}H(X_1, X_2, \ldots, X_n)$. In particolare, se la sorgente è senza memoria (DMS), risulta $\frac{1}{n}H(X_1, X_2, \ldots, X_n) = H(X_1)$, per cui:

$$H(X_1) \leq \overline{\ell}_n < H(X_1) + \frac{1}{n},$$

per cui la lunghezza media $\overline{\ell}_n$ può essere resa arbitrariamente prossima all'entropia per simbolo $H(X_1)$ (e quindi l'efficienza η può essere resa arbitrariamente prossima all'unità) aumentando la lunghezza del blocco n.

La (10.5) non si applica però solo alle sorgenti DMS, ma a qualunque sorgente per la quale si possa definire l'entropia di sorgente $H(\mathcal{S})$. Infatti, passando al limite per $n \to \infty$ nella (10.5), si ha che $\lim_{n\to\infty} \frac{1}{n} H(X_1, X_2, \ldots, X_n) = H(\mathcal{S})$, se tale limite esiste finito, per cui

$$\lim_{n\to\infty} \overline{\ell}_n = H(\mathcal{S}),$$

secondo la quale l'entropia di sorgente $H(\mathcal{S})$ rappresenta proprio la minima lunghezza media per simbolo di un codice che rappresenta in maniera non ambigua l'informazione emessa da tale sorgente. Tale risultato è di fondamentale importanza nella codifica di sorgente, e prende il nome di *primo teorema di Shannon*, che possiamo formulare sinteticamente come segue:

Teorema 10.5 (primo teorema di Shannon). Data una sorgente discreta $\mathcal{S}$, è sempre possibile costruire un codice binario a blocchi (di lunghezza n) a prefisso (e quindi univocamente decifrabile) tale che la sua lunghezza media per simbolo di sorgente $\overline{\ell}_n$ sia compresa tra i seguenti limiti:

$$\frac{1}{n} H(X_1, X_2, \ldots, X_n) \leq \overline{\ell}_n < \frac{1}{n} H(X_1, X_2, \ldots, X_n) + \frac{1}{n},$$

e quindi arbitrariamente prossima all'entropia media per simbolo della sorgente. Inoltre, se $H(\mathcal{S})$ è l'entropia di sorgente (supposta esistente), risulta

$$\lim_{n\to\infty} \overline{\ell}_n = H(\mathcal{S}).$$

Il primo teorema di Shannon fornisce al tempo stesso un'interpretazione "operativa" del concetto di entropia di sorgente e stabilisce il limite ultimo per l'efficienza di un codice binario e quindi per la compattazione dati. Il prezzo da pagare per migliorare l'efficienza del codice risiede in un incremento della complessità del codificatore e decodificatore di sorgente all'aumentare della lunghezza n dei blocchi da codificare. Tale incremento è in genere esponenziale con n, a meno di non fissare qualche vincolo strutturale sul codice, il che tuttavia non consente in genere di ottenere le prestazioni ottime.

10.5.3 Efficienza dei codici a lunghezza fissa

Molti dei risultati del paragrafo precedente si applicano, come caso particolare, anche ai codici a lunghezza fissa. Per questi ultimi, tuttavia, possiamo ragionare in maniera diretta e molto semplice. Infatti, supponiamo di effettuare la codifica simbolo a simbolo di una sorgente con K possibili simboli, utilizzando un codice binario di lunghezza fissa ℓ: le possibili parole codice sono 2^ℓ, per cui si richiede, per l'univoca decifrabilità, che (si ricordi che il logaritmo è in base 2):

$$2^\ell \geq K \Rightarrow \ell \geq \log K\,.$$

D'altra parte, $\log K$ rappresenta proprio il massimo valore dell'entropia $H(X)$ associata ad una sorgente con K simboli, valore che si ottiene quando i simboli sono equiprobabili, per cui $H(X) \leq \log K$. Concatenando tali disuguaglianze, si ha per codici a lunghezza fissa:

$$\ell \geq \log K \geq H(X)$$

e quindi, se $H(X) \ll \log K$, si ha un'efficienza $\eta \ll 1$, per cui non riusciremo ad ottenere prestazioni confrontabili a quelle dei codici a lunghezza variabile, salvo nel caso in cui la sorgente emetta simboli equiprobabili. D'altra parte, le cose non migliorano se pensiamo di estendere la codifica a blocchi di n simboli. Infatti, in questo caso avremo K^n differenti blocchi, per cui la lunghezza $\ell(n)$ del codice binario dovrà soddisfare alla seguente disuguaglianza:

$$2^{\ell(n)} \geq K^n \Rightarrow \ell(n) \geq \log K^n = n \log K\,.$$

Se allora calcoliamo la lunghezza media per simbolo $\ell_n = \ell(n)/n$, avremo:

$$\ell_n \geq \log K\,,$$

cioè la stessa limitazione che ottenevamo per la codifica simbolo a simbolo, per cui le cose non sono affatto migliorate, così come invece avveniva codificando blocchi di n simboli con un codice a lunghezza variabile.

Possiamo pertanto affermare che i codici a lunghezza fissa, almeno sulla base di queste semplici considerazioni, non sono competitivi con i codici a lunghezza variabile. Considerazioni più avanzate porterebbero a strategie di codifica a lunghezza fissa più sofisticate, nelle quali i simboli di sorgente vengono raccolti in blocchi molto lunghi e non a tutti i blocchi si associano parole codice (si accetta cioè la possibilità che alcuni blocchi possano non essere codificati). In questo modo si riesce ad ottenere una lunghezza del codice che approssima a piacere l'entropia della sorgente, a patto tuttavia di accettare una (piccola) probabilità di mancata codifica. Va detto tuttavia che tali tecniche, per la loro complessità, rivestono un interesse puramente teorico.

10.5.4 Codici di Huffmann

In questa sezione introdurremo una classe di codici a prefisso, noti come *codici di Huffmann*, che risultano *ottimi* e per i quali è possibile fornire una procedura di costruzione sistematica. L'ottimalità di tali codici non va intesa nel senso che essi presentano necessariamente lunghezza media pari al valore minimo possibile, cioè all'entropia, ma nel senso che, tra tutti i codici che operano su blocchi di sorgente di lunghezza prefissata, i codici di Huffmann presentano la *minima lunghezza media*.[8]

[8] Per una discussione più approfondita ed una prova dell'ottimalità di tali codici, si veda [12].

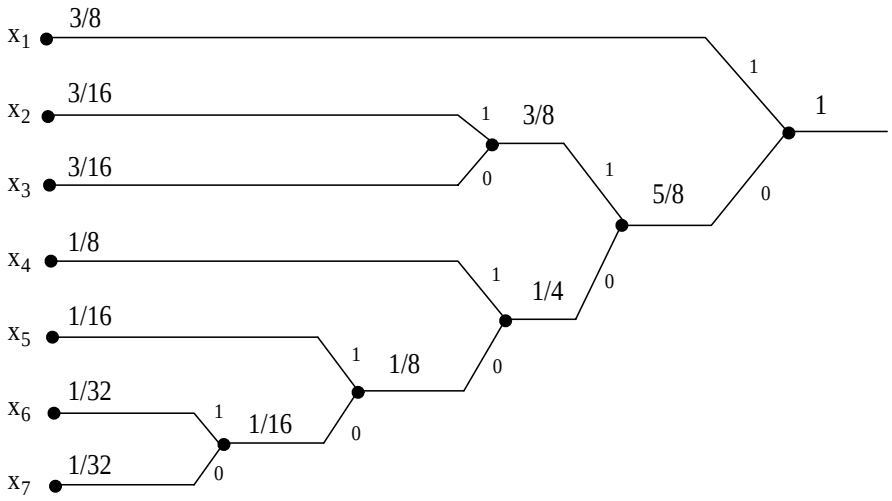


Fig. 10.4. Costruzione di un codice di Huffmann per una sorgente con $K = 7$ simboli.

Anziché definire formalmente le proprietà dei codici di Huffmann, nei seguenti esempi mostriamo direttamente come sia possibile costruire tali codici per determinate sorgenti $\mathcal{S}$.

► *Esempio* 10.15. Sia $\mathcal{S}$ una sorgente con alfabeto di $K = 7$ simboli $\mathcal{X} = \{x_1, x_2, \ldots, x_7\}$, caratterizzati dalle seguenti probabilità (che assumiamo ordinate in senso decrescente, senza ledere la generalità):

$$
\begin{aligned}
p_1 &= 3/8 \\
p_2 &= p_3 = 3/16 \\
p_4 &= 1/8 \\
p_5 &= 1/16 \\
p_6 &= p_7 = 1/32
\end{aligned}
$$

Codificare tale sorgente con un codice a lunghezza fissa richiederebbe $\lceil \log K \rceil = 3$ bit per simbolo. Se però calcoliamo l'entropia della sorgente (in bit), troviamo:

$$
\begin{aligned}
H(X) = -\sum_{k=1}^{7} p_k \log p_k = \\
&= -(3/8)\log(3/8) - 2(3/16)\log(3/16) + \\
&- (1/8)\log(1/8) - (1/16)\log(1/16) - 2(1/32)\log(1/32) = \\
&= 2.37 \text{ bit},
\end{aligned}
$$

per cui l'efficienza di un tale codice a lunghezza fissa è pari a $\eta = 2.37/3 = 79\%$, e pertanto relativamente modesta; ci aspettiamo di poter ottenere un risultato migliore costruendo un codice a lunghezza variabile.

La procedura sistematica per la costruzione di un codice di Huffmann si articola come segue: si costruisce un albero, partendo da sinistra dai simboli di sorgente ordinati secondo le loro probabilità in senso decrescente (vedi Fig. 10.4). Ad ogni passo, i due simboli con probabilità più *piccole* sono combinati in un nuovo simbolo, cui si assegna una probabilità pari alla somma delle due. L'albero in questo modo viene costruito a partire dai nodi terminali fino alla radice, procedendo da sinistra verso destra e combinando via via i simboli meno probabili (tenendo conto anche dei nuovi simboli che si formano per combinazione dei simboli meno probabili nei passi precedenti della procedura), fino ad esaurire i simboli a disposizione. A questo punto, il codice di Huffmann si ottiene ripercorrendo l'albero da destra verso sinistra ed associando a ciascun simbolo la stringa costituita dai contrassegni dei rami. Il codice che si ottiene con tale procedura è riportato in Tab. 10.3.

La lunghezza media $\overline{\ell}$ di tale codice di Huffmann è pari a 2.44 bit per simbolo di sorgente, il che confrontato con l'entropia, che è pari a 2.37 bit, mostra che siamo molto vicini al massimo livello di compattazione ottenibile (l'efficienza del codice è pari a $\eta = 2.37/2.44 \approx 97\%$). ◄

X	probabilità p_k	parola codice	lunghezza ℓ_k
x_1	3/8	1	1
x_2	3/16	011	3
x_3	3/16	010	3
x_4	1/8	001	3
x_5	1/16	0001	4
x_6	1/32	00001	5
x_7	1/32	00000	5

Tab. 10.3. Codice di Huffmann per una sorgente con $K = 7$ simboli.

