$\mathcal{CALCOLO}$ $\mathcal{DELLE}$ $\mathcal{PROBABILITA'}$ — $\mathcal{PARTE}$ $\mathcal{PRIMA}$

1. SPAZI DI PROBABILITA' e VARIABILI ALEATORIE I.

Vi sono più modi di definire la probabilità. Si può innanzitutto adottare un punto di vista assiomatico (Kolmogorov), ricorrendo in modo essenziale alla teoria della misura. Oppure si può far appello al rapporto tra il numero di casi favorevoli ed il numero di alternative (classico), o ancora alle frequenze relative (Von Mises). Infine si può considerare la probabilità come il grado soggettivo di fiducia nel verificarsi di un dato evento (de Finetti). Non ci addentreremo qui in una discussione dell'efficacia o della debolezza dei diversi punti di vista, facendo fin da subito una precisa scelta di campo: in queste note discuteremo alcuni aspetti di una teoria deduttiva della probabilità basata su una procedura assiomatica. In tale prospettiva, il concetto primario è quello di *spazio degli eventi elementari* Ω. I punti di Ω, indicati con il simbolo ω, sono detti eventi elementari e possono essere pensati come *risultati di un esperimento*.

Ad esempio tale esperimento può essere costituito da una successione di lanci di un dado o di una moneta, o dal conteggio del numero delle chiamate passanti per una data centrale telefonica in un dato intervallo si tempo, o da quello del numero di particelle di un dato gas presenti in un volume assegnato.

I sottoinsiemi $C \subseteq \Omega$ sono detti *eventi aleatori* o semplicemente *eventi*. L'evento $C = \Omega$ è detto *evento certo*, mentre $C = \emptyset$ è detto *evento impossibile*. Si possono formare unioni, intersezioni, complementi e differenze di tali eventi. In particolare, dato $C \subseteq \Omega$, l'evento $\overline{C} = \Omega \setminus C$, cioè l'evento in cui C non si verifica, si dice *evento complementare* a C. La collezione di tutti gli eventi C forma un'algebra che denoteremo con il simbolo $\mathcal{F}$.

(1.1) Esercizio. Mostrare che se $|\Omega| = N < \infty$, allora $\mathcal{F}$ consiste in 2^N eventi distinti.

La locuzione 'evento aleatorio' va intesa nel senso che il suo verificarsi nel corso di un esperimento, ovvero nel complesso delle circostanze che lo caratterizzano, non è in generale un dato certo. Similmente potremmo dire che è influenzato dal 'caso', quale che sia il significato preciso che vogliamo attribuire a tale nozione.

In particolare, dati due eventi C_1 e C_2, l'evento $C = C_1 \cup C_2$ sarà quello in cui si verifica almeno uno dei due, cioè l'evento C_1 o l'evento C_2, mentre $C = C_1 \cap C_2$ sarà quello in cui gli eventi C_1 e C_2 si verificano entrambi. Diremo a tal proposito che C_1 e C_2 sono *mutuamente esclusivi* se $C_1 \cap C_2 = \emptyset$. Diremo infine che una collezione $C_1, \ldots, C_r$ determina una *partizione* di Ω, o anche un *sistema completo di eventi* in Ω, se $C_i \cap C_j = \emptyset$ per $i \neq j$ e $\cup_{j=1}^{r} C_j = \Omega$.

(1.2) Esempio. Nel lancio di un dado con sei facce lo spazio Ω è costituito dai sei punti $\omega_i \equiv i$ $(i = 1, \ldots, 6)$, corrispondenti ai numeri mostrati dalla faccia superiore del dado dopo che è stato lanciato. $\mathcal{F}$ consiste in $2^6 = 64$ eventi. Ad esempio $C_1 = \{\text{pari}\}$, $C_2 = \{\text{dispari}\}$ o $C_3 = \{\text{minore di tre}\}$ sono eventi. Così $C_1 \cup C_3$ sarà l'evento $\{1 \,\text{o}\, 2 \,\text{o}\, 4 \,\text{o}\, 6\}$ mentre $C_1 \cap C_3$ sarà l'evento $\{2\}$. Evidentemente $C_1 \cap C_2 = \emptyset$ e $C_1 \cup C_2 = \Omega$, dunque C_1 e C_2 formano una partizione di Ω.

Nello sviluppo assiomatico della probabilità si parte dall'assunzione che ad ogni evento aleatorio C in un esperimento (detto altrimenti: ad ogni elemento C di un'algebra di eventi aleatori) sia assegnato un valore numerico $\mathrm{P}(C)$: la *probabilità* dell'evento in questione. Tale grandezza, come vedremo, costituisce una nozione matematica precisa. D'altra parte la sua relazione con i fenomeni reali è assai meno evidente e non può essere stabilita deduttivamente. Una possibile chiave per l'interpretazione empirica della nozione di probabilità è la seguente: immaginiamo l'esecuzione di un certo esperimento con un dato insieme Ω di risultati possibili, esecuzione che supponiamo 'ripetibile un numero arbitrario di volte'. Consideriamo allora un evento C che possa o meno verificarsi in base ai possibili risultati dell'esperimento. Se ad esempio l'esperimento consiste nel lancio di un dado, l'evento in questione può consistere nel fatto che la faccia superiore risultante dopo il lancio mostri un numero pari (l'evento C_1 in (1.2)). Affermare che un modello matematico del suddetto fenomeno è fornito dall'insieme Ω e da una funzione P che all'evento C assegna la probabilità $\mathrm{P}(C)$ consiste nel sostenere, sulla base di osservazioni empiriche ottenute ad esempio ripetendo molte volte l'esperimento, che se eseguiamo n *prove* dell'esperimento, cioè se riproduciamo n volte le condizioni in cui l'evento C può verificarsi, allora il numero di volte in cui C effettivamente si verifica sarà, quando n è molto grande, approssimativamente uguale a $n \cdot \mathrm{P}(C)$, a parte casi particolarmente sfortunati, che però 'agli effetti pratici' possono essere trascurati.

Ciò riconduce l'interpretazione empirica della probabilità alla *frequenza relativa* f_C di un evento C, ossia al rapporto k/n tra il numero di volte k in cui l'evento C effettivamente si verifica ed il numero n di prove. Vediamo preliminarmente alcune proprietà della frequenza relativa. Evidentemente si ha $0 \leq f_C \leq 1$. Inoltre la frequenza relativa dell'evento certo sarà uguale a uno, mentre quella dell'evento impossibile sarà sempre nulla. Inoltre, se C_1 e C_2 sono mutuamente esclusivi (cioè $C_1 \cap C_2 = \emptyset$) e se nelle n ripetizioni dell'esperimento C_1 si verifica k_1 volte e C_2 k_2 volte, allora l'evento $C_1 \cup C_2$ si verificherà esattamente $k_1 + k_2$ volte, e quindi $f_{C_1 \cup C_2} = f_{C_1} + f_{C_2}$. Evidentemente la frequenza relativa f_C di un evento è una quantità che 'dipende dal caso' e può fluttuare (in particolare quando il numero n di prove è finito, ma negli esperimenti reali n è sempre finito!). D'altra parte, la probabilità di un evento è una quantità fissata, che ovviamente dipende dal tipo di esperimento, ma non dalla particolare successione in cui tale esperimento viene ripetuto. Inoltre, nel senso visto sopra, dovrà costituire una 'predizione teorica' per la frequenza relativa. Assumeremo quindi le seguenti proprietà:

I. $\mathrm{P}(C) \geq 0$ per ogni $C \in \mathcal{F}$;

II. $\mathrm{P}(\Omega) = 1$;

III. se $C_1, C_2 \in \mathcal{F}$ e $C_1 \cap C_2 = \emptyset$ allora $\mathrm{P}(C_1 \cup C_2) = \mathrm{P}(C_1) + \mathrm{P}(C_2)$.

(1.3) Esercizio. Le seguenti proprietà sono conseguenze degli assiomi I, II, III:

i) se $C_1 \subseteq C_2$, allora $\mathrm{P}(C_2) \geq \mathrm{P}(C_1)$ e $\mathrm{P}(C_2 \setminus C_1) = \mathrm{P}(C_2) - \mathrm{P}(C_1)$;

ii) per ogni collezione di eventi $C_1, \ldots, C_s$ due a due mutuamente esclusivi si ha $\mathrm{P}(\cup_{i=1}^{s} C_i) = \sum_{i=1}^{s} \mathrm{P}(C_i)$; se, in particolare, tale collezione forma una partizione di Ω allora $\sum_{i=1}^{s} \mathrm{P}(C_i) = 1$;

iii) dati due eventi arbitrari C_1 e C_2 si ha $\mathrm{P}(C_1 \cup C_2) = \mathrm{P}(C_1) + \mathrm{P}(C_2) - \mathrm{P}(C_1 \cap C_2)$;

iv) più in generale, per un arbitraria collezione di eventi $C_1, \ldots, C_s$ vale la *formula di inclusione-esclusione*

$$\begin{aligned}\mathrm{P}(\cup_{i=1}^{s} C_i) = &\sum_{i=1}^{s} \mathrm{P}(C_i) - \sum_{i<j} \mathrm{P}(C_i \cap C_j) \\ &+ \sum_{i<j<k} \mathrm{P}(C_i \cap C_j \cap C_k) + \cdots + (-1)^{s-1} \mathrm{P}(C_1 \cap \cdots \cap C_s),\end{aligned}$$

in particolare si ha $\mathrm{P}(\cup_{i=1}^{s} C_i) \leq \sum_{i=1}^{s} \mathrm{P}(C_i)$;

L'assunzione III (e la sua generalizzazione (ii)) afferma che la funzione P è *additiva*. Tale proprietà è sufficiente per trattare problemi che coinvolgono un numero finito di eventi possibili. Tuttavia, per trattare problemi più generali e poter utilizzare efficacemente la teoria della misura sarà conveniente assumere che, non soltanto l'unione finita, ma anche l'unione numerabile di eventi, cioè di elementi dell'algebra $\mathcal{F}$, faccia parte dell'algebra stessa, ossia che $\mathcal{F}$ sia una σ-algebra. La corrispondente ulteriore assunzione sulla funzione P sarà allora quella di essere *σ-additiva*, cioè avremo

III'. se $C_i \in \mathcal{F}$, $i = 1, 2, \ldots$ e $C_i \cap C_j = \emptyset$, $i \neq j$, allora $\mathrm{P}(\cup_{i=1}^{\infty} C_i) = \sum_{i=1}^{\infty} \mathrm{P}(C_i)$.

Vediamo ora un criterio generale per garantire la III'. Premettiamo una definizione: si chiama *sequenza monotòna di eventi* una collezione $\{C_k\}$, $C_k \in \mathcal{F}$, tale che $C_k \subseteq C_{k+1}$ (e in questo caso $C_k \uparrow C = \cup_k C_k$) oppure $C_k \supseteq C_{k+1}$ (e in questo caso $C_k \downarrow C = \cap_k C_k$).

(1.4) Teorema. Una funzione P su $\mathcal{F}$ che soddisfa le proprietà I, II e III sopra elencate soddisfa anche la III' se e solo se per ogni sequenza monotòna di eventi $C_k \in \mathcal{F}$ si ha

$$\mathrm{P}\left(\lim_{k\to\infty} C_k\right) = \lim_{k\to\infty} \mathrm{P}(C_k).$$

Dimostrazione. Consideriamo, per fissare le idee, il caso $C_k \subseteq C_{k+1}$. Poniamo $C_0 = \emptyset$ e $B_i = C_i \setminus C_{i-1}$, $i \geq 1$, cosicchè $B_i \cap B_j = \emptyset$ per $i \neq j$ e $C_k = \cup_{i \leq k} B_i$.

Inoltre, evidentemente, $\cup_k C_k = \cup_i B_i$. Allora, se P soddisfa la III' si ha

$$\mathrm{P}\left(\lim_{k\to\infty} C_k\right) = \mathrm{P}\,(\cup_k C_k) = \mathrm{P}\,(\cup_i B_i) = \sum_{i=1}^{\infty} \mathrm{P}\,(B_i)$$
$$= \lim_{k\to\infty} \sum_{i=1}^{k} \mathrm{P}\,(B_i) = \lim_{k\to\infty} \mathrm{P}\,(\cup_{i\le k} B_i) = \lim_{k\to\infty} \mathrm{P}\,(C_k).$$

Inversamente, l'additività finita implica che $\mathrm{P}\,(\cup_i B_i) = \sum_{i=1}^{n} \mathrm{P}\,(B_i) + \mathrm{P}\,(\cup_k C_k) - \mathrm{P}\,(C_n)$. Per la condizione del teorema prendendo $n\to\infty$ si ha $\mathrm{P}\,(C_n) \to \mathrm{P}\,(\cup_k C_k)$ e dunque $\mathrm{P}\,(\cup_i B_i) = \sum_{i=1}^{\infty} \mathrm{P}\,(B_i)$. q.e.d.

(1.5) Definizione. Con le assunzioni I, II, III' la tripletta $(\Omega, \mathcal{F}, \mathrm{P})$ è detta *spazio di probabilità* (nel senso di Kolmogorov).

Un caso importante è quello in cui Ω è discreto, ossia è formato da una collezione (al più) numerabile di punti $\omega_1, \omega_2, \ldots, \omega_i, \ldots$. Indichiamo con $\{\omega_i\}$ l'evento consistente nel solo punto ω_i (evento elementare). Poniamo inoltre $\mathrm{P}\,(\{\omega_i\}) = p(\omega_i) \equiv p_i$ $(i = 1, 2, \ldots)$. Dalle proprietà I e II della probabilità segue che

$$p_i \ge 0 \qquad \text{e} \qquad \sum_{i\ge 1} p_i = 1. \tag{1.6}$$

Un vettore $p = (p_1, p_2, \ldots)$ le cui componenti p_i soddisfano le proprietà (1.5) si chiama *distribuzione di probabilità* o semplicemente *distribuzione*. Tale espressione trova l'interpretazione che la probabilità 1 dell'evento certo è 'distribuita' tra gli eventi elementari ω_i, in analogia con la distribuzione della massa totale di un sistema meccanico tra i suoi punti costituenti. Abbiamo dunque visto che ogni probabilità P genera una distribuzione p su Ω. Viceversa, ogni distribuzione p su uno spazio discreto Ω genera una misura di probabilità P sulla σ-algebra $\mathcal{F}$ dei sottoinsiemi di Ω attraverso la formula

$$\mathrm{P}\,(C) = \sum_{\omega_i \in C} p(\omega_i). \tag{1.7}$$

In effetti, le proprietà I e II seguono da (1.6), mentre la σ-additività (III') segue dal fatto che ogni serie convergente di termini non-negativi può essere riarrangiata in qualunque ordine e quindi

$$\mathrm{P}\,(\cup_l C_l) = \sum_{\omega_i \in \cup_l C_l} p(\omega_i) = \sum_l \sum_{\omega_i \in C_l} p(\omega_i) = \sum_l \mathrm{P}\,(C_l).$$

(1.8) Schema finito. L'esempio più semplice immaginabile di spazio di probabilità è dato da uno *schema finito*, ottenuto assegnando una distribuzione $p = (p_1, \ldots, p_N)$ ad un insieme finito $X = \{x_1, \ldots, x_N\}$. In questo caso si ha $\Omega = X$, $\mathcal{F}$ ha 2^N elementi e P è determinata dalla (1.7). In quanto segue adotteremo talvolta la locuzione *distribuzione* per indicare la coppia (X, p) (e non soltanto il vettore p). Se $N = 2$, cioè se gli eventi elementari si riducono a due alternative: 'testa' o 'croce', 'vincita' o 'perdita' etc., allora (X, p) si chiama distribuzione, o schema, di Bernoulli. Infine, un caso che si incontra spesso nelle applicazioni è quello in cui $p_i = 1/N$, si dice allora che (X, p) è una *distribuzione uniforme*.

(1.9) Esercizio. Supponendo che il dado in (1.2) sia perfetto, otteniamo uno schema finito con $X = \{1, 2, \ldots, 6\}$ e $p = (\frac{1}{6}, \ldots, \frac{1}{6})$. Verificare, usando la (1.6), che $\mathrm{P}(C_1) = \mathrm{P}(C_2) = \frac{1}{2}$, $\mathrm{P}(C_3) = \frac{1}{3}$, $\mathrm{P}(C_1 \cup C_3) = \frac{2}{3}$, $\mathrm{P}(C_1 \cap C_3) = \frac{1}{6}$ e $\mathrm{P}(C_1 \cap C_2) = 0$.

Una semplice generalizzazione di uno schema finito (X, p), di estrema utilità nelle applicazioni, si ottiene considerando lo spazio $\Omega = \Omega_n = X^{[1,n]}$, cioè lo spazio delle parole di lunghezza n della forma $\omega = (\omega_1 \omega_2 \ldots \omega_n)$ con $\omega_i \in X = \{x_1, \ldots, x_N\}$. Possiamo ad esempio pensare ad X come ad un *alfabeto* di N lettere e ad ω come ad una parola di lungezza n scritta in tale alfabeto. Su Ω_n assegnamo la distribuzione di probabilità $p(\omega) = \prod_{i=1}^{n} p(\omega_i)$, dove $p(\omega_i) = p_k$ se $\omega_i = x_k$, $k = 1, \ldots, N$. E' immediato mostrare che i numeri $p(\omega)$, $\omega \in \Omega_n$, determinano una distribuzione su $\Omega_n = X^{[1,n]}$.

(1.10) Definizione. Un evento $C \subset \Omega_n$ della forma

$$C = \{\omega \in \Omega_n \,:\, \omega_{i_1} \in B_{i_1}, \cdots, \omega_{i_k} \in B_{i_k}\}$$

dove $i_1, \ldots, i_k$ sono indici assegnati e $B_{i_1}, \cdots, B_{i_k}$ sono arbitrari sottoinsiemi (ad esempio punti) di X, si dice *cilindro* k-dimensionale ($k \leq n$).

La σ-algebra $\mathcal{F}_n$ dei sottoinsiemi di Ω_n sarà la più piccola σ-algebra che contiene la famiglia dei cilindri, ovvero la σ-algebra *generata* da tale famiglia, mentre la misura di probabilità P_n su $\mathcal{F}_n$ sarà la misura generata dalla distribuzione $p(\omega)$, $\omega \in \Omega_n$.

(1.11) Definizione. Lo spazio di probabilità $(\Omega_n, \mathcal{F}_n, \mathrm{P}_n)$ si chiama *sequenza di n eventi indipendenti*[1] *estratti con distribuzione* (X, p).

(1.12) Esempio. Sia $X = \{0, 1\}$, $p = (\frac{1}{2}, \frac{1}{2})$. Allora la sequenza di n eventi indipendenti estratti con distribuzione (di Bernoulli) (X, p) può essere assunto come modello matematico per una successione di n lanci di una moneta (non truccata), dove il simbolo 1 corrisponde all'occorrenza 'testa', mentre a 'croce' facciamo corrispondere il simbolo 0. Evidentemente $|\Omega_n| = 2^n$ e $p(\omega) = 1/2^n$. Ciò determina una distribuzione uniforme su Ω. Sia ora $C_{k,n} = \{k \text{ volte testa}\}$, cioè $C_{k,n}$ indica l'evento aleatorio in cui, in n lanci, il risultato è rappresentato esattamente k volte dal simbolo 1, in un ordine qualsiasi. Osserviamo che la collezione $\{C_{k,n}\}$ con $k \in \{0, 1, \ldots, n\}$ costituisce una partizione di Ω_n. Usando la (1.7) si trova

$$\mathrm{P}_n(C_{k,n}) = \sum_{\omega \in C_{k,n}} p(\omega) = \frac{1}{2^n} \cdot {}_nC_k$$

dove ${}_nC_k$ è uguale al numero di modi in cui k oggetti identici possono essere sistemati in n scatole se ogni scatola può ospitare al più un oggetto: il primo di essi potrà essere sistemato in n modi diversi, il secondo in $n-1$, fino al k-esimo che potrà trovar posto in una qualsiasi delle rimanenti $n-k+1$ scatole. D'altra parte, l'ordine in cui sistemiamo gli oggetti è irrilevante e vi sono $k!$ ordinamenti possibili. Pertanto si avrà

$$(1.13) \qquad {}_nC_k = \frac{n(n-1)\cdots(n-k+1)}{k!} = \frac{n!}{k!(n-k)!} = \binom{n}{k}$$

e quindi

$$(1.14) \qquad \mathrm{P}_n(C_{k,n}) = \frac{1}{2^n} \cdot \binom{n}{k}, \qquad \sum_{k=0}^{n} \mathrm{P}_n(C_{k,n}) = 1,$$

dove la seconda identità segue immediatamente dalla formula del binomio. Ora, se n è pari $\mathrm{P}_n(C_{k,n})$ è massima per $k = n/2$, mentre se n è dispari $\mathrm{P}_n(C_{k,n})$ è massima per $k = (n \pm 1)/2$. Consideriamo dunque per semplicità sequenze di $2n$

[1] La nozione generale di eventi indipendenti sarà discussa più avanti.

lanci e cerchiamo di stimare $\mathrm{P}_{2n}(C_{n,2n})$ quando n è grande. A tal fine faremo uso della formula asintotica di Stirling:

$$n! = e^{-n}\, n^n\, \sqrt{2\pi n}\, (1 + \mathcal{O}(\frac{1}{n})). \tag{1.15}$$

Allora si ha

$$\begin{aligned} \mathrm{P}_{2n}(C_{n,2n}) &= \frac{1}{2^{2n}} \cdot \frac{(2n)!}{(n!)^2} = \frac{1}{2^{2n}} \cdot \frac{e^{-2n}\,(2n)^{2n}\,\sqrt{4\pi n}}{(e^{-n}\, n^n\, \sqrt{2\pi n})^2}\,(1 + \mathcal{O}(\frac{1}{n})) \\ &= \frac{1}{\sqrt{\pi n}}\,(1 + \mathcal{O}(\frac{1}{n})). \end{aligned} \tag{1.16}$$

In modo simile, se $C^{(r)}_{n,2n}$ indica l'evento in cui, in $2n$ lanci, il simbolo 1 viene osservato un numero di volte che differisce da n per non più che una quantità fissata r, allora si ha (tralasciando i termini $\mathcal{O}(\frac{1}{n})$):

$$\mathrm{P}_{2n}(C^{(r)}_{n,2n}) = \sum_{n-r\le k\le n+r} \mathrm{P}_{2n}(C_{k,2n}) \le \frac{2r+1}{\sqrt{\pi n}} \to 0 \qquad (n \to \infty) \tag{1.17}$$

In conclusione, abbiamo trovato che al crescere del numero dei lanci la probabilità di osservare 'testa' *esattamente* la metà delle volte, o comunque un numero di volte che si mantiene vicino alla metà entro un margine fissato, tende a zero. Ciò non deve sorprendere. Quale che sia il contenuto di una eventuale 'legge dei grandi numeri', non sarà certamente sensato aspettarsi esattamente 1000 volte testa in 2000 lanci! Vedremo che la domanda corretta da porre riguarda la probabilità che la *proporzione* delle occorrenze 'testa' sia vicina a 1/2.

Osserviamo infine che lo spazio $\Omega = X^{[1,n]}$ con $|X| = 2$ si incontra in numerosi problemi di teoria della probabilità e di meccanica statistica. Possiamo immaginare un sistema costituito da n elementi identici, ad esempio dei magneti elementari (spins), ciascuno dei quali può trovarsi in due stati possibili: orientato verso l'alto (simbolo 1) o verso il basso (simbolo -1), cosicchè lo stato dell'intero sistema sarà identificato da una parola $\omega = (\omega_1 \dots \omega_n)$ con $\omega_i \in \{-1, 1\}$. Una generalizzazione di questo schema consiste nel dare un'interpretazione più generale dell'indice i dell'elemento ω_i. Ad esempio potremmo avere $i \in \Lambda$ con $\Lambda \subseteq \mathbb{Z}^d$ formato da punti della forma $i = (i_1, \dots, i_d)$ con $L_1^{(k)} \le i_k \le L_2^{(k)}$, $k = 1, \dots, d$. X poi potrà essere del tutto arbitrario, ad esempio un insieme numerabile o uno spazio metrico qualsiasi. Lo spazio $\Omega = X^\Lambda$ in tal modo ottenuto si incontra ad esempio nella teoria dei campi aleatori. D'altra parte, la situazione dell'esempio (1.11) non è la solo ad essere descrivibile come una sequenza di n eventi indipendenti estratti con una certa distribuzione.

(1.18) Esempio. Il problema delle coincidenze. Un'urna contiene N biglie numerate: $1, \ldots, N$. Consideriamo un esperimento consistente in n estrazioni con rimpiazzamento. Allora si ha $X = \{1, \ldots, N\}$ e $|\Omega_n| = N^n$. Assumeremo inoltre su X una distribuzione uniforme, cioè che le estrazioni siano tutte equiprobabili. Ciò implica, evidentemente, $p(\omega) = N^{-n}$, $\forall \omega \in \Omega_n$. Calcoliamo la probabilità dell'evento in cui non vi sono ripetizioni, ossia dell'evento

$$C = \{\omega \in \Omega_n : \omega_i \neq \omega_j, i \neq j\}.$$

E' evidente che $|C| = N(N-1)\cdots(N-n+1)$ e dunque, per la (1.7),

$$\mathrm{P}\,(C) = \left(1 - \frac{1}{N}\right)\left(1 - \frac{2}{N}\right)\cdots\left(1 - \frac{n-1}{N}\right).$$

Questo risultato può essere interpretato come segue. Supponiamo che in una classe vi siano n studenti. Qual'è la probabilità $P_{2,n}$ che almeno due di essi festeggino il loro compleanno nello stesso giorno? Se supponiamo che il compleanno di ogni studente cada in uno dei primi 365 giorni dell'anno e che ciascuno di questi giorni sia equiprobabile, allora il problema si riduce a quello precedente dell'urna con $N = 365$ (e dunque i compleanni degli n studenti sono interpretati come le biglie estratte dall'urna). Troviamo così:

$$P_{2,n} = 1 - \prod_{i=1}^{n-1}\left(1 - \frac{i}{365}\right).$$

Così, ad esempio, troviamo $P_{2,4} = 0.016$, $P_{2,16} = 0.284$, $P_{2,22} = 0.476$, $P_{2,23} = 0.507$, $P_{2,40} = 0.891$, $P_{2,64} = 0.997$. E' abbastanza sorprendente osservare che sono sufficienti 23 studenti per avere probabilità 1/2. In altre parole, in accordo con l'interpretazione empirica vista precedentemente, se il nostro modello probabilistico

$(\Omega_n, \mathcal{F}_n, \mathrm{P}_n)$ descrive bene la situazione, allora considerando un numero abbastanza grande di classi con 23 studenti ciascuna, e se non siamo troppo 'sfortunati', in circa la metà di esse vi saranno almeno due studenti con lo stesso compleanno.

(1.19) Esempio. Premi e lotterie. Consideriamo una lotteria in cui vi siano N biglietti, numerati da 1 a N, dei quali soltanto i primi n vincono dei premi ($N \geq 2n$). Ci chiediamo allora quale sia la probabilità P_1 di vincere almeno un premio comprando, a caso, n biglietti. Dal momento che l'ordine in cui i biglietti vengono acquistati non ha alcuno ruolo nella presenza o assenza di biglietti vincenti, avremo che $|\Omega_n| = \binom{N}{n}$. Sia

$$C_0 = \{\omega \in \Omega_n : \omega_i \in \{n+1, \ldots, N\}, \forall i = 1, \ldots, n\}$$

l'evento in cui non vi sia alcun biglietto vincente tra quelli acquistati. Evidentemente $|C_0| = \binom{N-n}{n}$ e dunque

$$\mathrm{P}(C_0) = \frac{\binom{N-n}{n}}{\binom{N}{n}} = \frac{(N-n)\cdots(N-2n+1)}{N(N-1)\cdots(N-n+1)} = \prod_{i=0}^{n-1}\left(1 - \frac{n}{N-i}\right).$$

Pertanto

$$P_1 = 1 - \mathrm{P}(C_0) = 1 - \prod_{i=0}^{n-1}\left(1 - \frac{n}{N-i}\right).$$

Se ad esempio $N = n^2$ e $n \to \infty$ allora $\mathrm{P}(C_0) \to e^{-1}$ e $P_1 \to 1 - e^{-1} \simeq 0.632$. La convergenza è piuttosto rapida: per $n = 10$ troviamo già $P_1 = 0.670$.

Estendiamo ora la nozione di distribuzione di probabilità ad ogni sequenza di numeri che rappresentano le probabilità associate ad un sistema completo di eventi. Abbiamo visto che gli eventi $\{C_{k,n}\}$ in (1.12) formano un sistema completo di eventi. Dunque, posto $p_{k,n} = \mathrm{P}_n(C_{k,n})$, la sequenza $(p_{0,n}, p_{1,n}, \ldots, p_{n,n})$ è una distribuzione di probabilità. Facendo ancora riferimento a (1.12), possiamo ulteriormente estendere l'uso di questo concetto introducendo la nozione di variabile aleatoria.

Consideriamo la funzione $\nu_n^{(1)} : \Omega_n \to I\!\!R$ che ad ogni $\omega \in \Omega_n$ associa il numero di volte in cui vi compare il simbolo 1. In formule

$$\nu_n^{(1)}(\omega) := \#\{i \in [1,n] : \omega_i = 1\}. \tag{1.20}$$

E' semplice verificare che $C_{k,n} = \{\omega \in \Omega : \nu_n^{(1)}(\omega) = k\}$, ovvero gli elementi del sistema di eventi $\{C_{k,n}\}$ sono gli insiemi di livello della funzione $\nu_n^{(1)}$. La funzione $\nu_n^{(1)}$ fornisce un esempio di *variabile aleatoria discreta*.

(1.21) Definizione. Sia $(\Omega, \mathcal{F}, \mathrm{P})$ uno spazio di probabilità. Una funzione $\xi = \xi(\omega)$ definita su Ω, e che assume valori in un insieme $\mathcal{Z}$ al più numerabile, si dice *variabile aleatoria discreta* se per ogni $x \in \mathcal{Z}$ si ha $\{\omega \in \Omega : \xi(\omega) = x\} \in \mathcal{F}$. L'ultima condizione si può esprimere dicendo che $\xi(\omega)$ deve essere una funzione *misurabile* su $(\Omega, \mathcal{F})$. Se Ω è discreto qualunque funzione reale $\xi = \xi(\omega)$ ha questa proprietà e dunque è una variabile aleatoria discreta.

Poniamo ora $\mathcal{Z} = \{x_k\}$ e consideriamo gli eventi $C_k = \{\omega \in \Omega : \xi(\omega) = x_k\}$. È chiaro che $C_r \cap C_s = \emptyset$ se $r \neq s$ (ξ assume un sol valore per ogni ω). Inoltre $\cup_k C_k = \Omega$ (ad ogni ω è associato un qualche valore $\xi(\omega) \in \mathcal{Z}$). In altre parole $\{C_k\}$ è un sistema completo di eventi. Il numero $\mathrm{P}(C_k)$ fornisce la probabilità che la variabile aleatoria ξ assuma il valore $x_k \in \mathcal{Z}$. Ciò conduce alla seguente definizione:

(1.22) Definizione. La sequenza di numeri $p_k = \mathrm{P}(C_k)$ definita su $\mathcal{Z}$ si chiama la *distribuzione di probabilità della variabile aleatoria discreta* ξ.

Così, la distribuzione $p_{k,n} = \mathrm{P}(C_{k,n}) = \frac{1}{2^n} \cdot \binom{n}{k}$ in (1.11) è la distribuzione di probabilità della variabile aleatoria $\nu_n^{(1)}$. Occorre fare attenzione al fatto che, ad eccezione del caso in cui Ω è discreto e $\xi(\omega) = \omega$, per cui $\mathcal{Z} \equiv \Omega$, la distribuzione sullo spazio $\mathcal{Z}$, definita qui sopra, non va confusa con la distribuzione definita in (1.6). Ciò detto, la distinzione apparirà chiara dal contesto.

Introduciamo ora due quantità numeriche associate ad una variabile aleatoria che avranno un ruolo importante in quanto segue: l'*aspettazione* e la *varianza*. Se immaginiamo di ripetere n prove dello stesso esperimento e di osservare ogni volta il valore assunto da ξ, allora se n è grande, in accordo con l'interpretazione empirica della probabilità (vedi sopra), in approssimativamente np_1 occasioni si avrà $\xi = x_1$, in approssimativamente np_2 occasioni si avrà $\xi = x_2$, e così via[1].

[1] Occorre tuttavia fare attenzione: è perfettamente possibile che una particolare

Prendendo la media aritmetica dei valori di ξ in tal modo ottenuti, otterremo approssimativamente il valore

$$\frac{np_1 \cdot x_1 + np_2 \cdot x_2 + \cdots}{n} = \sum_k x_k\, p_k$$

che fornisce allora il valore attorno al quale le medie artimetiche dei valori di ξ fluttuano.

