10.10 Teorema del limite centrale

Siamo ora giunti a quello che è il più importante teorema sulle distribuzioni di probabilità.

Nei paragrafi precedenti abbiamo imparato a valutare previsione e incertezza di previsione di combinazioni lineari di variabili casuali, indipendentemente dalla loro distribuzione. Ma per esprimere affermazioni probabilistiche su tali funzioni serve conoscere la distribuzione di probabilità e questo è un problema non banale (vedi paragrafi facoltativi 10.2 e 10.3). Comunque, l'esperienza su distribuzioni tipiche (vedi paragrafo 7.10) ci insegna che in genere c'è alta probabilità che in valore della variabile aleatoria cada entro alcune deviazioni standard dal valore atteso. Quando poi ci sono servite della considerazioni probabilistiche estreme, indipendentemente dalla distribuzione, abbiamo fatto uso della disuguaglianza di Cebicev.

In realtà le combinazioni lineari tendono ad avere una distribuzione di probabilità universale, indipendentemente dalle distribuzioni di partenza, come consequenza del *teorema del limite centrale*, che formularemo fra breve. Siccome consideriamo assolutamente facoltativa la dimostrazione del teorema e molto più importante una sua modellizzazione e rappresentazione visiva, cominciamo con degli esempi pratici.

Abbiamo visto nel capitolo 6 come costruire una variabile casuale sulla somma degli esiti di due dadi. Mentre la distribuzione dell'esito di un solo dado è uniforme, la distribuzione della somma mostra un addenzamento della probabilià al centro, ed un massimo in corrispondenza di $X = 7$. La figura 6.2 mostra chiaramente l'origine del cambiamento di forma: ci sono molte combinazioni che possono dare valori centrali e poche che possono dare valori laterali. Se si somma la variabile "due dadi" con un'altra variabile indipendente legata ad un solo dado, di nuovo sarà molto probabile formare combinazioni intorno al centro della distribuzione (vedi ad esempio figura 10.1).

Anche partendo da una distribuzione continua uniforme si ha lo stesso effetto: la somma di due variabili dà luogo ad una distribuzione triangolare (chi è interessato alla dimostrazione può consultare il paragrafo 10.3.3, ma anche la sola figura 10.3 è autoesplicativa e mostra l'analogia con il lancio dei dadi).

Osserviamo ora la figura 10.5. Essa mostra le distribuzioni simulate ottenute estraendo un certo numero di variabili casuali ($n = 1$, 2, 3, 5, 10, 20 e 100) e sommandole fra di loro. Questo processo è ripetuto 10000 volte. Si ottiene quindi, per ciascun caso, una distribuzione statistica che somiglia alla corrispondente distribuzione di probabilità[11]. Le distribuzioni di base sono una uniforme fra 0 e 1 e una distribuzione "strana" uniforme fra 0 e 0.25, fra 0.75 e 1, nulla altrove. La distribuzione "strana" è particolarmente istruttiva per capire l'effetto di addensamento al centro della probabilità dovuto alle combinazione dei diversi valori. Ad esempio, per $n = 2$, si nota un triangolo centrale doppio di ciascuno dei triangoli laterali. Esso è dovuto alla probabilità che un valore grande di una variabile si combini con un valore grande dell'altra variabile.

[11] Si ricorda che la previsione di una distribuzione statistica è uguale alla distribuzione di probabilità, con una incertezza che decresce con il numero di estrazioni. Per l'uso delle simulazioni per stimare distribuzioni di probabilità si veda il paragrafo 10.5