► *Esempio* 10.16. Mostriamo adesso con un esempio come sia generalmente più conveniente la codifica a blocchi rispetto a quella simbolo a simbolo. Consideriamo una sorgente $\mathcal{S}$ senza memoria con alfabeto di sorgente $\mathcal{X} = \{x_1, x_2, x_3\}$, aventi probabilità 3/4, 3/16, e 1/16. Per semplicità di notazione, poniamo $x_1 = A$, $x_2 = B$ e $x_3 = C$. L'entropia per simbolo di sorgente è pari a 1.012 bit, ed il codice di Huffmann costruito sulla base dell'albero in Fig. 10.5 e riportato in Tab. 10.4 ha lunghezza media $\overline{\ell}$ pari a 1.25, per un'efficienza pari a $\eta = 1.012/1.25 = 81\%$.

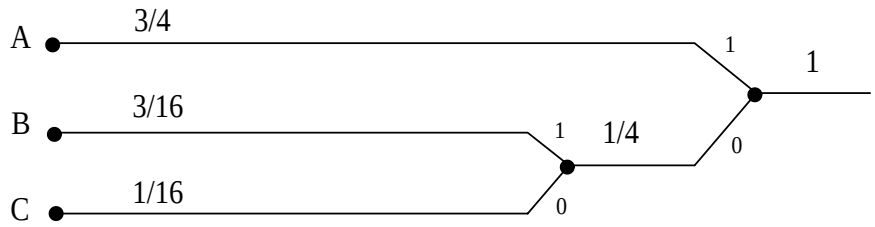


Fig. 10.5. Costruzione di un codice di Huffmann per una sorgente con $K = 3$ simboli (codifica simbolo a simbolo)

X	probabilità p_k	parola codice	lunghezza ℓ_k
A	3/4	1	1
B	3/16	01	2
C	1/16	00	2

Tab. 10.4. Codice di Huffmann per una sorgente con $K = 3$ simboli (codifica simbolo a simbolo).

Possiamo migliorare questo risultato codificando, anziché singoli simboli di sorgente, blocchi di lunghezza $n = 2$. In questo caso, tutto procede come se considerassimo una sorgente equivalente di blocchi, che emette i blocchi AA, AB, etc. Poichè la sorgente è senza memoria, e quindi i simboli successivamente emessi dalla sorgente sono indipendenti, le probabilità associate ai blocchi si ottengono semplicemente moltiplicando tra loro le probabilità dei simboli, e sono riportate in Tab. 10.5.

L'entropia di questa nuova sorgente è chiaramente doppia rispetto a quella della sorgente per $n = 1$, essendo i simboli indipendenti (sorgente senza memoria), e vale pertanto 2.024 bit; l'entropia per simbolo però non è cambiata, e vale ancora 1.012 bit. La costruzione del codice di Huffmann procede come mostrato in Fig. 10.6 (notiamo che non abbiamo ordinato i blocchi in ordine decrescente di probabilità), ed il codice relativo è riportato in Tab. 10.5.

Se si calcola la lunghezza media del codice, si trova $\overline{\ell}(n) = 2.074$, ma stavolta con tale codice si codificano $n = 2$ simboli di sorgente, per cui la lunghezza media per simbolo di sorgente $\overline{\ell}_n = \overline{\ell}(n)/n$ è pari a $2.074/2 = 1.037$ bit, inferiore al valore ottenuto con la codifica di un simbolo alla volta. Difatti, l'efficienza passa dal valore $\eta = 81\%$ a $\eta = 2.024/2.074 = 1.012/1.037 \approx 98\%$, mostrando il significativo vantaggio conseguito con tale strategia di codifica a blocchi. Notiamo che tale vantaggio della codifica a blocchi si è manifestato

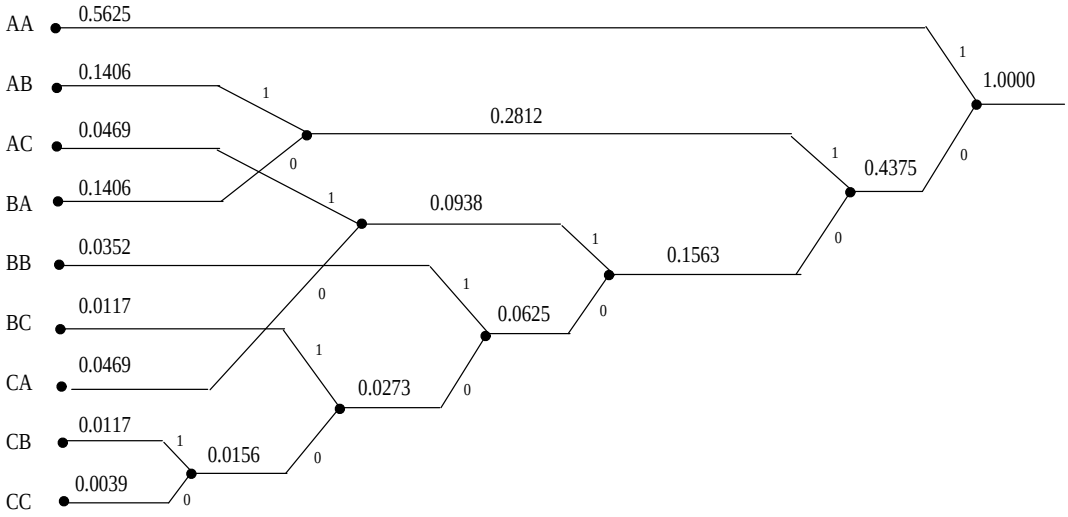


Fig. 10.6. Costruzione di un codice di Huffmann per una sorgente con $K = 3$ simboli (codifica a blocchi di $n = 2$ simboli).

X_1X_2	probabilità p_k	parola codice	lunghezza ℓ_k
AA	0.5625	1	1
AB	0.1406	011	3
AC	0.0469	0011	4
BA	0.1406	010	3
BB	0.0352	0001	4
BC	0.0117	00001	5
CA	0.0469	0010	4
CB	0.0117	000001	6
CC	0.0039	000000	6

Tab. 10.5. Codice di Huffmann per una sorgente con $K = 3$ simboli (codifica a blocchi di $n = 2$ simboli).

anche se di fatto la sorgente è senza memoria: vantaggi ancora maggiori si ottengono per sorgenti con memoria. Il prezzo da pagare per questa compattazione più spinta è la maggiore complessità realizzativa del codificatore e del decodificatore. Tale complessità è certamente proporzionale al numero K^n di parole codice e quindi cresce esponenzialmente con la lunghezza n del blocco da codificare. ◀

Vale la pena osservare che, per una fissata sorgente, la procedura per la costruzione di codici di Huffmann non porta ad un unico codice, in quanto in alcuni passi è possibile avere più di una scelta nel combinare le due probabilità più piccole. Quello che invece è certo è che i codici di Huffmann risultanti, pur diversi nella scelta delle parole codice, presenteranno tutti *la stessa* lunghezza media, che è la minima possibile.

10.6 Esercizi proposti

Esercizio 10.1. Calcolare l'entropia[9] associata ai seguenti esperimenti aleatori:

a) lancio di una moneta;

b) lancio di un dado;

c) lancio di una moneta e di un dado;

d) estrazione di un numero al lotto.

Esercizio 10.2. Calcolare l'entropia di un alfabeto composto da 4 simboli con probabilità $1/2$, $1/4$, $1/8$, $1/8$.

Esercizio 10.3. Calcolare l'entropia di un alfabeto composto da 5 simboli con probabilità $1/4$, $1/8$, $1/8$, $3/16$, $5/16$.

Esercizio 10.4. Dimostrare per differenziazione diretta rispetto a p che l'entropia binaria $H(p)$ è massima per $p = 1/2$.

★ *Esercizio* 10.5. Dimostrare utilizzando la tecnica dei *moltiplicatori di Lagrange* che l'entropia $H(\boldsymbol{p})$ è massima quando la distribuzione di probabilità $\boldsymbol{p} = [p_1, p_2, \ldots, p_K]$ è uniforme.

Esercizio 10.6. Il segnale vocale digitale (segnale PCM) consiste in *campioni* che si susseguono alla velocità di 8000 campioni al secondo. Se ogni campione può assumere 256 diversi valori, determinare (assumendo l'indipendenza tra i campioni):

a) l'entropia di un singolo campione;

b) il tasso di informazione R_s (in bit/s o multipli) del segnale PCM.

Esercizio 10.7. Il segnale audio digitale stereo (segnale CD) consiste in *due* successioni di *campioni* (uno per il canale sinistro e l'altro per il canale destro) che si susseguono alla velocità di 44100 campioni al secondo. Se ogni campione può assumere $2^{16} = 65536$ diversi valori, determinare (assumendo l'indipendenza tra i campioni e tra i canali):

a) l'entropia di un singolo campione (canale destro o sinistro);

b) l'entropia di una *coppia* di campioni (canale destro e sinistro);

c) il tasso di informazione R_s (in bit/s o multipli) del segnale CD.

Esercizio 10.8. Il segnale televisivo in bianco e nero consiste di *quadri* che si susseguono alla velocità di 30 al secondo. Se ogni quadro è composto da 525 linee, ed ogni linea è composta da 525 elementi di immagine (pixel), ed ogni pixel può assumere 256 diversi valori (livelli di grigio), determinare (assumendo l'indipendenza tra pixel, linee e quadri):

a) l'entropia di un singolo pixel;

b) l'entropia di una singola riga;

c) l'entropia di un singolo quadro;

d) il tasso di informazione R_s (in bit/s o multipli) del segnale televisivo in bianco e nero.

Esercizio 10.9. Sia X una variabile aleatoria che assume i valori -1, 0 ed 1 in maniera equiprobabile.

a) Calcolare l'entropia di X.

b) Calcolare l'entropia di $Y = 2X$, confrontare con quella di X e giustificare intuitivamente il risultato.

c) Calcolare l'entropia di $Y = X^2$, confrontare con quella di X e giustificare intuitivamente il risultato.