(1.23) Definizione. L'*aspettazione matematica*, o *valor medio*, di una variabile aleatoria discreta ξ è il numero

$$\mathrm{E}\,\xi = \sum_k x_k\, p_k \cdot$$

Vediamone alcune proprietà. L'aspettazione è la media aritmetica pesata dei valori x_k, cioè sullo spazio $\mathcal{Z}$, con pesi p_k ed è analoga alla nozione di centro di massa che si incontra in meccanica. Quando è definita, $\mathrm{E}\,\xi$ indica dunque vicino a quale numero i valori della variabile aleatoria discreta ξ si concentrano in media. Ovviamente tale numero è sempre compreso tra l'estremo inferiore e l'estremo superiore dei valori possibili di ξ. In particolare si ha

$$(1.24) \qquad \xi \geq 0 \quad \Longrightarrow \quad \mathrm{E}\,\xi \geq 0 \qquad \text{e} \qquad \xi \equiv 1 \quad \Longrightarrow \quad \mathrm{E}\,\xi = 1 \cdot$$

D'altra parte, affinchè $\mathrm{E}\,\xi$ abbia senso dovremo richiedere che la somma su k converga assolutamente, cioè $\sum_k |x_k|\, p_k < \infty$. In caso contrario, un riarrangiamento dei termini darebbe un valor medio diverso. Ciò detto, consideriamo la combinazione lineare $a\xi + b\eta$ di due variabili aleatorie ξ e η, i cui valori possibili sono dati da

---

sequenza di n prove dia risultati completamente diversi: ad esempio n volte lo stesso risultato $\xi = x_k$. Per la 'legge dei grandi numeri', che discuteremo più avanti, le sequenze di n prove che danno il risultato 'ragionevolmente atteso' di cui sopra, cioè quello 'predetto' dal modello matematico probabilistico che assegna probabilità p_k all'evento $\{\xi = x_k\}$, formano, per n grande, la stragrande maggioranza delle sequenze possibili.

$\{x_k\}$ e $\{y_k\}$, rispettivamente, e tali che $E\xi$ ed $\mathrm{E}\,\eta$ siano definiti. Sia $C_{kj} = \{\omega : \xi(\omega) = x_k,\ \eta(\omega) = y_j\}$. Si verifica facilmente che

$$\sum_k \mathrm{P}\,(C_{kj}) = \mathrm{P}\,(\eta = y_j) \quad \text{e} \quad \sum_j \mathrm{P}\,(C_{kj}) = \mathrm{P}\,(\xi = x_k). \tag{1.25}$$

Gli insiemi C_{kj} con $k, j = 1, 2, \ldots$ formano un sistema completo di eventi e la collezione di numeri $p_{kj} = \mathrm{P}\,(C_{kj})$ si dice *distribuzione di probabilità congiunta* delle variabili aleatorie ξ ed η. Se $\zeta = f(\xi, \eta)$ allora $\mathrm{E}\,\zeta = \sum_{k,j} f(x_k, y_j)\, p_{kj}$. Pertanto

$$\begin{aligned} E(a\xi + b\eta) &= \sum_k \sum_j (a x_k + b y_j)\, p_{kj} \\ &= a \sum_k x_k \mathrm{P}\,(\xi = x_k) + b \sum_j y_j \mathrm{P}\,(\eta = y_j) = a\,\mathrm{E}\,\xi + b\,\mathrm{E}\,\eta \end{aligned} \tag{1.26}$$

perchè la somma di due serie assolutamente convergenti è ancora assolutamente convergente. Le proprietà (1.24) e (1.26) implicano che E è un funzionale lineare, non-negativo e normalizzato sullo spazio vettoriale delle variabili aleatorie (discrete). Osserviamo che dalla (1.26) segue che se $\eta = \xi - \mathrm{E}\,\xi$ allora $\mathrm{E}\,\eta = 0$. Infine, le variabili aleatorie si possono moltiplicare tra loro (formando un anello commutativo) e si ha $E\xi \cdot \eta = \sum_{k,j} x_k\, y_j\, p_{kj}$.

Ora, se lo spazio degli eventi elementari Ω è discreto troviamo

$$\begin{aligned} \mathrm{E}\,\xi &= \sum_k x_k\, p_k = \sum_k x_k\, \mathrm{P}\,(C_k) = \sum_k x_k \sum_{\omega \in C_k} p(\omega) \\ &= \sum_k \sum_{\omega \in C_k} \xi(\omega) p(\omega) = \sum_{\omega \in \Omega} \xi(\omega) p(\omega) \cdot \end{aligned} \tag{1.27}$$

L'ultima espressione è ancora una media aritmetica pesata, ma questa volta sullo spazio Ω, e può essere presa come definizione alternativa di $\mathrm{E}\,\xi$ quando Ω è discreto.

(1.28) Prima disuguaglianza di Chebyshev. Se $\xi \geq 0$ e $\mathrm{E}\,\xi < \infty$, allora per ogni $\alpha > 0$ si ha

$$\mathrm{P}\,(\{\omega : \xi(\omega) \geq \alpha\}) \leq \frac{\mathrm{E}\,\xi}{\alpha}$$

Dimostrazione. Consideriamo per semplicità il caso in cui Ω è discreto. Allora si ha

$$\begin{aligned}
\mathrm{P}\left(\{\omega\,:\,\xi(\omega)\geq a\}\right) &= \sum_{\omega\,:\,\xi(\omega)\geq a} p(\omega) \leq \sum_{\omega\,:\,\xi(\omega)\geq a} \left(\frac{\xi(\omega)}{a}\right) p(\omega) \\
&= \frac{1}{a} \sum_{\omega\,:\,\xi(\omega)\geq a} \xi(\omega)\, p(\omega) \leq \frac{1}{a} \sum_{\omega\in\Omega} \xi(\omega)\, p(\omega) = \frac{\mathrm{E}\,\xi}{a}
\end{aligned}$$

dove nell'ultima uguaglianza si è fatto uso della (1.27). q.e.d.

Un'altra quantità numerica importante, che caratterizza le fluttuazioni (in media) di una variabile aleatoria rispetto al suo valor medio, è la varianza.

(1.29) Definizione. La varianza di una variabile aleatoria ξ tale che $\mathrm{E}\,\xi < \infty$ è il numero $\mathrm{Var}\,\xi = \mathrm{E}\,(\xi - \mathrm{E}\,\xi)^2$.

Vediamone alcune proprietà. Innanzitutto, sviluppando il quadrato e usando la linearità del valor medio si trova

$$\mathrm{Var}\,\xi = \mathrm{E}\,\xi^2 - 2(\mathrm{E}\,\xi)^2 + (\mathrm{E}\,\xi)^2 = \mathrm{E}\,\xi^2 - (\mathrm{E}\,\xi)^2 \tag{1.30}$$

da cui segue che $\mathrm{Var}\,\xi < \infty$ se e solo se $\mathrm{E}\,\xi^2 < \infty$. Consideriamo ora la variabile aleatoria $a\xi + b$. Si ha

$$\begin{aligned}
\mathrm{Var}\,(a\xi + b) &= \mathrm{E}\,(a\xi + b - \mathrm{E}\,(a\xi + b))^2 = \mathrm{E}\,(a^2(\xi - \mathrm{E}\,\xi)^2) \\
&= a^2\,\mathrm{E}\,(\xi - \mathrm{E}\,\xi)^2 = a^2\,\mathrm{Var}\,\xi\,\cdot
\end{aligned} \tag{1.31}$$

(1.32) Seconda disuguaglianza di Chebyshev. Se $\mathrm{Var}\,\xi < \infty$, allora per ogni $\alpha > 0$ si ha

$$\mathrm{P}\left(\{\omega\,:\,|\xi(\omega) - \mathrm{E}\,\xi| \geq \alpha\}\right) \leq \frac{\mathrm{Var}\,\xi}{\alpha^2}$$

Dimostrazione. Consideriamo la variabile aleatoria $\eta = (\xi - \mathrm{E}\,\xi)^2 \geq 0$. Si ha $\mathrm{E}\,\eta = \mathrm{Var}\,\xi$ e dunque usando la prima disuguaglianza di Chebyshev si trova

$$\mathrm{P}\left(\{\omega\,:\,|\xi(\omega) - \mathrm{E}\,\xi| \geq \alpha\}\right) = \mathrm{P}\left(\{\omega\,:\,\eta(\omega) \geq \alpha^2\}\right) \leq \frac{\mathrm{E}\,\eta}{\alpha^2} = \frac{\mathrm{Var}\,\xi}{\alpha^2}\,\cdot \qquad \text{q.e.d.}$$

(1.33) Esempio. Calcoliamo $\mathrm{E}\,\nu_n^{(1)}$ e $\mathrm{Var}\,\nu_n^{(1)}$, con $\nu_n^{(1)}$ definita in (1.20). Si ha

$$\begin{aligned}
\mathrm{E}\,\nu_n^{(1)} &= \sum_{k=0}^{n} k\mathrm{P}_n(C_{k,n}) = \sum_{k=1}^{n} k\binom{n}{k}\frac{1}{2^n} = \frac{1}{2^n}\sum_{k=1}^{n} n\binom{n-1}{k-1} \\
&= n\,\frac{1}{2^n}\sum_{s=0}^{n-1}\binom{n-1}{s} = n\,\frac{1}{2^n}\,2^{n-1} = \frac{n}{2}
\end{aligned} \tag{1.34}$$

dove nella terza uguaglianza si è fatto uso della proprietà $k\binom{n}{k} = n\binom{n-1}{k-1}$, mentre nella penultima si è usata la formula del binomio. Per la varianza si ha

$$\mathrm{Var}\, \nu_n^{(1)} = \mathrm{E}\,(\nu_n^{(1)})^2 - (\mathrm{E}\,\nu_n^{(1)})^2 = \sum_{k=1}^{n} k^2 \binom{n}{k} \cdot \frac{1}{2^n} - \frac{n^2}{4} \cdot$$

D'altra parte si ha

$$\begin{aligned} k^2\binom{n}{k} &= nk\binom{n-1}{k-1} = n(k-1)\binom{n-1}{k-1} + n\binom{n-1}{k-1} \\ &= n(n-1)\binom{n-2}{k-2} + n\binom{n-1}{k-1}, \end{aligned}$$

e pertanto

$$\mathrm{Var}\, \nu_n^{(1)} = \left(\, n\,(n-1)\; 2^{n-2} + n\; 2^{n-1} \,\right) \cdot \frac{1}{2^n} - \frac{n^2}{4} = \frac{n}{4} \cdot \tag{1.35}$$

Ora, usando la (1.20) e la (1.34), la (1.17) può essere riscritta nella forma

$$\mathrm{P}_{\,n}(\{\omega\ : |\nu_n^{(1)}(\omega) - \mathrm{E}\,\nu_n^{(1)}| \leq r\}) \leq \frac{2r+1}{\sqrt{\pi n/2}} \tag{1.36}$$

Nella direzione opposta, la disuguaglianza di Chebyshev dà

$$\mathrm{P}_{\,n}(\{\omega\ : |\, \nu_n^{(1)}(\omega) - \mathrm{E}\,\nu_n^{(1)}\,| \geq r\}) \leq \frac{n}{4\,r^2}, \tag{1.37}$$

che però non è di grande interesse. Possiamo tuttavia ottenere informazioni più utili se, invece di guardare a $\nu_n^{(1)}(\omega)$, consideriamo $f_n^{(1)}(\omega) := \frac{\nu_n^{(1)}(\omega)}{n}$, cioè la frequenza relativa dell'evento 'testa' in n lanci. Usando le proprietà del valor medio e della varianza troviamo

$$\mathrm{E}\, f_n^{(1)} = \frac{1}{2}, \qquad \mathrm{Var}\, f_n^{(1)} = \frac{1}{4n} \tag{1.38}$$

Allora dalla seconda disuguaglianza di Chebyshev si ottiene

$$\mathrm{P}_{\,n}(\{\omega\ : |\, f_n^{(1)}(\omega) - \mathrm{E}\, f_n^{(1)}\,| \geq r\}) \leq \frac{1}{4\,n\,r^2} \tag{1.39}$$

e quindi, per ogni $r > 0$,

$$\lim_{n\to\infty} \mathrm{P}_{\,n}\left(\, |\, f_n^{(1)}(\omega) - \mathrm{E}\, f_n^{(1)}\,| \leq r \right) = 1 \cdot \tag{1.40}$$

Osserviamo che, essendo P_n generata da una distribuzione uniforme su Ω_n, per un arbitrario evento $C \subseteq \Omega_n$ si ha

$$\mathrm{P}_n(C) = \sum_{\omega \in C} p(\omega) = \frac{\#\{\omega : \omega \in C\}}{|\Omega_n|}, \tag{1.41}$$

e dunque la (1.40) equivale ad affermare che la frazione delle parole ω su cui la frequenza relativa dell'evento 'testa' si discosta dal suo valore atteso $1/2$ per meno di una quantità r, positiva ma per altro arbitraria, tende a uno al crescere della lunghezza delle parole. Si tratta di un caso particolare della *legge dei grandi numeri*, che discuteremo nella sezione seguente.

In vista di questo risultato, poniamoci la seguente domanda: se consideriamo per semplicità il caso di sequenze di $2n$ lanci, quanto può essere grande una fluttuazione di $\nu_{2n}^{(1)}$ rispetto al valore atteso $\mathrm{E}\,\nu_{2n}^{(1)} = n$ per non divenire sorprendente? Ad esempio per non farci sospettare che la moneta sia truccata? Prendiamo nuovamente in esame la distribuzione $p_{k,2n} = 2^{-2n}\binom{2n}{k}$, $k = 0, 1, \ldots, 2n$, di $\nu_{2n}^{(1)}$. Ciò che vorremmo è una successione $r_n \uparrow \infty$ tale che

$$\lim_n \mathrm{P}_{2n}(\,|\nu_{2n}^{(1)} - n| \le r_n\,) = \gamma, \qquad 0 < \gamma < 1. \tag{1.42}$$

Osserviamo a questo proposito che $\sqrt{\mathrm{Var}\,\nu_{2n}^{(1)}} = \sqrt{n/2}$ fornisce una stima della 'larghezza' della distribuzione $p_{k,2n}$. In alternativa, basterà ricordare che il valore massimo di $p_{n,2n}$ è $\mathcal{O}(1/\sqrt{n})$ e che $\sum_k p_{k,2n} = 1$. Possiamo dunque ipotizzare che l'ampiezza di una fluttuazione non sospetta sia $r_n = \mathcal{O}(\sqrt{n})$. In effetti se poniamo ad esempio $r_n = \alpha_n \sqrt{n/2}$ allora dovrà essere $\underline{\lim}\,\alpha_n > 0$ altrimenti, per la (1.36), si avrebbe $\gamma = 0$ nella (1.42). D'altra parte, dalla (1.37) con $r = r_n$ si vede che deve essere $\overline{\lim}\,\alpha_n < \infty$, altrimenti si avrebbe $\gamma = 1$. Allora possiamo ad esempio richiedere che $\alpha_n \to \alpha$ con $0 < \alpha < \infty$, ma allora non v'è ragione per non provare $\alpha_n \equiv \alpha$. Vogliamo dunque calcolare

$$Q_{2n}(\alpha) \equiv \mathrm{P}_{2n}\left(\,|\nu_{2n}^{(1)} - n| \le \alpha\sqrt{\frac{n}{2}}\right) = \frac{1}{2^{2n}} \sum_{j \in R_n(\alpha)} \binom{2n}{n+j} \tag{1.43}$$

dove abbiamo posto $R_n(\alpha) = \{j : |j| \le \alpha\sqrt{n/2}\}$. Ora

$$\binom{2n}{n+j} = \frac{(2n)!}{(n+j)!(n-j)!} = \binom{2n}{n} \cdot \frac{n(n-1)\cdots(n-j+1)}{(n+j)\cdots(n+1)}. \tag{1.44}$$

Denotando l'ultimo fattore $D_{n,j}$ possiamo riscrivere la (1.44) nella forma

$$Q_{2n}(\alpha) = p_{n,2n} \cdot \sum_{j \in R_n(\alpha)} D_{n,j}. \tag{1.45}$$

Cerchiamo di stimare i termini $D_{n,j}$. Si ha

$$D_{n,j} = \left[\left(1 + \frac{j}{n}\right) \left(1 + \frac{j}{n-1}\right) \cdots \left(1 + \frac{j}{n-j+1}\right) \right]^{-1}$$

e dunque

$$\log D_{n,j} = -\sum_{k=0}^{j-1} \log\left(1 + \frac{j}{n-k}\right).$$

Sviluppiamo il logaritmo:

$$\log\left(1 + \frac{j}{n-k}\right) = \frac{j}{n-k} \left(1 + \mathcal{O}\left(\frac{j}{n-k}\right)\right).$$

D'altra parte, se $j \in R_n(\alpha)$ si ha $\mathcal{O}\left(\frac{j}{n-k}\right) = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$, e inoltre

$$\frac{j}{n-k} = \frac{j}{n} \cdot \frac{1}{1-k/n} = \frac{j}{n}\left(1 + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)\right).$$

In conclusione

$$\log D_{n,j} = -\frac{j^2}{n}\left(1 + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)\right)$$

e dunque, essendo $j^2/n < \alpha^2/2$ per $j \in R_n(\alpha)$,

$$D_{n,j} = e^{-j^2/n}\left(1 + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)\right).$$

Infine, mettendo insieme la (1.16) e la (1.45) troviamo

$$Q_{2n}(\alpha) = \left(1 + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)\right) \sum_{j \in R_n(\alpha)} \frac{1}{\sqrt{\pi n}} e^{-j^2/n} \tag{1.46}$$

A questo punto, se eseguiamo il cambiamento di variabili: $x_j = j\sqrt{2/n}$, $\Delta x = x_{j+1} - x_j = \sqrt{2/n}$, la condizione $j \in R_n(\alpha)$ corrisponde a $-\alpha \le x_j \le \alpha$ e la (1.46) diviene

$$Q_{2n}(\alpha) = \left(1 + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)\right) \sum_{-\alpha \le x_j \le \alpha} \frac{1}{\sqrt{2\pi}} e^{-x_j^2/2} \Delta x \tag{1.47}$$

Ma l'ultimo fattore altro non è che la somma che approssima l'integrale di Riemann della funzione $p(x) = e^{-x^2/2}/\sqrt{2\pi}$, tra $-\alpha$ e α. Quindi

$$\lim_{n\to\infty} Q_{2n}(\alpha) = \frac{1}{\sqrt{2\pi}} \int_{-\alpha}^{\alpha} e^{-x^2/2}\, dx \cdot \tag{1.48}$$

Il caso dispari, cioè con $2n+1$ lanci, può essere similmente trattato osservando che per ogni $\delta > 0$ ed ogni n abbastanza grande $(n > 1/2\delta^2)$

$$\begin{aligned}\left\{\omega \,:\, |\,\nu_{2n}^{(1)} - n\,| \le (\alpha-\delta)\sqrt{\frac{n}{2}}\right\} &\subset \left\{\omega \,:\, |\,\nu_{2n+1}^{(1)} - (n+\frac{1}{2})\,| \le \alpha\sqrt{\frac{(n+1/2)}{2}}\right\} \\ &\subset \left\{\omega \,:\, |\,\nu_{2n}^{(1)} - n\,| \le (\alpha+\delta)\sqrt{\frac{n}{2}}\right\} \cdot\end{aligned}$$

Pertanto si ha

$$\lim_n Q_{2n}(\alpha-\delta) \le \underline{\lim}\, Q_{2n+1}(\alpha) \le \overline{\lim}\, Q_{2n+1}(\alpha) \le \lim_n Q_{2n}(\alpha+\delta)\,.$$

Abbiamo dunque mostrato che

$$\lim_{n\to\infty} \mathrm{P}_n\left(|\,\nu_n^{(1)} - \frac{n}{2}\,| \le \alpha\,\frac{\sqrt{n}}{2}\right) = \frac{1}{\sqrt{2\pi}} \int_{-\alpha}^{\alpha} e^{-x^2/2}\, dx \cdot \tag{1.49}$$

In modo equivalente possiamo scrivere

$$\lim_{n\to\infty} \mathrm{P}_n\left(|\,f_n^{(1)} - \frac{1}{2}\,| \le \alpha\,\frac{1}{2\sqrt{n}}\right) = \frac{1}{\sqrt{2\pi}} \int_{-\alpha}^{\alpha} e^{-x^2/2}\, dx \tag{1.50}$$

ed osserviamo che $1/(2\sqrt{n}) = \sqrt{\mathrm{Var}\, f_n^{(1)}}$. In parole, abbiamo trovato che la distribuzione asintotica delle deviazioni della frequenza relativa delle teste da $1/2$, di ampiezza pari a una frazione della deviazione media $\sqrt{\mathrm{Var}\, f_n^{(1)}}$, cioè deviazioni $\mathcal{O}(1/\sqrt{n})$, è governata da un integrale gaussiano. Si tratta di un caso particolare del *teorema del limite centrale*, sul quale torneremo più volte in seguito. Osserviamo che un'interessante 'verifica sperimentale' di questo risultato può essere ottenuta per mezzo del cosiddetto 'tavolo di Galton'.

Concludiamo questa sezione con alcune osservazioni euristiche sul significato di 'variabile aleatoria continua'. A questo riguardo consideriamo la variabile aleatoria discreta

$$\eta_n^{(1)} := \frac{f_n^{(1)} - \mathrm{E}\, f_n^{(1)}}{\sqrt{\mathrm{Var}\, f_n^{(1)}}} \tag{1.51}$$

e notiamo che $\eta_n^{(1)}$ è *normalizzata*, ossia $\mathrm{E}\,\eta_n^{(1)} = 0$ e $\mathrm{Var}\,\eta_n^{(1)} = 1$. Inoltre, se $\nu_n^{(1)}$ varia in $R_n(\alpha)$ allora $\eta_n^{(1)}$ varia nell'intervallo $[-\alpha, \alpha]$. Per mezzo di $\eta_n^{(1)}$ la (1.34) può essere ulteriormente riscritta nella forma

$$(1.52) \qquad \lim_{n\to\infty} \mathrm{P}_{\,n}\left(\,|\,\eta_n^{(1)}\,| \le \alpha \right) = \frac{1}{\sqrt{2\pi}} \int_{-\alpha}^{\alpha} e^{-x^2/2}\,dx$$

Ora, il numero di valori assunti dalla variabile $\eta_n^{(1)}$ tende all'infinito ma, nello stesso tempo, la distanza tra due valori successivi $x_{k+1} - x_k$ tende a zero in modo tale che la quantità $\sum_{-\alpha \le x_k \le \alpha} \mathrm{P}_{\,n}(\eta_n^{(1)} = x_k)$ converge ad un valore finito, dato dall'integrale gaussiano scritto sopra. Più in generale, immaginiamo una variabile aleatoria discreta ξ_n su Ω_n che assume valori $x_0, \ldots, x_n$, per cui si abbia $p_k \equiv \mathrm{P}_{\,n}(\xi_n = x_k) = p(x_k)\Delta x_k + o(\Delta x_k)$ dove $p(x)$ è una funzione continua su $I\!R$. In tal caso la somma $\sum_{a \le x_k \le b} p_k$ si comporta come la somma di Riemann per l'integrale $\int_a^b p(x)dx$. Evidentemente deve essere $p(x) \ge 0$ e $\int_{-\infty}^{\infty} p(x)dx = 1$. Ogni funzione che soddisfa queste due proprietà si chiama *densità di probabilità*.

(1.53) Definizione. Sia $(\Omega, \mathcal{F}, \mathrm{P}\,)$ uno spazio di probabilità e $\xi : \Omega \to I\!R$ una funzione misurabile, ossia tale che per ogni $B \in \mathcal{B}(I\!R)$ si abbia $\{\omega : \xi(\omega) \in B\} \in \mathcal{F}$. Se per ogni intervallo $[a, b]$ si ha

$$\mathrm{P}\,(\{\omega \,:\, a \le \xi(\omega) \le b\}) = \int_a^b p(x)\,dx$$

allora diciamo che ξ ha distribuzione di probabilità con densità $p(x)$.

L'aspettazione e la varianza di una variabile aleatoria distribuita con densità di probabilità $p(x)$ sono date da

$$(1.54) \qquad \mathrm{E}\,\xi = \int_{-\infty}^{\infty} x\,p(x)\,dx, \quad \mathrm{Var}\,\xi = \int_{-\infty}^{\infty} x^2\,p(x)\,dx - \left(\int_{-\infty}^{\infty} x\,p(x)\,dx \right)^2 \cdot$$

(1.55) Esempi ed esercizi.

1) $p(x) = \frac{1}{\sqrt{2\pi}\sigma}\, e^{-(x-m)^2/2\sigma^2}$ si dice distribuzione gaussiana con parametri m, σ. Mostrare che per questa distribuzione si ha $\mathrm{E}\,\xi = m$ e $\mathrm{Var}\,\xi = \sigma^2$.
2) $p(x) = \frac{1}{b-a} \cdot 1_{[a,b]}$ si dice distribuzione uniforme sull'intervallo $[a, b]$. Mostrare che $\mathrm{E}\,\xi = (a + b)/2$ e $\mathrm{Var}\,\xi = (b - a)^2/12$.

3) $p(x) = \lambda e^{-\lambda x} \cdot 1_{[0,\infty]}$ si dice distribuzione esponenziale. Mostrare che $\mathrm{E}\,\xi = 1/\lambda$ e $\mathrm{Var}\,\xi = 1/\lambda^2$.

(1.56) Altre disuguaglianze utili.

Applicando la prima disuguaglianza di Chebyshev (1.28) alla variabile aleatoria $\eta = |\xi|^k$ si ottiene la *disuguaglianza di Markov*,

$$\mathrm{P}\left(\{\omega : |\xi(\omega)| \geq \alpha\}\right) \leq \frac{\mathrm{E}\,\eta}{\alpha^k}. \tag{1.57}$$

Più in generale, se $\xi \geq 0$ e $\lambda > 0$, la *disuguaglianza di Chebyshev esponenziale* afferma che

$$\mathrm{P}\left(\{\omega : \xi(\omega) \geq \alpha\}\right) = \mathrm{P}\left(\{\omega : e^{\lambda\xi(\omega)} \geq e^{\lambda\alpha}\}\right) \leq \mathrm{E}\,e^{\lambda(\xi-\alpha)}. \tag{1.58}$$

Inoltre, essendo λ positivo ma del tutto arbitrario, è evidente che

$$\mathrm{P}\left(\{\omega : \xi(\omega) \geq \alpha\}\right) \leq \inf_{\lambda>0} \mathrm{E}\,e^{\lambda(\xi-\alpha)}. \tag{1.59}$$

Una funzione ϕ definita su un intervallo I si dice *convessa* se soddisfa $\phi(px + (1-p)y) \leq p\,\phi(x) + (1-p)\,\phi(y)$, $0 < p < 1$, $x, y \in I$. Se ϕ è almeno C^2, una condizione sufficiente è $\phi''(x) \geq 0$, $\forall x \in I$. Per induzione si vede poi che $\phi(\sum_{i=1}^N p_i x_i) \leq \sum_{i=1}^N p_i \phi(x_i)$, ogni volta che $p_i \geq 0$, $\sum_{i=1}^N p_i = 1$ e $x_i \in I$. Se ora ξ assume i valori x_i con probabilità p_i, questa diviene la *disuguaglianza di Jensen*,

$$\phi(\mathrm{E}\,\xi) \leq \mathrm{E}\,(\phi(\xi)), \tag{1.60}$$

valida se ϕ è convessa su un intervallo che contiene i possibili valori di ξ. Siano ora $p > 1$ e $q > 1$ tali che $1/p + 1/q = 1$. Dati a e b positivi esisteranno s e t tali che $a = e^{s/p}$ e $b = e^{t/q}$. Essendo e^x convessa, si ha $e^{s/p+t/q} \leq p^{-1}e^s + q^{-1}e^t$, ovvero $ab \leq a^p/p + b^q/q$. Se ora poniamo $a = |\xi|/\mathrm{E}^{1/p}(|\xi|^p)$ e $b = |\eta|/\mathrm{E}^{1/q}(|\eta|^q)$ si ottiene

$$\left|\frac{\xi\eta}{\mathrm{E}^{1/p}(|\xi|^p)\,\mathrm{E}^{1/q}(|\eta|^q)}\right| \leq \frac{1}{p}\frac{|\xi|^p}{\mathrm{E}\,(|\xi|^p)} + \frac{1}{q}\frac{|\eta|^q}{\mathrm{E}\,(|\eta|^q)}.$$

Osserviamo che l'aspettazione del termine a destra è uguale a 1. Pertanto, prendendo l'aspettazione di ambo i membri si ottiene la *disuguaglianza di Hölder*,

$$\mathrm{E}\,(|\xi\,\eta|) \leq \mathrm{E}^{1/p}(|\xi|^p)\,\mathrm{E}^{1/q}(|\eta|^q) \tag{1.61}$$

La scelta $p = q = 2$ dà la *disuguaglianza di Schwartz*,

$$\mathrm{E}\,(|\xi\,\eta|) \le \mathrm{E}^{\,1/2}(\xi^2)\,\mathrm{E}^{\,1/2}(\eta^q). \tag{1.62}$$

Infine, dati due numeri $0 < \alpha < \beta$, si prendano nella (1.61) $p = \beta/\alpha$, $q = \beta/(\beta-\alpha)$, $\eta \equiv 1$ e si sostiuisca ξ con $|\xi|^\alpha$. Il risultato è la *disuguaglianza di Lyapunov*,

$$\mathrm{E}^{\,1/\alpha}(|\xi|^\alpha) \le \mathrm{E}^{\,1/\beta}(|\xi|^\beta). \tag{1.63}$$

2. SEQUENZE DI PROVE INDIPENDENTI: LEGGE DEI GRANDI NUMERI, TEOREMI LIMITE, ENTROPIA.

Prendiamo nuovamente in esame lo spazio di probabilità $(\Omega_n, \mathcal{F}_n, \mathrm{P}_{\,n})$ definito in (1.10) e studiamo più in dettaglio le proprietà della misura $\mathrm{P}_{\,n}$. Consideriamo in particolare il cilindro

$$C^{\,i_1\,,\,\dots\,,\,i_k}_{x_{j_1},\dots,x_{j_k}} = \{\omega\,:\,\omega_{i_1} = x_{j_1}, \cdots, \omega_{i_k} = x_{j_k}\} \tag{2.1}$$

dove $x_{j_1}, \dots, x_{j_k}$ sono elementi assegnati dell'insieme finito $X = \{x_1, \dots, x_N\}$, non necessariamente distinti. Si ha

$$\begin{aligned}
\mathrm{P}_{\,n}(C^{\,i_1\,,\,\dots\,,\,i_k}_{x_{j_1},\dots,x_{j_k}}) &= \sum_{\omega\,\in\,C^{\,i_1\,,\,\dots\,,\,i_k}_{x_{j_1},\dots,x_{j_k}}} \prod_{i=1}^{n} p(\omega_i) = \sum_{\substack{\omega_i,\; i\ne i_1,\dots,i_k \\ \omega_{i_l} = x_{j_l},\; l=1,\dots,k}} \prod_{i=1}^{n} p(\omega_i) \\
&= p_{j_1}\dots p_{j_k} \sum_{\omega_i,\; i\ne i_1,\dots,i_k} \;\prod_{i\ne i_1,\dots,i_k} p(\omega_i) \\
&= p_{j_1}\dots p_{j_k} \sum_{\omega\in\Omega_{n-k}} \prod_{i=1}^{n-k} p(\omega_i),
\end{aligned}$$

dove, lo ricordiamo $p(\omega_i) = p_j$ se $\omega_i = x_j \in X$. D'altra parte l'ultima somma è uguale a uno perchè i numeri $p(\omega)$ formano una distribuzione per ogni arbitraria lunghezza n delle parole ω. Abbiamo quindi mostrato il

(2.2) Lemma. $\mathrm{P}_{\,n}(C^{\,i_1\,,\,\dots\,,\,i_k}_{x_{j_1},\dots,x_{j_k}}) = p_{j_1}\dots p_{j_k}.$

Generalizzando la (1.20), introduciamo ora la variabile aleatoria

$$\nu_n^{(h)}(\omega) := \ \# \{i \in [1,n] \,:\, \omega_i = x_h\}. \tag{2.3}$$

Osserviamo che si può scrivere

$$\nu_n^{(h)}(\omega) = \sum_{i=1}^{n} 1_h^{(i)}(\omega), \quad \text{dove} \quad 1_h^{(i)}(\omega) = \begin{cases} 1, & \text{se } \omega_i = x_h \,, \\ 0, & \text{altrimenti.} \end{cases} \tag{2.4}$$

Evidentemente $\nu_n^{(h)}(\omega)$ assume i valori $0, 1, \ldots, n$.

(2.5) Teorema.

$$\mathrm{P}_{\,n}(\,\nu_n^{(h)} = k\,) = \binom{n}{k} p_h^k (1-p_h)^{n-k}, \quad \mathrm{E}\,\nu_n^{(h)} = n\,p_h, \quad \mathrm{Var}\,\nu_n^{(h)} = n\,p_h(1-p_h).$$

Dimostrazione. Dati k indici $1 \le i_1 < i_2 < \cdots < i_k \le n$ consideriamo il cilindro

$$C_{h\,;\,i_1,\ldots,i_k} = \{\omega \,:\, \omega_i = x_h, \quad \text{se} \quad i = i_j, \ \ j = 1, \ldots, k, \ \ \omega_i \neq x_h, \ \text{altrimenti}\,\}$$

E' evidente che $\omega \in C_{h\,;\,i_1,\ldots,i_k}$ implica $\nu_n^{(h)}(\omega) = k$. Ragionando come nella dimostrazione di (2.2) e tendendo conto del fatto che $\sum_{\omega_i \neq x_h} p(\omega_i) = 1 - p_h$ si ottiene facilmente che $\mathrm{P}_{\,n}(C_{h\,;\,i_1,\ldots,i_m}) = p_h^k (1-p_h)^{n-k}$. D'altra parte, come abbiamo visto in (1.11), vi sono $\binom{n}{k}$ modi diversi di realizzare l'evento $\{\nu_n^{(h)} = k\}$, corrispondenti ad altrettanti cilindri disgiunti della forma $C_{h\,;\,i_1,\ldots,i_k}$.
Facendo riferimento alla (2.4) e usando (2.2) si trova $\mathrm{E}\,1_h^{(i)} = \mathrm{P}_{\,n}(C^{\,i}_{x_h}) = p_h$. Dalla (2.4) si ha allora

$$\mathrm{E}\,\nu_n^{(h)} = \sum_{i=1}^{n} \mathrm{E}\,1_h^{(i)} = n\,p_h.$$

Inoltre, usando il fatto che $(1_h^{(i)})^2 = 1_h^{(i)}$, si trova

$$\begin{aligned}
\mathrm{Var}\,\nu_n^{(h)} &= \mathrm{E}\left(\sum_{i=1}^{n} 1_h^{(i)}\right)^2 - n^2\,p_h^2 = \sum_{i_1, i_2} \mathrm{E}\,1_h^{(i_1)} 1_h^{(i_2)} - n^2\,p_h^2 \\
&= \sum_i \mathrm{E}\,\left(1_h^{(i)}\right)^2 + \sum_{i_1 \neq i_2} \mathrm{E}\,1_h^{(i_1)} 1_h^{(i_2)} - n^2\,p_h^2 \\
&= n\,p_h + \sum_{i_1 \neq i_2} \mathrm{P}_{\,n}(C^{\,i_1,\,i_2}_{x_h, x_h}) - n^2\,p_h^2 \\
&= n\,p_h + n\,(n-1)\,p_h^2 - n^2\,p_h^2 = n\,p_h(1-p_h). \qquad \text{q.e.d.}
\end{aligned}$$

(2.6) Corollario. Sia $f_n^{(h)} = \nu_n^{(h)}/n$. Allora

$$\mathrm{E}\, f_n^{(h)} = p_h, \qquad \mathrm{Var}\, f_n^{(h)} = \frac{p_h(1-p_h)}{n}$$

(2.7) Definizione. Dato un numero $0 \le p \le 1$, la distribuzione di una variabile aleatoria ξ, che assume i valori $k = 0, 1, \ldots, n$, data da

$$\mathrm{P}\,(\xi = k) = \binom{n}{k} p^k (1-p)^{n-k}$$

si chiama *distribuzione binomiale* con parametri $p, q = 1 - p$.

(2.8) Legge dei grandi numeri. Per ogni $\delta > 0$ si ha

$$\mathrm{P}_n(\{\omega \,:\, |\, f_n^{(h)}(\omega) - \mathrm{E}\, f_n^{(h)}\,| \ge \delta \text{ per almeno un } h,\ 1 \le h \le N\}) \to 0,\ n \to \infty\,.$$

Dimostrazione. Poniamo $A_h = \{\,|\, f_n^{(h)} - \mathrm{E}\, f_n^{(h)}\,| \ge \delta, \}$. Allora vogliamo stimare $\mathrm{P}_n\left(\cup_{h=1}^N A_h\right)$. Da (1.32) e (2.6) si ha $\mathrm{P}_n(A_h) \le p_h(1-p_h)/n\delta^2$. Dunque, usando la (1.3-iv), troviamo

$$\mathrm{P}_n\left(\cup_{h=1}^N A_h\right) \le \sum_{h=1}^N \mathrm{P}_n(A_h) \le \frac{N\, \max_h\, p_h(1-p_h)}{n\,\delta^2} \to 0, \quad n \to \infty \qquad \text{q.e.d.}$$

(2.9) Definizione. Sia $(\Omega_n, \mathcal{F}_n, \mathrm{P}_n)$ come sopra e ξ_n una variabile aleatoria definita su Ω_n, $n = 1, 2, \ldots$. Si dirà che la successione $\{\xi_n\}_{n \in \mathbb{Z}_+}$ converge *in probabilità* a un limite $l \in \mathbb{R}$ per $n \to \infty$ se $\forall \delta > 0$ si ha

$$\lim_{n \to \infty} \mathrm{P}_n(\{\omega \in \Omega_n \,:\, |\xi_n(\omega) - l| \ge \delta\}) = 0\,.$$

Ora osserviamo che dalla (2.4) si ha

$$f_n^{(h)}(\omega) = \frac{1}{n} \sum_{i=1}^n 1_h^{(i)}(\omega) \tag{2.10}$$

ovvero $f_n^{(h)}$ è la media aritmetica delle variabili aleatorie $1_h^{(i)}$, $i = 1, \ldots, n$. La (2.8) può allora essere espressa dicendo che le successioni delle medie aritmetiche (2.10) convergono in probabilità ai limiti p_h quando $n \to \infty$. Si tratta di un caso particolare di un *teorema ergodico* generale, sul comportamento asintotico delle medie aritmetiche di variabili aleatorie, valido sotto ipotesi molto più ampie.

(2.11) Applicazione alla teoria dei numeri. Una semplice applicazione della (2.8) alla teoria dei numeri è la seguente. Dato un numero reale $x \in [0,1]$ consideriamo il suo sviluppo in base N:

$$x = \omega_1 \cdot N^{-1} + \omega_2 \cdot N^{-2} + \cdots \tag{2.12}$$

dove ad esempio $N = 2$ o $N = 10$. La (2.12) stabilisce una corrispondenza $\pi : \Omega_\infty \to [0,1]$ tra lo spazio Ω_∞ delle sequenze infinite $\omega = \omega_1\omega_2 \ldots$, $\omega_i \in X = \{0, 1, \ldots N-1\}$ e i punti di $[0,1]$. Tale corrispondenza non è biunivoca: vi sono numeri, i 'razionali' N-adici, della forma r/N^s con r, s interi, che possono essere rappresentati da due sequenze. Ad esempio se $N = 10$ e $x = 1/10$ allora se $\omega^{(1)} = 1000\ldots$ e $\omega^{(2)} = 0999\ldots$ si ha $\pi(\omega^{(1)}) = \pi(\omega^{(2)}) = x$. Possiamo tuttavia definire π^{-1} anche su tale insieme scegliendo sempre, ad esempio, la sequenza che non termina (così $\pi^{-1}(1/10) = 0999\ldots$). Fissiamo ora n simboli (cifre) $a_1, \ldots, a_n$, $a_i \in X$ e consideriamo tutti i numeri $x \in [0,1]$ con la proprietà di avere i primi n simboli nello sviluppo (2.12) coincidenti con gli a_i:

$$I_{a_1,\ldots,a_n} = \{x \in [0,1] \, : \, x = \pi(\omega), \; \omega_i = a_i, 1 \le i \le n\} \tag{2.13}$$

Se $x \in I_{a_1,\ldots,a_n}$ allora i casi estremi si hanno quando $\omega_i = 0$ oppure $\omega_i = N-1$, $i > n$. Il primo caso lo scartiamo, in accordo con la convenzione stabilita sopra. Così

$$\sum_{i=1}^{n} \frac{a_i}{N^i} < x \le \sum_{i=1}^{n} \frac{a_i}{N^i} + \sum_{i=n+1}^{\infty} \frac{N-1}{N^i}. \tag{2.14}$$

È facile rendersi conto che l'ultimo termine della (2.14) è uguale a $1/N^n$ e quindi

$$I_{a_1,\ldots,a_n} = \left(\sum_{i=1}^{n} \frac{a_i}{N^i} \, , \, \sum_{i=1}^{n} \frac{a_i}{N^i} + \frac{1}{N^n} \right] . \tag{2.15}$$

Osserviamo inoltre che, con la notazione (2.1), si ha $I_{a_1,\ldots,a_n} = \pi(C^{1,\,\ldots,\,n}_{a_1,\ldots,a_n})$ con $C^{1,\,\ldots,\,n}_{a_1,\ldots,a_n} \in \Omega_n$. Se ora assegnamo una distribuzione uniforme su X, cioè un vettore $p = (\frac{1}{N}, \ldots, \frac{1}{N})$, allora da (2.2) otteniamo

$$\mathrm{P}_n(C^{1,\,\ldots,\,n}_{a_1,\ldots,a_n}) = N^{-n} = |I_{a_1,\ldots,a_n}|. \tag{2.16}$$

L'evento $\{\omega \in \Omega_n : \nu_n^{(h)}(\omega) = k\}$ corrisponde all'insieme dei numeri $x \in [0,1]$ con la proprietà che, tra i primi n simboli nello sviluppo (2.12), la cifra h compare esattamente k volte. Fissato $\delta > 0$ ma per altro arbitrario, poniamo

$$\Delta_{n,\delta} = \left\{ x \in [0,1] \,:\, x = \pi(\omega),\ \left| \frac{\nu_n^{(h)}(\omega)}{n} - \frac{1}{N} \right| \le \delta,\ h = 1, \ldots, N \right\} \tag{2.17}$$

allora, da (2.8) e (2.17) segue che

$$\lim_{n \to \infty} |\Delta_{n,\delta}| = 1\,. \tag{2.18}$$

Se ad esempio $N = 10$, ciò significa che nello sviluppo decimale di un numero 'tipico', tutte le cifre da 0 a 9 si incontrano approssimativamente con la stessa frequenza 1/10. Vedremo più avanti che si può dare un risultato anche più forte.

(2.19) Problema. Quale tra i seguenti risultati è più probabile: 1) che lanciando 6 volte un dado risulti almeno una volta 'sei', 2) che lanciandolo 12 volte risulti almeno due volte 'sei', 3) che lanciandolo 18 volte risulti almeno tre volte 'sei'? Sia $X = \{1, 2, \ldots, 6\}$, $p = (\frac{1}{6}, \ldots, \frac{1}{6})$ e $\Omega_n = X^{[1,n]}$ (n lanci di un dado). Consideriamo gli eventi $A_{k,n} = \{\omega \in \Omega_n \,:\, \nu_n^{(6)} \ge k\}$. Allora dobbiamo calcolare $\mathrm{P}_n(A_{k,n})$ per $(k,n) = (1,6), (2,12), (3,18)$. Da (2.5) si ha $\mathrm{E}\,\nu_6^{(6)} = 1$, $\mathrm{E}\,\nu_{12}^{(6)} = 2$ e $\mathrm{E}\,\nu_{18}^{(6)} = 3$. Si potrebbe allora pensare che le probabilità in questione siano tutte uguali, ad esempio uguali a 1/2. Vediamo le cose più in profondità. È evidente che $A_{k,n} = \cup_{j=k}^n C_{j,n}$ dove $C_{j,n} = \{\omega \in \Omega_n \,:\, \nu_n^{(6)} = j\}$ e $C_{i,n} \cap C_{j,n} = \emptyset$ se $i \neq j$. Pertanto, usando ancora (2.5) otteniamo

$$\begin{aligned} \mathrm{P}_n(A_{k,n}) &= \sum_{j=k}^{n} \mathrm{P}_n(C_{j,n}) = \sum_{j=k}^{n} \binom{n}{j} \left(\frac{1}{6}\right)^j \left(\frac{5}{6}\right)^{n-j} \\ &= 1 - \sum_{j=0}^{k-1} \binom{n}{j} \left(\frac{1}{6}\right)^j \left(\frac{5}{6}\right)^{n-j} \end{aligned}$$

Vediamo alcuni valori numerici. Poniamo $p(k,n) \equiv \mathrm{P}_n(A_{k,n})$, allora si trova: $p(1,6) = 0.665$, $p(2,12) = 0.619$, $p(3,18) = 0.597$, $p(4,24) = 0.584$, $p(5,30) = 0.576$, $p(16,96) = 0.542$, $p(100,600) = 0.517$, $p(150,900) = 0.514$.
Dunque, in particolare, tra risultati sopra elencati il primo risulta il più probabile.
L'interpretazione di questo fatto si basa sull'osservazione che se $p \neq q$, la distribuzione

binomiale (2.7) non è simmetrica rispetto al punto np (nel nostro caso $p = 1/6$). Lo diviene solo quando $n \to \infty$ come ora andiamo a vedere.

Consideriamo dunque la variabile aleatoria $\nu_n^{(h)}$ definita in (2.3). Fissato $h \in \{1, \ldots, N\}$ adottiamo la notazione semplificata:

$$p = p_h, \qquad q = 1 - p_h, \qquad W_k = \binom{n}{k} p^k (1-p)^{n-k}. \tag{2.20}$$

Da (2.5) e (2.6) segue allora che $\mathrm{P}_n(\,\nu_n^{(h)} = k\,) = W_k$, $\mathrm{E}\,\nu_n^{(h)} = n\,p$ e $\mathrm{Var}\,\nu_n^{(h)} = n\,p\,q$. Generalizzando quanto fatto in (1.11) e (1.33) per il caso particolare $p = q = 1/2$, studiamo come si comportano le probabilità W_k in funzione di k nel limite $n \to \infty$. Vedremo che esiste un dominio relativamente piccolo di valori di k in cui le W_k sono confrontabilmente grandi e un rimanente dominio in cui le W_k sono trascurabilmente piccole. Innnazitutto osserviamo che

$$\frac{W_{k+1}}{W_k} = \frac{\binom{n}{k+1}}{\binom{n}{k}} \cdot \frac{p}{q} = \frac{(n-k)}{(k+1)} \cdot \frac{p}{q} \tag{2.21}$$

che è > 1 o < 1 a seconda che $k < np - q$ o $k > np - q$, come è facile verificare. Dunque ci aspetteremo che il massimo della distribuzione W_k si trovi attorno al punto np. Usando la (1.15) si ha

$$\begin{aligned} W_k &= \frac{n^n\, e^{-n}\, \sqrt{2\pi n}}{k^k\, e^{-k}\, \sqrt{2\pi k}\, (n-k)^{n-k}\, e^{-n+k}\, \sqrt{2\pi (n-k)}}\, p^k\, q^{n-k}\, (\,1 + \epsilon(n,k)\,) \\ &= \frac{1}{\sqrt{2\pi n \frac{k}{n} \left(\frac{n-k}{n}\right)}} \left(\frac{k}{n}\right)^{-k} \left(\frac{n-k}{n}\right)^{-n+k} p^k\, q^{n-k}\, (\,1 + \epsilon(n,k)\,) \\ &= \frac{1}{\sqrt{2\pi n t\,(1-t)}}\, e^{-nH(t)}\, (\,1 + \epsilon(n,k)\,) \end{aligned} \tag{2.22}$$

con $\epsilon(n,k) = \mathcal{O}(1/n) + \mathcal{O}(1/k) + \mathcal{O}(1/(n-k))$ e dove si è posto $t = k/n$ e

$$H(t) = t \log \frac{t}{p} + (1-t) \log \frac{1-t}{q}. \tag{2.23}$$

Osserviamo che $H(p) = 0$. In particolare prendendo $k = np$ (supponendo $np \in \boldsymbol{Z}$, altrimenti dovremmo prendere $k = [np]$) si trova

$$W_{np} = \frac{1}{\sqrt{2\pi npq}} \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right)$$

che generalizza la (1.16). Osserviamo ora che $\sqrt{\mathrm{Var}\,\nu_n^{(h)}} = \sqrt{n\,p\,q}$. In accordo con quanto fatto al punto (1.33), introduciamo la variabile aleatoria normalizzata

$$\eta_n^{(h)} := \frac{f_n^{(h)} - \mathrm{E}\, f_n^{(h)}}{\sqrt{\mathrm{Var}\, f_n^{(h)}}} \tag{2.24}$$

per cui si ha

$$W_k = \mathrm{P}_n(\,\nu_n^{(h)} = k\,) = \mathrm{P}_n(\,\eta_n^{(h)} = x\,) \quad \text{con} \quad x = \frac{k-np}{\sqrt{n\,p\,q}}, \tag{2.25}$$

e inoltre, dato $\alpha > 0$, consideriamo i valori di t corrispondenti a $|x| \le \alpha$, ossia $|t-p| \le \alpha\sqrt{p\,q/n}$. In tali circostanze, poniamo $t = p + x\sqrt{p\,q/n}$ e $1-t = q - x\sqrt{p\,q/n}$ cosicché

$$t(1-t) = pq\left(1 + \frac{(q-p)x}{\sqrt{npq}} - \frac{x^2}{n}\right)$$

e dunque per il primo fattore della (2.22) otteniamo:

$$\frac{1}{\sqrt{2\pi n t\,(1-t)}} = \frac{1}{\sqrt{2\pi npq}}\,\left(1 - \frac{(q-p)x}{2\sqrt{npq}} + \mathcal{O}\left(\frac{1}{n}\right)\right).$$

Ora se $0 < t < 1$ si ha

$$H'(t) = \log\frac{t}{p} - \log\frac{1-t}{q}, \quad H''(t) = \frac{1}{t} + \frac{1}{1-t}, \quad H'''(t) = -\frac{1}{t^2} + \frac{1}{(1-t)^2},$$

da cui otteniamo

$$H'(p) = 0, \quad H''(p) = \frac{1}{pq}, \quad H'''(p) = \frac{p-q}{p^2q^2}.$$

Dunque, ponendo $t = p + x\sqrt{p\,q/n}$ come sopra e usando la formula di Taylor in $t = p$ (cioè $x = 0$) troviamo

$$H(t) = \frac{x^2}{2n} - \left(\frac{q-p}{\sqrt{npq}}\right)\frac{x^3}{6n} + \mathcal{O}(|t-p|^4)$$

Da cui,

$$e^{-nH(t)} = e^{-\dfrac{x^2}{2}}\left(1 + \frac{(q-p)x^3}{6\sqrt{npq}} + \mathcal{O}\left(\frac{1}{n}\right)\right).$$

Inserendo queste espressioni nella (2.22) ricaviamo il *teorema limite locale* (De-Moivre-Laplace):

(2.26) Teorema. Se $|x| = \left|(k-np)/\sqrt{npq}\right| \leq \alpha$ allora

$$W_k = \frac{e^{-x^2/2}}{\sqrt{2\pi npq}} \left(1 + \frac{(q-p)(x^3-3x)}{6\sqrt{npq}} + \mathcal{O}\left(\frac{1}{n}\right)\right).$$

In altre parole, le probabilità W_k, con k che varia entro un intervallo di ampiezza $\mathcal{O}\left(\sqrt{n}\right)$ attorno al valor medio np, sono approssimate dai valori della distribuzione gaussiana $p(k) = e^{-(k-m)^2/2\sigma^2}/\sqrt{2\pi\sigma}$ con $m = np$ e $\sigma = \sqrt{npq}$. Un'immediata conseguenza è il seguente *teorema limite integrale* (De-Moivre-Laplace):

(2.27) Teorema. Sia $R_{np}(\alpha) = \{k : |k-np| \leq \alpha\sqrt{n\,p\,q}\}$, allora

$$\lim_{n\to\infty} \sum_{k\in R_{np}(\alpha)} W_k = \frac{1}{\sqrt{2\pi}} \int_{-\alpha}^{\alpha} e^{-x^2/2}\, dx\,.$$

Dimostrazione. Poniamo $x_k = (k-np)/\sqrt{npq}$, cosicchè $\Delta x = x_{k+1} - x_k = 1/\sqrt{npq}$. Allora da (2.27) si ha

$$\sum_{k\in R_{np}(\alpha)} W_k = \sum_{|x|\leq\alpha} \frac{e^{-x^2/2}}{\sqrt{2\pi}}\, \Delta x \left(1 + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)\right)$$

e il risultato segue dalle considerazioni già svolte nella sezione precedente. q.e.d.

(2.28) Osservazione. Dalle (2.24), (2.25) e (2.27) si evince che la variabile $\eta_n^{(h)}$ non ha un limite definito quando $n \to \infty$, ma ammette piuttosto una distribuzione limite, e tale distribuzione limite è quella con densità normale, cioè gaussiana con parametri $0, 1$. Ciò si può esprimere dicendo che $\eta_n^{(1)}$ converge *in distribuzione* a una variabile aleatoria distribuita con densità normale.

(2.29) Osservazione. Nella dimostrazione di (2.26) e (2.27) non abbiamo fatto uso della legge dei grandi numeri (2.10). Inversamente, è facile vedere che (2.27) implica (2.10). Sia $B \in X$ uno dei possibili risultati di un esperimento e poniamo $\mathrm{P}\,(B) = p$. Sia f_B la frequenza relativa dell'evento B in una sequenza di n prove indipendenti dell'esperimento. Mostriamo che dati $\epsilon > 0$ e $\delta > 0$, si può trovare un $n_0 = n_0(\epsilon, \delta)$ tale che per ogni $n \geq n_0$ si abbia $\mathrm{P}_{\,n}(|f_B - p| \leq \delta) \geq 1 - \epsilon$. In effetti, si ha $\mathrm{P}_n(|f_B - p| \leq \delta) = \sum_{|k-np|\leq n\delta} W_k$. Scegliamo un numero $\beta > 0$ tale che $\frac{1}{\sqrt{2\pi}}\int_{-\beta}^{\beta} e^{-x^2/2}dx \geq 1 - \epsilon/2$. Allora, per $n \geq \beta^2 pq/\delta^2 \equiv n_1$ si ha $\beta\sqrt{npq} < n\delta$ e dunque $\mathrm{P}_n(|f_B - p| \leq \delta) \geq \sum_{|k-np|\leq\beta\sqrt{npq}} W_k$. Da (2.27) segue che quest'ultima tende a $\frac{1}{\sqrt{2\pi}}\int_{-\beta}^{\beta} e^{-x^2/2}dx$. Dunque esiste un n_2 tale che per $n \geq n_2$ la somma in questione è $\geq \frac{1}{\sqrt{2\pi}}\int_{-\beta}^{\beta} e^{-x^2/2}dx - \epsilon/2$, e dunque basta scegliere $n_0 = \max\{n_1, n_2\}$ per ottenere l'asserto.

(2.30) Problema. Con riferimento al punto (2.29) risolviamo ora il seguente problema, tipico della statistica matematica: qual'è il più piccolo numero di prove necessarie affinchè

$$\mathrm{P}_n(|f_B - p| \leq \delta) \geq 1 - \epsilon$$

dove ϵ e δ sono numeri assegnati? Sappiamo che $P_n(f_B = \frac{k}{n}) = \binom{n}{k}p^k(1-p)^{n-k}$ cosicchè $\mathrm{E}\, f_B = p$ e $\mathrm{Var}\, f_B = pq/n$. La seconda disuguaglianza di Chebyshev ci dà quindi

$$\mathrm{P}_n(|f_B - p| \leq \delta) \geq 1 - \frac{pq}{n\delta^2} \geq 1 - \frac{1}{4n\delta^2}$$

dove nell'ultima disuguaglianza (utile quando non si conosca il valore di p) si è usato il fatto che $4pq \leq 1$. Pertanto dovrà essere

$$n \geq \frac{\epsilon^{-1}}{4\delta^2}.$$

Se ad esempio $\delta = 0.02$ e $\epsilon = 0.05$ troviamo $n \geq 12.500$.
D'altra parte dal teorema di De-Moivre-Laplace abbiamo

$$\mathrm{P}_n(|f_B - p| \leq \delta) = \mathrm{P}_n\left(\left|\frac{f_B - p}{\sqrt{pq/n}}\right| \leq \delta\sqrt{\frac{n}{pq}}\right) \simeq \frac{1}{\sqrt{2\pi}}\int_{-\delta\sqrt{\frac{n}{pq}}}^{\delta\sqrt{\frac{n}{pq}}} e^{-x^2/2}dx$$

Procedendo come in (2.29) sia $\beta = \beta(\epsilon) > 0$ tale che $\frac{1}{\sqrt{2\pi}}\int_{-\beta}^{\beta} e^{-x^2/2}dx = 1 - \epsilon$. Ora, essendo $\delta\sqrt{\frac{n}{pq}} \geq 2\delta\sqrt{n}$, se definiamo n come l'intero più piccolo per cui $2\delta\sqrt{n} \geq \beta(\epsilon)$ otteniamo la disuguaglianza cercata. Si ha dunque la condizione

$$n \geq \frac{\beta^2(\epsilon)}{4\delta^2}.$$

Se ancora $\delta = 0.02$ e $\epsilon = 0.05$ troviamo $\beta(\epsilon) = 1.960$ e dunque $n \geq 2.500$, che è decisamente più piccolo di 12.500! Le quantità da confrontare sono $1/\epsilon$ e $\beta(\epsilon)$

(ottenibile dalle tavole). Ad esempio si trova

ϵ	$1/\epsilon$	$\beta(\epsilon)$
0.5	2	0.675
0.3173	3.1516	1
0.1	10	1.645
0.01	100	2.576

(2.31) Passeggiate aleatorie I. Sia $X = \{-1, 1\}$, $p(1) = p$, $p(-1) = q$, $\Omega_n = X^{[1,n]}$. Un punto di Ω_n sarà dunque una parola $\omega = \omega_1 \dots \omega_n$ con $\omega_i = \pm 1$. Consideriamo la sequenza di variabili aleatorie ξ_k, $0 \le k \le n$, definite su Ω_n come segue:

$$\xi_0(\omega) = 0, \qquad \xi_k(\omega) = \sum_{i=1}^{k} \omega_i \tag{2.32}$$

Dato $\omega \in \Omega_n$ tracciamo sul piano (t, y) la spezzata $y(t)$ passante per i punti $(k, \xi_k(\omega))$, $0 \le k \le n$. Da $y(0) = 0$, $y(k) - y(k-1) = \omega_k$ segue che $y(t)$ definisce ω in modo unico. E' inoltre evidente che $y(k)$ ha la stessa parità di k. La funzione $y(t)$ prende il nome di *passeggiata aleatoria* o anche *cammino aleatorio*. In effetti possiamo considerare $y(t)$ come la traiettoria di un oggetto o di una persona che si muove di moto rettilineo ma che, ad intervalli di tempo regolari, ad esempio ogni secondo, cambia la sua direzione in modo arbitrario, formando angoli di $\pm 45°$ con l'orizzontale. Il cambiamento di direzione può essere pensato come il risultato del lancio di una moneta, o in modo più pittoresco come l'effetto di una sbornia (passeggiata dell'ubriaco). Un'altra interpretazione è la seguente: immaginiamo un giocatore, che inizi a giocare contro una casa da gioco (con risorse illimitate) partendo con una cifra iniziale unitaria, perdendo o vincendo un'unità di denaro, con probabilità p e q rispettivamente, ad ogni giocata. Allora $y(n)$ è l'ammontare della vincita (o della perdita) dopo n giocate. Definiamo una distribuzione di probabilità sui cammini aleatori ponendo

$$\mathrm{P}_n(\{y(t),\, 0 \le t \le n\}) = p(\omega) = \prod_{i=1}^{n} p(\omega_i) = p^{\nu_n^{(1)}(\omega)} q^{\nu_n^{(-1)}(\omega)}. \tag{2.33}$$

Le funzioni $\nu_n^{(1)}$ e $\nu_n^{(-1)}$ forniscono in questo caso il numero di tratti elementari della spezzata formanti un angolo di $+45°$ con l'orizzontale, e il numero di quelli formanti un angolo di $-45°$, rispettivamente. Evidentemente si ha $\nu_n^{(1)}(\omega) + \nu_n^{(-1)}(\omega) = n$ e dunque

$$y(n) = \nu_n^{(1)}(\omega) - \nu_n^{(-1)}(\omega) = 2\nu_n^{(1)}(\omega) - n. \tag{2.34}$$

Allora si ha

$$\mathrm{E}\, y(n) = \mathrm{E}\,(2\nu_n^{(1)} - n) = 2\mathrm{E}\,\nu_n^{(1)} - n = (2p-1)n = (p-q)\,n \tag{2.35}$$

e, usando (1.31),

$$\mathrm{Var}\, y(n) = \mathrm{Var}\,(2\nu_n^{(1)} - n) = 4\,\mathrm{Var}\,\nu_n^{(1)} = 4\,n\,p\,q. \tag{2.36}$$

In particolare, se $p = q = 1/2$ si ha una *passeggiata aleatoria simmetrica* per cui si ha $\mathrm{E}\, y(n) = 0$ e $\mathrm{Var}\, y(n) = n$. Osserviamo che $\sqrt{\mathrm{Var}\, y(n)} = \sqrt{n}$. Dunque, sulla base di quanto visto sopra, ci aspetteremo che i valori tipici di $y(n)$ si trovino in un dominio $\mathcal{O}(\sqrt{n})$. Più precisamente, usando la (2.34) e (2.26), troviamo

$$\begin{aligned} \mathrm{P}_n(|y(n)| \le \alpha\sqrt{n}) &= \mathrm{P}_n(|2\nu_n^{(1)} - n| \le \alpha\sqrt{n}) \\ &= \mathrm{P}_n(|\nu_n^{(1)} - \frac{n}{2}| \le \alpha\frac{\sqrt{n}}{2}) \\ &= \mathrm{P}_n(|\nu_n^{(1)} - np| \le \alpha\sqrt{npq}) \\ &\longrightarrow \frac{1}{\sqrt{2\pi}} \int_{-\alpha}^{\alpha} e^{-x^2/2}\,dx\,. \end{aligned} \tag{2.37}$$

In altre parole, durante una *lunga* passeggiata aleatoria (simmetrica), la probabilità di trovarci a una distanza $\mathcal{O}(\sqrt{n})$ dall'origine è strettamente positiva. Si tratta di una caratteristica tipica dei *processi diffusivi*.

(2.38) Passeggiate aleatorie II. Consideriamo ancora la passeggiata aleatoria simmetrica. Definiamo innanzitutto il tempo τ_{2n} di primo ritorno in zero nei primi $2n$ passi:

$$\tau_{2n} = \min\{\, 1 \le k \le 2n \,:\, y(k) = 0 \,\} \tag{2.39}$$

ponendo $\tau_{2n} = \infty$ se $y(k) \neq 0$ for $1 \leq k \leq 2n$. Inoltre, per $0 \leq k \leq n$, poniamo

$$u_{2k} = \mathrm{P}_{2n}(\{y(2k) = 0\}), \qquad f_{2k} = \mathrm{P}_{2n}(\{\tau_{2n} = 2k\}).$$

Evidentemente si ha $u_0 = 1$ e $u_{2k} = \binom{2k}{k} 2^{-2k}$. Una relazione interessante che lega u_{2k} a f_{2k} si ottiene usando un argomento di 'primo passaggio': per andare da 0 a 0 in $2k$ passi (per la prima volta o no), sarà necessario raggiungere 0 per la prima volta dopo $2r$ passi e quindi andare da 0 a 0 in $2(k-r)$ passi, e ciò potrà accadere in k modi diversi (cioè per $1 \leq r \leq k$). Ciò conduce all'espressione

$$u_{2k} = \sum_{r=1}^{k} f_{2r} \cdot u_{2(k-r)}, \tag{2.40}$$

la cui derivazione rigorosa richiede la nozione di probabilità condizionata[1].
Mostriamo ora che, inversamente,

$$f_{2k} = u_{2(k-1)} - u_{2k} = \frac{1}{2k} u_{2(k-1)}. \tag{2.41}$$

Innanzitutto osserviamo che

$$\begin{aligned} u_{2k} &= \binom{2k}{k} 2^{-2k} = \frac{2k}{k} \binom{2k-1}{k-1} \frac{2^{-2(k-1)}}{4} \\ &= \frac{(2k-1)}{2k} \binom{2(k-1)}{k-1} 2^{-2(k-1)} = \frac{(2k-1)}{2k} u_{2(k-1)} \end{aligned}$$

da cui segue la seconda identità in (2.41). Inoltre si ha

$$\begin{aligned} \{\tau_{2n} = 2k\} &= \{\tau_{2n} > 2(k-1)\} \setminus \{\tau_{2n} > 2k\}, \\ \{\tau_{2n} > 2l\} &= \{y(1) \neq 0, \ldots, y(2l) \neq 0\} \end{aligned}$$

e dunque

$$\{\tau_{2n} = 2k\} = \{y(1) \neq 0, \ldots, y(2(k-1)) \neq 0\} \setminus \{y(1) \neq 0, \ldots, y(2k) \neq 0\},$$

ovvero

$$f_{2k} = \mathrm{P}_{2n}(\{y(1) \neq 0, \ldots, y(2(k-1)) \neq 0\}) - \mathrm{P}_{2n}(\{y(1) \neq 0, \ldots, y(2k) \neq 0\}).$$

[1] Vedi più avanti, (4.50).