Esercizio 10.10. Stabilire, utilizzando la rappresentazione ad albero, se i seguenti codici sono a prefisso.

a) $\mathcal{C}_1 = \{0, 101\}$

[9] In tutti gli esercizi, le entropie vanno calcolate in bit, se non diversamente specificato.

b) $\mathcal{C}_2 = \{1, 101\}$

c) $\mathcal{C}_3 = \{0, 10, 110, 111\}$

d) $\mathcal{C}_4 = \{00, 01, 10, 11\}$

e) $\mathcal{C}_5 = \{0, 01, 011, 111\}$

Esercizio 10.11. Costruire un codice di Shannon simbolo a simbolo per un alfabeto di $K = 13$ simboli di sorgente, aventi la seguente distribuzione di probabilità:

$$0.2 \quad 0.18 \quad 0.1 \quad 0.1 \quad 0.1 \quad 0.061 \quad 0.059 \quad 0.04 \quad 0.04 \quad 0.04 \quad 0.04 \quad 0.03 \quad 0.01$$

e confrontare la lunghezza media delle parole codice con l'entropia dell'alfabeto $H(X)$ (considerare sia il codice di Shannon senza "potatura" dell'albero, che quello con potatura). Confrontare il risultato ottenuto con quello di un codice a lunghezza fissa.

Esercizio 10.12. Ripetere l'esercizio 10.11 utilizzando un codice di Huffmann.

Esercizio 10.13. Costruire un codice di Shannon ed uno di Huffmann, simbolo a simbolo, per un alfabeto di $K = 7$ simboli di sorgente, aventi la seguente distribuzione di probabilità:

$$0.3 \quad 0.2 \quad 0.15 \quad 0.15 \quad 0.1 \quad 0.06 \quad 0.04$$

e confrontare la lunghezza media delle parole codice con l'entropia dell'alfabeto $H(X)$ (considerare sia il codice di Shannon senza "potatura" dell'albero, che quello con potatura). Confrontare il risultato ottenuto con quello di un codice a lunghezza fissa.

Esercizio 10.14. Si consideri una sorgente discreta senza memoria, binaria, con probabilità dei simboli $q = 0.1$ e $p = 0.9$.

a) Costruire un codice di Huffmann *simbolo a simbolo* e calcolare l'efficienza di codifica.

b) Costruire un codice di Huffmann *blocco a blocco* per $n = 2, 3, 4$ e calcolare l'efficienza di codifica nei diversi casi.

Esercizio 10.15. Ripetere l'esercizio 10.14 per una sorgente con $q = 0.4$ e $p = 0.6$.

★ *Esercizio* 10.16. Costruire un codice di Huffman per le lettere dell'alfabeto italiano, utilizzando i dati riportati in Tab. 10.1, e confrontare il risultato con l'entropia dell'alfabeto $H(X)$.

Appendice A

Fattoriale e coefficiente binomiale

In questa appendice si richiamano brevemente le definizioni e le proprietà del fattoriale e del coefficiente binomiale.

A.1 Fattoriale

Il fattoriale $n!$ di un numero $n \in \mathbb{N} \cup \{0\}$ è definito come:

$$n! \triangleq n\,(n-1)\,(n-2)\cdots 3\cdot 2\cdot 1\,.$$

Ad esempio, si ha $3! = 3\cdot 2\cdot 1 = 6$ e $5! = 5\cdot 4\cdot 3\cdot 2\cdot 1 = 120$. Convenzionalmente, si pone $0! = 1$. Nel calcolo combinatorio, il fattoriale rappresenta il numero di differenti *permutazioni* di n elementi. A volte si utilizza anche il simbolo $n!!$ (doppio fattoriale), che rappresenta il prodotto dei soli numeri *dispari* fino ad n. Ad esempio, $5!! = 5\cdot 3\cdot 1 = 15$.

In Matlab, il fattoriale si può calcolare come `prod(1:n)`. Il fattoriale è una funzione che cresce molto rapidamente, ed un'approssimazione valida per valori elevati di n è la cosiddetta *formula di Stirling*:

$$n! \approx \sqrt{2\pi}\, n^{n+1/2}\, e^{-n}\,.$$

A.2 Coefficiente binomiale

Il coefficiente binomiale di parametri n e $k \leq n$ è definito come:

$$\binom{n}{k} \triangleq \frac{n\,(n-1)\cdots(n-k+2)\,(n-k+1)}{k!} = \frac{n!}{k!\,(n-k)!}\,. \tag{A.1}$$

Nel calcolo combinatorio, il coefficiente binomiale di parametri n e k rappresenta il numero di disposizioni non ordinate e senza sostituzioni di n oggetti su k posti (vedi Appendice B).

Valgono le seguenti identità notevoli:

$$\binom{n}{0} = 1; \quad \binom{n}{1} = n; \quad \binom{n}{k} = \binom{n}{n-k}.$$

nonché la seguente:

$$\binom{n}{k} + \binom{n}{k+1} = \binom{n+1}{k+1}.$$

In Matlab, il coefficiente binomiale si può calcolare con il comando `nchoosek(n,k)`.

A.3 Espansioni binomiali

Il coefficiente binomiale compare nell'espansione della potenza n-esima di un binomio, come enunciato dal seguente *teorema binomiale*:

$$(a+b)^n = \sum_{k=0}^{n} \binom{n}{k} a^k b^{n-k}, \tag{A.2}$$

valido per ogni $n \in \mathbb{N}$ e per ogni $a, b \in \mathbb{R}$.

Il teorema può essere generalizzato al caso di elevazione a potenza qualsiasi, ricorrendo allo sviluppo in serie di Mc-Laurin di $(1+x)^\alpha$. Si ha:

$$(1+x)^\alpha = \sum_{k=0}^{\infty} \binom{\alpha}{k} x^k \tag{A.3}$$

dove $\alpha \in \mathbb{R}$, e l'espansione vale se $|x| < 1$. Nella (A.3), la definizione di coefficiente binomiale è una semplice generalizzazione della (A.1):

$$\binom{\alpha}{k} \triangleq \frac{\alpha\,(\alpha-1)\cdots(\alpha-k+2)\,(\alpha-k+1)}{k!}. \tag{A.4}$$

Per sviluppare $(a+b)^\alpha$, allora, ci si riconduce al caso della (A.3), mettendo in evidenza il maggiore tra a e b.

La relazione (A.3) può essere applicata, in particolare, al caso in cui $\alpha = -n$, con $n \in \mathbb{N}$. Si ha in tal caso:

$$(1+x)^{-n} = \sum_{k=0}^{\infty} \binom{-n}{k} (-x)^k = \sum_{k=0}^{\infty} \binom{-n}{k} (-1)^k x^k.$$

Ma essendo, per la (A.4),

$$\begin{aligned}\binom{-n}{k} &= \frac{(-n)(-n-1)\cdots(-n-k+1)}{k!} = (-1)^k \frac{n(n+1)\cdots(n+k-1)}{k!} \\ &= (-1)^k \frac{(n+k-1)!}{k!\,(n-1)!} = (-1)^k \binom{n+k-1}{k}\end{aligned}$$

si ottiene

$$(1+x)^{-n} = \sum_{k=0}^{\infty} \binom{n+k-1}{k} (-1)^k x^k, \qquad |x| < 1.$$

Ponendo $-x$ in luogo di x, si ottiene poi la formula più compatta:

$$(1-x)^{-n} = \sum_{k=0}^{\infty} \binom{n+k-1}{k} x^k, \qquad |x| < 1, \tag{A.5}$$

che va sotto il nome di *espansione binomiale negativa*.

Appendice B

Elementi di calcolo combinatorio

Questa appendice fornisce un'introduzione snella ai principi ed alla terminologia del calcolo combinatorio, con un'applicazione al calcolo delle probabilità del gioco del poker.

B.1 Introduzione

Per determinare il numero di elementi di un insieme, il modo più diretto è quello di contarli; ovviamente ciò è possibile solo se tale numero non è eccessivamente elevato. Una possibile alternativa è ricorrere al calcolo combinatorio, che fornisce una serie di regole per determinare il numero degli elementi di un insieme (ovvero la sua cardinalità) senza doverli effettivamente contare. Nel calcolo della probabilità, il calcolo combinatorio serve per determinare le probabilità in spazi campione Ω discreti con un numero finito di risultati equiprobabili, in accordo alla definizione *classica* o *laplaciana*:

$$P(A) \triangleq \frac{\text{card}(A)}{\text{card}(\Omega)}$$

Come primo risultato, enunciamo e dimostriamo il fondamentale:

Teorema B.1 (teorema fondamentale del conteggio). Se ho n_1 oggetti di tipo 1, ed n_2 oggetti di tipo 2,esistono $n_1\, n_2$ modi distinti di scegliere un oggetto di tipo 1 *ed* un oggetto di tipo 2.

Prova. Basta costruire una tabella con n_1 righe ed n_2 colonne, nella quale ogni riga corrisponde ad un oggetto di tipo 1, ed ogni colonna ad un oggetto di tipo 2. Poiché il numero totale di posti della tabella è $n_1\, n_2$, tale sarà anche il numero di modi distinti di scegliere un oggetto di tipo 1 ed un oggetto di tipo 2. □

Un modo alternativo di interpretare il risultato precedente è quello che, se Ω_1 è un insieme di cardinalità n_1, e Ω_2 è un insieme di cardinalità n_2, la cardinalità del *prodotto cartesiano* $\Omega_1 \times \Omega_2$ è pari a $n_1\, n_2$.

B.2 Schema fondamentale del conteggio

Sebbene non tutti i problemi di conteggio e calcolo combinatorio abbiano a che fare con estrazioni, la gran parte di essi possono essere ricondotti al seguente schema fondamentale.

> **Schema fondamentale del conteggio:** abbiamo una scatola che contiene n oggetti distinti, che denotiamo con $\omega_1, \omega_2, \omega_3, \ldots, \omega_n$, ed estraiamo k oggetti; vogliamo contare il quante differenti estrazioni (ovvero quante differenti k-ple) possono essere effettuate.

Osserviamo subito che, per come è formulato, lo schema fondamentale contiene ben *due* elementi di ambiguità. In primo luogo, dobbiamo chiarire se l'estrazione degli oggetti dalla scatola avvenga *con sostituzione* oppure *senza sostituzione*. Nel primo caso, si immagina che una volta estratto un oggetto, nelle successive estrazioni lo stesso oggetto possa essere nuovamente estratto; si può pensare che tale schema descriva una delle due seguenti situazioni:

- l'estrazione dei k oggetti avviene in successione, e dopo ogni estrazione l'oggetto viene inserito nuovamente nella scatola (estrazione con *reimmissione* o con *rimpiazzo*);
- l'estrazione di ciascuno dei k oggetti avviene da k scatole uguali, ciascuna delle quali contiene tutti gli oggetti che è possibile estrarre.

È chiaro allora che in una estrazione con sostituzione la coppia (ω_1, ω_1) è ammissibile, mentre non lo è in un'estrazione senza sostituzione.

Un secondo elemento di ambiguità dello schema fondamentale è se noi consideriamo *diverse* oppure *uguali* coppie come (ω_1, ω_2) e (ω_2, ω_1) che differiscono solo per l'ordine in cui compaiono gli elementi: nel primo caso parliamo di conteggio di coppie *con ordinamento*, nel secondo caso parliamo di coppie *senza ordinamento*.

Notiamo che prima di iniziare a contare dobbiamo capire con chiarezza gli aspetti precedentemente menzionati: se cioè l'esperimento sia con sostituzione oppure senza sostituzione, il che è legato al meccanismo di estrazione ed è pertanto un fatto "oggettivo"; e se poi il conteggio considera le coppie ordinate oppure non ordinate, il che non è legato direttamente all'esperimento, ma al nostro modo di interpretarne i risultati, ed è pertanto un fatto "soggettivo", e quindi talvolta meno chiaro. Premettiamo che in molti casi un medesimo problema può essere risolto correttamente considerando le coppie "soggettivamente" ordinate oppure no.

In conclusione, abbiamo 2 modalità di estrazione (con o senza sostituzione) e 2 modalità di conteggio (con o senza ordinamento), per cui abbiamo 4 situazioni possibili (sì, questo è il vostro primo conteggio corretto!):

- con sostituzione, con ordinamento;
- senza sostituzione, con ordinamento;
- con sostituzione, senza ordinamento;
- senza sostituzione, senza ordinamento.

Negli esempi che seguono presentiamo le quattro differenti situazioni in un caso particolarmente semplice, di una scatola con $n = 4$ oggetti A, B, C, D. Effettueremo il conteggio per enumerazione

delle coppie di oggetti ($k = 2$) che è possibile estrarre, e cercheremo poi di ricavare le leggi generali, per valori di n e k arbitrari.

► *Esempio* B.1 *(estrazione con sostituzione, conteggio con ordinamento).* In questo caso, le possibili coppie estratte sono le seguenti:

AA	AB	AC	AD
BA	BB	BC	BD
CA	CB	CC	CD
DA	DB	DC	DD

Notiamo che le coppie AA, BB, CC, DD sono ammissibili (l'estrazione è con sostituzione) e le coppie AB e BA sono considerate distinte (il conteggio è con ordinamento).

Per enumerazione, notiamo che il numero di coppie in questo caso è pari a 16; questo perchè il primo elemento della coppia possiamo sceglierlo in 4 modi differenti, ed il secondo elemento della coppia possiamo sceglierlo in 4 modi differenti; per cui il numero totale è $4 \times 4 = 16$. ◄

È allora facile generalizzare questo risultato: infatti, se dobbiamo contare le k-ple di n oggetti con sostituzione e con ordinamento, basta osservare che il primo oggetto della k-pla possiamo sceglierlo in n modi differenti, il secondo elemento della k-pla in n modi differenti, e così via, fino all'ultimo; allora il numero totale di k-ple distinte sarà:

$$\boxed{\underbrace{n \times n \times n \times \cdots \times n}_{k \text{ volte}} = n^k} \tag{B.1}$$

Questa è la prima formula fondamentale del calcolo combinatorio, e conta il numero di k-ple *con sostituzione e con ordinamento.*

► *Esempio* B.2 *(estrazione senza sostituzione, conteggio con ordinamento).* Se non ammettiamo la sostituzione, dobbiamo eliminare dal conteggio precedente le coppie con due elementi uguali, cioè AA, BB, CC, DD. La tabella delle possibili coppie estratte si modifica come segue:

AB	AC	AD
BA	BC	BD
CA	CB	CD
DA	DB	DC

Per enumerazione, notiamo che il numero di coppie in questo caso è pari a 12; infatti, il primo elemento della coppia possiamo sceglierlo in 4 modi differenti, ma il secondo elemento della coppia possiamo sceglierlo in 3 modi differenti (perché non possiamo scegliere nuovamente il primo); per cui il numero totale è $4 \times 3 = 12$. ◄

Anche questo risultato può essere facilmente generalizzato: infatti, se dobbiamo contare le k-ple di n oggetti senza sostituzione e con ordinamento, basta osservare che il primo oggetto della k-pla possiamo sceglierlo in n modi differenti, il secondo elemento della k-pla in $n-1$ modi differenti, il terso elemento della k-pla in $n-2$ modi differenti, e così via, fino al k-esimo, che potremo scegliere in $n-k+1$ modi differenti; allora il numero totale di k-ple distinte sarà:

$$\boxed{\underbrace{n \times (n-1) \times (n-2) \times \cdots \times (n-k+1)}_{k \text{ termini}} = n(n-1)(n-2)\cdots(n-k+1) = \frac{n!}{(n-k)!}} \tag{B.2}$$

Questa formula conta il numero di k-ple *senza sostituzione e con ordinamento.* Notiamo che nel caso $k = n$ questa formula restituisce il numero $n!$ delle *permutazioni* di n elementi.

► *Esempio* B.3 *(estrazione senza sostituzione, conteggio senza ordinamento).* Se adesso non teniamo conto dell'ordinamento, dovremo considerare coincidenti due coppie come AB e BA che differiscono soltanto per il numero degli elementi. La tabella delle possibili coppie estratte si ottiene modificando quella precedente ed riportando una sola volta le coppie che differiscono per l'ordine degli elementi.

AB	AC	AD
BC	BD	
CD		

Per enumerazione, notiamo che rispetto al caso con ordinamento il numero di coppie si è ridotto a 6; infatti, ricordiamo che nel caso con ordinamento avevamo 12 coppie, ma ciascuna di esse ne aveva un'altra con gli elementi scambiati, per cui il numero di coppie senza ordinamento è pari a $12/2 = 6$. ◄

Sebbene non proprio banale come i precedenti, anche questo risultato ammette una generalizzazione quasi immediata a valori arbitrari di n e k. Infatti, partiamo dal numero di k-ple senza sostituzione e con ordinamento, pari a $n(n-1)\cdots(n-k+1)$; se prendiamo una qualsiasi di queste k-ple, essa appartiene ad un gruppo di $k!$ $k-ple$ che differiscono solo per una *permutazione* dei k elementi che la compongono; se allora voglio contare il numero di k-ple senza ordinamento, devo dividere il numero $n(n-1)\cdots(n-k+1)$ proprio per $k!$, ottenendo così:

$$\boxed{\frac{n(n-1)\ldots(n-k+1)}{k!} = \frac{n!}{k!\,(n-k)!} = \binom{n}{k}} \tag{B.3}$$

Questa formula conta il numero di k-ple *senza sostituzione e senza ordinamento*, che è uno dei casi che più spesso si presenta in pratica.

► *Esempio* B.4 *(estrazione con sostituzione, conteggio senza ordinamento).* Consideriamo quest'ultimo caso, che è quello più difficile da trattare. Partiamo dalla tabella che abbiamo ricavato nel caso di estrazione con sostituzione e conteggio con ordinamento, ed eliminiamo le confugurazioni che differiscono solo per l'ordine degli elementi. Si ha:

AA	AB	AC	AD
BB	BC	BD	
CC	CD		
DD			

In questo caso, il numero delle coppie risulta pari a 10, ma non sappiamo subito giustificare questo risultato (notiamo che non risulta banalmente pari a $16/2! = 8$, ovvero dividendo il numero nel caso con ordinamento per il numero delle permutazioni. Il ragionamento è allora un po' più complesso, e possiamo portarlo avanti nel caso specifico, per poi ricavare la legge generale. Immaginiamo che i quattro oggetti si possano rappresentare nel seguente modo, separati da $4 + 1 = 5$ *pareti*

A	B	C	D

Allora la scelta delle coppie si può interpretare come segue; abbiamo due *segnaposti*, che denotiamo con S1 ed S2, e possiamo piazzarli arbitrariamente tra le pareti. Ad esempio, la coppia AA corrisponde a collocare i due segnaposti in A, come segue:

S1 S2 A	B	C	D

e così la coppia AB corrisponde alla seguente configurazione dei segnaposti:

$$\begin{array}{|c|c|c|c|} \text{S1} & \text{S2} & & \\ \text{A} & \text{B} & \text{C} & \text{D} \end{array}$$

e così via. Notiamo allora che possiamo anche eliminare completamente gli elementi A, B, C, D nella descrizione e sintetizzare le coppie con:

$$\text{AA} \quad \Leftrightarrow \quad \begin{array}{|c|c|c|c|} \text{S1 S2} & & & \end{array}$$

$$\text{AB} \quad \Leftrightarrow \quad \begin{array}{|c|c|c|c|} \text{S1} & \text{S2} & & \end{array}$$

Possiamo ottenere una descrizione ancora più sintetica eliminando la prima e l'ultima parete, che in effetti sono ridondanti, perché occupano sempre la stessa posizione. Quindi avremo:

$$\text{AA} \quad \Leftrightarrow \quad \begin{array}{|c|c|} \text{S1 S2} & \end{array}$$

$$\text{AB} \quad \Leftrightarrow \quad \begin{array}{|c|c|} \text{S1} & \text{S2} \end{array}$$

e quindi mi sono ricondotto ad un problema con 3 pareti e 2 segnaposti, che possono occupare $3+2=5$ posizioni. A questo punto avrò che la descrizione di una coppia di oggetti (AA, AB, etc) è equivalente specificare in quale tra le 5 posizioni a disposizione si trovano i 2 segnaposti. Notiamo che le configurazioni che differiscono solo per l'ordine dei segnaposti corrispondono alla stessa coppia; ad esempio le configurazioni

$$\begin{array}{|c|c|} \text{S1 S2} & \end{array}$$

$$\begin{array}{|c|c|} \text{S2 S1} & \end{array}$$

corrispondono entrambi alla coppia BB. Allora abbiamo ricondotto il problema di contare le coppie AA, AB etc. a quello di contare in quanti modi possiamo collocare i 2 segnaposti sulle 5 posizioni a disposizione, in un'estrazione senza sostituzione e senza tener conto dell'ordinamento: tale numero è dato dalla (B.3), ed è pari a

$$\binom{5}{2} = \frac{5 \times 4}{2} = 10$$

che è lo stesso risultato che avevamo trovato per enumerazione. ◄

La tecnica di conteggio basata sui segnaposti si può generalizzare al caso di k ed n arbitrari; in tal caso, avrò un totale di $n+k-1$ posizioni (eliminando le due estreme) e di k segnaposti, per cui si avrà:

$$\boxed{\binom{n+k-1}{k} = \frac{(n+k-1)!}{k!\,(n-1)!}} \tag{B.4}$$

Questa formula conta il numero di k-ple *con sostituzione e senza ordinamento*, che è uno dei casi che più raramente si presenta in pratica.

I risultati ottenuti, relativi alle due modalità di estrazione, con o senza sostituzione, e alle due interpretazioni dei risultati, con o senza ordinamento, sono sinteticamente riportati in Tab. B.1 per una rapida consultazione. Nel paragrafo precedente applicheremo le nozioni apprese di calcolo combinatorio al problema del calcolo delle probabilità associate ai vari punti del gioco del poker.