Basterà quindi mostrare che

$$N_{2k}(\, y(1) \neq 0, \ldots, y(2k) \neq 0\,) = N_{2k}(\, y(2k) = 0\,),$$

dove $N_l(E)$ indica il numero di passeggiate di lunghezza l che verificano la proprietà E. Evidentemente si ha che

$$N_{2k}(\, y(1) \neq 0, \ldots, y(2k) \neq 0\,) = 2N_{2k}(\, y(1) > 0, \ldots, y(2k) > 0\,).$$

D'altra parte esiste un corrispondenza biunivoca tra i $2k$-cammini positivi e quelli non-negativi che almeno una volta si annullano (cioè $y(i) = 0$ per almeno un $1 \leq i \leq 2k$). Infatti dato un cammino di quest'ultimo tipo $A = (y(1), \ldots, y(2k))$, basta prendere il primo indice i tale che $y(i) = 0$ ed associarvi il cammino $B = (y(1), \ldots, y(i-1), y(i) + 2, \ldots, y(2k) + 2)$, che è evidentemente un cammino positivo. Inversamente, dato un cammino positivo $B = (y(1), \ldots, y(2k))$, basterà prendere l'ultimo indice i tale che $y(i) = 1$ ed associarvi il cammino non-negativo $A = (y(1), \ldots, y(i), y(i+1) - 2, \ldots, y(2k) - 2)$. Ciò mostra che

$$\begin{aligned}2N_{2k}(\, y(1) > 0, \ldots, y(2k) > 0\,) &= N_{2k}(\, y(1) > 0, \ldots, y(2k) > 0\,) + \\ &+ N_{2k}(\, y(1) \geq 0, \ldots, y(2k) \geq 0 \;\text{ e }\; \exists i,\; 1 \leq i \leq 2k, \text{ tale che } y(i) = 0\,).\end{aligned}$$

Ora, supponiamo di avere un cammino $(y(0), y(1), \ldots, y(2k))$ tale che $y(2k) = 0$ e avente un unico minimo (assoluto), diciamo nel punto $(l, -m)$ per qualche $1 \leq l \leq 2k$. Riflettiamo il tratto $(y(0), \ldots, y(l))$ rispetto alla retta verticale passante per $t = l$ e trasliamo il tratto risultante verso destra e verso l'alto in modo da farlo partire dal punto $(2k, 0)$. A questo punto spostiamo l'origine in $(l, -m)$. Il cammino in tal modo ottenuto è positivo (con $y(2k) = 2m$). Inversamente, sia $(y(0), y(1), \ldots, y(2k))$ un cammino positivo con $y(2k) = 2m$. Sia q l'ultimo punto in cui $y(q) = m$. Riflettiamo il tratto $(y(q), \ldots, y(2k))$ rispetto alla retta verticale $t = q$ e trasliamo il tratto risultante verso sinistra e verso il basso in modo da farlo terminare nell'origine $(0, 0)$. Spostiamo ora l'origine in $(q - 2k, m)$. Il cammino in tal modo ottenuto ha un unico minimo in $(2k - q, -m)$ e soddisfa $y(2k) = 0$. Vi è dunque una corrispondenza biunivoca tra i $2k$-cammini che terminano in 0 e con un solo minimo assoluto ed i cammini positivi. Ragionando allo stesso modo si stabilisce una corrispondenza biunivoca tra i cammini che terminano in 0 e con almeno due minimi assoluti e i cammini non-negativi che almeno una volta si annullano. Ciò conclude la dimostrazione di (2.41).

L'aspettazione del tempo di primo di ritorno in zero nei primi $2n$ passi è dunque

$$\mathrm{E}\,\tau_{2n} = \sum_{k=1}^{n} 2k\, f_{2k} = \sum_{k=1}^{n} u_{2(k-1)}. \tag{2.42}$$

Usando l'espressione esplicita di u_{2k} e applicando Stirling otteniamo che $u_{2k} \sim 1/\sqrt{\pi k}$ e dunque $\mathrm{E}\,\tau_{2n}$ cresce come $\sqrt{n}$. In particolare si ha $\sum_{k=1}^{\infty} u_{2(k-1)} = \infty$ e quindi il valor medio del tempo di primo ritorno in 0 (in un numero arbitrario di passi) è ∞. Questo fatto sta alla base di alcune interessanti quanto inaspettate proprietà delle passeggiate aleatorie simmetriche. In particolare, da ciò segue il fatto assai poco intuitivo che una 'tipica' passeggiata aleatoria $(y(0), y(1), \dots, y(n))$ non avrà un carattere oscillante (cosicchè grosso modo la metà del tempo si troverebbe nel semipiano positivo e metà in quello negativo), ma piuttosto avrà l'aspetto di una sorta di onda allungata. La formulazione precisa di questo fatto costituisce la *legge dell'arcoseno*, che ora discuteremo brevemente.

Innanzitutto diremo che durante la sua passeggiata il punto si trova nel semipiano positivo nell'intervallo $[i-1, i]$ se almeno uno dei valori $y(i-1)$ e $y(i)$ è positivo. Introduciamo quindi la probabilità $\mathrm{P}_{2n,2k}$ che durante l'intervallo $[0, 2n]$ il punto trascorra $2k$ istanti di tempo nel semipiano positivo. Mostriamo innanzitutto che

$$\mathrm{P}_{2n,2k} = u_{2k} \cdot u_{2n-2k}, \qquad 0 \le k \le n. \tag{2.43}$$

Se $k = 0$ e $k = n$ la (2.43) si riduce a $\mathrm{P}_{2n,2n} = \mathrm{P}_{2n,0} = u_{2n}$, che segue da quanto visto precedentemente. Sia ora $1 \le k \le n-1$. Se il punto si trova nel semipiano positivo per esattamente $2k$ passi, allora prima o poi deve attraversare lo 0. Sia $2r$ il tempo del primo passaggio attraverso lo 0 (cosicchè $\tau_{2n} = 2r$). Vi sono due possibilità: $y(i) \ge 0$ oppure $y(i) \le 0$, per $i \le 2r$. Il numero di cammini del primo tipo sarà dato da

$$\left(\frac{1}{2} 2^{2r} f_{2r}\right) \cdot 2^{2(n-r)} \mathrm{P}_{2(n-r),2(k-r)} = \frac{1}{2} \cdot 2^{2n} \cdot f_{2r} \cdot \mathrm{P}_{2(n-r),2(k-r)},$$

mentre quelli del secondo tipo saranno

$$\frac{1}{2} \cdot 2^{2n} \cdot f_{2r} \cdot \mathrm{P}_{\,2(n-r),2k}.$$

Pertanto avremo

$$\mathrm{P}_{\,2n,2k} = \frac{1}{2} \sum_{r=1}^{k} f_{2r} \cdot \mathrm{P}_{\,2(n-r),2(k-r)} + \frac{1}{2} \sum_{r=1}^{k} f_{2r} \cdot \mathrm{P}_{\,2(n-r),2k}.$$

Supponendo ora che $\mathrm{P}_{\,2m,2k} = u_{2k} \cdot u_{2m-2k}$ sia valida per $m = 1, \ldots, n-1$ e usando la (2.40) si ha

$$\begin{aligned}
\mathrm{P}_{\,2n,2k} &= \frac{1}{2} u_{2n-2k} \sum_{r=1}^{k} f_{2r} \cdot u_{2k-2r} + \frac{1}{2} u_{2k} \sum_{r=1}^{k} f_{2r} \cdot u_{2n-2r-2k} \\
&= \frac{1}{2} u_{2n-2k} \cdot u_{2k} + \frac{1}{2} u_{2k} \cdot u_{2n-2k} = u_{2k} \cdot u_{2n-2k},
\end{aligned}$$

che è l'espressione cercata. Per regioni di simmetria, un'identica espressione vale per la probabilità che durante l'intervallo $[0, 2n]$ il punto trascorra un tempo $2k$ nel semipiano negativo.

Sia ora $N(2n)$ il *tempo di occupazione* del semipiano positivo, ossia numero di passi (o unità di tempo) in cui il punto si trova nel semipiano positivo nell'intervallo $[0, 2n]$. Per quanto detto avremo evidentemente che

$$E\left(\frac{N(2n)}{2n}\right) = \frac{1}{2}.$$

Inoltre si avrà, per $1/2 < x \leq 1$,

$$\mathrm{P}_{\,2n}\left(\left\{\frac{1}{2} < \frac{N(2n)}{2n} \leq x\right\}\right) = \sum_{\{k\,:\,1/2<(2k/2n)\leq x\}} \mathrm{P}_{\,2n,2k}.$$

Ora, essendo $u_{2k} \sim 1/\sqrt{\pi k}$ quando $k \to \infty$, avremo

$$\mathrm{P}_{\,2n,2k} = u_{2k} \cdot u_{2n-2k} \sim \frac{1}{\pi\sqrt{k(n-k)}}$$

quando $k \to \infty$ e anche $n - k \to \infty$. Pertanto, quando $n \to \infty$ avremo

$$\sum_{\{k\,:\,1/2<(2k/2n)\leq x\}} \mathrm{P}_{\,2n,2k} - \sum_{\{k\,:\,1/2<(2k/2n)\leq x\}} \frac{1}{\pi n} \cdot \left[\frac{k}{n}\left(1 - \frac{k}{n}\right)\right]^{-1/2} \to 0,$$

e dunque,

$$\sum_{\{k\,:\,1/2<(2k/2n)\le x\}} \mathrm{P}_{\,2n,2k} - \frac{1}{\pi}\int_{1/2}^{x} \frac{dt}{\sqrt{t(1-t)}} \to 0,$$

e l'integrale vale $(2/\pi)\arcsin\sqrt{x} - 1/2$. Tenendo conto del fatto che, per simmetria $\sum_{\{k\,:\,k/n\le 1/2\}} \mathrm{P}_{\,2n,2k} \to 1/2$ quando $n\to\infty$ abbiamo mostrato il seguente teorema limite

Teorema. (Legge dell'arcoseno) *La probabilità che la frazione di tempo spesa dal punto nel semipiano positivo sia al più x tende a $2\pi^{-1}\arcsin\sqrt{x}$:*

$$\sum_{\{k\,:\,k/n\le x\}} \mathrm{P}_{\,2n,2k} \to \frac{2}{\pi}\arcsin\sqrt{x}. \tag{2.44}$$

In particolare, ricordando la forma del grafico della funzione integranda in

$$\frac{1}{\pi}\int_{0}^{x} \frac{dt}{\sqrt{t(1-t)}},$$

si vede che è più probabile che la frazione di tempo spesa nel semipiano positivo sia vicina a 0 o a 1 piuttosto che al valore intuitivo 1/2. Detto in altri termini, in un gioco alla pari è più probabile che la vincita del giocatore sia positiva (o negativa) per la maggior parte del tempo, piuttosto che una situazione di equilibrio tra perdita e vincita.

(2.45) Grandi deviazioni. Sia ξ una variabile aleatoria come la (2.10), cioè definita da una media aritmetica $\xi = S_n/n$ dove $S_n = \sum_{k=1}^{n}\eta_k$, e le η_i formano una sequenza di prove indipendenti con $\mathrm{P}(\eta_i = 1) = p$ e $\mathrm{P}(\eta_i = 0) = q$, $\forall i \ge 1$, cosicché $\mathrm{E}\,\xi = p$. Il teorema integrale di De Moivre-Laplace consente di stimare le probabilità degli eventi $\{\xi - p \le \alpha\sqrt{pq}/\sqrt{n}\}$, che caratterizzano le deviazioni 'standard' (di ordine $1/\sqrt{n}$) di ξ da p. Per deviazioni più grandi, di ordine 1, la sola stima che abbiamo a disposizione è quella, piuttosto brutale, fornita dalla disuguaglianza di Chebyshev. Per ottenere stime più accurate possiamo usare la disuguaglianza esponenziale (1.59). A questo scopo, poniamo $\varphi(\lambda) = \mathrm{E}\,e^{\lambda\eta_1}$. Allora si ha $\varphi(\lambda) = 1 - p + pe^{\lambda}$, e data l'indipendenza delle η_i,

$$\mathrm{E}\,e^{\lambda S_n} = [\varphi(\lambda)]^n \quad\Longrightarrow\quad \mathrm{E}\,e^{\lambda\xi} = \left[\varphi(\frac{\lambda}{n})\right]^n. \tag{2.46}$$

Usando la (1.59), otteniamo

$$\begin{aligned}\mathrm{P}\,(\xi \geq \alpha) \leq \inf_{\lambda>0} \mathrm{E}\, e^{\lambda(\xi-\alpha)} &= \inf_{\lambda>0} e^{-n[\alpha\lambda/n - \log\varphi(\lambda/n)]} \\ &= \inf_{s>0} e^{-n[\alpha s - \log\varphi(s)]} = e^{-n\, \sup_{s>0}[\alpha s - \log\varphi(s)]}\end{aligned}$$

In modo simile si ottiene $\mathrm{P}\,(\xi \leq \alpha) \leq e^{-n\,\sup_{s<0}[\alpha s - \log\varphi(s)]}$. Ora, la funzione $g(s) = \alpha s - \log\varphi(s) = \alpha s - \log(1-p+pe^s)$ raggiunge il suo massimo per $p \leq \alpha \leq 1$ nel punto $s_0 (g'(s_0) = 0)$, ovvero $e^{s_0} = \alpha(1-p)/p(1-\alpha)$, e dunque

$$\sup_{s>0} g(s) =: H(\alpha) = \alpha \log\frac{\alpha}{p} + (1-\alpha)\log\frac{1-\alpha}{1-p} \tag{2.47}$$

è la funzione già introdotta in (2.23). Pertanto, se $p \leq \alpha \leq 1$ si ha $\mathrm{P}\,(\xi \geq \alpha) \leq e^{-nH(\alpha)}$. Ora poniamo $\alpha = p + \epsilon$ con $\epsilon > 0$ ed osserviamo che, essendo $4p(1-p) \leq 1$ se $0 \leq p \leq 1$,

$$H(p+\epsilon) = \frac{\epsilon^2}{2p(1-p)} + \mathcal{O}(\epsilon^3) \geq 2\epsilon^2,$$

da cui otteniamo

$$\mathrm{P}\,(\xi - p \geq \epsilon) \leq e^{-2n\epsilon^2}, \tag{2.48}$$

che insieme ad un'analoga disuguaglianza in senso opposto (ottenuta come sopra ponendo $\alpha = p - \epsilon$) dà la seguente disuguaglianza per la *probabilità delle grandi deviazioni*

$$\mathrm{P}\,(|\xi - p| \geq \epsilon) \leq 2\, e^{-2n\epsilon^2}, \tag{2.49}$$

che si può leggere: i valori della distribuzione binomiale al di fuori del dominio di validità del teorema di De-Moivre-Laplace tendono a zero con velocità esponenziale quando $n \to \infty$.

(2.50) La distribuzione di Poisson. In determinate condizioni la distribuzione binomiale può essere approssimata dalla distribuzione di Poisson. Per introdurre questa distribuzione consideriamo un problema concreto. Supponiamo che n palline vengano distribuite a caso tra N urne. Qual'è la probabilità che scegliendo un'urna a caso, essa contenga esattamente k palline? Ora, essendovi N possibilità egualmente probabili per ogni pallina, la probabilità cercata è data dalla distribuzione binomiale con parametri $1/N$, $1-1/N$, ossia

$$W_k = \binom{n}{k}\left(\frac{1}{N}\right)^k\left(1-\frac{1}{N}\right)^{n-k} \tag{2.51}$$

Supponiamo ora di far tendere n ed N all'infinito in modo tale che il rapporto n/N si mantenga uguale a una costante positiva λ. Allora si ha

$$\begin{aligned} W_k &= \frac{n(n-1)\cdots(n-k+1)}{k!}\left(\frac{\lambda}{n}\right)^k\left(1-\frac{\lambda}{n}\right)^{n-k} \\ &= \frac{\lambda^k}{k!}\left(1-\frac{\lambda}{n}\right)^{n-k}\prod_{j=1}^{k-1}\left(1-\frac{j}{n}\right). \end{aligned} \tag{2.52}$$

Ora, qui k è fissato e $\lim_{n\to\infty}\left(1-\frac{\lambda}{n}\right)^n = e^{-\lambda}$. Allora si ha

$$\lim_{\substack{n\to\infty \\ n=\lambda N}} W_k = \mathrm{P}_k = \frac{\lambda^k}{k!}e^{-\lambda}, \qquad (k=0,1,\ldots) \tag{2.53}$$

Dalla serie di potenze per e^{λ} si ha $\sum_{k\geq 0}\mathrm{P}_k = e^{-\lambda}\sum_{k\geq 0}\frac{\lambda^k}{k!} = 1$, dunque i numeri P_k formano una distribuzione di probabilitità su $\mathbb{N}$, detta *distribuzione di Poisson con parametro* λ. Si tratta del primo esempio incontrato finora di una distribuzione discreta concentrata su insieme numerabile di punti. Il significato di λ nel problema discusso sopra è evidentemente quello di numero medio di palline presenti in un'urna. Più in generale, se $\mathrm{P}_k = \mathrm{P}(\xi = k)$ si ha

$$\mathrm{E}\,\xi = e^{-\lambda}\sum_{k\geq 0} k\frac{\lambda^k}{k!} = \lambda e^{-\lambda}\sum_{k\geq 1} k\frac{\lambda^{k-1}}{k!} = \lambda e^{-\lambda}\frac{de^{\lambda}}{d\lambda} = \lambda \tag{2.54}$$

Inoltre

$$\mathrm{E}(\xi^2) = e^{-\lambda}\sum_{k\geq 0} k^2\frac{\lambda^k}{k!} = \lambda + e^{-\lambda}\sum_{k\geq 2} k(k-1)\frac{\lambda^k}{k!} = \lambda + \lambda^2 e^{-\lambda}\frac{d^2e^{\lambda}}{d\lambda^2} = \lambda + \lambda^2$$

e dunque

$$\mathrm{Var}\,\xi = \lambda + \lambda^2 - \lambda^2 = \lambda = \mathrm{E}\,\xi. \tag{2.55}$$

Ora, nella derivazione (2.52)-(2.53), non abbiamo fatto alcun uso del fatto che la probabilità di entrare in una data urna è $p = 1/N$ con $N \in \mathbb{Z}_+$. Inoltre, la condizione $np = \lambda$ può ovviamente essere sostituita da $np \to \lambda$. Sia allora $X = \{0,1\}$ e $\omega = \omega_1 \dots \omega_n$, $\omega_i \in X$. Supponiamo che $p = p(1)$ dipenda da n in modo tale che $\lim_n np = \lambda > 0$. Allora, da (2.5) e da quanto visto sopra, si ha il

(2.56) Teorema limite di Poisson.

$$\mathrm{P}_n(\nu_n^{(1)} = k) \to \frac{\lambda^k}{k!} e^{-\lambda}, \quad \mathrm{E}\, \nu_n^{(1)} = np \to \lambda, \quad \mathrm{Var}\, \nu_n^{(1)} = np(1-p) \to \lambda$$

La condizione $p \sim \lambda/n$ implica che la probabilità di 1 (successo, vincita, etc.) tende a zero come $\mathcal{O}(1/n)$. Inoltre $\mathrm{E}\, \nu_n^{(1)} \sim \lambda$ significa che il numero medio di 1 in una parola di lunghezza n resta finito, in particolare non cresce con n. Per questa ragione (2.56) viene talvolta chiamato *teorema limite per eventi rari.*

(2.57) Distribuzione di Poisson in meccanica statistica. Vediamo ora una interessante generalizzazione dell'esempio delle palline nelle urne. Uno dei modelli matematici più semplici studiati in meccanica statistica è quello del *gas perfetto*. Si tratta di un sistema di n particelle non interagenti poste in una certa regione V, ad esempio un cubo d-dimensionale centrato nell'origine e di lato L, cosicchè $\mathrm{vol}\, V = L^d$. Il *limite termodinamico* per questo sistema consiste nel prendere $L \to \infty$ e $n \to \infty$ in modo tale che il numero di particelle per unità di volume resti finito, cioè $n/L^d \to \lambda > 0$. L'assenza d'interazione tra le particelle si esprime con l'assunzione che la posizione di ogni particella del gas è una variabile uniformemente distribuita in V, indipendentemente dalla posizione delle altre particelle. Più precisamente poniamo $X = V$ e $\Omega_n = X^{[1,n]}$. Ad ogni sottoinsieme $Q \subset V$ associamo la probabilità $p(Q) = \mathrm{vol}\, Q/L^d$. In altre parole, su X è definita una distribuzione con densità costante, che è un'immediata estensione a $\mathbb{R}^d$ della distribuzione introdotta al punto (1.55)-(2), mentre su Ω_n è assegnata la misura prodotto P_n definita come segue: fissati k indici $i_1, \dots, i_k$, definiamo il cilindro

$$C_{i_1,\dots,i_k} = \{\omega : \omega_{i_l} \in Q, 1 \le l \le k, \omega_j \notin Q, j \ne i_1, \dots, i_k\}. \tag{2.58}$$

Allora si ha

$$\mathrm{P}_n(C_{i_1,\dots,i_k}) = \left(\frac{\mathrm{vol}\, Q}{L^d}\right)^k \left(1 - \frac{\mathrm{vol}\, Q}{L^d}\right)^{n-k}. \tag{2.59}$$

Ora è evidente che se $\nu_n^{(Q)}(\omega)$ fornisce il numero di particelle della configurazione ω che si trovano in Q, allora l'evento $\{\omega : \nu_n^{(Q)}(\omega) = k\}$ sarà dato dall'unione degli $\binom{n}{k}$ cilindri della forma (2.58). Troviamo dunque una distribuzione binomiale in cui la probabilità di successo (= particella in Q) è data da $p = p(Q) = \mathrm{vol}\, Q/L^d = n\, \mathrm{vol}\, Q/(nL^d) \sim \lambda\ \mathrm{vol}\, Q/n$, nel limite termodinamico. Da (2.56) abbiamo allora il

(2.60) Teorema. Il numero di particelle di un gas perfetto presenti in un dominio assegnato Q segue, nel limite termodinamico, una distribuzione di Poisson con parametro $\lambda\ \mathrm{vol}\, Q$:

$$\mathrm{P}_n(\nu_n^{(Q)} = k) \to \frac{(\lambda\ \mathrm{vol}\, Q)^k}{k!} e^{-\lambda\, \mathrm{vol}\, Q}, \quad \mathrm{E}\, \nu_n^{(Q)} = \mathrm{Var}\, \nu_n^{(Q)} \to \lambda\ \mathrm{vol}\, Q.$$

(2.61) Osservazione. Questo risultato può essere usato anche in astronomia: dietro l'ipotesi (certamente molto approssimativa) che la densità di stelle nella Via Lattea sia uguale a una costante λ, si ha che la probabilità che una regione Q dello spazio contenga k stelle è data dall'espressione scritta sopra. Vi sono poi altre situazioni, come i processi di disintegrazione radioattiva o le chiamate passanti per una centralina telefonica, che sono descrivibili con una distribuzione poissoniana con parametro proporzionale al tempo t. Ad esempio, non è difficile mostrare che sotto certe ipotesi (non troppo irrealistiche nelle 'ore di punta'), la probabilità $P_k(t)$ che per una data centralina telefonica passino k telefonate nell'intervallo di tempo $t = t_2 - t_1$ è data da $(\lambda\, t)^k\, e^{-\lambda\, t}/k!$ dove $\lambda > 0$ rappresenta il numero medio di chiamate per unità di tempo. Un modo di vedere questo fatto che fa uso degli argomenti già visti è il seguente: suddividiamo l'intervallo di tempo t in n intervallini di tempo di durata τ, cosicchè $t = n\tau$. Supponiamo che le telefonate giungano in modo indipendente le une dalle altre e che durante ciascun intervallino di tempo possa giungere al più una telefonata (dal momento che prenderemo il limite $\tau \to 0$ ciò corrisponde ad escludere più telefonate simultanee). La probabilità di ricevere una telefonata nell'intervallino di tempo τ sarà $p = \lambda\, \tau$.

Dunque, ragionando come nelle (2.52)-(2.53), troviamo

$$P_k(t) = \lim_{\substack{\tau \to 0 \\ t = n\tau}} \binom{n}{k} (\lambda\tau)^k (1-\lambda\tau)^{n-k} = \lim_{n\to\infty} \binom{n}{k} \left(\frac{\lambda t}{n}\right)^k \left(1 - \frac{\lambda t}{n}\right)^{n-k}$$
$$= \frac{(\lambda\, t)^k}{k!}\, e^{-\lambda\, t}.$$

(2.62) Entropia e informazione.

Uno schema finito (X,p) è ottenuto assegnando una distribuzione $p=(p_1,\ldots,p_N)$ ad un insieme finito di possibilità $X = \{x_1,\ldots,x_N\}$. Dunque descrive una situazione di *incertezza*. Se ad esempio abbiamo un esperimento con due risultati possibili, ad esempio $X = \{0,1\}$, allora dei due casi $p = (\frac{1}{2},\frac{1}{2})$ e $p = (\frac{1}{100},\frac{99}{100})$, il primo rappresenta una situazione molto più incerta. Nel secondo caso, è 'quasi certo' che il risultato dell'esperimento sarà 1, mentre nel primo saremo naturalmente riluttanti a fare qualunque previsione. Il caso $p = (\frac{1}{3},\frac{2}{3})$, sarà poi in un qualche modo intermedio tra i due precedenti. Per varie applicazioni è utile introdurre una quantità numerica che *misuri* il grado d'incertezza associato ad un dato schema finito. Ora, la teoria dell'informazione, ispirata dalla meccanica statistica, associa ad un evento di probabilità p una quantità: l'informazione apportata dal verificarsi di tale evento, $I(p)$, data dalla formula

$$I(p) = \log\frac{1}{p} = -\log p. \tag{2.63}$$

Come si vede, l'informazione $I(p)$ cresce quando l'evento diviene più improbabile, ossia quando p decresce. Possiamo allora identificare l'incertezza sul verificarsi di un dato evento, *prima* di eseguire l'esperimento, con l'informazione apportata dall'eventuale verificarsi di tale evento *dopo* l'esecuzione dell'esperimento. In altre parole, l'informazione che otteniamo sulla base del risultato dell'esperimento annulla l'incertezza presente prima di eseguirlo. Il grado d'incertezza di uno schema finito sarà allora misurato dall'informazione media o *entropia*,

$$H(p_1,\ldots,p_N) = \sum_{i=1}^{N} p_h\, I(p_h) = -\sum_{h=1}^{N} p_h\, \log p_h \tag{2.64}$$

con la posizione $x\log x = 0$ se $x=0$. Verifichiamo che $H(p_1,\ldots,p_N)$ ha i requisiti che ci aspettiamo debba soddisfare in quanto misura dell'incertezza. Innanzitutto, dalle condizioni $p_h \geq 0$ e $\sum_h p_h = 1$ segue immediatamente che $H(p_1, \ldots, p_N) = 0$ se e solo se $p_k = 1$ per un certo k e $p_h = 0$ per $h \neq k$, cioè nel caso in cui un risultato elementare sia certo e gli altri impossibili. In tutti gli altri casi si ha $H > 0$. Osserviamo poi che $\phi(x) = x \log x$ è una funzione convessa e pertanto soddisfa la disuguaglianza

$$\phi\left(\frac{1}{N}\sum_{i=1}^{N} a_i\right) \leq \frac{1}{N}\sum_{i=1}^{N} \phi(a_i) \tag{2.65}$$

con $a_1, \ldots, a_N$ non negativi. Se $a_h = p_h$ otteniamo

$$\frac{1}{N}\log\frac{1}{N} = \phi\left(\frac{1}{N}\sum_{h=1}^{N} p_h\right) \leq \frac{1}{N}\sum_{h=1}^{N} p_h \log p_h = -\frac{1}{N}H(p_1, \ldots, p_N) \tag{2.66}$$

da cui segue

$$H(p_1, \ldots, p_N) \leq \log N = H\left(\frac{1}{N}, \ldots, \frac{1}{N}\right). \tag{2.67}$$

La (2.67) afferma che lo schema con maggiore incertezza è quello corrispondente alla distribuzione uniforme. Un'altra proprietà importante dell'entropia è la seguente: dati due schemi (X, p) e (Y, π), consideriamo lo schema prodotto, cioè lo schema i cui eventi elementari hanno la forma $x_i \cdot y_j$ con le rispettive probabilità date da $p_i \cdot \pi_j$. Allora si ha

$$\begin{aligned}
\sum_{i,j} p_i \cdot \pi_j \, \log(p_i \cdot \pi_j) &= \sum_{i,j} p_i \cdot \pi_j \, (\log p_i + \log \pi_j) \\
&= \sum_i p_i \log p_i \sum_j \pi_j + \sum_j \pi_j \log \pi_j \sum_i p_i \\
&= \sum_i p_i \log p_i + \sum_j \pi_j \log \pi_j .
\end{aligned}$$

In altre parole, l'entropia del prodotto di due schemi è la somma delle rispettive entropie. Ora, per ogni n fissato, l'insieme Ω_n delle parole di lunghezza n e la distribuzione di probabilità su di esso data da $p(\omega) = \prod_{i=1}^{n} p(\omega_i)$ con $p(\omega_i) = p_k$ se $\omega_i = x_k$, determinano, nel senso visto sopra, uno schema prodotto, ottenuto moltiplicando (X, p) per se stesso n volte. La sua entropia sarà dunque data da $n \cdot H(p_1, \ldots, p_N)$. D'altra parte, vorremmo definire un'entropia associata a $(\Omega_n, \mathcal{F}_n, \mathrm{P}_n)$ come una quantità definita anche quando $n \to \infty$.

(2.68) Definizione. L'entropia di una sequenza di eventi indipendenti estratti con distribuzione (X, p) è la quantità

$$h = \lim_{n \to \infty} -\frac{1}{n} \sum_{\omega \in \Omega_n} p(\omega) \log p(\omega) = -\sum_{h=1}^{N} p_h \, \log p_h .$$

Una conseguenza della legge dei grandi numeri è il seguente

(2.69) Teorema di Shannon-McMillan. Dati $\epsilon > 0$ e $\delta > 0$, esiste $n_0 = n_0(\epsilon, \delta)$ tale che $\forall n \geq n_0$ è possibile dividere le parole di lunghezza n, cioè gli elementi di Ω_n, in due classi C_n^{rare} e $C_n^{\text{frequenti}}$ in modo che

1. $\mathrm{P}_n(C_n^{\text{rare}}) < \epsilon$;
2. se $\omega \in C_n^{\text{frequenti}}$ allora $e^{-n(h+\delta)} \leq p(\omega) \leq e^{-n(h-\delta)}$;
3. se $e^{n(h-\delta)} \leq |C_n^{\text{frequenti}}| \leq e^{n(h+\delta)}$.

Dimostrazione. Sia $\alpha > 0$, il cui valore verrà specificato tra breve e poniamo

$$C_n^{\text{frequenti}} = \{\omega \,:\, |f_n^{(h)}(\omega) - p_h| \leq \alpha,\ 1 \leq h \leq N\}.$$

Dalla legge dei grandi numeri segue che $\mathrm{P}_n(C_n^{\text{frequenti}}) \to 1$ quando $n \to \infty$. Allora se $n \geq n_1(\alpha, \epsilon)$ si ha $\mathrm{P}_n(C_n^{\text{frequenti}}) \geq 1 - \epsilon$ e dunque $\mathrm{P}_n(C_n^{\text{rare}}) < \epsilon$. Quanto alla proprietà 2, data $\omega = \omega_1 \ldots \omega_n$ si ha

$$\begin{aligned} p(\omega) &= \prod_{i=1}^{n} p(\omega_i) = \prod_{h=1}^{N} p_h^{\nu_n^{(h)}(\omega)} \\ &= \exp\left[\sum_{h=1}^{N} \nu_n^{(h)}(\omega) \log p_h\right] \\ &= \exp\left[-n\left(-\sum_{h=1}^{N} \frac{\nu_n^{(h)}(\omega)}{n} \log p_h\right)\right] \\ &= \exp\left[-n\left(h - \sum_{h=1}^{N} \left(\frac{\nu_n^{(h)}(\omega)}{n} - p_h\right) \log p_h\right)\right] \end{aligned}$$

Ora, se $\omega \in C_n^{\text{frequenti}}$ allora

$$\left|\sum_{h=1}^{N} \left(\frac{\nu_n^{(h)}(\omega)}{n} - p_h\right) \log p_h\right| \leq \alpha \sum_{h=1}^{N} |\log p_h|.$$

La 2 segue allora scegliendo α in modo tale che $\alpha \sum_{h=1}^{N} |\log p_h| \leq \delta/2$. Infine osserviamo che, per quanto visto,

$$e^{-n(h+\delta/2)}|C_n^{\text{frequenti}}| \leq \mathrm{P}_n(C_n^{\text{frequenti}}) \leq e^{-n(h-\delta/2)}|C_n^{\text{frequenti}}|$$

e dunque, da un parte,

$$|C_n^{\text{frequenti}}| \leq e^{n(h+\delta/2)}\mathrm{P}_n(C_n^{\text{frequenti}}) \leq e^{n(h+\delta)}$$

e, dall'altra, se $n \geq n_2(\epsilon, \delta)$ si ha

$$|C_n^{\text{frequenti}}| \geq e^{n(h-\delta/2)}\mathrm{P}_n(C_n^{\text{frequenti}}) \geq e^{n(h-\delta/2)}(1-\epsilon) \geq e^{n(h-\delta)}$$

e l'asserto segue con $n_0 = \max\{n_1, n_2\}$. q.e.d.

Il risultato precedente ha un ruolo significativo in numerosi problemi di teoria dell'informazione e teoria ergodica. Nella sostanza, possiamo argomentare come segue: il numero totale di parole di lunghezza n è N^n. Tra queste, ve n'è un sottoinsieme, con cardinalità dell'ordine di e^{nh}, che ha probabilità arbitrariamente prossima a 1. Inoltre; gli elementi di tale insieme, presi singolarmente, hanno approssimativamente la stessa probabilità, seppure nel senso piuttosto debole della 2, dell'ordine di e^{-nh}.

Ora, se ad esempio $h = \log N$, cioè se la distribuzione di probabilità delle lettere dell'alfabeto X è quella uniforme, allora tale sottoinsieme coincide con l'intero spazio Ω_n e non si dà riduzione possibile. Se invece h è 'significativamente' più piccola di $\log N$ (distribuzione non-uniforme) allora la riduzione dallo spazio Ω_n al sottoinsieme $C_n^{\text{frequenti}}$ corrisponde ad una significativa riduzione dell'informazione necessaria a specificare l'insieme delle parole 'effettivamente osservate'. Non è troppo difficile immaginare la applicazioni che tale risultato può avere quando ad esempio si tratti di comprimere un testo molto lungo allo scopo di trasmetterlo. Ad esempio, se l'alfabeto X contiene N simboli, allora le parole di lunghezza n estratte dal 'testo infinito' $\omega \in \Omega$, possono essere riscritte, salvo 'casi rari', cioè ad eccezione delle parole in C_n^{rare}, in un altro alfabeto X' con N' simboli, purché continui a valere $h \leq \log N'$. Alternativamente, si possono riscrivere le parole di lunghezza n (sempre con le eccezioni di cui sopra) nello stesso alfabeto ma in modo che siano più corte, diciamo di lunghezza n', con la massima compressione possibile data da $n'/n = h/\log N$.