B.3 Applicazione al calcolo delle probabilità nel gioco del poker

Il gioco del poker è un gioco di scommessa originario del sud degli Stati Uniti e diffuso in tutto il mondo; in Italia ha assunto una sua particolare regolamentazione che si discosta da quella americana, e a cui faremo riferimento nella trattazione che segue.[1] Si gioca con un mazzo di carte

[1] Le informazioni sul gico del poker sono tratte, più che dalla scarsa esperienza personale dell'autore, da E. Fantini e C.E. Santelia "I giochi di carte", Rizzoli, Milano, 1985.

	senza sostituzione	con sostituzione
ordinate	$\frac{n!}{(n-k)!}$	n^k
non ordinate	$\binom{n}{k}$	$\binom{n+k-1}{k}$

Tab. B.1. Numero di possibili estrazioni di k oggetti da una scatola contenente n oggetti.

francesi, composto da 52 carte, divise in quattro *semi*: ♡ (cuori), ♢ (quadri), ♣ (fiori), ♠ (picche), con 13 *valori* per seme: $2, 3, 4, 5, 6, 7, 8, 9, 10, J, Q, K, A$. Si gioca con il mazzo completo o incompleto a seconda del numero dei giocatori, che va da 2 fino a 9; in particolare, se i giocatori sono 9 si gioca con il mazzo completo, per ogni giocatore in meno si tolgono dal mazzo le quattro carte di minor valore di ogni seme (in pratica detto N il numero dei giocatori, il valore della carta più bassa di ogni seme è pari a $V = 11 - N$). Ogni giocatore riceve in una prima distribuzione cinque carte, ed effettua delle scommesse sulla base del punto che ha totalizzato nelle cinque carte. Successivamente, i giocatori possono sostituire fino a quattro carte della mano, ed effettuare un nuovo giro di scommesse. Vince il giocatore che al termine delle scommesse è in possesso del punto migliore oppure quello che, avendo formulato la scommessa più alta, non trova nessuno disposto ad accettarla.

Per semplicità, consideriamo il gioco con il mazzo completo di 52 carte; le tecniche di conteggio introdotte, tuttavia, possono essere facilmente generalizzate al caso di un mazzo incompleto. Preliminarmente dobbiamo contare il numero di mani che si possono avere distribuendo 5 carte da un mazzo di 52 carte. Si tratta di un esperimento senza sostituzione e senza ordinamento (il punto ottenuto in una mano non dipende dall'ordine delle carte), con $n = 52$ e $k = 5$. Pertanto, per la (B.3), tale numero è pari a:

$$\binom{n}{k} = \binom{52}{5} = \frac{52 \times 51 \times 50 \times 49 \times 48}{5 \times 4 \times 3 \times 2 \times 1} = 2\,598\,960\,.$$

Nel seguito introduciamo i punti del poker, in ordine decrescente di importanza (e crescente di probabilità), e calcoliamo per ciascuno di essi il numero di diversi modi in cui si può ottenere e quindi la sua probabilità utilizzando la definizione classica. Notiamo che dal punto di vista del calcolo combinatorio abbiamo a che fare sempre con un esperimento senza sostituzione (se nessuno bara!) e senza ordinamento.

Scala reale: è una mano con cinque carte dello stesso seme in sequenza.

Per un fissato seme, una scala reale può avere come valore più alta $5, 6, 7, 8, 9, 10, J, Q, K, A$, per cui il numero di *scale reali* per ogni seme è pari a 10, e quindi il numero totale di *scale reali* è pari (per il teorema fondamentale del conteggio) a

$$4 \times 10 = 40$$

La probabilità di una *scala reale* è allora

$$P(\textit{scala reale}) = \frac{40}{2\,598\,960} \approx 1.54 \cdot 10^{-5}$$

Poker: è una mano con quattro carte dello stesso valore.

Ho 13 differenti scelte di quattro carte uguali (corrispondenti ai 13 diversi valori delle carte di un seme); la restante quinta carta può essere scelta in $52 - 4 = 48$ modi differenti; pertanto, il numero totale di *poker* è pari a

$$13 \times 48 = 624$$

La probabilità di un *poker* è allora

$$P(poker) = \frac{624}{2\,598\,960} \approx 2.40 \cdot 10^{-4}$$

Full: è una mano con tre carte dello stesso valore e due carte dello stesso valore.

Consideriamo le tre carte uguali; poiché ho 13 differenti valori, potrò avere 13 diverse triplette, ciascuna delle quali si ottiene in $\binom{4}{3} = 4$ modi diversi; per le due carte uguali, un ragionamento analogo mi porta a dire che se considero le due carte uguali, posso sceglierle adesso in 12 modi differenti, ed ogni coppia si può presentare in $\binom{4}{2} = 6$ modi differenti. Il numero totale di *full* che posso avere è:

$$13 \times 4 \times 12 \times 6 = 3744$$

La probabilità di un *full* è allora

$$P(full) = \frac{3744}{2\,598\,960} \approx 1.44 \cdot 10^{-3}$$

Colore: è una mano con cinque carte dello stesso colore.

Considerando un solo seme, il numero di *colore* differenti si può ottenere estraendo senza sostituzione e senza ordinamento 5 carte tra le 13 dello stesso seme. Pertanto, il numero totale di *colore* che posso avere è:

$$4 \times \binom{13}{5} = 4 \times \frac{13 \times 12 \times 11 \times 10 \times 9}{5 \times 4 \times 3 \times 2 \times 1} = 4 \times 1287 = 5148$$

In questo calcolo, non ho tenuto conto del fatto che devo sottrarre dal numero $\binom{13}{5} = 1287$ di *colore* per ogni seme le 10 combinazioni che danno luogo ad una *scala reale* (vedi), per cui il numero di *colore* per seme (escluse le scale reali) è pari a $1287 - 10 = 1277$. In conclusione, allora, il numero totale di *colore* sarà

$$1277 \times 4 = 5108$$

La probabilità di un *colore* è allora

$$P(colore) = \frac{5108}{2\,598\,960} \approx 1.97 \cdot 10^{-3}$$

Scala: è una mano con cinque carte in sequenza.

Poiché una scala può avere come carta più alta $5, 6, 7, 8, 9, 10, J, Q, K, A$, ho 10 differenti scale; ciascuna carta della scala può essere scelta tra i 4 semi, quindi ognuna delle 10 scale può essere scelta in $4 \times 4 \times 4 \times 4 \times 4 = 4^5 = 1024$ modi differenti, da cui devo togliere 4 scale reali (una per ciascun seme), ottenendo 1020. Il numero totale di *scale* è allora pari a

$$10 \times 1020 = 10\,200$$

La probabilità di una *scala* è allora

$$P(scala) = \frac{10\,200}{2\,598\,960} \approx 3.92 \cdot 10^{-3}$$

Tris: è una mano con tre carte dello stesso valore.

Consideriamo le tre carte uguali; poiché ho 13 valori, potrò avere 13 diverse triplette, ciascuna delle quali si può ottenere in $\binom{4}{3} = 4$ modi diversi; rimangono 12 valori di carte, corrispondenti a 48 carte, che potrò disporre sui rimanenti due posti in $\binom{48}{2} = 1128$ modi diversi; devo però escludere tutte le combinazioni con due carte uguali, che darebbero luogo ad un *full*, e sono in numero pari a $12 \times \binom{4}{2} = 72$. Pertanto il numero totale di *tris* è pari a

$$13 \times 4 \times (1128 - 72) = 54\,912$$

La probabilità di un *tris* è allora

$$P(tris) = \frac{54\,912}{2\,598\,960} \approx 2.11 \cdot 10^{-2}$$

Doppia coppia: è una mano con una coppia di carte dello stesso valore, ed un'altra coppia di carte dello stesso valore.

Ho 13 possibili valori per ciascuna delle due coppie, il numero totale delle combinazioni dei due valori è pari a $\binom{13}{2} = 78$; ciascuna coppia si potrà poi ottenere in $\binom{4}{2} = 6$ modi differenti. Rimangono $11 \times 4 = 44$ scelte per la quinta carta, per cui il numero totale di *doppie coppie* è

$$78 \times 6 \times 6 \times 44 = 123\,552$$

La probabilità di una *doppia coppia* è allora

$$P(doppia\ coppia) = \frac{123\,552}{2\,598\,960} \approx 4.75 \cdot 10^{-2}$$

Coppia: è una mano con due carte dello stesso valore.

Ho 13 possibili valori per la coppia, e ciascuna coppia si potrà poi ottenere in $\binom{4}{2} = 6$ modi. Rimangono 12 valori, da piazzare (senza ripetizione) sui tre posti rimanenti, e questo si può fare in $\binom{12}{3} = 220$ modi differenti. Ciascuno dei valori, poi, si può ottenere in 4 modi differenti. In totale, il numero di *coppie* è allora pari a

$$13 \times 6 \times 220 \times 4 \times 4 \times 4 = 1\,098\,240$$

La probabilità di una *coppia* è allora

$$P(coppia) = \frac{1\,098\,240}{2\,598\,960} \approx 0.42$$

Nessun punto: è una mano con nessuno dei punti precedentemente specificati.

È chiaro che potrei ottenere il risultato per differenza. Proviamo però a calcolarlo direttamente, così possiamo fare un'utile verifica. Dobbiamo contare il numero di modi in cui si possono avere cinque carte tutte diverse tra loro che non formano una *scala/scala reale* o un *colore*. Poichè ho 13 differenti valori per seme, ho $\binom{13}{5} = 1287$ modi di combinarli insieme senza ripetizioni; da questi, dovrò sottrarre 10 combinazioni corrispondenti alle possibili

scale o scale reali, ottenendo 1277 combinazioni. Ogni valore può essere di 4 semi differenti, per cui avrò $4^5 = 1024$ possibili combinazioni, da cui dovrò sottrarre le 4 combinazioni delle carte di dello stesso colore, che darebbero un *colore*, ottenendo 1020. In definitiva, il numero totale di *nessun punto* è

$$1277 \times 1020 = 1\,302\,540$$

La probabilità di *nessun punto* è allora

$$P(\textit{nessun punto}) = \frac{1\,302\,540}{2\,598\,960} \approx 0.50$$

Come verifica, notiamo che la somma delle probabilità calcolate dà effettivamente 1, in quanto si ha:

$$\frac{40 + 624 + 3744 + 5108 + 10\,200 + 54\,912 + 123\,552 + 1\,098\,240 + 1\,302\,540}{2\,598\,960} = \frac{2\,598\,960}{2\,598\,960} = 1$$

In Tab. B.2, abbiamo riassunto i risultati determinati in precedenza, calcolando le probabilità con tre cifre significative.