(2.70) Esercizi e complementi.

1. Distribuzione geometrica. Se ripetiamo un esperimento dove in ogni prova si ha probabilità p di successo finchè il successo non arriva, il numero di prove N necessarie ha distribuzione geometrica

$$P(N = n) = (1-p)^{n-1}p.$$

Sia $X = \{-1, 1\}$ con $p(1) = p$ e $p(-1) = q$. Consideriamo ancora lo spazio di probabilità $(\Omega_n, \mathcal{F}_n, \mathrm{P}_n)$ con $\Omega_n = X^{[1,n]}$. Sia $\xi : \Omega_n \to I\!N \cup \{0\}$ la variabile aleatoria definita da $\xi(\omega) = \inf\{1 < k \leq n : \omega_k \neq \omega_1\}$, ove si ponga $\xi(\omega) = 0$ se $\inf\{1 < k \leq n : \omega_k \neq \omega_1\} = \emptyset$. Mostrare che

$$\mathrm{P}_n(\xi = 0) = q^n + p^n \quad \text{e} \quad \mathrm{P}_n(\xi = k) = p\,q^{k-1} + q\,p^{k-1}, \quad k = 2, 3, \ldots, n.$$

2. Distribuzione multinomiale. Sia $(\Omega_n, \mathcal{F}_n, \mathrm{P}_n)$ come in (1.10). La probabilità di una singola parola $\omega \in \Omega_n$ è

$$p(\omega) = \prod_{i=1}^{n} p(\omega_i) = \prod_{h=1}^{N} p_h^{\nu_n^{(h)}(\omega)}$$

e dunque la probabilità dell'evento certo può essere decomposta come

$$\sum_{\omega \in \Omega_n} p(\omega) = \sum_{\substack{n_1 \geq 0, \ldots, n_N \geq 0, \\ n_1 + \cdots + n_N = n}} C_n(n_1, \ldots, n_N) p_1^{n_1} \cdots p_N^{n_N}$$

dove $C_n(n_1, \ldots, n_N)$ è il numero di sequenze (ordinate) $(\omega_1, \ldots, \omega_n)$ in cui x_1 occorre n_1 volte, $\ldots$, x_N occorre n_N volte. Se poniamo ${}_nC_k = \frac{n!}{k!(n-k)!}$ si ha

$$\begin{aligned} C_n(n_1, \ldots, n_N) &= {}_nC_{n_1} \cdot {}_{n-n_1}C_{n_2} \cdots {}_{n-(n_1+\cdots+n_{N-1})}C_{n_N} \\ &= \frac{n!}{n_1!(n-n_1)!} \frac{(n-n_1)!}{n_2!(n-(n_1+n_2))!} \cdots \frac{(n-(n_1+\cdots+n_{N-1}))!}{n_N!} \\ &= \frac{n!}{n_1! n_2! \cdots n_N!} \end{aligned}$$

Pertanto

$$\sum_{\omega \in \Omega_n} p(\omega) = \sum_{\substack{n_1 \geq 0, \dots, n_N \geq 0, \\ n_1 + \cdots + n_N = n}} \frac{n!}{n_1! n_2! \cdots n_N!} p_1^{n_1} \cdots p_N^{n_N} = (p_1 + \cdots + p_N)^n = 1$$

e ciò fornisce un sistema accettabile di assegnare probabilità agli elementi di una partizione: se

$$A_{n_1, \dots, n_N} = \{\omega \in \Omega_n \,:\, \nu_n^{(1)} = n_1, \dots, \nu_n^{(N)} = n_N\}$$

allora $\Omega_n = \cup A_{n_1,\dots,n_N}$ e la distribuzione $P_n(A_{n_1,\dots,n_N})$ si chiama distribuzione multinomiale.

Nel contesto di (2.50) si consideri il seguente problema: in un quartiere con n abitanti il ristorante Mario è aperto dalle 12:00 alle 13:00. Supponendo che ciascun individuo decida indipendentemente di andare o meno da Mario tra le 12:00 e le 13:00 con probabilità λ/n, e che ogni persona che ci va, lo faccia in un momento qualunque (cioè scelto a caso con distribuzione uniforme) nel suddetto intervallo, la probabilità che esattamente k persone mangino da Mario è data dalla (2.52). Mostrare che la probabilità che, tra le k persone che vanno a mangiare da Mario, l ci vadano tra le 12:00 e le 12:20, mentre $m = k - l$ ci vadano tra le 12:20 e le 13:00 è data da

$$\frac{n(n-1)\cdots(n-l-m+1)}{l!m!} \left(\frac{\lambda}{3n}\right)^l \left(\frac{2\lambda}{3n}\right)^m \left(1 - \frac{\lambda}{n}\right)^{n-l-m}$$
$$= \frac{(\lambda/3)^l}{l!} \frac{(2\lambda/3)^m}{m!} \left(1 - \frac{\lambda}{n}\right)^{n-k} \prod_{j=1}^{k-1} \left(1 - \frac{j}{n}\right) .$$

Quando $n \to \infty$ questa converge a

$$\frac{(\lambda/3)^l}{l!} \frac{(2\lambda/3)^m}{m!} e^{-\lambda} = \frac{(\lambda/3)^l}{l!} e^{-\lambda/3} \cdot \frac{(2\lambda/3)^m}{m!} e^{-2\lambda/3}.$$

Ciò si può interpretare con il fatto che il numero di individui che vanno al ristorante in due intervalli di tempo disgiunti sono variabili indipendenti.

3. PROBABILITA' E ASPETTAZIONE CONDIZIONATE.

Per introdurre la probabilità condizionata, ricorriamo ancora all'intuizione fornitaci dall'esame della frequenza relativa in un esempio concreto. Consideriamo due 'filtri' F_B e F_A posti in successione attraverso i quali si dovranno cercare di far passare 'sassolini' di varie dimensioni. Siano B, A e $A \cap B$ gli eventi 'passare da F_B', 'passare da F_A' e 'passare da entrambi', rispettivamente. Ci chiediamo quale sia la probabilità che un sassolino, scelto a caso *tra quelli che hanno passato* F_B, passi anche da F_A. Indicheremo tale evento con $A|B$. Supponiamo ora che di n sassolini iniziali, soltanto m siano passati da F_B, ossia che $f_B = m/n$. Supponiamo inoltre che, tra questi m, soltanto k sassolini passino anche attraverso F_A. Allora avremo $f_{A\cap B} = k/n$ e $f_{A|B} = k/m$, da cui si vede che $f_{A|B} = f_{A\cap B}/f_B$. Ciò giustifica le seguente definizione.

(3.1) Definizione. Sia $(\Omega, \mathcal{F}, \mathrm{P}\,)$ uno spazio di probabilità e $A, B \in \mathcal{F}$ due eventi, con $\mathrm{P}\,(B) > 0$. La probabilità condizionata di un evento A dato l'evento B è data da

$$\mathrm{P}\,(A|B) = \frac{\mathrm{P}\,(A \cap B)}{\mathrm{P}\,(B)}.$$

In modo analogo, se ξ è una variabile aleatoria su Ω per cui $\mathrm{E}\,\xi$ è definita, l'aspettazione condizionata di ξ rispetto all'evento B è data da

$$\mathrm{E}\,(\xi|B) = \frac{\mathrm{E}\,(\xi \cdot 1_B)}{\mathrm{P}\,(B)}.$$

(3.2) Problema. Due dadi, uno rosso e uno bianco, vengono lanciati. Qual'è la probabilità di ottenere due 'sei', condizionata al fatto che il dado bianco mostri un 'sei'? Risposta: $(1/36)/(1/6) = 1/6$.

L'importanza del concetto di probabilità condizionata risiede nel fatto che in un gran numero di problemi i dati iniziali consistono appunto in probabilità condizionate, a partire dalle quali si cercherà di ottenere informazioni sulle probabilità ordinarie. La quantità $\mathrm{P}\,(A|B)$ dipende evidentemente sia da A che da B, ma in modo assai diverso. Come funzione di A la probabilità condizionata soddisfa le stesse proprietà della probabilità ordinaria:

I. $\mathrm{P}(A|B) \geq 0$;

II. $\mathrm{P}(\Omega|B) = 1$;

III. se $\{A_i\}$ è una sequenza di eventi disgiunti t.c. $A = \cup_i A_i$, allora

$$\mathrm{P}(A|B) = \sum_i \mathrm{P}(A_i|B)$$

.

Dunque, fissato $B \in \mathcal{F}$ con $\mathrm{P}(B) > 0$, lo spazio $(\Omega, \mathcal{F}, \mathrm{P}(\cdot|B))$ è uno spazio di probabilità.

Il tipo di dipendenza da B può essere compreso come segue. Sia $\{D_i\}$ una partizione finita o numerabile di Ω, tale che $\mathrm{P}(D_i) > 0$ per ogni i. Dato $A \in \mathcal{F}$, si ha

$$\mathrm{P}(A) = \sum_i \mathrm{P}(A \cap D_i) = \sum_i \mathrm{P}(A|D_i)\,\mathrm{P}(D_i). \tag{3.3}$$

L'identità tra il primo e l'ultimo termine si chiama *formula della probabilità totale* ed è analoga all'espressione che, in meccanica, definisce il centro di massa di un corpo pesante. Da $\sum_i \mathrm{P}(D_i) = 1$ segue che l'ultimo termine della (3.3) è la media pesata delle probabilità condizionate $\mathrm{P}(A|D_i)$ presa ciascuna con peso $\mathrm{P}(D_i)$. In particolare, si ha

$$\inf_i \mathrm{P}(A|D_i) \leq \mathrm{P}(A) \leq \sup_i \mathrm{P}(A|D_i). \tag{3.4}$$

(3.5) Esempio. Un'urna contiene n palline, m rosse e $n-m$ bianche. Consideriamo una successione di estrazioni senza rimpiazzamento, così che dopo la k-esima estrazione nell'urna vi sono $n-k$ palline. Sia C_k l'evento in cui alla k-esima estrazione otteniamo una pallina rossa. Allora si ha $\mathrm{P}(C_1) = m/n$. Inoltre, dalla (3.3) si ottiene $\mathrm{P}(C_2) = \mathrm{P}(C_2|C_1)\,\mathrm{P}(C_1) + \mathrm{P}(C_2|\overline{C}_1)\,\mathrm{P}(\overline{C}_1)$ e dunque

$$\mathrm{P}(C_2) = \frac{m-1}{n-1} \cdot \frac{m}{n} + \frac{m}{n-1} \cdot \frac{n-m}{n} = \frac{m}{n}.$$

Procedendo induttivamente si ottiene $\mathrm{P}(C_k) = m/n$ per ogni $1 \leq k \leq n$.

(3.6) Esempio. Consideriamo il problema di una 'passeggiata aleatoria con una barriera assorbente'. Il problema può essere descritto come segue. Un punto, che si trova inizialmente nel punto $x = 1$ si muove lungo una retta orizzontale a passi interi, con probabilità p di andare a destra, $1-p$ di andare a sinistra. Se raggiunge la posizione $x = 0$ viene assorbito e il processo termina. Questo modello può essere usato per descrivere alcune situazioni più o meno idealizzate. Ad esempio: durante il difficile ritorno a casa dopo una serata trascorsa alla taverna, un uomo, completamente ubriaco, perde la sua strada e si ritrova pericolosamente vicino ($x = 1$) al bordo di un burrone ($x = 0$): se fa un passo a sinistra precipita nel burrone, se lo fa a destra se ne allontana, accrescendo le sue possibilità di sopravvivere. Una domanda naturale sarà allora: data $0 < p < 1$, qual'è la probabilità di restare vivo?

Un'altra situazione è quella già considerata al punto (2.32): a un giocatore d'azzardo non gli è rimasto altro che un biglietto (unità di denaro) in tasca e decide di tentare per l'ultima volta la fortuna giocando, ad oltranza, contro un Casinó dalle risorse illimitate, un gioco alla pari ($p = 1/2$). Qual'è la probabilità di non perdere prima o poi tutti i suoi soldi?

Torniamo dunque al nostro punto che si muove sulla retta e indichiamo con $C_{k,j}$ l'evento in cui il punto, partendo da $x = k$, giunga prima o poi in $x = j$. In particolare, $C_{k,0}$ è l'evento in cui il punto viene (prima o poi) assorbito partendo dalla posizione $x = k$. Usando la formula della probabilità totale (3.3) si trova, per ogni $k \geq 1$, $\mathrm{P}(C_{k,0}) = (1-p)\,\mathrm{P}(C_{k-1,0}) + p\,\mathrm{P}(C_{k+1,0})$. In particolare, essendo $\mathrm{P}(C_{0,0}) = 1$ si ha $\mathrm{P}(C_{1,0}) = 1 - p + p\,\mathrm{P}(C_{2,0})$. D'altra parte, i cammini che portano da $x = 2$ a $x = 0$ si possono dividere in due parti: uno che porta da $x = 2$ a $x = 1$ (non necessariamente in un passo) e uno che porta da $x = 1$ a $x = 0$ (ancora non necessariamente in un passo). In altre parole si ha $C_{2,0} = C_{2,1} \cap C_{1,0}$. Evidentemente si ha $\mathrm{P}(C_{2,1}) = \mathrm{P}(C_{1,0})$, poiché la struttura del processo è identica alla situazione iniziale, soltanto traslata di un passo verso destra (invarianza per traslazione). Inoltre gli eventi $C_{2,1}$ e $C_{1,0}$ sono indipendenti. Dunque $\mathrm{P}(C_{2,0}) = (\mathrm{P}(C_{1,0}))^2$. Poniamo per semplicità $\mathrm{P}(C_{1,0}) \equiv P_1$. Allora si trova $P_1 = 1-p+p\,P_1^2$ che ammette le soluzioni $P_1 = 1$ e $P_1 = (1-p)/p$. Vediamo quale dobbiamo scegliere. Se $p = 1/2$ le due soluzioni coincidono: $P_1 = 1$. Se $p = 1$ allora $P_1 = 0$

perchè il punto si muove sempre verso destra. Se $p < 1/2$ la seconda soluzione è impossibile perchè $(1-p)/p > 1$. Dunque la soluzione da considerare è, assumendo $P_1(p)$ continua,

$$P_1(p) = \begin{cases} 1, & \text{se } 0 \leq p \leq 1/2; \\ \frac{(1-p)}{p}, & \text{se } 1/2 \leq p \leq 1. \end{cases}$$

Così, ad esempio, se $p = 1/2$ il nostro ubriacone avrà l'assoluta certezza di precipitare e il nostro povero giocatore d'azzardo avrà l'assoluta certezza di perdere tutti i suoi soldi, prima o poi. Mentre avranno entrambi uguali probabilità di farcela se $p = 2/3$. Ciò, a prima vista, può apparire strano. Saremmo tentati di supporre che se, ad esempio, le giocate non sono truccate (cioè la perdità media = vincita media = zero), allora l'intero gioco sarà bilanciato. E ciò è in generale corretto. Se immaginiamo questo gioco, con $p = 1/2$, giocato infinite volte, allora dopo n giocate l'ammontare medio di denaro rimasto al giocatore è di un'unità, per ogni n finito. Infatti, abbiamo visto, con le notazioni del (2.31), che $x(n) = y(n) + 1 = 2\nu_n^{(1)} - n + 1$ e dunque $\mathrm{E}\, x(n) = 1$. Così, lo sbilanciamento che abbiamo ottenuto, cioè la certezza di perdere prima o poi, può essere visto come uno dei paradossi dell'infinito. Un altro fatto che può apparire soprendente (e che abbiamo già osservato, vedi (2.42)) è che se $p = 1/2$ il numero medio di prove (passi, giocate) necessarie per l'assorbimento (caduta, bancarotta) non è finito.

Torniamo ora alla nozione generale di probabilità condizionata. Se $A, B \in \mathcal{F}$ con $\mathrm{P}(A) > 0$ e $\mathrm{P}(B) > 0$, allora dai valori di $\mathrm{P}(A)$, $\mathrm{P}(B)$ e $\mathrm{P}(A|B)$ possiamo ottenere anche $\mathrm{P}(B|A)$. In effetti

$$\mathrm{P}(B|A) = \frac{\mathrm{P}(A|B)\,\mathrm{P}(B)}{\mathrm{P}(A)}. \tag{3.7}$$

Se poi $\mathcal{D} = \{D_i\}$ è una partizione come sopra, allora usando la (3.3) si ottiene la *formula di Bayes*:

$$\mathrm{P}(D_i|A) = \frac{\mathrm{P}(A|D_i)\,\mathrm{P}(D_i)}{\sum_i \mathrm{P}(A|D_i)\,\mathrm{P}(D_i)}. \tag{3.8}$$

(3.9) Osservazione. In statistica matematica gli eventi D_i vengono talvolta chiamati *ipotesi* (o cause) e le probabilità $\mathrm{P}(D_i)$ *probabilità a priori* (cioè date prima dell'esperimento). Sulla base del fatto che in una prova dell'esperimento un particolare evento A si è verificato, vorremmo stimare quale delle ipotesi sia la più probabile. Tale stima può in linea di principio essere fatta calcolando con la (3.8) le *probabilità a posteriori* $\mathrm{P}(D_i|A)$ o, per usare una locuzione evocativa ma che può dar luogo a fraintendimenti, le 'probabilità delle cause'. Ed è proprio questo tipo di interpretazione che ha dato adito a numerose controversie riguardo all'uso che della (3.8) è lecito fare. In effetti, le probabilità a priori $\mathrm{P}(D_i)$ raramente sono note e allora si ricorre all'espediente di assegnar loro un valore più o meno arbitrario. In altri casi una stima attendibile di $\mathrm{P}(D_i)$ può dare risultati di grande utilità.

Esempio. Supponiamo che una data popolazione abbia il 40% di fumatori. Supponiamo poi che l'incidenza di una certa malattia respiratoria M sia del 25% nei fumatori e del 7% nei non-fumatori. Problema diretto: qual'è la probabilità $\mathrm{P}(M)$ che una persona scelta a caso sia affetta dalla malattia M? Siano F e $N = \overline{F}$ gli eventi corrispondenti all'essere fumatore o no. Questi formano evidentemente una partizione. Si ha $\mathrm{P}(F) = 0.4$, $\mathrm{P}(N) = 0.6$ e inoltre $\mathrm{P}(M|F) = 0.25$ e $\mathrm{P}(M|N) = 0.07$. Allora, usando la formula della probabilità totale, si trova

$$\mathrm{P}(M) = \mathrm{P}(M|F)\mathrm{P}(F) + \mathrm{P}(M|N)\mathrm{P}(N) = 0.142$$

Problema inverso: qual'è la probabilità che una persona malata sia un fumatore? Questa sarà data evidentemente da $\mathrm{P}(F|M)$ e dunque, per la formula di Bayes,

$$\mathrm{P}(F|M) = \frac{\mathrm{P}(M|F)\mathrm{P}(F)}{\mathrm{P}(M)} = 0.704.$$

Problema. Ci sono tre mobili identici, ciascuno con due cassetti. Il primo mobile ha una moneta d'oro in ciascun cassetto, il secondo ne ha una d'oro in un cassetto e una d'argento nell'altro, il terzo una moneta d'argento in ciascun cassetto. Sia A l'evento in cui aprendo a caso un cassetto vi troviamo una moneta d'oro e calcoliamo la probabilità che anche l'altro cassetto contenga una moneta di uguale valore. A questo scopo indichiamo con D_i l'evento in cui il cassetto prescelto appartenga all'i-esimo mobile, con $i = 1, 2, 3$. Evidentemente avremo

$\mathrm{P}(A|D_1) = 1$, $\mathrm{P}(A|D_2) = 1/2$, $\mathrm{P}(A|D_3) = 0$ e $\mathrm{P}(D_i) = 1/3$ per ogni i, cosicchè $\mathrm{P}(A) = \sum_{i=1}^{3} \mathrm{P}(A|D_i)\mathrm{P}(D_i) = 1/2$. La probabilità cercata sarà allora $\mathrm{P}(D_1|A)$ la formula di Bayes dà $(1/3)/(1/2) = 2/3$.

Osserviamo che dalle (3.1) e (3.7) si ha

$$\frac{\mathrm{P}(A|B)}{\mathrm{P}(A)} = \frac{\mathrm{P}(B|A)}{\mathrm{P}(B)} = \frac{\mathrm{P}(A \cap B)}{\mathrm{P}(A)\,\mathrm{P}(B)}. \tag{3.10}$$

Se dunque, ad esempio, il verificarsi dell'evento B accresce la probabilità di A, allora il verificarsi di A accresce dello stesso fattore la probabilità di B, e non è in alcun modo possibile inferire la direzione della relazione causale a partire soltanto dai valori delle probabilità condizionate.

D'altra parte può accadere che il verificarsi dell'evento B non abbia alcuna influenza sulla probabilità di A, cioè che $\mathrm{P}(A|B) = \mathrm{P}(A)$. Si dice in tal caso che A è *indipendente* da B. Per la (3.10), se A è indipendente da B, allora anche B è indipendente da A e dunque possiamo dire che A e B sono eventi tra loro indipendenti. Usando ancora la (3.10) formuliamo la seguente

(3.11) Definizione. Due eventi A e B si dicono indipendenti se $\mathrm{P}(A \cap B) = \mathrm{P}(A)\,\mathrm{P}(B)$.

Osserviamo che

$$\mathrm{P}(\overline{A} \cap B) = \mathrm{P}((\Omega \setminus A) \cap B) = \mathrm{P}(B) - \mathrm{P}(A \cap B) = \mathrm{P}(B)(1 - \mathrm{P}(A \cap B)/\mathrm{P}(B))$$

e, se A e B sono indipendenti, l'ultimo fattore è uguale a $\mathrm{P}(\overline{A})$. Espressioni analoghe si ottengono per $\mathrm{P}(A \cap \overline{B})$ e $\mathrm{P}(\overline{A} \cap \overline{B})$. Vale pertanto il seguente risultato.

(3.12) Lemma. Gli eventi A e B sono indipendenti se e solo se $(\overline{A}, B)$, $(A, \overline{B})$ o $(\overline{A}, \overline{B})$ sono indipendenti.

Più in generale, se $A_1, \ldots, A_n \in \mathcal{F}$ e $\mathrm{P}(A_1) > 0$, $\mathrm{P}(A_1 \cap A_2) > 0$, ... , $\mathrm{P}(A_1 \cap \cdots \cap A_{n-1}) > 0$, possiamo scrivere $\mathrm{P}(A_1 \cap \cdots \cap A_n)$ nella forma

(3.13)

$$\mathrm{P}(A_1 \cap \cdots \cap A_n) = \mathrm{P}(A_1)\,\mathrm{P}(A_2|A_1)\,\mathrm{P}(A_3|A_1 \cap A_2) \cdots \mathrm{P}(A_n|A_1 \cap \cdots \cap A_{n-1})$$

e dunque diremo che gli eventi $A_1, \ldots, A_n$ sono *congiuntamente indipendenti* se $\mathrm{P}(A_1 \cap \cdots \cap A_n) = \mathrm{P}(A_1) \cdots \mathrm{P}(A_n)$. Evidentemente ciò implica che gli elementi di ciascuna sotto-collezione $A_{i_1}, \ldots, A_{i_l}$ estratta da $A_1, \ldots, A_n$ sono indipendenti, ma non è necessariamente vero il contrario.

(3.14) Esempio. Nel contesto dell'Esempio 1.12 (successione di n lanci di una moneta) consideriamo gli eventi $B_{jk} = \{\omega : \omega_j = \omega_k\}$. Evidentemente $P(B_{jk}) = P(\{\omega_j = 0, \omega_k = 0\}) + P(\{\omega_j = 1, \omega_k = 1\}) = 1/4 + 1/4 = 1/2$. Inoltre si ha che B_{12}, B_{13} e B_{23} sono indipendenti due a due, ma essendo $B_{12} \cap B_{13} \subset B_{23}$ non sono congiuntamente indipendenti.

Possiamo dare a quanto sopra una formulazione un po' diversa. Diamo innanzitutto la seguente definizione.

(3.15) Definizione. Sia $\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_n$ una collezione finita di σ-sottoalgebre di $\mathcal{F}$. Diremo che le $\mathcal{F}_i$, $1 \le i \le n$, sono (congiuntamente) indipendenti se per ogni scelta di $C_1 \in \mathcal{F}_1, \ldots, C_n \in \mathcal{F}_n$, si ha

$$\mathrm{P}(C_1 \cap \cdots \cap C_n) = \mathrm{P}(C_1) \cdots \mathrm{P}(C_n).$$

Abbiamo già discusso la nozione di σ-algebra $\mathcal{F}(\mathcal{A})$ *generata* da una sottoclasse $\mathcal{A}$ di eventi, cioè l'intersezione di tutte le σ-algebre che contengono $\mathcal{A}$. Consideriamo ora due partizioni $\mathcal{A} = (A, \overline{A})$ e $\mathcal{B} = (B, \overline{B})$ di Ω. A tali partizioni corrispondono in modo unico a sotto-algebre finite, in questo caso $\mathcal{F}(\mathcal{A}) = \{\emptyset, A, \overline{A}, \Omega\}$ e $\mathcal{F}(\mathcal{B}) = \{\emptyset, B, \overline{B}, \Omega\}$. Ora, se A e B sono indipendenti, allora è facile verificare usando (3.12) che $\mathrm{P}(C_1 \cap C_2) = \mathrm{P}(C_1)\mathrm{P}(C_2)$ per ogni scelta di $C_1 \in \mathcal{F}(\mathcal{A})$ e $C_2 \in \mathcal{F}(\mathcal{B})$. In altre parole, se A e B sono indipendenti, allora le σ-sottoalgebre generate dalle partizioni $\mathcal{A}$ e $\mathcal{B}$ sono anch'esse indipendenti. D'altra parte, ciò non è necessariamente vero se si considerano σ-sottoalgebre generate da collezioni arbitrarie di eventi indipendenti.

(3.16) Esempio. Nella situazione dell'Esempio 3.14, sia $\mathcal{A}_1 = \{B_{12}, B_{13}\}$ e $\mathcal{A}_2 = \{B_{23}\}$. È ovvio che $\mathcal{A}_1$ e $\mathcal{A}_2$ non costituiscono partizioni di Ω_n. Abbiamo visto che B_{12}, B_{13} e B_{23} sono indipendenti due a due. D'altra parte $\overline{B}_{23} =$

$(B_{12} \cup B_{13}) \setminus (B_{12} \cap B_{13})$ e dunque B_{23} fa parte anche di $\mathcal{F}(\mathcal{A}_1)$. Ciò implica che $\mathcal{F}(\mathcal{A}_1)$ e $\mathcal{F}(\mathcal{A}_2)$ non sono indipendenti.

(3.17) Osservazione. Se $\mathcal{A} = \{A_1, \ldots, A_r\}$ è una partizione finita, allora ogni $C \in \mathcal{F}(\mathcal{A})$ può essere espresso con una 'formula' del tipo $C = A_1 \cap \overline{A}_3$ o $C = (A_2 \cap A_9) \cup (A_3 \cap \overline{A}_5)$. Se invece $\mathcal{A}$ è infinita, gli elementi di $\mathcal{F}(\mathcal{A})$ possono essere molto complicati; il modo di rendere precisa l'idea che gli elementi di $\mathcal{A}$ 'determinano' C non è quello di richiedere una formula, ma semplicemente quello di richiedere che C faccia parte di $\mathcal{F}(\mathcal{A})$. Nel seguito, date due partizioni $\mathcal{A}_1$ e $\mathcal{A}_2$, diremo che $\mathcal{A}_2$ è più *fine* di $\mathcal{A}_1$, indicato $\mathcal{A}_1 \preceq \mathcal{A}_2$, se $\mathcal{F}(\mathcal{A}_1) \subseteq \mathcal{F}(\mathcal{A}_2)$.

(3.18) Definizione. Sia $\xi_1, \xi_2, \ldots \xi_n$ una sequenza di variabili aleatorie discrete definite su uno spazio Ω. Per ogni $1 \leq i \leq n$ la collezione di eventi $\{C_k^{(i)}\}$, con $C_k^{(i)} = \{\omega : \xi_i(\omega) = x_k^{(i)}\}$, forma una partizione $\mathcal{C}_i$ di Ω, finita o numerabile, la quale a sua volta genera una σ-sottoalgebra che denoteremo con $\mathcal{F}_{\xi_i}$. Le variabili aleatorie $\xi_1, \xi_2, \ldots \xi_n$ si dicono indipendenti se le σ-sottoalgebre $\mathcal{F}_i \equiv \mathcal{F}_{\xi_i}$, $1 \leq i \leq n$, sono indipendenti. Diremo poi che sono variabili aleatorie indipendenti e identicamente distribuite (v.a.i.i.d) se, in aggiunta, le loro distribuzioni di probabilità coincidono.

In particolare, per ogni collezione $C_1, \ldots, C_n$ di sottoinsiemi dello spazio discreto $\mathcal{Z}$, si ha

$$\mathrm{P}\,(\xi_1 \in C_1, \cdots, \xi_n \in C_n) = \prod_{i=1}^{n} \mathrm{P}\,(\xi_i \in C_i). \tag{3.19}$$

Se inoltre sono identicamente distribuite e μ indica la comune distribuzione di probabilità su $\mathcal{Z}$, allora l'espressione scritta sopra è uguale a $\prod_{i=1}^{n} \mu(C_i)$.

Un semplice calcolo basato sulla (3.19) mostra che

(3.20) Lemma. Siano $\xi_1, \xi_2, \ldots \xi_n$ indipendenti e tali che $\mathrm{E}\,\xi_i$ esista per ogni $1 \leq i \leq n$. Allora $\mathrm{E}\,(\xi_1 \cdots \xi_n)$ esiste e vale $\mathrm{E}\,(\xi_1 \cdots \xi_n) = \mathrm{E}\,\xi_1 \cdots \mathrm{E}\,\xi_n$. Se inoltre $\mathrm{Var}\,\xi_i$ è definita allora $\mathrm{Var}\,(\xi_1 + \cdots + \xi_n) = \mathrm{Var}\,\xi_1 + \cdots + \mathrm{Var}\,\xi_n$. Se poi $\xi_1, \xi_2, \ldots \xi_n$ sono identicamente distribuite, con $m = \mathrm{E}\,\xi_i$ e $\sigma^2 = \mathrm{Var}\,\xi_i$, allora $\mathrm{E}\,(\xi_1 \cdots \xi_n) = m^n$ e $\mathrm{Var}\,(\xi_1 + \cdots + \xi_n) = n \cdot \sigma^2$.

Inoltre, data una sequenza di variabili aleatorie $\xi_1, \xi_2, \dots \xi_n$ su Ω e una sequenza di indici interi $(i_1, i_2, \dots, i_k)$, denoteremo con $\mathcal{F}_{\xi_{i_1}, \xi_{i_2}, \dots, \xi_{i_k}}$ la σ-algebra generata dagli insiemi della forma $\{\xi_{i_1} \in C_1, \xi_{i_2} \in C_2, \cdots, \xi_{i_k} \in C_k\}$. Un'altra conseguenza interessante della (3.19) è che se $(i_1, i_2, \dots, i_k)$, $(j_1, j_2, \dots, j_l)$ sono due insiemi disgiunti di indici interi, allora $\mathcal{F}_1 = \mathcal{F}_{\xi_{i_1}, \xi_{i_2}, \dots, \xi_{i_k}}$ e $\mathcal{F}_2 = \mathcal{F}_{\xi_{j_1}, \xi_{j_2}, \dots, \xi_{j_l}}$ sono indipendenti.