Punto	*Numero di combinazioni*	*Probabilità*
Scala reale	40	0.0000154
Poker	624	0.000240
Full	3744	0.00144
Colore	5108	0.00197
Scala	10 200	0.00392
Tris	54 912	0.0211
Doppia coppia	123 552	0.0475
Coppia	1 098 240	0.423
Nessun punto	1 302 540	0.501

Tab. B.2. Numero di combinazioni e probabilità dei punti del gioco del poker.

Appendice C

La funzione $\mathbb{G}(x)$

Questa appendice contiene la definizione e le principali proprietà della funzione $\mathbb{G}(x)$ (CDF di una variabile aleatoria gaussiana standard). Di particolare utilità negli esercizi la Tab. C.1, contenente i valori di $\mathbb{G}(x)$ per $0 \leq x \leq 3.29$.

C.1 La funzione $\mathbb{G}(x)$

La funzione $\mathbb{G}(x)$ è definita dall'integrale:

$$\mathbb{G}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{u^2}{2}}\, du\,.$$

Le principali proprietà della $\mathbb{G}(x)$ sono le seguenti:

1. $\mathbb{G}(-\infty) = 0$, $\mathbb{G}(+\infty) = 1$, $\mathbb{G}(0) = \frac{1}{2}$;
2. $\mathbb{G}(x)$ è una funzione monotona strettamente crescente;
3. $\mathbb{G}(-x) = 1 - \mathbb{G}(x)$;
4. per valori grandi di x, si ha
$$\mathbb{G}(x) \approx 1 - \frac{1}{x\sqrt{2\pi}}\, e^{-\frac{x^2}{2}}\,. \tag{C.1}$$

Inoltre la funzione $\mathbb{G}(x)$ può essere espressa in termini della *funzione di errore*:

$$\mathrm{erf}(x) \triangleq \frac{2}{\sqrt{\pi}} \int_{0}^{x} e^{-u^2}\, du\,.$$

Infatti si ha:

$$\mathbb{G}(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-\frac{u^2}{2}}\, du = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \int_{0}^{x} e^{-\frac{u^2}{2}}\, du\,,$$

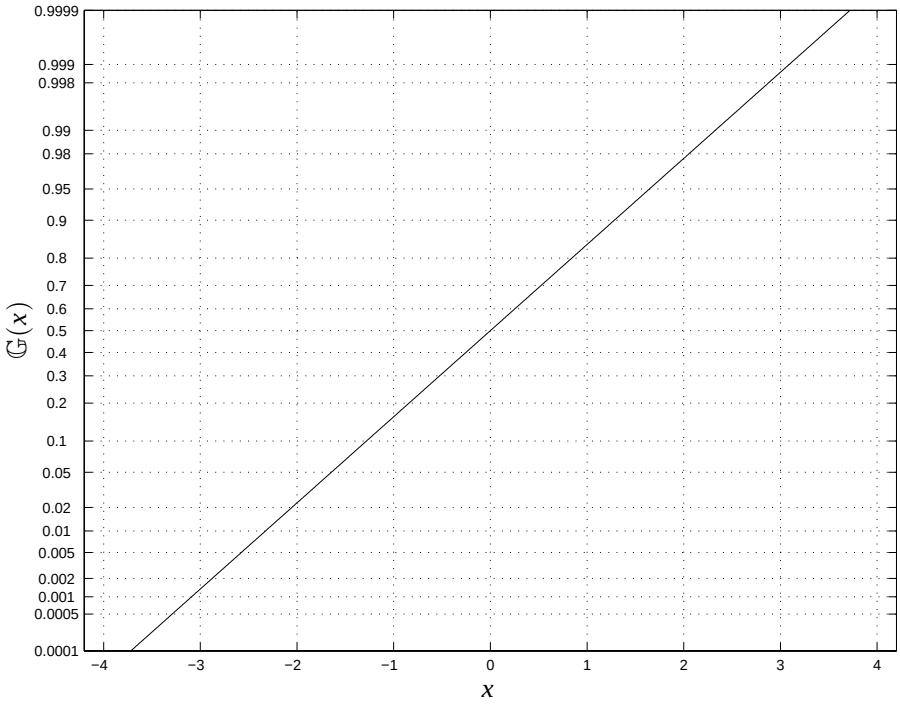


Fig. C.1. Grafico in scala gaussiana della funzione $\mathbb{G}(x)$.

e con il cambio di variabile $u/\sqrt{2} = v$ nell'integrale si ha:

$$\mathbb{G}(x) = \frac{1}{2} + \frac{1}{\sqrt{\pi}} \int_0^{\frac{x}{\sqrt{2}}} e^{-v^2}\, dv = \frac{1}{2}\left[1 + \mathrm{erf}\left(\frac{x}{\sqrt{2}}\right)\right] .$$

Tale espressione è conveniente quando si voglia implementare la funzione $\mathbb{G}(x)$ al calcolatore, in quanto quasi tutti i linguaggi di programmazione (Matlab tra essi) possiedono la $\mathrm{erf}(x)$ tra le funzioni di libreria.

Ad esempio, riportiamo di seguito una function Matlab per il calcolo della $\mathbb{G}(x)$, che può essere salvata nel file `G.m`.

```
function y = G(x);
%G  Funzione G.
%   G(X) Calcola la funzione G nel valore x.
y = (1/2) * (1+erf(x/sqrt(2)));
```

Ad esempio, per ottenere il grafico di Fig. 3.23, si possono utilizzare i comandi

```
>> x = [-4:0.01:4];
>> plot(x,G(x));
```

Se non si dispone di un calcolatore, un grafico in scala gaussiana (Fig. C.1) della $\mathbb{G}(x)$, nel quale la funzione appare come una retta, consente di determinare abbastanza precisamente i valori della funzione. Per una valutazione ancora più accurata, è tuttavia indispensabile utilizzare una tabella dei valori della $\mathbb{G}(x)$. In Tab. C.1, tratta da [1, pp. 176–177], sono riportati i valori di $\mathbb{G}(x)$ con quattro cifre decimali per $0 \leq x \leq 3.29$. Per valori di $x < 0$, si usi la relazione $\mathbb{G}(-x) = 1 - \mathbb{G}(x)$, per valori di $x > 3.29$ si usi l'approssimazione (C.1). La tabella va letta come

segue: sulle righe sono riportati i valori di x con passo 0.1, spostandosi poi lungo una riga si ottengono i valori con passo 0.01. Ad esempio, la terza colonna della terza riga corrisponde a $x = 0.22$.

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5159	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7518	0.7549
0.7	0.7580	0.7612	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8016	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8380
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8718	0.8729	0.8749	0.8770	0.8790	0.8810	0.8836
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9083	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9430	0.9441
1.6	0.9452	0.9463	0.9474	0.9485	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9509	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9758	0.9762	0.9767
2.0	0.9773	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9865	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9989	0.9980	0.9980	0.9981
2.9	0.9981	0.9982	0.9983	0.9984	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9986	0.9987	0.9987	0.9988	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995

Tab. C.1. Valori della funzione $\mathbb{G}(x)$ (adattata da [1, pp. 176–177]).

Appendice D

L'impulso di Dirac

In questa appendice viene introdotto, con approccio intuitivo, l'impulso di Dirac e vengono presentate le sue principali proprietà.

D.1 Impulso di Dirac

L'impulso di Dirac $\delta(x)$ non è una funzione ordinaria, ma una funzione *generalizzata* o, più precisamente, una *distribuzione*. Proviamo a darne una definizione formale, anche se un maggior rigore matematico richiederebbe l'uso della *teoria delle distribuzioni*:

Definizione (impulso di Dirac). Sia $\varphi(x)$ una qualsiasi funzione continua in $x = 0$. L'impulso di Dirac $\delta(x)$ è definito dalla seguente condizione:

$$\int_a^b \varphi(x)\,\delta(x)\,dx = \begin{cases} \varphi(0), & \text{se } 0 \in]a,b[\\ 0, & \text{se } 0 \notin [a,b] \end{cases} \tag{D.1}$$

Notiamo che l'impulso di Dirac "campiona" il valore della funzione $\varphi(x)$ nel punto 0. È chiaro che non esiste nessuna funzione *ordinaria* che ha questa proprietà; tuttavia, una buona approssimazione di $\delta(x)$ è una funzione "stretta" ed "alta" di area unitaria, ad esempio:

$$\delta_T(x) = \begin{cases} \frac{1}{T}, & |x| \le T/2\,; \\ 0, & |x| > T/2\,; \end{cases}$$

con $T \ll 1$ (Fig. D.1). Infatti, se l'intervallo $]-T/2, T/2[$ è contenuto in $]a,b[$, e se la funzione $\varphi(x)$ è lentamente variabile nell'intervallo $]-T/2, T/2[$, per cui si può porre $\varphi(x) \approx \varphi(0)$ per $|x| \le T/2$, si ha:

$$\int_a^b \varphi(x)\,\delta_T(x)\,dx = \frac{1}{T}\int_{-T/2}^{T/2} \varphi(x)\,dx \approx \varphi(0)\,.$$

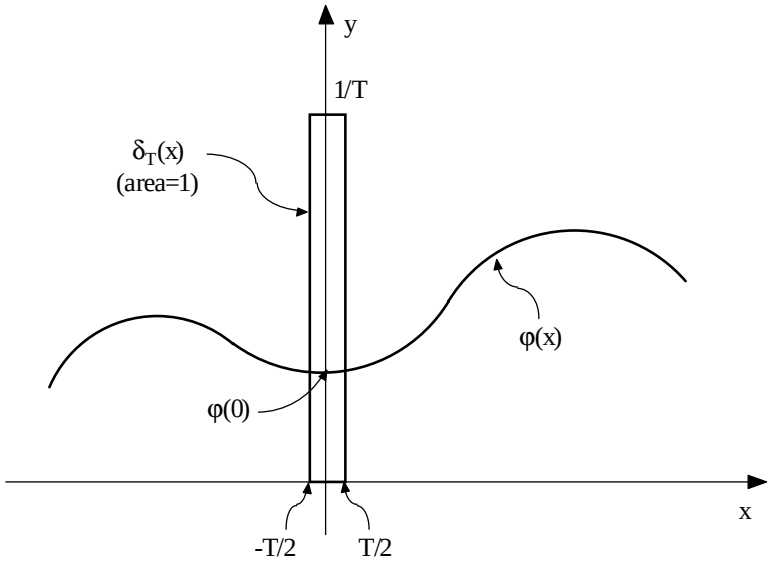


Fig. D.1. La funzione $\delta_T(x)$ rappresenta, al diminuire della durata T, un'approssimazione della delta di Dirac $\delta(x)$. Osserviamo che per T sufficientemente piccolo la funzione $\varphi(x) \approx \varphi(0)$ nell'intervallo $]-T/2, T/2[$.

In realtà, la precedente uguaglianza approssimata diventa esatta se si passa al limite per $T \to 0$:

$$\lim_{T\to 0} \int_a^b \varphi(x)\, \delta_T(x)\, dx = \lim_{T\to 0} \frac{1}{T} \int_{-T/2}^{T/2} \varphi(x)\, dx = \varphi(0)\,.$$

nell'ipotesi che $\varphi(x)$ sia continua in $x = 0$. Questo consente di interpretare l'impulso di Dirac come il limite di una famiglia di funzioni $\delta_T(x)$ con le seguenti proprietà:

- per $T \to 0$, le funzioni diventano sempre più "strette";
- per $T \to 0$, le funzioni diventano sempre più "alte";
- l'area di tali funzioni vale 1 indipendentemente da T.

Tale interpretazione dell'impulso di Dirac, tuttavia, per quanto intuitivamente chiara, non è matematicamente rigorosa. Infatti, con riferimento alla famiglia di funzioni $\delta_T(x)$ considerata precedentemente, basta osservare, che essa converge, in senso ordinario, alla funzione $\mu(x)$ identicamente nulla per ogni $x \neq 0$, mentre per $x = 0$ non converge affatto (diverge); è chiaro poi che la funzione $\mu(x)$ quasi ovunque nulla non soddisfa la definizione (D.1), poiché risulta

$$\int_a^b \varphi(x)\, \mu(x)\, dx = 0\,.$$

La conclusione è che, a stretto rigore matematico, non possiamo considerare $\delta(x)$ come il limite per $T \to 0$ della famiglia di funzioni $\delta_T(x)$; tuttavia tale interpretazione, anche se imprecisa, può rappresentare un valido aiuto all'intuizione.

L'impulso di Dirac gode delle seguenti proprietà, che si possono facilmente dimostrare sulla base della definizione (D.1):

1. Area unitaria: $\displaystyle\int_{-\infty}^{\infty} \delta(x)\, dx = 1$;

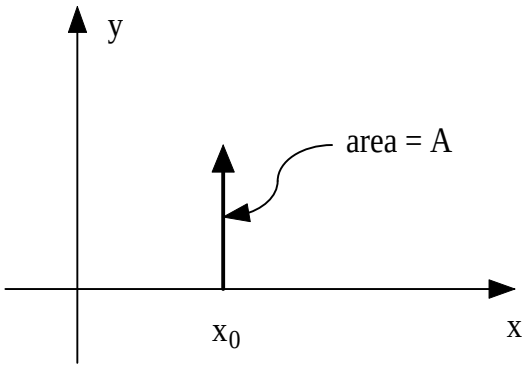


Fig. D.2. La rappresentazione grafica di un impulso di Dirac $A\,\delta(x-x_0)$ è una freccia centrata in x_0, la cui altezza è proporzionale all'area dell'impulso; si suole indicare il valore dell'area A a lato dell'impulso.

2. Campionamento o prodotto: $f(x)\,\delta(x) = f(0)\,\delta(x)$;
3. Traslazione: $f(x)\,\delta(x-x_0) = f(x_0)\,\delta(x-x_0)$;
4. Cambiamento di scala: $\delta(ax) = \dfrac{1}{|a|}\,\delta(x)$;
5. Derivazione: $\delta(x) = \dfrac{d}{dx}\,u(x)$;
6. Integrazione: $u(x) = \displaystyle\int_{-\infty}^{x} \delta(u)\,du$.

Sulla base delle proprietà 2 e 3, è possibile considerare il caso più generale di un impulso $A\,\delta(x-x_0)$, che rappresenta un impulso di Dirac *centrato* in x_0 e di *area* pari ad A (Fig. D.2).

Una importante conseguenza della proprietà 5 è la proprietà di derivazione di una funzione discontinua: se la funzione $f(x)$ presenta una discontinuità di prima specie del punto x_0, la sua derivata *generalizzata* presenterà un impulso di Dirac nel punto x_0 di area pari al valore del salto di discontinuità $f(x_0^+) - f(x_0^-)$ nel punto in questione; ovvero, detta $h(x)$ la derivata convenzionale, si avrà:

$$\frac{d}{dx} f(x) = h(x) + [f(x_0^+) - f(x_0^-)]\,\delta(x-x_0)\,.$$

Appendice E

Richiami di algebra lineare

In questa appendice sono richiamate le principali definizioni e proprietà delle matrici e dei vettori, con riferimento principalmente a quelle utilizzate nel testo. La trattazione non ha alcuna pretesa di completezza nè di originalità; per una trattazione approfondita si rimanda il lettore interessato ai testi specifici, quali [13] e [14].

E.1 Definizioni ed operazioni fondamentali

E.1.1 Matrici e vettori

Una *matrice* $\boldsymbol{A}$, di dimensioni $m \times n$, è una griglia (array) rettangolare di numeri reali[1] disposti su m righe ed n colonne:

$$\boldsymbol{A} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \cdots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{pmatrix}$$

che può essere indicata in forma abbreviata anche con $\boldsymbol{A} = \{a_{ij}\}, i = 1, 2 \ldots, m; j = 1, 2 \ldots n$. Per denotare le matrici, useremo simboli maiuscoli in grassetto. Una matrice con ugual numero di righe e colonne ($m = n$) prende il nome di *matrice quadrata* di ordine m. Una matrice quadrata di particolare importanza è la *matrice identità*:

$$\boldsymbol{I} = \begin{pmatrix} 1 & 0 & 0 & \ldots & 0 \\ 0 & 1 & 0 & \ldots & 0 \\ 0 & 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & & \vdots \\ 0 & 0 & 0 & \ldots & 1 \end{pmatrix}$$

[1] Per semplicità tratteremo solo il caso di matrici e vettori reali, sebbene gran parte delle definizioni e delle proprietà si possano estendere al caso complesso senza grosse difficoltà.

Un *vettore colonna* ad m elementi è un caso particolare di matrice $m \times 1$, avente cioè una sola colonna (ed m righe):

$$\boldsymbol{a} = \begin{pmatrix} a_1 \\ a_2 \\ \vdots \\ a_m \end{pmatrix}$$

mentre un *vettore riga* ad n elementi è un caso particolare di matrice $1 \times n$, avente cioè una sola riga (ed n colonne):

$$\boldsymbol{a} = (a_1, a_2, \ldots, a_n)$$

Per denotare vettori riga e colonna utilizzeremo usualmente simboli minuscoli in grassetto, come $\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{c}$. Nel testo si utilizzano esclusivamente *vettori colonna*; per denotare un vettore riga si usa la notazione $\boldsymbol{a}^T$, dove l'apice T indica l'operazione di *trasposizione* di una matrice (vedi dopo). Notiamo che un vettore colonna si può interpretare anche come un punto dello spazio $\mathbb{R}^m$, per cui scriveremo anche $\boldsymbol{a} \in \mathbb{R}^m$.

E.1.2 Somma di due matrici e prodotto per uno scalare

La somma tra due matrici $\boldsymbol{A}$ e $\boldsymbol{B}$ si può effettuare solo se le due matrici hanno lo stesso numero di righe e di colonne, e si effettua sommando tra loro gli elementi di posto corrispondente:

$$\boldsymbol{C} = \boldsymbol{A} + \boldsymbol{B} \quad \Longleftrightarrow \quad c_{ij} = a_{ij} + b_{ij} \qquad i = 1, 2, \ldots, m; \quad j = 1, 2, \ldots, n$$

Il prodotto di una matrice $\boldsymbol{A}$ per uno scalare reale μ si ottiene moltiplicando ciascun elemento della matrice per lo scalare

$$\boldsymbol{B} = \mu \boldsymbol{A} \quad \Longleftrightarrow \quad b_{ij} = \mu\, a_{ij} \qquad i = 1, 2, \ldots, m; \quad j = 1, 2, \ldots, n$$

Le operazioni di somma e prodotto per uno scalare godono delle proprietà commutativa, associativa e distributiva.

E.1.3 Prodotto di due matrici (righe per colonne)

Il prodotto *righe per colonne* di due matrici $\boldsymbol{A}$ e $\boldsymbol{B}$, di dimensioni $m \times r$ e $r \times n$ si può effettuare se e solo se il numero di colonne della prima matrice è uguale al numero di righe della seconda, ed è una matrice $\boldsymbol{C}$ di dimensioni $m \times n$, data da:

$$\boldsymbol{C} = \boldsymbol{A}\,\boldsymbol{B} \quad \Longleftrightarrow \quad c_{ij} = \sum_{k=1}^{r} a_{ir}\, b_{rj} \qquad i = 1, 2, \ldots, m; \quad j = 1, 2, \ldots, n$$

Il prodotto di due matrici gode della proprietà associativa e distributiva, ma *non* della proprietà commutativa; in altri termini, il prodotto $\boldsymbol{B}\,\boldsymbol{A}$ può essere privo di significato, e anche quando è possibile effettuarlo non restituisce lo stesso risultato di $\boldsymbol{A}\,\boldsymbol{B}$.

E.1.4 Trasposizione

La *trasposta* della matrice $\boldsymbol{A}$ di dimensioni $m \times n$ è la matrice $\boldsymbol{A}^T = \{a_{ji}\}$ di dimensioni $n \times m$ ottenuta scambiando gli indici di riga con gli indici di colonna. L'operazione di trasposta gode

delle seguenti proprietà (si suppone che le somme e moltiplicazioni abbiamo senso):

$$(\boldsymbol{A}^T)^T = \boldsymbol{A}$$
$$(\boldsymbol{A} + \boldsymbol{B})^T = \boldsymbol{A}^T + \boldsymbol{B}^T$$
$$(\boldsymbol{A}\,\boldsymbol{B})^T = \boldsymbol{B}^T\,\boldsymbol{A}^T$$

E.2 Operazioni e proprietà delle matrici quadrate

In questo paragrafo supporremo che tutte le matrici siano quadrate di ordine $m = n$.