4. CATENE DI MARKOV.

Sia X un insieme finito e $\xi_0, \xi_1, \dots \xi_n$ una sequenza di variabili aleatorie su uno spazio di probabilità $(\Omega, \mathcal{F}, \mathrm{P}\,)$ che assumono valori in X. Nel caso di una sequenza di v.a.i.i.d. si avrà per ogni $0 \le k < n$

$$\mathrm{P}\,(\xi_{k+1} = x_j | \xi_0 = x_l, \xi_1 = x_m, \dots, \xi_k = x_i) = \mathrm{P}\,(\xi_{k+1} = x_j) = \mu_j, \tag{4.1}$$

dove (μ_i) è una distribuzione di probabilità assegnata su X. Supponiamo ora che ad ogni coppia x_i e x_j di elementi di X sia assegnato un numero non-negativo p_{ij} ed inoltre che tali numeri soddisfino la condizione

$$\sum_{j : x_j \in X} p_{ij} = 1, \quad x_i \in X. \tag{4.2}$$

Allora diremo che la sequenza $\xi_0, \xi_1, \dots \xi_n$ è una *catena di Markov* su $(\Omega, \mathcal{F}, \mathrm{P}\,)$ se

$$\begin{aligned} &\mathrm{P}\,(\xi_{k+1} = x_j | \xi_0 = x_l, \xi_1 = x_m, \dots, \xi_k = x_i) \\ &\qquad = \mathrm{P}\,(\xi_{k+1} = x_j | \xi_k = x_i) = p_{ij}. \end{aligned} \tag{4.3}$$

Gli elementi di X vanno pensati come i possibili *stati* di un sistema e ξ_k è dunque lo stato al 'tempo' k. La sequenza $\xi_0, \xi_1, \dots \xi_n$ (detta anche *processo*) rappresenta un tratto di storia del sistema, che si evolve in accordo con la legge probabilità (4.3), la quale richiede che la distribuzione condizionata dello stato 'futuro' ξ_{k+1} dato lo stato 'presente' ξ_k, non debba dipendere anche dal 'passato' $\xi_0, \xi_1, \dots, \xi_{k-1}$. I numeri p_{ij} prendono il nome di *probabilità di transizione* e formano, per definizione, gli elementi di una *matrice stocastica* $\mathbb{P}$, cioè $p_{ij} = (\mathbb{P})_{\mathrm{ij}}$. Cosa succede se lo stato inziale è aleatorio? Supponiamo che $\mathrm{P}\,(\xi_k = x_j) = \mu_j$, cioè assegnamo una *distribuzione di probabilità iniziale* su X allora usando la (3.13) e la (4.3) si ha

(4.4)

$$\begin{aligned} &\mathrm{P}\,(\xi_0 = x_{i_0}, \xi_1 = x_{i_1}, \dots, \xi_k = x_{i_k}) \\ &= \mathrm{P}\,(\xi_0 = x_{i_0})\, \mathrm{P}\,(\xi_1 = x_{i_1} | \xi_0 = x_{i_0}) \cdots \mathrm{P}\,(\xi_k = x_{i_k} | \xi_0 = x_{i_0}, \dots, \xi_{k-1} = x_{i_{k-1}}) \\ &= \mu_{i_0}\, p_{i_0 i_1} \cdots p_{i_{k-1} i_k}\,. \end{aligned}$$

E' evidente che se le righe della matrice stocastica $\mathbb{P}$ sono tutte uguali, cioè se $p_{ij} = \mu_j$ allora riotteniamo la sequenza di v.a.i.i.d. (4.1).

La situazione di riferimento sarà quella in cui $\xi_i(\omega) = \omega_i$, la catena di Markov è definita dallo spazio di probabilità $(\Omega, \mathcal{F}, \mathrm{P}\,)$ stesso: $\Omega_n = X^{[0,n]}$, e P è la misura su $\mathcal{F}$ generata dalla distribuzione su Ω che ad una parola finita $\omega = (\omega_0, \ldots, \omega_k) \in \Omega_k$ associa la probabilità $p(\omega) = \mu_{\omega_0}\, p_{\omega_0\omega_1} \cdots p_{\omega_{k-1}\omega_k}$.

(4.5) Esercizio. Usando (4.12) e $\sum_j \mu_j = 1$ mostrare che $\sum_{\omega\in\Omega_n} p(\omega) = 1$.

(4.6) Esempio. *Modello di Laplace-Bernoulli della diffusione.* Immaginiamo r palle bianche e r palle rosse distribuite tra due urne, con il vincolo che ciascuna urna debba contenere esattamente r palle. Lo stato del sistema è specificato ad esempio dal numero di palle bianche nella prima urna e così $X = \{0, 1, \ldots, r\}$. Il meccanismo di transizione è il seguente: ad ogni passo viene estratta una palla da ciascuna urna e le due palle vengono scambiate. Allora, se lo stato presente è i, la probabilità di passare a $i - 1$ è data dalla probabilità i/r di estrarre una palla bianca dalla prima urna, per la probabilità i/r di estrarre una palla rossa dalla seconda urna. Usando analogo argomento per le altre possibilità si trovano le probabilità di transizione

$$p_{ij} = \left(\frac{i}{r}\right)^2 \delta_j^{i-1} + \left(\frac{r-i}{r}\right)^2 \delta_j^{i+1} + 2\frac{i(r-i)}{r^2}\,\delta_j^i.$$

Si tratta dunque di un modello probabilistico molto semplice per descrivere il processo di mescolamento di due liquidi diversi posti in due contenitori comunicanti.

(4.7) Esempio. Passeggiata aleatorie con due barriere assorbenti. Siano a e b due interi con $a \le 0 \le b$ e poniamo $X = \{a, a+1, \ldots, b-1, b\}$. Per ogni $a < i, j < b$ poniamo $p_{ij} = p\,\delta_j^{i+1} + q\,\delta_j^{i-1}$, con $\sum_j p_{ij} = p + q = 1$, e $p_{aa} = p_{bb} = 1$. Lo stato del sistema si può ad esempio riguardare come la fortuna di un giocatore: l'assorbimento in $x = a$ rappresenta la rovina, l'assorbimento in $x = b$ rappresenta, ad esempio, la rovina del suo avversario. La distribuzione iniziale può essere ad esempio data da $\mu_j = \delta_j^x$ (posizione iniziale in x) per qualche $x \in [a, b]$. Il valore di $\xi_k(\omega) = \omega_k$ indica dunque la posizione su X dopo il k-esimo salto (giocata). Prima dell'assorbimento in a o in b, le variabili aleatorie definite da $\eta_k(\omega) = \xi_k(\omega) - \xi_{k-1}(\omega)$ formano una sequenza di v.a. indipendenti bernoulliane con $\mathrm{P}\,(\eta_k = 1) = p$ e $\mathrm{P}\,(\eta_k = -1) = q$.

Inversamente, se $\eta_1, \ldots, \eta_n$ è una sequenza di v.a. indipendenti bernoulliane, la sequenza $S_{k+1} = S_k + \eta_{k+1}$, $0 \le k < n$, forma una catena di Markov con $X = [-n, n]$, $p_{ij} = \mathrm{P}(S_{k+1} = j | S_k = i) = \mathrm{P}(\eta_{k+1} = j - i)$ e $\mu_j = \delta_j^0$. Più in generale, se $\xi_0, \eta_1, \ldots, \eta_n$ sono v.a. indipendenti la sequenza $\xi_0, \ldots, \xi_n$ data da

$$\xi_{k+1} = f_k(\xi_k, \eta_{k+1}), \quad 0 \le k < n, \tag{4.8}$$

forma ancora una catena di Markov. In connessione con la (4.8), osserviamo che una catena di Markov può essere considerata un analogo probabilistico di una sequenza deterministica $x_0, \ldots, x_n$ generata dalle equazioni ricorsive

$$x_{k+1} = f_k(x_k). \tag{4.9}$$

Il prossimo esempio ha origine dalla teorie delle code.

(4.10) Esempio. Supponiamo che ad una stazione di taxi, questi ultimi giungano ad intervalli di tempo regolari ed uno alla volta. Se non c'è nessuno ad aspettare, il taxi riparte immediatamente. Sia η_k il numero di passeggeri che arrivano alla stazione al tempo k e supponiamo che $\eta_1, \ldots, \eta_k$ siano v.a. indipendenti. Sia ξ_k la lunghezza della coda d'attesa al tempo k, con $\xi_0 = 0$. Ora, se $\xi_k = i$ avremo

$$\xi_{k+1} = \begin{cases} \eta_{k+1}, & \text{se } i = 0, \\ i - 1 + \eta_{k+1}, & \text{se } i > 0. \end{cases}$$

che si può anche scrivere nella forma (4.8): $\xi_{k+1} = (\xi_k - 1)^+ + \eta_{k+1}$, dove $a^+ = \max(a, 0)$. Quindi la sequenza $\xi_0, \ldots, \xi_n$ è una catena di Markov.

(4.11) Esempio. Questo esempio provene dalla teoria dei *processi di ramificazione*. Un processo siffatto è una sequenza di v.a. $\xi_0, \ldots, \xi_n$ dove ξ_k fornisce ad esempio il numero di particelle presenti al tempo k in un dato materiale, o anche il numero di individui alla k-esima generazione di una data popolazione. Si suppone che $\xi_0 = 1$. Se al tempo k vi sono ξ_k individui (o particelle), numerati da 1 a ξ_k, allora ξ_{k+1} è dato dalla somma aleatoria di variabili aleatorie

$$\xi_{k+1} = \eta_1^{(k)} + \cdots + \eta_{\xi_k}^{(k)},$$

dove $\eta_i^{(k)}$ è il numero di individui generati dall'i-esimo individuo della k-esima generazione. Se supponiamo che tutte le v.a. $\eta_i^{(k)}$, $k \geq 0$, siano mutuamente indipendenti otteniamo

$$\begin{aligned} P(\xi_{k+1} = i_{k+1} | \xi_k = i_k, \xi_{k-1} = i_{k-1}, \ldots) &= P(\xi_{k+1} = i_{k+1} | \xi_k = i_k) \\ &= P(\eta_1^{(k)} + \cdots + \eta_{i_k}^{(k)} = i_{k+1}) \end{aligned}$$

da cui segue che la sequenza $\xi_0, \ldots, \xi_n$ forma una catena di Markov.

Una catena di Markov può essere riguardata come una distribuzione di probabilità sui cammini su un grafo: i vertici del grafo saranno i punti di X, e due vertici, diciamo l'i-esimo e il j-esimo, saranno connessi da una freccia orientata se $p_{ij} > 0$.

(4.12) Esercizio. Sia $X = \{0, 1, 2\}$ e

$$\mathbb{P} = \begin{pmatrix} 1 & 0 & 0 \\ \frac{1}{2} & 0 & \frac{1}{2} \\ \frac{1}{3} & \frac{2}{3} & 0 \end{pmatrix}$$

Determinare il grafo corrispondente. Osserviamo che una volta giunti nello stato 0 non se ne esce più, poiché $p_{00} = 1$. In tal caso si dice che lo stato 0 è *assorbente*. Così, gli stati a e b dell'esempio (4.7) sono assorbenti.

Una sequenza finita di stati $\omega = \omega_0 \omega_1 \ldots \omega_n$ con $p(\omega) > 0$ sarà allora rappresentata da un cammino di lunghezza n che parte dal punto ω_0 e $p(\omega)$ è la probabilità di tale cammino. In generale, dati due stati x_i e x_j in X e un intero n, vi saranno più cammini ammissibili (cioè con probabilità non nulla) che connettono x_i e x_j in esattamente n passi. Sommando sugli stati intermedi si ottiene che le probabilità di transizione di ordine n sono date da

$$p_{ij}^{(n)} = \mathrm{P}\left(\xi_{k+n} = x_j | \xi_k = x_i\right) = \sum_{k_1, \ldots, k_{n-1}} p_{ik_1} \cdots p_{k_{n-1}j} = (\mathbb{P}^n)_{ij}, \tag{4.13}$$

da cui, posto $p_{ij}^{(0)} = \delta_i^j$, si ottiene l'equazione di Chapman-Kolmogorov

$$p_{ij}^{(n+m)} = \sum_k p_{ik}^{(n)} p_{kj}^{(m)}, \qquad \sum_j p_{ij}^{(n)} = 1, \tag{4.14}$$

o, in forma matriciale,

$$\mathbb{P}^{\mathrm{n+m}} = \mathbb{P}^{\mathrm{n}} \cdot \mathbb{P}^{\mathrm{m}}. \tag{4.15}$$

Evidentemente, sarà possibile partendo da x_i raggiungere x_j se e solo se $p_{ij}^{(n)} > 0$ per almeno un n. Se ciò accade per ogni coppia di stati x_i e x_j allora la catena si dice *irriducibile*. Dato uno stato $x_j \in X$, diremo che x_j ha *periodo* t_j se $p_{jj}^{(n)} > 0$ implica che t_j divide n e se t_j è l'intero più grande con questa proprietà. In altre parole t_j è il m.c.d. dell'insieme

$$\{n \geq 1 : p_{jj}^{(n)} > 0\}. \tag{4.16}$$

Ora, se la catena è irriducibile allora per ogni coppia di stati x_i e x_j esistono due interi n_1 e n_2 tali che $p_{ij}^{(n_1)} > 0$ e $p_{ji}^{(n_2)} > 0$, e inoltre da (4.13) si ha

$$p_{ii}^{(n+n_1+n_2)} \geq p_{ij}^{(n_1)} p_{jj}^{(n)} p_{ji}^{(n_2)}.$$

Siano ora t_i e t_j i periodi di x_i e x_j. Prendendo $n = 0$ la disuguaglianza scritta sopra mostra che t_i divide $n_1 + n_2$; d'altra parte, sempre da questa disuguaglianza segue che se $p_{jj}^{(n)} > 0$ allora t_i divide anche $n + n_1 + n_2$, e dunque divide n. Dunque t_i divide ogni intero nell'insieme in (4.16) e dunque si ha $t_i \leq t_j$. Scambiando i e j otteniamo dunque che se la catena è irriducibile x_i e x_j hanno lo stesso periodo. Possiamo dunque in questo caso parlare di periodo della catena stessa. Se tale periodo è 1 la catena si dirà *aperiodica*.

(4.17) Lemma. Una catena di Markov finita è irriducibile e aperiodica se e solo se esiste un intero n_0 tale che per ogni i e j, $p_{ij}^{(n)} > 0$ per ogni $n > n_0$.

Dimostrazione. Essendo $p_{jj}^{(n+m)} \geq p_{jj}^{(n)} p_{jj}^{(m)}$, se n ed m appartengono all'insieme in (4.16), allora anche la somma $n + m$ vi appartiene. Ora, è un fatto di teoria dei numeri che se un insieme di interi positivi è additivamente chiuso ed ha m.c.d. uguale a 1, allora tale insieme contiene tutti gli interi più grandi di un certo k. Dati i e j, sia $r = r(i,j)$ tale che $p_{ij}^{(r)} > 0$. Sia inoltre $s = s(i,j)$ dato da $s = r + k$. Allora per $n > s$ si ha $p_{ij}^{(n)} \geq p_{ij}^{(r)} p_{jj}^{(n-r)} > 0$. Dunque, prendendo $n_0 = \max_{i,j} s(i,j)$, vediamo che se $\mathbb{P}$ è irriducibile e aperiodica allora vale la proprietà del lemma. L'implicazione inversa è ovvia. q.e.d.

(4.18) Esempio. Sia $X = \{0, 1\}$ e

$$\mathbb{P} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

allora si ha

$$\mathbb{P}^{2n} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \qquad \mathbb{P}^{2n+1} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Evidentemente questa catena è irriducibile ma non aperiodica (ha periodo 2). Cosa si può dire per la catena dell'esercizio (4.12)?

(4.19) Esempio. Sia ancora $X = \{0, 1\}$ e

$$\mathbb{P} = \begin{pmatrix} p_{00} & p_{01} \\ p_{10} & p_{11} \end{pmatrix}$$

Si trova facilmente che

$$\mathbb{P}^2 = \begin{pmatrix} p_{00}^2 + p_{01}p_{10} & p_{01}(p_{00} + p_{11}) \\ p_{10}(p_{00} + p_{11}) & p_{11}^2 + p_{01}p_{10} \end{pmatrix}.$$

Assumendo $\mathrm{tr}\,\mathbb{P} = p_{00} + p_{11} < 2$ e procedendo per induzione otteniamo l'espressione generale

$$\mathbb{P}^{\mathrm{n}} = \frac{1}{\rho} \begin{pmatrix} 1 - p_{11} & 1 - p_{00} \\ 1 - p_{11} & 1 - p_{00} \end{pmatrix} + \frac{(1-\rho)^{\mathrm{n}}}{\rho} \begin{pmatrix} 1 - p_{00} & -1 + p_{00} \\ -1 + p_{11} & 1 - p_{11} \end{pmatrix},$$

dove $\rho = 2 - \mathrm{tr}\,\mathbb{P}$. Perciò si vede che se $0 < \rho < 2$ allora quando $n \to \infty$,

$$\mathbb{P}^{\mathrm{n}} \to \frac{1}{\rho} \begin{pmatrix} 1 - p_{11} & 1 - p_{00} \\ 1 - p_{11} & 1 - p_{00} \end{pmatrix}$$

e, in particolare, se poniamo $\pi_0 = (1 - p_{11})/\rho$ e $\pi_1 = (1 - p_{00})/\rho$, si ha $\lim_n p_{ij}^{(n)} = \pi_j$. Questa catena di Markov, se $0 < \rho < 2$, ha quindi un comportamento asintoticamente regolare, nel senso che l'influenza dello stato di partenza sulla probabilità di trovarsi in uno stato o nell'altro diviene trascurabile (con velocità esponenziale), tendendo a un limite che dipende solo dallo stato di arrivo. Notiamo inoltre che nelle ipotesi fatte $\pi_i > 0$, $i = 0, 1$, e $\pi_0 + \pi_1 = 1$, ovvero i π_i formano una distribuzione di probabilità. Infine, il vettore $\pi = (\pi_0, \pi_1)$ soddisfa $\pi = \mathbb{P}\pi$, come è immediato verificare.

Supponiamo che una data catena abbia una distribuzione di probabilità iniziale $\pi = (\pi_i)$ che soddisfa $\sum_i \pi_i = 1$ e $\pi = \mathbb{P}\pi$, ovvero

$$\pi_j = \sum_{i\,:\,x_i \in X} \pi_i \, p_{ij}, \qquad x_j \in X. \tag{4.20}$$

Allora segue per induzione che

$$\pi_j = \sum_{i \,:\, x_i \in X} \pi_i \, p_{ij}^{(n)}, \qquad x_j \in X. \tag{4.21}$$

Pertanto se π_i è la probabilità che $\xi_0 = x_i$, allora la somma a destra in (4.21) è la probabilità che $\xi_n = x_j$, e dunque dal confronto di (4.20) e (4.21) segue che la probabilità dell'evento $\{\xi_n = j\}$ è la stessa per tutti gli n. Per questa ragione, una distribuzione di probabilità che soddisfa la (4.20) si chiama *distribuzione stazionaria* o anche *distribuzione di equilibrio*. Una catena che ammette una distribuzione stazionaria tale che $\lim p_{ij}^{(n)} = \pi_j > 0$ per ogni i, j si dice *ergodica*.

(4.22) Teorema ergodico. Una catena di Markov finita, irriducibile e aperiodica è ergodica, e si possono trovare due costanti $C > 0$ e $0 \le \theta < 1$ tali che

$$|p_{ij}^{(n)} - \pi_j| \le C\,\theta^n, \qquad \forall x_i, x_j \in X.$$

Dimostrazione. Poniamo $\alpha_j^{(n)} = \min_i p_{ij}^{(n)}$ e $\beta_j = \max_i p_{ij}^{(n)}$. Allora si ha

$$\begin{aligned} \alpha_j^{(n+1)} &= \min_i \sum_l p_{il} p_{lj}^{(n)} \ge \alpha_j^{(n)} \sum_l p_{il} = \alpha_j^{(n)} \\ \beta_j^{(n+1)} &= \max_i \sum_l p_{il} p_{lj}^{(n)} \le \beta_j^{(n)} \sum_l p_{il} = \beta_j^{(n)}, \end{aligned}$$

da cui otteniamo

$$0 \le \alpha_j^{(1)} \le \alpha_j^{(2)} \le \cdots \beta_j^{(2)} \le \beta_j^{(1)} \le 1.$$

Supponiamo inizialmente che $\min_{ij} p_{ij} = \delta > 0$. Evidentemente dovrà essere $\delta \le N^{-1}$. Indichiamo poi con $\sum_k^+$ ($\sum_k^-$) la somma rispetto agli indici k per i cui i termini sono positivi (negativi). Allora si ha

$$\begin{aligned} \beta_j^{(n+1)} - \alpha_j^{(n+1)} &\le p_{ij}^{(n+1)} - p_{lj}^{(n+1)} \\ &= \sum_k (p_{ik} - p_{lk})\, p_{kj}^{(n)} \\ &\le \sum_k {}^{+}(p_{ik} - p_{lk})\, \beta_j^{(n)} + \sum_k {}^{-}(p_{ik} - p_{lk})\, \alpha_j^{(n)}, \\ &= \sum_k {}^{+}(p_{ik} - p_{lk})\, (\beta_j^{(n)} - \alpha_j^{(n)}) \\ &\le (1 - N\delta)(\beta_j^{(n)} - \alpha_j^{(n)}) \end{aligned}$$

dove nell'ultima disuguaglianza abbiamo usato il fatto che $\sum_k^+ p_{lk} + \sum_k^- p_{ik} \geq N\delta$ e dunque $\sum_k^+(p_{ik} - p_{lk}) = 1 - \sum_k^- p_{ik} - \sum_k^+ p_{lk} \geq 1 - N\delta$. Da ciò segue che $\beta_j^{(n)} - \alpha_j^{(n)} \leq (1 - N\delta)^n$. Pertanto $\beta_j^{(n)}$ e $\alpha_j^{(n)}$ ammettono un limite comune π_j e si ha $|p_{ij}^{(n)} - \pi_j| \leq (1 - N\delta)^n$. Ovviamente si ha $\pi_j \geq \inf \alpha_j^{(n)} = \alpha_j^{(1)} > \delta$. L'asserto segue prendendo $C = 1$, $\theta = 1 - N\delta$ e passando al limite in $p_{ij}^{(n+1)} = \sum_k p_{ik} p_{kj}^{(n)}$, da cui si vede che i numeri π_j formano una distribuzione stazionaria. E' facile infine rendersi conto che la distribuzione π è necessariamente unica. Se infatti vi fosse un'altra distribuzione stazionaria π', allora

$$\pi'_j = \sum_i \pi'_i p_{ij} = \cdots = \sum_i \pi'_i p_{ij}^{(n)}$$

e dal momento che $p_{ij}^{(n)} \to \pi_j$ si avrebbe

$$\pi'_j = \sum_i \pi'_i \pi_j = \pi_j.$$

Concludiamo osservando che, nel caso generale abbiamo, in virtù del Lemma (4.27), $\min_{ij} p_{ij}^{(\ell)} = \delta > 0$ per $\ell > n_0$. Ripetendo l'argomento usato più sopra otteniamo che, se $n = m\ell + k$ con $0 \leq k < \ell$, allora $\beta_j^{(n)} - \alpha_j^{(n)} \leq (1 - N\delta)^m$, per cui basterà scegliere $C = (1 - N\delta)^{-1}$ e $\theta = (1 - N\delta)^{1/\ell}$. q.e.d.

(4.23) Osservazione. Nelle ipotesi del Teorema (4.22) la matrice $\mathbb{P}^n$ tende alla matrice le cui righe sono date da $(\pi_1, \pi_2, \ldots, \pi_N)$ (assumiamo $|X| = N$). Da ciò segue che data un'arbitraria distribuzione iniziale $\mu = (\mu_1, \mu_2, \ldots)$ su X si ha che $\mu\mathbb{P}^n \to \pi$ con $\pi\mathbb{P}^n = \pi$ per ogni n. In altre parole, la distribuzione di probabilità degli stati del sistema al tempo n, diviene, quando $n \to \infty$, indipendente dalla distribuzione iniziale. Da qui il nome 'ergodico' per un processo siffatto.

(4.24) Osservazione. Osserviamo che una distribuzione stazionaria (anche unica) può esistere anche per catene non ergodiche. Un esempio è dato da (4.28), ovvero da (4.29) con $\rho = 2$, dove evidentemente il limite $\lim p_{ij}^{(n)}$ non esiste. D'altra parte l'equazione (4.20) si riduce alla coppia di equazioni $\pi_0 = \pi_1$ e $\pi_1 = \pi_0$ la cui unica soluzione che soddisfa $\pi_0 + \pi_1 = 1$ è $\pi = (1/2, 1/2)$. Notiamo che se anche $\lim p_{ij}^{(n)}$ non esiste, le probabilità π_j possono in questo caso essere ottenute come limiti delle medie aritmetiche:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} p_{ij}^{(k)} = \pi_j.$$

Più in generale, in riferimento a (4.29), ogni matrice stocastica 2×2 tale che $\rho = 2 - \mathrm{tr}\,\mathbb{P} > 0$ ammette la distribuzione stazionaria:

$$\pi_0 = \frac{1 - p_{11}}{\rho}, \qquad \pi_1 = \frac{1 - p_{00}}{\rho}.$$

Osserviamo che, in questo caso, la condizione $\rho > 0$ equivale all'irriducibilità, mentre l'aperiodicità è soddisfatta se, in aggiunta, $0 < \rho < 2$. Se invece prendiamo $\rho = 0$ otteniamo la matrice

$$\mathbb{P} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

che non è neppure irriducibile. In questo caso l'equazione (4.20) si riduce alla coppia di equazioni $\pi_0 = \pi_0$ e $\pi_1 = \pi_1$ che ammettono infinite soluzioni della forma $\pi = (p, 1-p)$ con $p \in [0,1]$. Ciascuna di queste soluzioni si può esprimere come una combinazione convessa $\pi = p\pi^{(1)} + (1-p)\pi^{(2)}$ delle due soluzioni 'estreme' $\pi^{(1)} = (1,0)$ e $\pi^{(2)} = (0,1)$ (decomposizione ergodica).

(4.25) Esempio. Sia

$$\mathbb{P} = \begin{pmatrix} p_0 & p_1 & \cdots & p_{N-1} \\ p_{N-1} & p_0 & \cdots & p_{N-2} \\ \vdots & \vdots & \vdots & \vdots \\ p_1 & p_2 & \cdots & p_0 \end{pmatrix}$$

Le righe di $\mathbb{P}$ sono permutazioni cicliche della prima riga: $p_{ij} = p_{j-i}$ con $j - i$ valutato $(\mathrm{mod}\, N)$. Osserviamo che non solo la somma degli elementi di ogni riga ma anche quelli degli elementi di ogni colonna dà 1. Si dice in questo caso che $\mathbb{P}$ è una matrice *bistocastica*. Da ciò segue facilmente che l'equazione (4.20) ha l'unica soluzione $\pi_j = 1/N$, cioè la distribuzione uniforme. Se $p_i > 0$, dal Teorema (4.22) si ha che $p_{ij}^{(n)}$ converge a $1/N$ con velocità esponenziale. Se allora $\xi_0, \eta_1, \ldots, \eta_n$ sono v.a.i.i.d. che prendono valori in $X = \{0, 1, \ldots, N-1\}$ con distribuzione $p = (p_0, p_1, \ldots p_{N-1})$, e se $\xi_n = \xi_0 + \eta_1 + \cdots + \eta_n \,(\mathrm{mod}\, N)$, allora $\mathrm{P}\{\xi_n = j\} \to 1/N$ quando $n \to \infty$. Le v.a. ξ_n descrivono una passeggiata aleatoria su un cerchio di N punti (condizioni periodiche al contorno), e qualunque sia la distribuzione iniziale, la posizione diviene nel limite equidistribuita.

Vediamo ora più in dettaglio il significato probabilistico della distribuzione π. A questo proposito, siano $f_n^{(i)}(\omega)$ e $f_n^{(ij)}(\omega)$ le variabili aleatorie che forniscono la frequenza relativa dei $0 \leq k < n$ per cui $\xi_k(\omega) = x_i$, e di quelli per cui $\xi_k(\omega) = x_i$ e $\xi_{k+1}(\omega) = x_j$, rispettivamente. Allora si ha la seguente legge (debole) dei grandi numeri:

(4.26) Teorema. Per una catena finita ed ergodica si ha per ogni $\delta > 0$ e per $n \to \infty$

$$\mathrm{P}\,\Big(\{\omega\,:\,|f_n^{(i)}(\omega) - \pi_i| \geq \delta\}\Big) \to 0, \qquad 1 \leq i \leq N,$$

e

$$\mathrm{P}\,\Big(\{\omega\,:\,|f_n^{(ij)}(\omega) - \pi_i\, p_{ij}| \geq \delta\}\Big) \to 0, \qquad 1 \leq i, j \leq N.$$

Dimostrazione. Poniamo, con ovvio significato dei simboli,

$$f_n^{(i)} = \frac{1}{n}\sum_{k=0}^{n-1} 1_i^{(k)}, \qquad f_n^{(ij)} = \frac{1}{n}\sum_{k=0}^{n-1} 1_{ij}^{(k)}.$$

Osserviamo inoltre che, se μ è la distribuzione iniziale su X, allora per l'ergodicità si ha $\mathrm{E}\,1_i^{(k)} = \sum_{l=1}^{N} \mu_l p_{li}^{(k)} \to \pi_i$ e $\mathrm{E}\,1_{ij}^{(k)} = \sum_{l=1}^{N} \mu_l p_{li}^{(k-1)} p_{ij} \to \pi_i p_{ij}$ quando $k \to \infty$. Di conseguenza

$$\mathrm{E}\, f_n^{(i)} \to \pi_i, \qquad \mathrm{E}\, f_n^{(ij)} \to \pi_i\, p_{ij}.$$

Ma allora, se n è grande abbastanza, valgono le inclusioni

$$\{\omega\,:\,|f_n^{(i)}(\omega) - \pi_i| \geq \delta\} \subseteq \{\omega\,:\,|f_n^{(i)}(\omega) - \mathrm{E}\, f_n^{(i)}| \geq \delta/2\},$$

e

$$\{\omega\,:\,|f_n^{(ij)}(\omega) - \pi_i p_{ij}| \geq \delta\} \subseteq \{\omega\,:\,|f_n^{(ij)}(\omega) - \mathrm{E}\, f_n^{(ij)}| \geq \delta/2\}.$$

Possiamo ora usare la seconda disuguaglianza di Chebyshev per ottenere

$$\mathrm{P}\,\Big(\{\omega\,:\,|f_n^{(i)}(\omega) - \mathrm{E}\, f_n^{(i)}| \geq \delta/2\}\Big) \leq \frac{4\mathrm{Var}\, f_n^{(i)}}{\delta^2},$$

e similmente

$$\mathrm{P}\left(\{\omega\,:\,|f_n^{(ij)}(\omega)-\mathrm{E}\,f_n^{(ij)}|\geq\delta/2\}\right)\leq\frac{4\mathrm{Var}\,f_n^{(ij)}}{\delta^2}.$$

Ci resta dunque da stimare $\mathrm{Var}\,f_n^{(i)}$ e $\mathrm{Var}\,f_n^{(ij)}$. Poniamo dunque $m_i^{(k)}=\mathrm{E}\,1_i^{(k)}=\sum_{l=1}^N\mu_l p_{li}^{(k)}$ così che

$$\begin{aligned}\mathrm{Var}\,f_n^{(i)}&=\mathrm{E}\left(\frac{\sum_{k=0}^{n-1}(1_i^{(k)}-m_i^{(k)})}{n}\right)^2\\&=\frac{\sum_{k=0}^{n-1}\mathrm{E}\,(1_i^{(k)}-m_i^{(k)})^2}{n^2}+2\,\frac{\sum_{k_1<k_2}\mathrm{E}\,(1_i^{(k_1)}-m_i^{(k_1)})(1_i^{(k_2)}-m_i^{(k_2)})}{n^2}\end{aligned}$$

Ora $0\leq 1_i^{(k)}\leq 1$ implica $-1\leq 1_i^{(k)}-m_i^{(k)}\leq 1$, $(1_i^{(k)}-m_i^{(k)})^2\leq 1$ e dunque $\sum_{k=0}^{n-1}\mathrm{E}\,(1_i^{(k)}-m_i^{(k)})^2\leq n$. Inoltre si ha

$$\begin{aligned}\mathrm{E}\,(1_i^{(k_1)}-m_i^{(k_1)})(1_i^{(k_2)}-m_i^{(k_2)})&=\mathrm{E}\,1_i^{(k_1)}1_i^{(k_2)}-m_i^{(k_1)}m_i^{(k_2)}\\&=\sum_{l=1}^N\mu_l p_{li}^{(k_1)}p_{ii}^{(k_2-k_1)}-m_i^{(k_1)}m_i^{(k_2)}=R_{k_1,k_2}.\end{aligned}$$

Ora, per il teorema ergodico si ha $m_i^{(k)}=\pi_i+A_i^{(k)}$ e $p_{ij}^{(k)}=\pi_i+B_{ij}^{(k)}$ con $\max_i|A_i^{(k)}|\leq C\theta^k$ e $\max_{i,j}|B_{ij}^{(k)}|\leq C\theta^k$. Pertanto

$$\begin{aligned}|R_{k_1,k_2}|&=\left|\sum_{l=1}^N\mu_l(\pi_i+B_{li}^{(k_1)})(\pi_i+B_{ii}^{(k_2-k_1)})-(\pi_i+A_i^{(k_1)})(\pi_i+A_i^{(k_2)})\right|\\&\leq(1+\pi_i)\left(|B_{li}^{(k_1)}|+|B_{ii}^{(k_2-k_1)}|+|A_i^{(k_1)}|+|A_i^{(k_2)}|\right)\\&\leq C_1\left(\theta^{k_1}+\theta^{k_2-k_1}+\theta^{k_1}+\theta^{k_2}\right).\end{aligned}$$

Da cui $|\sum_{k_1<k_2}R_{k_1,k_2}|\leq C_2\,n$. Mettendo tutto insieme si ottiene allora $\mathrm{Var}\,f_n^{(i)}=\mathcal{O}(n^{-1})$. In modo simile si stima $\mathrm{Var}\,f_n^{(ij)}$. q.e.d.