E.2.1 Determinante

Il *determinante* di una matrice quadrata $\boldsymbol{A}$ è:

$$\det(\boldsymbol{A}) \triangleq \sum_{\text{permutazioni}} (-1)^{f(j_1,j_2,\ldots,j_n)} \prod_{i=1}^{n} a_{ij_i}$$

dove la somma è effettuata su tutte le $m!$ distinte permutazioni $\{j_1, j_2, \ldots, j_n\}$ dell'insieme degli interi $\{1, 2, \ldots, n\}$, e $f(j_1, j_2, \ldots, j_p)$ è il numero di *trasposizioni* richiesto per trasformare la n-pla $(1, 2, \ldots, n)$ in $(j_1, j_2, \ldots, j_n)$.[2]

Si può dare una definizione *ricorsiva* di determinante, utile per il calcolo (espansione di Laplace). Definita con $\boldsymbol{A}_{ij}$ la *sottomatrice* (quadrata) ottenuta da $\boldsymbol{A}$ eliminando la riga i-esima e la colonna j-esima, si ha:

$$\det(\boldsymbol{A}) = \sum_{i=1}^{n} a_{ij}\,(-1)^{i+j}\det(\boldsymbol{A}_{ij}) = \sum_{j=1}^{n} a_{ij}\,(-1)^{i+j}\det(\boldsymbol{A}_{ij})$$

nella quale il determinante di $\boldsymbol{A}$ è espresso ricorsivamente in funzione dei determinanti $\det(\boldsymbol{A}_{ij})$, denominati *minori* degli elementi a_{ij}. Per inizializzare la ricorsione, basta definire il determinante di una matrice 1×1 come l'unico elemento della matrice. Notiamo che lo sviluppo del determinante si può effettuare su una riga o colonna a scelta, per cui in genere per semplificare il calcolo si sceglie una riga od una colonna contenente numerosi zeri.

Valgono lo seguenti proprietà del determinante:

$$\det(\boldsymbol{A}^T) = \det(\boldsymbol{A})$$
$$\det(\boldsymbol{A}\,\boldsymbol{B}) = \det(\boldsymbol{A})\det(\boldsymbol{B})$$

E.2.2 Inversa

Una matrice quadrata si dice *non singolare* se e solo se $\det(\boldsymbol{A}) \neq 0$. Una matrice non singolare è dotata di *inversa* $\boldsymbol{A}^{-1}$, che è l'unica matrice tale che $\boldsymbol{A}\,\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\,\boldsymbol{A} = \boldsymbol{I}$.

Per determinare l'espressione analitica esplicita dell'inversa, con A_{ij} il *complemento algebrico* dell'elemento a_{ij}, pari a $(-1)^{i+j}\det(\boldsymbol{A}_{ij})$ Se denotiamo con b_{ij} l'elemento di posto (i, j) di $\boldsymbol{A}^{-1}$, si ha:

$$a^{ij} = \frac{A_{ji}}{\det(\boldsymbol{A})} \tag{E.1}$$

[2] Una trasposizione consiste nello scambiare due numeri; si può mostrare che, sebbene si possa trasformare $(1, 2, \ldots, n)$ in $(j_1, j_2, \ldots, j_n)$ mediante trasposizioni in più modi differenti, il numero di trasposizioni richiesto è sempre pari o sempre dispari, cosicché la quantità $(-1)^{f(j_1,j_2,\ldots,j_n)}$ è definita in maniera non ambigua.

ovvero l'inversa di $\boldsymbol{A}$ si calcola costruendo la matrice dei complementi algebrici degli elementi di $\boldsymbol{A}$, effettuando la trasposizione (si noti lo scambio degli indici nelle (E.1)), e dividendo per il determinante di $\boldsymbol{A}$.

► *Esempio* E.1 *(inversa di una matrice* 2×2*).* Il calcolo della matrice inversa è particolamente semplice per una matrice 2×2:

$$\boldsymbol{A} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

Si ha, adoperando la (E.1),

$$\boldsymbol{A}^{-1} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

Si verifica facilmente mediante moltiplicazione diretta che $\boldsymbol{A}\,\boldsymbol{A}^{-1} = \boldsymbol{A}^{-1}\,\boldsymbol{A} = \boldsymbol{I}$. ◄

L'operazione di inversa gode delle seguenti proprietà:

$$\begin{aligned} (\boldsymbol{A}\,\boldsymbol{B})^{-1} &= \boldsymbol{B}^{-1}\,\boldsymbol{A}^{-1} \\ (\boldsymbol{A}^T)^{-1} &= (\boldsymbol{A}^{-1})^T \end{aligned} \tag{E.2}$$

Inoltre, per le proprietà del determinante, si ha:

$$\det(\boldsymbol{A}^{-1}) = \frac{1}{\det \boldsymbol{A}}$$

E.2.3 Matrici diagonali

Una matrice quadrata si dice *diagonale* se $a_{ij} = 0, \forall i \neq j$, e si indica:

$$\boldsymbol{A} = \mathrm{diag}(a_{11}, a_{22}, \ldots, a_{mm})$$

o più sinteticamente con $\boldsymbol{A} = \mathrm{diag}(a_{ii})$. La matrice identica $\boldsymbol{I}$ è un esempio di matrice diagonale.

Le matrici diagonali godono di particolari semplificazioni nel calcolo. Si ha, in particolare:

$$\begin{aligned} \mathrm{diag}(a_{ii})\,\mathrm{diag}(b_{ii}) &= \mathrm{diag}(a_{ii}\,b_{ii}) \\ \det[\mathrm{diag}(a_{ii})] &= \prod_{i=1}^{m} a_{ii} \\ [\mathrm{diag}(a_{ii})]^{-1} &= \mathrm{diag}\left(\frac{1}{a_{ii}}\right) \end{aligned}$$

L'ultima proprietà vale se la matrice diagonale è non singolare, il che accade se e solo se $a_{ii} \neq 0, \forall i$.

E.2.4 Matrici simmetriche e forme quadratiche

Una matrice quadrata si dice *simmetrica* se $\boldsymbol{A} = \boldsymbol{A}^T$, ovvero se $a_{ij} = a_{ji}$. Le matrici diagonali sono esempi di matrici simmetriche.

Data una matrice simmetrica $\boldsymbol{A}$ di ordine m, ed un vettore colonna $\boldsymbol{x} = [x_1, x_2, \ldots, x_m]^T$, la *forma quadratica* associata ad $\boldsymbol{A}$ è un polinomio omogeneo di secondo grado nelle variabili $x_1, x_2, \ldots, x_n$:

$$\boldsymbol{x}^T\,\boldsymbol{A}\,\boldsymbol{x} = \sum_{i,j=1}^{m} a_{ij}\,x_i\,x_j$$

La matrice $\boldsymbol{A}$ (e la forma quadratica ad essa associata) si dice *semidefinita positiva* se

$$\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} \geq 0 \qquad \forall \boldsymbol{x} \in \mathbb{R}^m$$

Se $\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x} > 0$ per ogni $\boldsymbol{x} \in \mathbb{R}^m - \{\boldsymbol{0}\}$, la matrice $\boldsymbol{A}$ si dice *definita positiva.* Talvolta per indicare sinteticamente che una matrice è semidefinita [definita] positiva si scrive $\boldsymbol{A} \geq 0$ [$\boldsymbol{A} > 0$].

Si dimostra che una matrice $\boldsymbol{A}$ semidefinita positiva presenta $\det(\boldsymbol{A}) \geq 0$. Inoltre, una matrice $\boldsymbol{A}$ semidefinita positiva è definita positiva se e solo se è non singolare, vale a dire $\det(\boldsymbol{A}) \neq 0$; ne segue che una forma quadratica definita positiva presenta $\det(\boldsymbol{A}) > 0$.

Appendice F

Identità matematiche notevoli

In questa appendice sono raccolte alcune identità matematiche che possono risultare utili per le derivazioni analitiche e per la risoluzione degli esercizi.

F.1 Sommatorie e serie

F.1.1 Sommatorie di potenze di interi

Le seguenti formule riguardano somme finite ($N \in \mathbb{N}$) di potenze di numeri interi e possono essere dimostrate per induzione:

$$\sum_{n=1}^{N} n = \frac{N(N+1)}{2}$$
$$\sum_{n=1}^{N} n^2 = \frac{N(N+1)(2N+1)}{6}$$
$$\sum_{n=1}^{N} n^3 = \left[\frac{N(N+1)}{2}\right]^2$$
$$\sum_{n=1}^{N} n^4 = \frac{N(N+1)(2N+1)(3N^2+3N-1)}{30}$$

F.1.2 Somma dei primi n termini di una serie geometrica

La seguente formula riguarda la somma dei primi n termini di una serie geometrica:

$$\sum_{i=0}^{n-1} z^i = \frac{1-z^n}{1-z}$$

con $z \in \mathbb{C}$ numero complesso qualsiasi.

F.1.3 Serie geometrica

Dalla relazione precedente, se $|z| < 1$, passando al limite per $n \to \infty$ si ottiene la formula per la somma di una serie geometrica:

$$\sum_{i=0}^{\infty} z^i = \frac{1}{1-z}\,.$$

F.2 Formula di Leibnitz

La formula di Leibnitz serve a derivare le funzioni definite mediante un integrale e dipendenti da un parametro. Sia

$$F(x) = \int_{\alpha(x)}^{\beta(x)} f(x,y)\,dy$$

una funzione di x definita mediante integrale. Si ha:

$$F'(x) = \int_{\alpha(x)}^{\beta(x)} \frac{\partial}{\partial x} f(x,y)\,dy + f[x,\beta(x)]\,\beta'(x) - f[x,\alpha(x)]\,\alpha'(x)$$

Per le condizioni di validità della formula si veda un qualunque testo di analisi.

Bibliografia

Probabilità (trattazione elementare)

[1] W. Feller *An Introduction to Probability Theory and Its Applications. Volume I.* John Wiley & Sons, 1950.

[2] B. V. Gnedenko, *Teoria della probabilità.* Editori Riuniti, 1979.

[3] A. Papoulis. *Probability, Random Variables, and Stochastic Processes. Third edition*. McGraw Hill International Editions, 1991.

[4] D. Stirzaker. *Elementary Probability.* Cambridge University Press, Cambridge, UK, 1994.

Probabilità (trattazione avanzata)

[5] G. Casella and R. L. Berger. *Statistical Inference*. Duxbury Press, Belmont, California, USA, 1990.

[6] W. Feller, *An Introduction to Probability Theory and Its Applications. Volume II.* John Wiley & Sons, 1966.

Teoria della misura

[7] H.L. Royden, *Real Analysis.* McMillan Publ. Co., seconda edizione, 1968.

Generazione di numeri casuali

[8] D. E. Knuth. *The Art of Computer Programming. Volume 2: Seminumerical Algorithms*. Addison-Wesley, Reading, Massachusetts, USA, 1971.

[9] S. K. Park e K. W. Miller, "Random number generators: Good ones are hard to find," *Communications of the ACM*, vol. 31, n. 10, pp. 1192-1201, 1988.

[10] B. D. Ripley. *Stochastic Simulation.* John Wiley & Sons, New York, 1987.

[11] R. Y. Rubinstein. *Simulation and the Monte Carlo Method.* John Wiley & Sons, New York, 1981.

Teoria dell'informazione

[12] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* John Wiley & Sons, New York, 1991.

Algebra lineare

[13] F. R. Gantmacher. *The Theory of Matrices.* Chelsea Publishing Company, New York, 1959.

[14] R. A. Horn and C. R. Johnson, *Matrix Analysis,* Cambridge University Press, Cambridge, 1990.

Indice analitico