Sia $C\subseteq X$ un sottoinsieme di stati. La frazione di tempo spesa nell'insieme C sarà data da

$$f_n^{(C)}=\frac{1}{n}\sum_{k=0}^{n-1}1_C(\xi_k).\tag{4.27}$$

Osserviamo che

$$\mathrm{E}\,(1_C(\xi_k)|\xi_0=x_i)=\mathrm{P}\,(\xi_k\in C|\xi_0=x_i)=\sum_{j\in C}p_{ij}^{(k)},\tag{4.28}$$

e dunque

$$(4.29)\qquad \mathrm{E}\,(f_n^{(C)}|\xi_0 = x_i) = \frac{1}{n}\sum_{k=0}^{n-1}\sum_{j\in C} p_{ij}^{(k)}.$$

Dal teorema ergodico segue che

$$(4.30)\qquad \mathrm{E}\,(f_n^{(C)}|\xi_0 = x_i) \to \pi_C = \sum_{j\in C}\pi_j,$$

Un immediato corollario del Teorema (4.26) è che, in effetti,

$$(4.34)\qquad \mathrm{P}\,\Big(\{\omega\,:\,|f_n^{(C)}(\omega) - \pi_C| \geq \delta\}\Big) \to 0, \qquad n\to\infty,$$

per ogni $\delta > 0$ e ogni distribuzione iniziale.

(4.35) Entropia di una catena di Markov. Nel caso di sequenze di prove indipendenti abbiamo visto (teorema di Shannon-McMillan) che per sequenze 'tipiche', l'entropia h è uguale al limite per $n\to\infty$ di $-(1/n)\log p(\omega)$. Ora, nel caso di una catena di Markov con distribuzione iniziale (μ_i) si ha

$$-(1/n)\log p(\omega) = -(1/n)\log \mu_{\omega_0} - \sum_{ij} f_n^{(ij)}(\omega)\,\log p_{ij}.$$

Dalla legge dei grandi numeri si ha che per sequenze tipiche $f_n^{(ij)}(\omega) \sim \pi_i\, p_{ij}$. Per tali sequenze si ha dunque

$$-(1/n)\log p(\omega) \sim -\sum_{ij}\pi_i\, p_{ij}\,\log p_{ij} =: h.$$

L'ultimo termine definisce dunque l'entropia per una catena di Markov finita, irriducibile ed aperiodica.

(4.36) Proprietà markoviana forte. La (4.13) prende il nome di proprietà markoviana della sequenza $\xi_0,\dots,\xi_n$. Mostreremo ora una versione più forte di questa proprietà, ossia il fatto che essa rimane valida quando il tempo k venga sostituito da un tempo aleatorio. A tale scopo, diremo (informalmente) che una v.a. τ è un *tempo di arresto* rispetto al processo $\xi_0,\xi_1,\dots$ se il verificarsi (o il non

verificarsi) dell'evento $\{\tau = n\}$ può essere determinato osservando unicamente i valori del processo fino a quel tempo: $\xi_0, \ldots, \xi_n$.
Ad esempio la variabile che dà il tempo di primo ritorno in uno stato $x_i \in X$, ossia

$$\tau_i = \min\{n \geq 1 : \xi_n = x_i\}$$

è un tempo di arresto perchè

$$\{\tau_i = n\} = \{\xi_1 \neq x_i, \ldots \xi_{n-1} \neq x_i, \xi_n = x_i\}$$

Dal momento che il verificarsi dell'evento $\{\tau = n\}$ dipende solo dai valori di $\xi_0, \ldots, \xi_n$ e in una catena di Markov la distribuzione del futuro dipende dal passato solo attraverso lo stato presente, è abbastanza ragionevole aspettarsi che la proprietà markoviana (4.13) debba valere quando il tempo k è sostituito con un tempo di arresto τ.

(4.37) Teorema. Sia τ un tempo di arresto. Se $\tau = n$ e $\xi_\tau = x_i$, ogni altra informazione su $\xi_0, \ldots, \xi_\tau$ è irrilevante per la predizione del futuro, e $\xi_{\tau+k}$, $k \geq 0$, si comporta come una catena di Markov con stato inziale x_i. In particolare si ha

$$\mathrm{P}\left(\xi_{\tau+k} = x_j \,|\, \xi_\tau = x_i, \tau = n\right) = p_{ij}^{(k)}.$$

Dimostrazione. Consideriamo per semplicità transizioni ad un passo, cioè $k = 1$. Il caso generale segue facilmente per induzione. Sia C_n l'insieme delle parole $\omega = \omega_0\omega_1 \ldots \omega_n \in X^{[0,n]}$ tali che se $\xi_0 = \omega_0, \ldots, \xi_n = \omega_n$ allora $\tau = n$ e $\xi_\tau = x_i$. Si ha

$$\begin{aligned}
&\mathrm{P}\left(\xi_{\tau+1} = x_j, \xi_\tau = x_i, \tau = n\right) \\
&\quad = \sum_{\omega \in C_n} \mathrm{P}\left(\xi_{n+1} = x_j, \xi_n = \omega_n, \ldots, \xi_0 = \omega_0\right) \\
&\quad = \sum_{\omega \in C_n} \mathrm{P}\left(\xi_{n+1} = x_j | \xi_n = \omega_n, \ldots, \xi_0 = \omega_0\right) \mathrm{P}\left(\xi_n = \omega_n, \ldots, \xi_0 = \omega_0\right) \\
&\quad = p_{ij} \sum_{\omega \in C_n} \mathrm{P}\left(\xi_n = \omega_n, \ldots, \xi_0 = \omega_0\right) = p_{ij}\, \mathrm{P}\left(\xi_\tau = x_i, \tau = n\right). \qquad \text{q.e.d.}
\end{aligned}$$

(4.37) Probabilità tabù e ulteriore classificazione degli stati.

Introduciamo le quantità (dette probabilità *tabù*)

$$f_{ij}^{(k)} = \mathrm{P}\left(\{\xi_k = x_j, \xi_l \neq x_j, 1 \leq l < k | \xi_0 = x_i\}\right). \tag{4.38}$$

In particolare, $f_{ii}^{(k)}$ fornisce la probabilità di osservare il primo ritorno allo stato x_i al tempo k. Se poniamo

$$\tau_{ii} = \min\{n \geq 1 : \xi_n = x_i \,|\, \xi_0 = x_i\}$$

allora la probabilità r_{ii} di tornare in x_i almeno una volta è data da

$$r_{ii} = \mathrm{P}\left(\tau_{ii} < \infty\right) = \sum_{k \geq 0} f_{ii}^{(k)}.$$

Per la proprietà markoviana forte, la probabilità di tornare in x_i ancora una volta, dopo che ci si è ritornati $n-1$ volte, è ancora r_{ii} e quindi la probabilità di tornarvi n volte è r_{ii}^n. A questo punto vi sono due possibilità:

1. $r_{ii} < 1$. In tal caso la probabilità di tornare n volte in x_i tende a zero. Lo stato x_i si dice in questo caso *transiente* in quanto dopo un certo tempo non riapparirà più nella catena di Markov;
2. $r_{ii} = 1$. In tal caso la probabilità di tornare n volte in x_i è sempre uguale a 1 e quindi la catena vi ritornerà infinite volte. In questo caso lo stato x_i si dice *ricorrente*. È evidente che una catena di Markov irriducibile ha tutti gli stati ricorrenti.

Esempio. Per comprendere più da vicino questa classificazione riprendiamo l'Esempio 4.7 con $X = \{0, 1, \ldots, N-1, N\}$, $p_{ij} = p\,\delta_j^{i+1} + q\,\delta_j^{i-1}$, per $0 < i, j < N$ e $p_{00} = p_{NN} = 1$. È chiaro che $\mathrm{P}\left(\tau_{00} < \infty\right) = \mathrm{P}\left(\tau_{00} = 1\right) = 1$ e dunque $r_{00} = 1$. Similmente si vede che $r_{NN} = 1$ e dunque 0 e N sono stati ricorrenti. Al contrario gli stati $1, \ldots, N-1$ sono tutti transienti. Consideriamo ad esempio lo stato iniziale 1. È evidente che $P(\tau_{11} = \infty) \geq q$ e dunque $P(\tau_{11} < \infty) \leq 1 - q < 1$. In modo simile si vede per gli altri stati. Osserviamo che l'equazione (4.21) ha un'infinità di soluzioni della forma $\pi = (\alpha, 0, \ldots, 0, \beta)$ con $\alpha + \beta = 1$ e gli elementi di matrice $p_{ij}^{(n)}$ non ammettono limite se $0 < i, j < N$. Si possono facilmente verficare queste affermazioni nel caso $N = 2$ per cui si ha

$$\mathbb{P}^n = \begin{pmatrix} 1 & 0 & 0 \\ q & 0 & p \\ 0 & 0 & 1 \end{pmatrix}, \qquad \forall n \geq 1.$$

Se ora poniamo

$$\tau_{ij} = \min\{n \geq 1 : \xi_n = x_j \,|\, \xi_0 = x_i\}$$

allora la probabilità r_{ij} di visitare x_j almeno una volta (partendo da x_i) è data da

$$r_{ij} = \mathrm{P}\,(\tau_{ij} < \infty) = \sum_{k \geq 0} f_{ij}^{(k)}.$$

Diremo che gli stati x_i e x_j sono *comunicanti* se $r_{ij} = 1$. Osserviamo che se x_i comunica con x_j allora esiste un $n > 0$ tale che $p_{ij}^{(n)} > 0$. Scegliendo il più piccolo n con questa proprietà avremo in particolare

$$\mathrm{P}\,(\tau_{ii} = \infty) = 1 - r_{ii} \geq p_{ik_1} p_{k_1 k_2} \cdots p_{k_{n-1} j}(1 - r_{ji})$$

per qualche sequenza $k_1, \ldots, k_{n-1}$ con $k_l \neq j$. D'altra parte se x_i è ricorrente il termine a sinistra è nullo e dunque deve essere $r_{ji} = 1$, cioè x_j comunica con x_i. Vale dunque la seguente proprietà: *se x_i comunica con x_j ma x_j non comunica con x_i allora x_i è transiente.*

Sia ν_j il numero di visite di ξ_n allo stato x_j, con $n = 1, 2, \ldots$, e poniamo

$$\mathrm{E}\,_i \nu_j \equiv \mathrm{E}\,(\nu_j | \xi_0 = x_i).$$

Per calcolare questà quantità osserviamo che per una v.a. discreta ξ a valori interi positivi si ha

$$\mathrm{E}\,\xi = \sum_{k=1}^{\infty} k\,\mathrm{P}\,(\xi = k) = \sum_{k=1}^{\infty} \sum_{l \geq k} \mathrm{P}\,(\xi = l) = \sum_{k=1}^{\infty} \mathrm{P}\,(\xi \geq k).$$

Applicando questa proprietà alla variabile ν_j e osservando che

$$\mathrm{P}\,_i(\nu_j \geq k) \equiv \mathrm{P}\,(\nu_j \geq k | \xi_0 = x_i) = r_{ij} r_{jj}^{k-1}$$

si trova la seguente identità:

$$\mathrm{E}\,_i \nu_j = \frac{r_{ij}}{1 - r_{jj}},$$

da cui segue che *x_j è ricorrente se e solo se* $\mathrm{E}\,_j \nu_j = \infty$.

Concludiamo osservando che una condizione sufficiente per l'ergodicità di una catena finita è che tutti gli stati comunichino tra loro e che ne esista uno, diciamo x_h, tale che $p_{hh} > 0$. Infatti, sotto la prima ipotesi, per ogni i, j si può trovare $n = n(i,j)$ tale che $p_{ij}^{(n)} > 0$. Allora se $m = \max_{i,j} n(i,j)$ allora

$$p_{ij}^{(2m)} \geq p_{ih}^{(n(i,h))} p_{hh}^{(2m-n(i,h)-n(h,j))} p_{hj}^{(n(h,j))} > 0$$

e dunque $\mathbb{P}^{2m}$ ha tutti gli elementi positivi.

(4.39) Rappresentazione rigenerativa delle probabilità di transizione e della distribuzione stazionaria.

Lemma. Le probabilità *tabù* sono legate alle probabilità di transizione dalla formula

$$p_{ij}^{(n)} = \sum_{k=1}^{n} f_{ij}^{(k)} p_{jj}^{(n-k)}, \quad \text{con} \quad p_{jj}^{(0)} = 1. \tag{4.40}$$

Il significato intuitivo della (4.40) è chiaro: per andare dallo stato x_i allo stato x_j in n passi è necessario raggiungere lo stato x_j per la prima volta in k passi $(1 \leq k \leq n)$ e dunque ritornare allo stato x_j in altri $n-k$ passi. Diamone ora un derivazione rigorosa. Sia x_j uno stato assegnato e poniamo

$$\tau_j = \min\{1 \leq k \leq n : \xi_k = x_j\}, \quad \text{con} \quad \min \emptyset = n+1.$$

Allora si ha $f_{ij}^{(k)} = \mathrm{P}(\tau_j = k | \xi_0 = x_i)$ e

$$\begin{aligned}
p_{ij}^{(n)} &= \mathrm{P}(\xi_n = x_j | \xi_0 = x_i) \\
&= \sum_{k=1}^{n} \mathrm{P}(\xi_n = x_j, \tau_j = k | \xi_0 = x_i) \\
&= \sum_{k=1}^{n} \mathrm{P}(\xi_{\tau_j+n-k} = x_j, \tau_j = k | \xi_0 = x_i) \\
&= \sum_{k=1}^{n} \mathrm{P}(\xi_{\tau_j+n-k} = x_j | \xi_0 = x_i, \tau_j = k) \mathrm{P}(\tau_j = k | \xi_0 = x_i) \\
&= \sum_{k=1}^{n} \mathrm{P}(\xi_{\tau_j+n-k} = x_j | \xi_0 = x_i, \tau_j = k) f_{ij}^{(k)}
\end{aligned}$$

dove la terza identità segue dal fatto che $\xi_{\tau_j+n-k} = \xi_n$ sull'insieme $\{\tau_j = k\}$. D'altra parte si ha per definizione $\{\tau_j = k\} = \{\tau_j = k, \xi_{\tau_j} = x_j\}$ e dunque, dal Teorema (4.37) si ha che, se $\mathrm{P}(\xi_0 = x_i, \tau_j = k) > 0$,

$$\mathrm{P}(\xi_{\tau_j+n-k} = x_j | \xi_0 = x_i, \tau_j = k) = \mathrm{P}(\xi_{\tau_j+n-k} = x_j | \xi_{\tau_j} = x_j, \tau_j = k) = p_{jj}^{(n-k)}$$

e dunque la (4.39).

Supponiamo ora che la catena sia finita, irriducibile e aperiodica e denotiamo con ρ_i il numero medio di passi necessari a ritornare per la prima volta nello stato x_i cioè

$$\rho_i = \sum_{k\geq 1} k \cdot f_{ii}^{(k)}. \tag{4.41}$$

Ora, il numero $\nu_n^{(i)}$ di ritorni a x_i nei primi n passi sarà circa n/ρ_i se n è abbastanza grande, ovvero $\nu_n^{(i)}/n \approx 1/\rho_i$. D'altra parte, l'aspettazione di $\nu_n^{(i)}/n$ è data da $n^{-1}\sum_{k=0}^{n} p_{ii}^{(k)}$ che per il teorema ergodico converge a π_i. Pertanto dovremmo aspettarci che

$$\pi_i = \frac{1}{\rho_i}, \qquad \forall i. \tag{4.42}$$

Limitiamoci a verificare questa relazione per l'Esempio (4.29), sotto l'ipotesi che $\rho > 0$. Supponiamo dapprima che la matrice $\mathbb{P}$ abbia tutti gli elementi positivi. Allora tracciando il grafo associato a $\mathbb{P}$ è facile ricavare le seguenti espressioni per le probabilità tabù:

$$\begin{aligned} f_{ii}^{(1)} = p_{ii}, \quad f_{ii}^{(k)} &= p_{ij}p_{jj}^{(k-2)}p_{ji}, \quad k > 1, \\ f_{ij}^{(k)} &= p_{ii}^{(k-1)}p_{ij}, \quad k \geq 1, \end{aligned} \tag{4.43}$$

dove si suppone $i \neq j$. È immediato verificare che

$$\sum_{k\geq 1} f_{ij}^{(k)} = 1, \quad \forall i, j \in \{0, 1\}. \tag{4.44}$$

La serie in (4.43) fornisce la probabilità che lasciando lo stato i si raggiunga prima o poi lo stato j (in particolare, se $i = j$ avremo la probabilità di tornare prima o poi in i). L'identità (4.44) non è dunque sorprendente essendo per ipotesi la catena irriducibile. Inoltre si trova

$$
\begin{aligned}
\rho_i &= \sum_{k\geq 1} k\cdot f_{ii}^{(k)} = p_{ii} + p_{ij}p_{ji}\sum_{k\geq 2} k\cdot p_{jj}^{(k-2)} \\
&= 1 + (1-p_{ii})(1-p_{jj})\sum_{k\geq 2}(k-1)\cdot p_{jj}^{(k-2)} \qquad (4.45)\\
&= 1 + \frac{1-p_{ii}}{1-p_{jj}} = \frac{1}{\pi_i},
\end{aligned}
$$

che è la relazione cercata. Osserviamo infine che nel caso $\rho = 2$ (cioè la catena dell'Esempio (4.28) che è irriducibile ma ha periodo 2), le relazioni (4.43) divengono $f_{ii}^{(k)} = \delta_k^2$ e $f_{ij}^{(k)} = \delta_k^1$, per cui la (4.44) continua a valere (banalmente) e inoltre $\rho_i = 2 = 1/\pi_i$. Che cosa distingue dunque il caso aperiodico da quello periodico? Nel caso aperiodico, cioè quando $0 < \rho < 2$ con $\rho = 2 - p_{00} - p_{11}$, segue dal teorema ergodico (e lo abbiamo verificato eseplicitamente) che $p_{ij}^{(n)} \to 1/\rho_i = \pi_j$. D'altra parte, se $\rho = 2$, il $\lim p_{ij}^{(n)}$ non esiste ma $\lim p_{jj}^{(2n)} = \lim p_{ij}^{(2n+1)} = 1 = 2/\rho_i$.

(4.46) Bilancio dettagliato.

La condizione di bilancio dettagliato:

$$\pi_i\, p_{ij} = \pi_j\, p_{ji},$$

esprime una proprietà di *reversibilità* della catena di Markov, più forte dell'invarianza $\pi\mathbb{P} = \pi$. Infatti, sommando sul primo indice si ottiene

$$\sum_i \pi_i\, p_{ij} = \pi_i \sum_i p_{ji} = \pi_j.$$

In modo figurato, pensiamo a π_i come alla quantità di sabbia in x_i e alla transizione $x_i \to x_j$ della catena come al trasporto della frazione p_{ij} (della sabbia in x_i) da x_i a x_j. La condizione di bilancio dettagliato afferma allora che la quantità di sabbia che va da x_i a x_j viene esattamente bilanciata da quella che va da x_j a x_i. Invece, la condizione $\pi\mathbb{P} = \pi$ afferma che una volta effettuati tutti gli scambi di sabbia, la quantità *totale* di sabbia che va a finire in uno stato è uguale a quella che se ne è dipartita. In generale, una condizione necessaria per avere il bilancio dettagliato è che $p_{ij} > 0$ implica $p_{ji} > 0$ (altrimenti si avrebbe $\pi_i = 0$, $\forall i$).

Esempio. Riprendiamo l'Esempio 4.6. L'equazione di bilancio dettagliato afferma in questo caso che

$$\pi_i \left(\frac{r-i}{r}\right)^2 = \pi_{i+1} \left(\frac{i+1}{r}\right)^2 .$$

Posto $\pi_0 = c$ si trova

$$\pi_k = c \prod_{i=0}^{k-1} \left(\frac{r-i}{i+1}\right)^2 = c \binom{r}{k}^2, \qquad k = 0, \ldots, r.$$

Pertanto, posto $1/c = \sum_{i=0}^{r} \binom{r}{k}^2$, la distribuzione π scritta sopra soddisfa il bilancio dettagliato, e in particolare è una distribuzione stazionaria.

(4.47) L'algoritmo di Metropolis. Sia ora $\mathbb{Q}$ una matrice di transizione irriducibile e simmetrica (e quindi bistocastica) e π una distribuzione su X tale che $\pi_i > 0$ per ogni $x_i \in X$. Poniamo

$$p_{ij} = \begin{cases} q_{ij}, & \text{se } \pi_j \geq \pi_i \\ q_{ij}\pi_j/\pi_i, & \text{se } \pi_j < \pi_i \\ 1 - \sum_{j \neq i} p_{ij}, & \text{se } j = i \end{cases} \tag{4.48}$$

In particolare, se $i \neq j$ allora $p_{ij} \leq q_{ij}$. È immediato verificare che anche $\mathbb{P} = (p_{ij})$ è una matrice stocastica. Inoltre π è reversibile (e dunque invariante) per $\mathbb{P}$. Infatti, supponendo che $\pi_j < \pi_i$, si ha

$$\pi_i \, p_{ij} = \pi_i \, q_{ij} \frac{\pi_j}{\pi_i} = \pi_j \, q_{ij} = \pi_j \, q_{ji} = \pi_j \, p_{ji},$$

e similmente si ottiene l'indentità di bilancio dettagliato per $\pi_j \geq \pi_i$. In più, se la distribuzione π non è uniforme allora $\mathbb{P}$ è ergodica. Per mostrare questo fatto, osserviamo innanzitutto che esistono due stati x_{i_0} e x_{j_0} tali che $q_{i_0 j_0} > 0$ e $\pi_{i_0} < \pi_{j_0}$: basta considerare il sottoinsieme $Y \subset X$ degli stati x_i tali che $\pi_i = \max_l \pi_l$ e usare l'irriducibilità di $\mathbb{Q}$ per trovare $i_0 \in Y$ e $j_0 \in X \setminus Y$ tali che $q_{i_0 j_0} > 0$. Inoltre si ha

$$\begin{aligned} p_{i_0 i_0} &= 1 - \sum_{j \neq i_0} p_{i_0 j} = 1 - \sum_{j \neq i_0, j_0} p_{i_0 j} - p_{i_0 j_0} \\ &\geq 1 - \sum_{j \neq i_0, j_0} q_{i_0 j} - q_{i_0 j_0} \frac{\pi_{j_0}}{\pi_{i_0}} \\ &= 1 - \sum_{j \neq i_0} q_{i_0 j} + q_{i_0 j_0} \left(1 - \frac{\pi_{j_0}}{\pi_{i_0}}\right) \\ &= q_{i_0 i_0} + q_{i_0 j_0} \left(1 - \frac{\pi_{j_0}}{\pi_{i_0}}\right) \geq q_{i_0 j_0} \left(1 - \frac{\pi_{j_0}}{\pi_{i_0}}\right) > 0 \end{aligned}$$

da cui segue l'asserto (in base all'osservazione fatta alla fine del punto (4.37)). Dal teorema ergodico segue che $\mu \mathbb{P}^n \to \pi$ per ogni distribuzione iniziale μ. La (4.48) consente quindi di costruire, a partire da una qualunque matrice di transizione $\mathbb{Q}$ simmetrica e irriducibile, una nuova matrice di transizione $\mathbb{P}$ la cui catena di Markov associata sia ergodica e abbia una prefissata distribuzione limite. Questo metodo prende il nome di *algoritmo di Metropolis* ed ha importanti applicazioni.

Supponiamo di voler simulare la scelta casuale di uno stato $x_i \in X$ con assegnata distribuzione (non uniforme) π. Se il numero degli stati è elevato questo compito può rivelarsi assai arduo usando sequenze di v.a. indipendenti. Possiamo però ottenere una v.a. con distribuzione approssimativamente uguale a π costruendo a partire da una matrice di transizione $\mathbb{Q}$ su X, simmetrica e irriducibile ma per altro arbitraria, la matrice $\mathbb{P}$ usando la (4.48) e generando la catena di Markov $\xi_0, \ldots, \xi_n$ associata. Se ad esempio $\xi_n = x_i$ basta scegliere a caso x_j con probabilità q_{ij} e quindi porre $\xi_{n+1} = x_j$ se $\pi_j \geq \pi_i$; se invece $\pi_j < \pi_i$ allora si pone $\xi_{n+1} = x_j$ con probabilità π_j/π_i, mentre con probabilità $1 - \pi_j/\pi_i$ si rifiuta la transizione e si lascia $\xi_{n+1} = x_i$. In questo modo, la probabilità della transizione $x_i \to x_i$ si ottiene sommando su tutti i 'rifiuti' possibili:

$$\begin{aligned} p_{ii} &= \sum_{j:\pi_j < \pi_i} q_{ij} \left(1 - \frac{\pi_j}{\pi_i}\right) \\ &= 1 - \sum_{j:\pi_j \geq \pi_i} q_{ij} - \sum_{j:\pi_j < \pi_i} q_{ij} \frac{\pi_j}{\pi_i} \\ &= 1 - \sum_{j \neq i} p_{ij} \end{aligned}$$

come prescritto.

Esempio: Ricottura simulata. Se, nell'algoritmo di Metropolis usiamo la distribuzione (di Boltzmann)

$$\pi_i(t) = \frac{e^{-H(x_i)/t}}{Z_t}$$

dove H è una funzione su X (*hamiltoniana*), t è un parametro positivo (*temperatura*) e Z_t è la costante di normalizzazione (*funzione di partizione*) definita da

$$Z_t = \sum_{x_i \in X} e^{-H(x_i)/t}$$

La regola di transizione (4.48) diviene

$$p_{ij}(t) = \begin{cases} q_{ij}, & \text{se } H(x_j) \leq H(x_i) \\ q_{ij} e^{-(H(x_j)-H(x_i))/t}, & \text{se } H(x_j) > H(x_i) \\ 1 - \sum_{j \neq i} p_{ij}(t), & \text{se } j = i \end{cases} \tag{4.49}$$

La regola è quindi la seguente: si sceglie un nuovo stato x_j secondo la matrice $\mathbb{Q}$. Se 'l'energia non cresce', cioè se $H(x_j) \leq H(x_i)$, allora si effettua la transizione; se invece $H(x_j) > H(x_i)$ allora ciò avviene solo con probabilità $e^{-(H(x_j)-H(x_i))/t}$, mentre la transizione viene rifiutata ed il processo resta in x_i con probabilità $1 - e^{-(H(x_j)-H(x_i))/t}$. Osserviamo che in questa procedura non occorre conoscere il valore della costante Z_t. Per quanto visto sopra, quando n è grande la catena ha una distribuzione approssimativamente uguale a $\pi(t)$. Ora, se t è piccolo questa distribuzione si concentra su quegli stati su cui H è piccola. È abbastanza immediato verificare che se $x_{i_1}, \ldots, x_{i_m}$ sono gli stati che corrispondono ai punti di minimo assoluto di H allora quando $t \to 0$ la distribuzione $\pi(t)$ converge alla distribuzione uniforme su tali punti. In particolare se x_i è l'unico minimo assoluto di H allora $\pi_j(t) \to 0$ per $j \neq i$ e $\pi_i(t) \to 1$.

Ciò suggerisce il seguente algoritmo di ottimizzazione globale: se si deve trovare il minimo assoluto di una funzione H su un insieme X, a partire da una qualunque matrice $\mathbb{Q}$ irriducibile e simmetrica su X si costruisce la matrice $\mathbb{P}(t) = (p_{ij}(t))$ con la regole (4.49) e poi si genera la catena di Markov corrispondente. Per quanto visto, se t è piccolo e n grande la catena si troverà con grande probabilità in uno stato x_i in cui H prende un valore molto vicino al suo valore di minimo assoluto. Questa procedura è di effettiva utilità quando l'insieme X ha grande cardinalità e la procedura di verifica sistematica di valori di H su tutti gli stati di X per vedere quale è più piccolo diviene proibitiva anche per le macchine più veloci.

Si può poi migliorare ulteriormente questo algoritmo costruendo una catena di Markov *non omogenea* in cui le probabilità di transizione sono calcolate ad ogni passo prendendo valori di t sempre più piccoli. Concretamente si determina un successione $t_n \searrow 0$ e si pone $p_{ij}(n) = p_{ij}(t_n)$. È un fatto ormai stabilito che se la successione t_n tende a zero abbastanza lentamente, la catena di Markov così ottenuta converge ad una distribuzione concentrata sui punti di minimo assoluto di H. Questa procedura, tra le più potenti per risolvere difficili problemi di ottimizzazione e di meccanica statistica, prende il nome di ricottura simulata (*simulated annealing*) (in inglese il termine 'annealing' indica un procedimento usato in metallurgia con il quale si ottengono materiali particolarmente resistenti per mezzo di un lentissimo raffreddamento, in opposizione a 'quenched' in cui il materiale viene raffreddato rapidamente 'congelando' le impurità presenti in posizioni definite).

5. ASPETTAZIONE CONDIZIONATA E MARTINGALE.

(5.1) Probabilità ed aspettazione condizionate rispetto a una partizione. Sia $(\Omega, \mathcal{F}, \mathrm{P})$ uno spazio di probabilità e $\mathcal{D} = \{D_i\}$ una partizione finita o numerabile di Ω, tale che $\mathrm{P}(D_i) > 0$ per ogni i. Dato $A \in \mathcal{F}$ definiamo la probabilità condizionata $\mathrm{P}(A|\mathcal{D})$ di A rispetto alla partizione $\mathcal{D}$ come la variabile aleatoria che è uguale a $\mathrm{P}(A|D_i)$ se $\omega \in D_i$:

$$\mathrm{P}(A|\mathcal{D})(\omega) = \sum_i \mathrm{P}(A|D_i)\, 1_{D_i}(\omega). \tag{5.2}$$

In modo analogo, sia ξ è una variabile aleatoria discreta per cui $\mathrm{E}\,\xi$ è definita. Poniamo $A_j = \{\omega : \xi(\omega) = x_j\}$. Allora l'aspettazione condizionata $\mathrm{E}(\xi|\mathcal{D})$ di ξ rispetto alla partizione $\mathcal{D}$ è data dalla variabile aleatoria che prende il valore $\sum_j x_j\, \mathrm{P}(A_j|D_i)$ se $\omega \in D_i$, ovvero:

$$\mathrm{E}(\xi|\mathcal{D}) = \sum_j x_j\, \mathrm{P}(A_j|\mathcal{D}) \tag{5.3}$$

Un'altro modo di definire $\mathrm{E}(\xi|\mathcal{D})$ è il seguente: si definisce innanzitutto, come in (3.1), l'aspettazione condizionata di ξ rispetto a D_i,

$$\mathrm{E}(\xi|D_i) = \frac{\mathrm{E}(\xi\, 1_{D_i})}{\mathrm{P}(D_i)} = \sum_j x_j\, \mathrm{P}(A_j|D_i) \tag{5.4}$$

e quindi

$$\mathrm{E}(\xi|\mathcal{D})(\omega) = \sum_i \mathrm{E}(\xi|D_i)\, 1_{D_i}(\omega). \tag{5.5}$$

Le seguenti proprietà seguono immediatamente dalle definizioni:

$$\begin{aligned} \mathrm{E}(a\xi + b\eta|\mathcal{D}) &= a\mathrm{E}(\xi|\mathcal{D}) + b\mathrm{E}(\xi|\mathcal{D}), \\ \mathrm{E}(\xi\Omega) &= \mathrm{E}\,\xi, \\ \mathrm{E}(c|\mathcal{D}) &= c, \quad c \text{ costante}; \end{aligned} \tag{5.6}$$

inoltre se $\xi = 1_A(\omega)$ si ha

$$\mathrm{E}(\xi|\mathcal{D}) = \mathrm{P}(A|\mathcal{D}),$$

da cui segue che le proprietà delle probabilità condizionate possono essere dedotte da quelle delle aspettazioni condizionate.

(5.7) Esercizio. Mostrare che la formula della probabilità totale (3.3) può essere riscritta nella forma compatta

$$\mathrm{E}\,\mathrm{P}\,(A|\mathcal{D}) = \mathrm{P}\,(A).$$

Da cui si ottiene l'importante relazione

$$\mathrm{E}\,\mathrm{E}\,(\xi|\mathcal{D}) = \sum_j x_j\, \mathrm{E}\,\mathrm{P}\,(A_j|\mathcal{D}) = \sum_j x_j\, \mathrm{P}\,(A_j) = \mathrm{E}\,\xi. \tag{5.8}$$

Sia ora $\eta = \eta(\omega)$ una variabile aleatoria che assume valori $y_1, y_2, \ldots$ con probabilità positive:

$$\eta(\omega) = \sum_j y_j\, 1_{D_j}(\omega) \tag{5.9}$$

dove $D_j = \{\omega : \eta(\omega) = y_j\}$. La probabilità condizionata $\mathrm{P}\,(A|\mathcal{D}_\eta)$ di A rispetto alla partizione $\mathcal{D}_\eta = \{D_j\}$ di Ω indotta da η si dirà *probabilità condizionata di A rispetto alla* v.a. η e si denoterà $\mathrm{P}\,(A|\eta)$. In modo analogo si definisce la probabilità condizionata $\mathrm{P}\,(A|\eta_1, \eta_2, \ldots, \eta_n)$ di A rispetto alle v.a. $\eta_1, \eta_2, \ldots, \eta_n$.

(5.10) Esempio. Date due v.a.i.i.d. ξ e η, ciascuna che prende i valori 0 e 1 con probabilità q e p, determiniamo le probabilità condizionate $\mathrm{P}\,(\xi + \eta = k|\eta)$ degli eventi $\xi + \eta = k$, $k = 0, 1, 2$, rispetto a η. A questo scopo, sarà conveniente osservare il seguente fatto generale: se ξ e η sono v.a. indipendenti che assumono valori x e y rispettivamente, allora

$$\begin{aligned}\mathrm{P}\,(\xi + \eta = z|\eta = y) &= \frac{\mathrm{P}\,(\xi + \eta = z, \eta = y)}{\mathrm{P}\,(\eta = y)} = \frac{\mathrm{P}\,(\xi + y = z, \eta = y)}{\mathrm{P}\,(\eta = y)}\\ &= \frac{\mathrm{P}\,(\xi + y = z)\mathrm{P}\,(\eta = y)}{\mathrm{P}\,(\eta = y)} = \mathrm{P}\,(\xi + y = z).\end{aligned}$$

Pertanto si avrà

$$\begin{aligned}\mathrm{P}\,(\xi + \eta = k|\eta) &= \mathrm{P}\,(\xi + \eta = k|\eta = 0)\, 1_{\eta=0}(\omega) + \mathrm{P}\,(\xi + \eta = k|\eta = 1)\, 1_{\eta=1}(\omega)\\ &= \mathrm{P}\,(\xi = k)\, 1_{\eta=0}(\omega) + \mathrm{P}\,(\xi = k - 1)\, 1_{\eta=1}(\omega),\end{aligned}$$

e dunque, essendo $1_{\eta=0} = 1-\eta$ e $1_{\eta=1} = \eta$,

$$\mathrm{P}\,(\xi+\eta=0|\eta) = q(1-\eta), \quad \mathrm{P}\,(\xi+\eta=1|\eta) = p(1-\eta)+q\eta, \quad \mathrm{P}\,(\xi+\eta=2|\eta) = p\eta.$$

Osserviamo che per la probabilità ordinaria $\mathrm{P}\,(\xi+\eta=k)$ si ha

$$\mathrm{P}\,(\xi+\eta=k) = \sum_{(x,y):x+y=k} \mathrm{P}\,(\xi=x)\,\mathrm{P}\,(\eta=y) = \sum_{x=0}^{1} \mathrm{P}\,(\xi=x)\,\mathrm{P}\,(\eta=k-x)$$

e dunque

$$\mathrm{P}\,(\xi+\eta=0) = q^2, \quad \mathrm{P}\,(\xi+\eta=1) = 2pq, \quad \mathrm{P}\,(\xi+\eta=2) = p^2.$$

(5.11) Misurabilità di una v.a. rispetto a una partizione. Sia $\mathcal{D} = \{D_j\}$ una partizione finita o numerabile di Ω. Diremo che un v.a. discreta η è misurabile rispetto alla partizione $\mathcal{D}$, o anche che è $\mathcal{D}$-misurabile, se $\mathcal{D}_\eta \preceq \mathcal{D}$ ($\mathcal{D}$ è più fine di $\mathcal{D}_\eta$), cioè se η può essere rappresentata nella forma

$$\eta(\omega) = \sum_j y_j 1_{D_j}(\omega),$$

dove alcuni y_j possono coincidere. In altre parole, una v.a. discreta è $\mathcal{D}$-misurabile se e solo se assume valori costanti sugli atomi di $\mathcal{D}$. Così, ad esempio, se $\mathcal{D}$ è la partizione banale $\mathcal{D} = \{\Omega\}$, allora η sarà $\mathcal{D}$-misurabile se e solo se $\eta \equiv c$, dove c è una costante. Inoltre, ogni v.a. η è misurabile rispetto a $\mathcal{D}_\eta$. Due proprietà ulteriori dell'aspettazione condizionata (la cui dimostrazione, che lasciamo per esercizio, consiste in un semplice calcolo basato sulle definizioni) sono le seguenti: se η è $\mathcal{D}$-misurabile, allora

$$\mathrm{E}\,(\xi\,\eta|\mathcal{D}) = \eta\,\mathrm{E}\,(\xi|\mathcal{D}), \tag{5.12}$$

e, in particolare,

$$\mathrm{E}\,(\eta|\mathcal{D}) = \eta, \qquad (\mathrm{E}\,(\eta|\mathcal{D}_\eta) = \eta). \tag{5.13}$$

Se $\mathcal{D}$ è la partizione indotta dalle v.a. $\eta_1, \ldots, \eta_n$, allora $\mathrm{E}\,(\eta|\mathcal{D})$ si dirà *aspettazione condizionata di* η *rispetto a* $\eta_1, \ldots, \eta_n$ e si denoterà con $\mathrm{E}\,(\eta|\eta_1, \ldots, \eta_n)$. Dalla definizione segue immediatamente che se ξ e η sono *indipendenti* allora

$$\mathrm{E}\,(\xi|\eta) = \mathrm{E}\,\xi, \tag{5.14}$$

e viceversa. Inoltre, dalla (5.12) segue che $\mathrm{E}(\xi|\xi) = \xi$. Più in generale, se η è indipendente da $\mathcal{D}$, cioè se η e 1_{D_j} sono indipendenti $\forall j$, allora

$$\mathrm{E}(\eta|\mathcal{D}) = \mathrm{E}\,\eta. \tag{5.15}$$

L'altra proprietà è la seguente: date due partizioni $\mathcal{D}_1 \preceq \mathcal{D}_2$, si ha

$$\mathrm{E}\,(\mathrm{E}(\xi|\mathcal{D}_2)|\mathcal{D}_1) = \mathrm{E}(\xi|\mathcal{D}_1). \tag{5.16}$$

(5.17) Esempio. Per le v.a. ξ e η dell'esempio (5.10) si trova, usando la prima delle (5.6) e (5.14),

$$\mathrm{E}(\xi + \eta|\eta) = \mathrm{E}\,\xi + \eta = p + \eta,$$

che si poteva ottenere anche usando il risultato dell'esempio (5.10):

$$\mathrm{E}(\xi + \eta|\eta) = \sum_{k=0}^{2} k\,\mathrm{P}(\xi + \eta = k|\eta) = p(1-\eta) + q\eta + 2p\eta = p + \eta.$$

(5.18) Esercizio. Mostrare che se ξ e η sono v.a.i.i.d. allora

$$\mathrm{E}(\xi|\xi + \eta) = \mathrm{E}(\eta|\xi + \eta) = \frac{\xi + \eta}{2}.$$

(5.19) Esempio. Siano $\eta_1, \eta_2 \dots \eta_n$, τ v.a. indipendenti, con $\eta_1, \eta_2 \dots \eta_n$ v.a.i.i.d. e τ prende i valori $1, \dots, n$. Se $S_\tau = \eta_1 + \dots + \eta_\tau$ è la somma di un numero aleatorio di variabili aleatorie, allora si ha, usando (5.4), (5.15) e (3.20),

$$\begin{aligned}\mathrm{E}(S_\tau|\tau) &= \sum_{i=1}^{n} \mathrm{E}(S_\tau|\tau = i)\,1_{\{\tau=i\}} = \sum_{i=1}^{n} \frac{\mathrm{E}(S_i\,1_{\{\tau=i\}})}{\mathrm{P}(\tau = i)}\,1_{\{\tau=i\}} \\ &= \sum_{i=1}^{n} \frac{\mathrm{E}\,S_i \cdot \mathrm{E}\,1_{\{\tau=i\}}}{\mathrm{P}(\tau = i)}\,1_{\{\tau=i\}} = (\mathrm{E}\,\eta_1) \sum_{i=1}^{n} i \cdot 1_{\{\tau=i\}} = (\mathrm{E}\,\eta_1) \cdot \tau\,.\end{aligned}$$

Inoltre, usando la (5.8), si ottiene

$$\mathrm{E}\,S_\tau = \mathrm{E}\,\mathrm{E}(S_\tau|\tau) = \mathrm{E}\,\tau \cdot \mathrm{E}\,\eta_1.$$

Similmente si trova,

$$\mathrm{E}(S_\tau^2|\tau) = \sum_{i=1}^{n} \mathrm{E}\,S_i^2 \cdot 1_{\{\tau=i\}} = (\mathrm{E}\,\eta_1)^2\,(\tau^2 - \tau) + \mathrm{E}\,\eta_1^2 \cdot \tau,$$

da cui si ottiene

$$\mathrm{E}\,(S_\tau^2) = (\mathrm{E}\,\eta_1)^2 \cdot (\mathrm{E}\,\tau^2 - \mathrm{E}\,\tau) + \mathrm{E}\,\eta_1^2 \cdot \mathrm{E}\,\tau = \mathrm{Var}\,\eta_1 \cdot \mathrm{E}\,\tau + (\mathrm{E}\,\eta_1)^2 \cdot \mathrm{E}\,\tau^2.$$

Mettendo insieme questa espressione e quella per $\mathrm{E}\,S_\tau$ otteniamo

$$\mathrm{Var}\,S_\tau = \mathrm{E}\,(S_\tau^2) - (\mathrm{E}\,S_\tau)^2 = \mathrm{E}\,\tau \cdot \mathrm{Var}\,\eta_1 + \mathrm{Var}\,\tau \cdot (\mathrm{E}\,\eta_1)^2\,.$$

Queste relazioni si chiamano *identità di Wald.*

(5.20) Osservazione. Fino ad ora abbiamo considerato partizioni i cui elementi (atomi) hanno probabilità positiva. D'altra parte, in un contesto più generale potrà essere necessario considerare probabilità condizionate rispetto a eventi che hanno probabilità *zero.* Consideriamo ad esempio il seguente esperimento. Sia ξ una v.a. distribuita uniformemente su $[0,1]$. Se $\xi = x$, lanciamo una moneta per la quale la probabilità di osservare 'testa' sia x, e 'croce' $1-x$. Sia ν il numero di occorrenze dell'evento 'testa' in n prove (lanci) indipendenti. Qual'è la "probabilità condizionata $\mathrm{P}\,(\nu = k|\xi = x)$"? Dal momento che $\mathrm{P}\,(\xi = x) = 0$, tale quantità non è definita, sebbene sia intuitivamente plausibile che "debba essere $\binom{n}{k}x^k(1-x)^{n-k}$". Vedremo come questa necessità venga soddisfatta introducendo la nozione più generale di aspettazione condizionata rispetto a una σ-algebra $\mathcal{G}$, $\mathcal{G} \subseteq \mathcal{F}$, denotata con $\mathrm{E}\,(\xi|\mathcal{G})$. Osserviamo per inciso che nel caso in cui $\mathcal{G}$ sia generata da una partizione *finita* (cioè $\mathcal{D} = \{D_1, \ldots, D_n\}$) tale definizione coincide con quella di $\mathrm{E}\,(\xi|\mathcal{D})$.

MARTINGALE. Il contesto naturale in cui introdurre le martingale è quello delle scommesse.

(5.21) Esempio. Sia $\eta_1, \eta_2 \ldots \eta_n$ una sequenza di prove (ad es. lanci) indipendenti, dove $\eta_i = +1$ o -1 con probabilità p e $1-p$. Supponiamo, prima del $(k+1)$-esimo lancio, di voler scommettere una certa somma b, in modo da vincere b se $\eta_{k+1} = 1$, perderla se $\eta_{k-1} = -1$. Una strategia di gioco è una regola che ci dice quanto scommettere al $(k+1)$-esimo lancio. Ovviamente, tale regola, per essere interessante, dovrà dipendere dai risultati delle prime k giocate. In generale dunque una strategia sarà una sequenza di funzioni $b_k : \{-1, +1\}^{(k)} \to [0, \infty)$ tale che $b_k(\eta_1, \ldots, \eta_k)$ è quanto dobbiamo scommettere sulla $(k+1)$-esima giocata. Se partiamo con una somma iniziale ξ_0, la somma posseduta dopo k giocate, ξ_k, è una variabile aleatoria definita da

$$\xi_{k+1} = \xi_k + \eta_{k+1}\, b_k(\eta_1, \ldots, \eta_k),$$

dove, ad esempio per questioni di solvibilità, richiederemo che $b_k(\eta_1, \ldots, \eta_k) \leq \xi_k$. Si ha la seguente proprietà

Lemma. *Se $p = 1/2$ allora*

$$\mathrm{E}\,(\xi_{k+1} \,|\, \xi_1, \ldots, \xi_k) = \xi_k, \qquad 1 \leq k < n,$$

(e la strategia si dirà equilibrata). Se invece $p \leq 1/2$ allora

$$\mathrm{E}\,(\xi_{k+1} \,|\, \xi_1, \ldots, \xi_k) \leq \xi_k, \qquad 1 \leq k < n,$$

(e la strategia si dirà sfavorevole).

Dimostrazione. Si ha

$$\begin{aligned} \mathrm{E}\,(\xi_{k+1} \,|\, \xi_1, \ldots, \xi_k) &= \mathrm{E}\,(\xi_k + \eta_{k+1}\, b_k(\eta_1, \ldots, \eta_k) \,|\, \xi_1, \ldots, \xi_k) \\ &= \xi_k + b_k(\eta_1, \ldots, \eta_k)\, \mathrm{E}\,(\eta_{k+1}), \end{aligned}$$

dove, nell'ultimo passaggio si è usata l'indipendenza delle η_i e il fatto che ξ_k è una funzione di $\eta_1, \ldots, \eta_k$. Ora, se $p = 1/2$ si ha $\mathrm{E}\,(\eta_{k+1}) = 0$, se invece $p < 1/2$ si ha $\mathrm{E}\,(\eta_{k+1}) < 0$. Inoltre b_k è non negativa. Pertanto avremo che se $p = 1/2$, $\mathrm{E}\,(\xi_{k+1} \,|\, \xi_1, \ldots, \xi_k) = \xi_k$, mentre se $p \leq 1/2$ allora $\mathrm{E}\,(\xi_{k+1} \,|\, \xi_1, \ldots, \xi_k) \leq \xi_k$. q.e.d.

Sia ora $(\Omega, \mathcal{F}, \mathrm{P}\,)$ uno spazio di probabilità (discreto) e $\mathcal{D}_1 \preceq \mathcal{D}_2 \preceq \ldots\ldots \preceq \mathcal{D}_n$ una sequenza di partizioni.

(5.22) Definizione. Una sequenza di v.a. $\xi_1, \ldots, \xi_n$ è detta *martingala* (rispetto alla sequenza di partizioni $\mathcal{D}_1 \preceq \mathcal{D}_2 \preceq \ldots \preceq \mathcal{D}_n$) se

1) ξ_k è $\mathcal{D}_k$-misurabile, per ogni $1 \leq k \leq n$
2) $\mathrm{E}\,(\xi_{k+1} | \mathcal{D}_k) = \xi_k$, $1 \leq k < n$.

In particolare, avremo che l'aspettazione $\mathrm{E}\,\xi_k$ è la stessa per ogni $1 \leq k \leq n$. Infatti, dalla definizione (5.2) e la (5.8) si ha:

$$\mathrm{E}\,\mathrm{E}\,(\xi_{k+1}|\mathcal{D}_k) = \mathrm{E}\,\xi_{k+1} = \mathrm{E}\,\xi_k = \cdots = \mathrm{E}\,\xi_1. \tag{5.23}$$

Inoltre, usando la (5.16) si ha,

(5.24)

$$\xi_k = \mathrm{E}\,(\xi_{k+1}|\mathcal{D}_k) = \mathrm{E}\,(\mathrm{E}\,(\xi_{k+2}|\mathcal{D}_{k+1})|\mathcal{D}_k) = \mathrm{E}\,(\xi_{k+2}|\mathcal{D}_k) = \cdots = \mathrm{E}\,(\xi_{k+j}|\mathcal{D}_k)$$

per ogni $1 \leq j \leq n-k$. In generale, per enfatizzare il sistema di partizioni rispetto al quale le v.a. formano una martingala scriveremo $(\xi_k, \mathcal{D}_k)_{1 \leq k \leq n}$. Se poi $\mathcal{D}_k$ è la partizione indotta da $\xi_1, \ldots, \xi_k$ allora diremo semplicemente che $(\xi_k)_{1 \leq k \leq n}$ è una martingala.

(5.25) Esempio. Sia η una v.a., $\mathcal{D}_1 \preceq \mathcal{D}_2 \preceq \cdots \preceq \mathcal{D}_n$ e $\xi_k = \mathrm{E}\,(\eta|\mathcal{D}_k)$. Mostriamo che $(\xi_k, \mathcal{D}_k)$ è una martingala. In primo luogo è evidente che $\mathrm{E}\,(\eta|\mathcal{D}_k)$ è $\mathcal{D}_k$-misurabile. Inoltre, usando la (5.16), si ha

$$\mathrm{E}\,(\xi_{k+1}|\mathcal{D}_k) = \mathrm{E}\,(\mathrm{E}\,(\eta|\mathcal{D}_{k+1})|\mathcal{D}_k) = \mathrm{E}\,(\eta|\mathcal{D}_k) = \xi_k.$$

(5.26) Esempio. Nell'esempio (4.1) la sequenza di partizioni è data da

$$\mathcal{D}_1 = \mathcal{D}_{\eta_1} = \{D^+, D^-\}$$

dove

$$D^+ = \{\omega : \eta_1(\omega) = +1\}, \quad D^- = \{\omega : \eta_1(\omega) = -1\};$$

inoltre

$$\mathcal{D}_2 = \mathcal{D}_{\eta_1 \eta_2} = \{D^{++}, D^{+-}, D^{-+}, D^{--}\}$$

dove

$$D^{++} = \{\omega : \eta_1(\omega) = +1, \eta_2(\omega) = +1\},\ D^{+-} = \{\omega : \eta_1(\omega) = -1; \eta_2(\omega) = -1\};$$

e così via, fino a $\mathcal{D}_n = \mathcal{D}_{\eta_1 \ldots \eta_n}$. Osserviamo che nel caso particolare in cui $b \equiv 1$ si ottiene la classica passeggiata aleatoria: $\xi_k \equiv S_k = \eta_1 + \cdots + \eta_k$. Per quanto visto sopra quest'ultima forma una martingala se e solo se $p = 1/2$. Più in generale, data una sequenza di v.a.i.i.d. $\eta_1, \eta_2 \ldots \eta_n$, la sequenza $S_k = \eta_1 + \cdots + \eta_k$ forma una martingala rispetto alla sequenza di partizioni $(\mathcal{D}_{\eta_1 \ldots \eta_k})_{1 \le k \le n}$, se e solo se $\mathrm{E}\,\eta_1 = 0$ (altrimenti avremo una *sub-martingala* se $\mathrm{E}\,\eta_1 \le 0$, una *super-martingala* se $\mathrm{E}\,\eta_1 \ge 0$).

(5.27) Esempio. Sia $S_k = \eta_1 + \cdots + \eta_k$ come all'esempio precedente. Mostriamo che, indipendentemente dal valore di $\mathrm{E}\,\eta_1$, la sequenza $(\xi_k, \mathcal{D}_k)_{1 \le k \le n}$ con $\xi_k = S_{n-k+1}/(n-k+1)$ e $\mathcal{D}_k = \mathcal{D}_{S_n, S_{n-1}, \ldots, S_{n-k+1}}$ è una martingala. In effetti si ha $\mathcal{D}_{k+1} \succeq \mathcal{D}_k$. Inoltre ξ_k è $\mathcal{D}_k$-misurabile e dunque, usando la (3.33), abbiamo

$$\xi_k = \frac{\mathrm{E}\,(S_{n-k+1}|\mathcal{D}_k)}{n-k+1} = \frac{1}{n-k+1} \sum_{j=1}^{n-k+1} \mathrm{E}\,(\eta_j|\mathcal{D}_k) = \mathrm{E}\,(\eta_1|\mathcal{D}_k)$$

dove nell'ultima uguaglianza abbiamo usato il fatto che per ogni $j \le n-k+1$ si ha $\mathrm{E}\,(\eta_j|\mathcal{D}_k) = \mathrm{E}\,(\eta_1|\mathcal{D}_k)$. L'asserto segue ora dall'esempio (4.5). Talvolta questa proprietà viene espressa dicendo che $(S_k/k)_{1 \le k \le n}$ forma una *martingala invertita.*

Mostriamo ora che la proprietà (4.3) persiste qualora il tempo k venga rimpiazzato da un tempo aleatorio τ. Ma prima ci serve una definizione.

(5.28) Definizione. Una v.a. $\tau = \tau(\omega)$ che assume valori $1, 2 \ldots, n$ si dice *tempo d'arresto* rispetto alla sequenza di partizioni $\mathcal{D}_1 \preceq \mathcal{D}_2 \preceq \cdots \preceq \mathcal{D}_n$ se la v.a. $1_{\{\tau = k\}}(\omega)$ è $\mathcal{D}_k$-misurabile.

Se riguardiamo $\mathcal{D}_k$ come la partizione generata dall'osservazione dei primi k passi di un qualche processo, ad esempio se $\mathcal{D}_k = \mathcal{D}_{\eta_1 \ldots \eta_k}$, la partizione indotta dalle v.a. $\eta_1, \ldots, \eta_k$, allora la $\mathcal{D}_k$-misurabilità di $1_{\{\tau = k\}}$ significa che il realizzarsi o il non realizzarsi dell'evento $\{\tau = k\}$ è determinato unicamente dall'osservazione dei primi k passi, ed è dunque indipendente da ciò che può accadere nel 'futuro'. Il tempo τ_{2n} di primo ritorno in zero nei primi $2n$ passi per una passeggiata aleatoria introdotto in (2.39) è un'esempio di tempo d'arresto.

(5.29) Teorema. Sia $(\xi_k, \mathcal{D}_k)_{1 \le k \le n}$ una martingala e τ un tempo d'arresto (rispetto a $(\mathcal{D}_n)_{1 \le k \le n}$). Allora si ha $\mathrm{E}\,\xi_\tau = \mathrm{E}\,\xi_1$, dove $\xi_\tau = \sum_{k=1}^n \xi_k \, 1_{\{\tau = k\}}(\omega)$.

Dimostrazione. Se $D \in \mathcal{D}_1$ si ha, usando la (4.4),

$$\begin{aligned}
\mathrm{E}\,(\xi_\tau | D) &= \frac{\mathrm{E}\,(\xi_\tau \cdot 1_D)}{\mathrm{P}\,(D)} = \frac{1}{\mathrm{P}\,(D)} \sum_{k=1}^{n} \mathrm{E}\,(\xi_k \cdot 1_{\{\tau = k\}} \cdot 1_D) \\
&= \frac{1}{\mathrm{P}\,(D)} \sum_{k=1}^{n} \mathrm{E}\,(\mathrm{E}\,(\xi_n | \mathcal{D}_k) \cdot 1_{\{\tau = k\}} \cdot 1_D) \\
&= \frac{1}{\mathrm{P}\,(D)} \sum_{k=1}^{n} \mathrm{E}\,(\mathrm{E}\,(\xi_n \cdot 1_{\{\tau = k\}} \cdot 1_D | \mathcal{D}_k)) \\
&= \frac{1}{\mathrm{P}\,(D)} \sum_{k=1}^{n} \mathrm{E}\,(\xi_n \cdot 1_{\{\tau = k\}} \cdot 1_D) = \frac{1}{\mathrm{P}\,(D)} \mathrm{E}\,(\xi_n \cdot 1_D) = \mathrm{E}\,(\xi_n | D).
\end{aligned}$$

Pertanto $\mathrm{E}\,(\xi_\tau | \mathcal{D}_1) = \mathrm{E}\,(\xi_n | \mathcal{D}_1) = \xi_1$ e l'asserto segue prendendo l'aspettazione di $\mathrm{E}\,(\xi_\tau | \mathcal{D}_1)$. q.e.d.

(5.30) Osservazione. L'interesse del teorema precedente nel contesto delle scommesse e delle strategie di gioco risiede nella seguente osservazione. Supponiamo di voler usare la v.a. τ come una *regola d'arresto*, cioè una regola che basandosi sui risultati delle prime k giocate, ci dice se abbandonare o meno il gioco dopo la k-esima giocata. Un esempio semplice di tale regola è: 'fermati appena hai guadagnato qualcosa', per cui si avrebbe, con le notazioni del punto (4.1), $\tau(\omega) = \min\{k : \xi_k > \xi_0\}$. Allora (4.7) ci dice che, se il gioco è equilibrato, usando una regola siffatta non v'è modo di accrescere la nostra attesa di guadagno (e dunque è inutile usarla).

(5.31) Ancora sulle passeggiate aleatorie. Riprendiamo l'esempio (4.6) della passeggiata aleatoria: $S_0 = 0$ e $S_k = \eta_1 + \cdots + \eta_k$ per $k \geq 1$, dove $\mathrm{P}\,(\eta_i = +1) = p$ e $\mathrm{P}\,(\eta_i = -1) = 1 - p$. Sia τ un tempo d'arresto (rispetto alla sequenza di partizioni discussa in (4.6)). Abbiamo visto che S_k forma una martingala se e solo se $p = 1/2$. In questo caso il teorema (4.8) afferma che $\mathrm{E}\,S_\tau = 0$. Ragionando allo stesso modo, non è difficile mostrare che $\mathrm{E}\,S_\tau^2 = \mathrm{E}\,\tau$. Osserviamo che queste relazioni costituiscono casi particolari delle identità di Wald ottenute in (3.39). Siano ora a e b due interi, con $a \leq 0 \leq b$. Un problema interessante è quello di determinare la probabilità che dopo n passi il punto abbia lasciato l'intervallo (a, b). In aggiunta, possiamo anche chiederci quale sia la probabilità che il punto lasci il suddetto intervallo in a o in b. Che queste siano domande naturali su cui interrogarsi risulta forse più chiaramente se le interpretiamo nei termini del gioco d'azzardo. Consideriamo due giocatori che in partenza possiedono rispettivamente le cifre $-a$ e b da scommettere al gioco. Supponiamo che se $\eta_i = +1$ il secondo giocatore paghi un'unità di denaro al primo; se invece $\eta_i = -1$ il primo paga il secondo. Allora S_k può essere interpretato come l'ammontare di denaro vinto dal primo giocatore al secondo (se $S_k < 0$ sarà l'ammontare perso dal primo a vantaggio del secondo) dopo k giocate. Nell'istante $k \leq n$ in cui per la prima volta $S_k = b$ ($S_k = a$) la somma a disposizione del secondo (primo) giocatore si sarà ridotta a zero, in altre parole il giocatore in questione è rovinato. Per formalizzare meglio questo problema consideriamo un intero x nell'intervallo $[a, b]$ (la 'condizione inziale') e poniamo $S_k^x = x + S_k$. Introduciamo quindi la sequenza di tempi d'arresto

$$\tau_k^x = \min\{0 \leq l \leq k : S_l^x = a \;\text{ o }\; b\}, \qquad 0 \leq k \leq n,$$

con la posizione $\tau_k^x = k$ se $a < S_l^x < b$ per ogni $0 \leq l \leq k$. Siano inoltre $\alpha_k(x)$ e $\beta_k(x)$ le probabilità che, partendo da x, il punto lasci (a, b) durante l'intervallo di tempo $[0, k]$ attraversando a e, rispettivamente, b. In formule

$$\alpha_k(x) = \mathrm{P}\left(\cup_{l=0}^{k}\{\omega : \tau_k^x = l, S_l^x = a\}\right), \quad \beta_k(x) = \mathrm{P}\left(\cup_{l=0}^{k}\{\omega : \tau_k^x = l, S_l^x = b\}\right).$$

Usando la formula della probabilità totale come nell'esempio (3.14) otteniamo la relazione ricorsiva

$$\alpha_k(x) = \frac{\alpha_{k-1}(x+1) + \alpha_{k-1}(x-1)}{2}, \quad \alpha_1(a) = 1, \quad \alpha_l(b) = 0, \quad 0 \leq l \leq n.$$

Una formula identica vale per $\beta_k(x)$ con le condizioni al contorno $\beta_1(b) = 1$, $\beta_l(a) = 0$, $0 \leq l \leq n$ (formule analoghe valgono nel caso asimmetrico $p \neq 1/2$). Osserviamo inoltre che dalla definizione segue immediatamente che $\alpha_k(x) \leq \alpha_{k+1}(x) \leq 1$ (e analoga disuguaglianza per $\beta_k(x)$). Pertanto $\alpha_k(x)$ e $\beta_k(x)$ ammettono dei valori limite $\alpha(x)$ e $\beta(x)$ i quali a loro volta soddisfano le equazioni

$$\alpha(x) = \frac{\alpha(x+1) + \alpha(x-1)}{2}, \quad \alpha(a) = 1, \quad \alpha(b) = 0,$$

e analoga per $\beta(x)$ con $\beta(b) = 1$, $\beta(a) = 0$, le cui (uniche) soluzioni continue sono

$$\alpha(x) = \frac{b-x}{b-a}, \quad \beta(x) = \frac{x-a}{b-a}, \quad \alpha(x) + \beta(x) = 1.$$

Le quantità $\alpha(x)$ e $\beta(x)$ si diranno *probabilità di rovina del primo e secondo giocatore*, rispettivamente, quando il primo inzia a giocare con la somma $x - a$ e il secondo con $b - x$. In particolare, se $a = 0$, allora $\beta(x) = x/b$ dà la probabilità che il punto che parte in $x \in [0, b]$ giunga in b prima che in 0; mentre $\alpha(x) = 1 - x/b$, da cui si vede, mandando $b \to \infty$, che in un gioco alla pari contro un avversario dalle risorse illimitate la probabilità di perdere prima o poi tutti i soldi è uguale a 1, come avevamo già osservato in (3.14). Consideriamo ora la *durata media* del gioco, o della passeggiata aleatoria, ossia la quantità $t(x) = \lim_{n\to\infty} \mathrm{E}\,\tau_n^x$. Consideriamo per semplicità il caso $x = 0$ e poniamo $\tau_n^0 = \tau_n$, $\alpha_n(0) = \alpha_n$, $\beta_n(0) = \beta_n$, $r(0) = r$, $\alpha(0) = \alpha$ etc. Per quanto visto sopra nel caso simmetrico (gioco equilibrato) si ha $\mathrm{E}\,\tau_n = \mathrm{E}\,S_{\tau_n}^2$ e dunque

$$\min(a^2, b^2) \leq \mathrm{E}\,\tau_n \leq \max(a^2, b^2).$$

Più precisamente si ha

$$\mathrm{E}\,\tau_n = \mathrm{E}\,S_{\tau_n}^2 = a^2 \cdot \alpha_n + b^2 \cdot \beta_n + \mathrm{E}\,(S_n^2 \cdot 1_{\{a<S_n<b\}}).$$

Poniamo $\gamma_n = \alpha + \beta - (\alpha_n + \beta_n) = 1 - (\alpha_n + \beta_n)$. Evidentemente si ha

$$\gamma_n = \mathrm{P}\,\{a < S_k < b, 0 \leq k \leq n\},$$

e inoltre tutte le quantità dipendenti da n nell'espressione di $\mathrm{E}\,\tau_n$ scritta sopra si possono stimare per mezzo di γ_n:

$$0 \leq \alpha - \alpha_n \leq \gamma_n, \quad 0 \leq \beta - \beta_n \leq \gamma_n,$$

e

$$\mathrm{E}\,(S_n^2 \cdot 1_{\{a<S_n<b\}}) \leq \max(a^2, b^2)\,\gamma_n.$$

Studiamo dunque l'andamento di γ_n. A questo scopo, fissiamo un intero m e poniamo $n = mr$. Scriviamo poi S_n nella forma $S_n = \zeta_1 + \cdots + \zeta_r$ con $\zeta_1 =$

$\eta_1 + \cdots + \eta_m$ e, per $j > 1$, $\zeta_j = \eta_{m(j-1)+1} + \cdots + \eta_{mj}$. Se ora poniamo $c = |a| + b$ allora si ha che

$$\begin{aligned}\{a < S_k < b, 0 \leq k \leq n\} &= \bigcap_{0 \leq k \leq n} \{|S_k| < c\} \subseteq \bigcap_{l=1}^{r} \{|S_{lm}| < c\} \\ &= \{|\zeta_1 + \ldots + \zeta_l| < c, 1 \leq l \leq r\} \subseteq \{|\zeta_1| < c, \ldots, |\zeta_r| < c\}.\end{aligned}$$

Pertanto, essendo le ζ_i indipendenti e identicamente distribuite la (3.19) dà

$$\gamma_n \leq \mathrm{P}\{|\zeta_1| < c, \ldots, |\zeta_r| < c\} = \prod_{i=1}^{r} \mathrm{P}\{|\zeta_i| < c\} = (\mathrm{P}\{|\zeta_1| < c\})^r.$$

Osserviamo ora che dalla definizione di ζ_1 segue che $|\zeta_1| \leq m$, $\mathrm{E}\,\zeta_1 = 0$ e $\mathrm{Var}\,\zeta_1 = m$. Dunque, se fosse $\mathrm{P}\{|\zeta_1| < c\} = 1$, si avrebbe

$$\begin{aligned} m &= \mathrm{Var}\,\zeta_1 = \sum_{\omega \in \{-1,1\}^m} \zeta_1^2(\omega) \cdot p(\omega) \\ &= \sum_{\omega : |\zeta_1(\omega)| < c} \zeta_1^2(\omega) \cdot p(\omega) + \sum_{\omega : |\zeta_1(\omega)| \geq c} \zeta_1^2(\omega) \cdot p(\omega) \\ &\leq c^2 \cdot \mathrm{P}\{|\zeta_1| < c\} + m^2 \cdot \{|\zeta_1| \geq c\} = c^2. \end{aligned}$$

Scegliendo m abbastanza grande vediamo dunque che deve essere $\mathrm{P}\{|\zeta_1| < c\} \leq \epsilon_1 < 1$. Pertanto, se n è abbastanza grande, si ha

$$\gamma_n \leq \epsilon^n, \quad \text{con} \quad \epsilon = \epsilon_1^{1/m} < 1.$$

Infine, mettendo tutto insieme, abbiamo ottenuto che quando $n \to \infty$ la durata media del gioco $\mathrm{E}\,\tau_n$ converge con velocità esponenziale al limite

$$t = a^2 \cdot \alpha + b^2 \cdot \beta = a^2 \cdot \frac{b}{b-a} - b^2 \cdot \frac{a}{b-a} = |ab|.$$