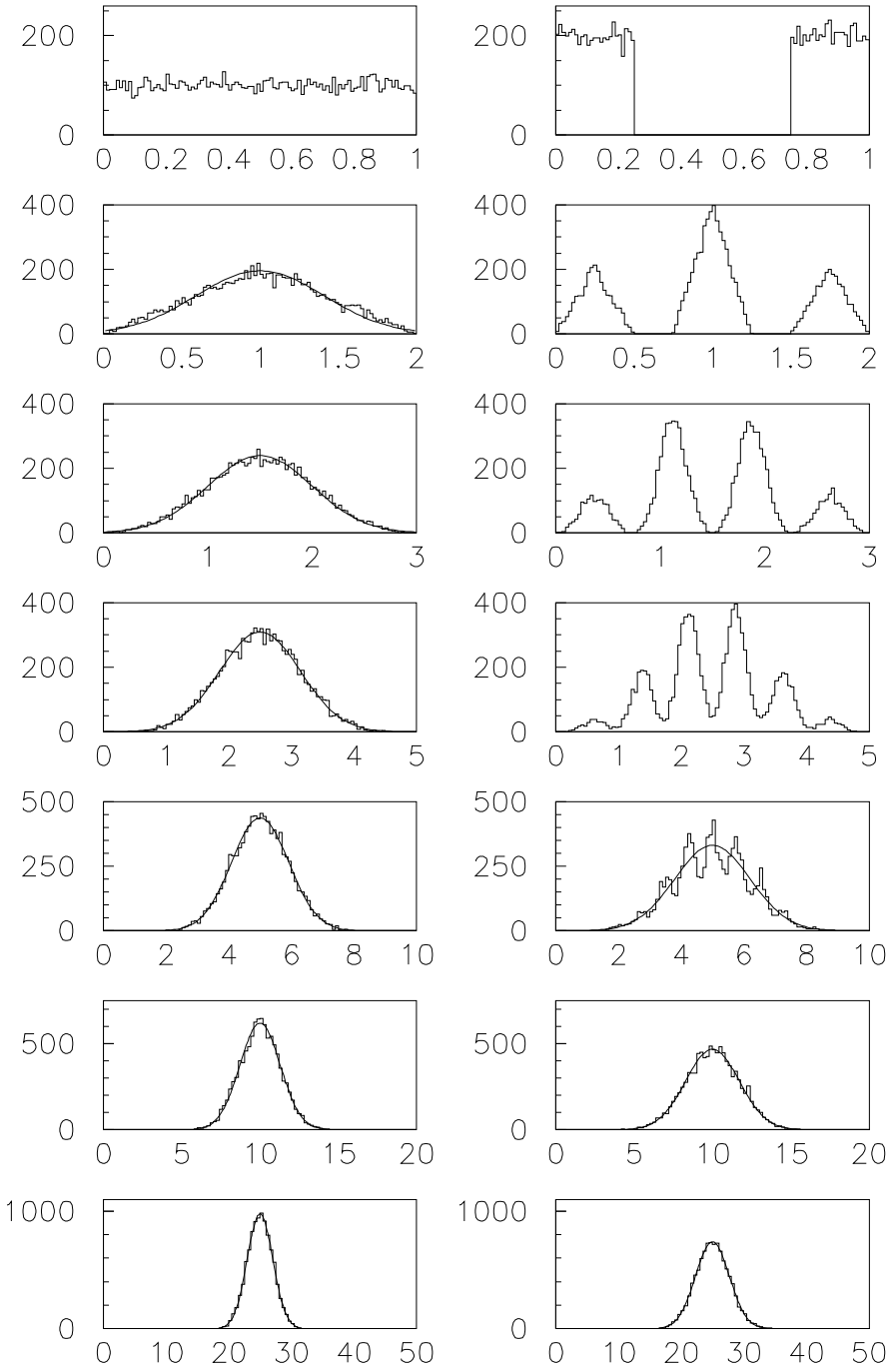

Figura 10.5: Teorema del limite centrale all'opera. Le due figure in alto mostrano la distribuzione 10000 eventi generati secondo una distribuzione uniforme (sinistra) e una distribuzione a "onda quadra", diversa da zero e costante negli intervalli fra 0 e 0.25 e fra 0.75 e 1. Successivamente (dall'alto verso il basso) sono mostrate le somme di n variabili casuali estratte dalle due distribuzioni. n vale, nell'ordine, 1, 2, 3, 5, 10, 20 e 100. In alcuni casi sono anche mostrate le gaussiane aventi la stessa media e varianza delle distribuzioni e normalizzate a 10000 eventi.

Si nota che, al crescere dei termini della sommatoria, la distribuzione risultante ha una forma regolare a campana indipendentemente dalla forma iniziale. Si noti come le distribuzioni asintotiche ($n = 100$) sono centrate sullo stesso valore in entrambi i casi, mentre sono diverse le larghezze. Questa è conseguenza che le due distribuzioni iniziali avessero stessa media, ma diversa deviazione standard (quella "strana" indica una previsione del quadrato degli scarti dalla media più grande dell'altro caso). La curva riportata sui vari istogrammi è una gaussiaana. Si vede come, da un certo n in poi (diverso per i due casi!) la distribuzione è ben approssimata da una normale.

Questo comportamento è dovuto al *teorema del limite centrale*:

> *date n variabili casuali indipendenti X_i, anche descritte da distribuzioni di probabilità diverse (purché aventi valore medio e varianza finiti), al crescere di n ("nel limite di $n \to \infty$") la combinazione lineare*
>
> $$Y = \sum_{i=1}^{n} \alpha_i X_i$$
>
> *ha distribuzione di probabilità normale di parametri*
>
> $$\begin{cases} \mu = \sum_i \alpha_i \mu_i \\ \sigma^2 = \sum_i \alpha_i^2 \sigma_i^2 \end{cases}$$
>
> *purché sia*
>
> $$\alpha_i^2 \sigma_i^2 \ll \sigma^2$$
>
> *per ogni variabile X_i non distribuita normalmente.*

È importante aggiungere delle note esplicative.

- La condizione $\alpha_i^2 \sigma_i^2 \ll \sigma^2$ può essere espressa dicendo che nessun termine deve dominare le fluttuazioni, altrimenti esso sarà determinante ai fini della distribuzione.

- La deroga a tale condizioni per le variabili distribuite normalmente è dovuto al fatto che le combinazioni lineari di gaussiane sono sempre gaussiane indipendentemente dal loro numero e dai valori di σ.

- Non c'è invece nessuna condizione sul valori attesi, in quanto questo teorema descrive le fluttuazioni, mentre le posizioni possono essere variate a piacere (purché i valori siano finiti, naturalmente).

- Il teorema non dice da quale valore di n l'approssimazione è valida. Dipende dal tipo di distribuzione. Dalla figura 10.5 si vede che, a partire da distribuzioni uniformi di larghezza confrontabile, già per $n = 4$ o 5 l'approssimazione è ragionevole. Partendo da una distribuzione triangolare (equivalente a 2 uniformi) sono sufficienti già 2 o 3 termini. In effetti per i casi pratici, con distribuzioni piccate al centro la convergenza è rapidissima.

- Per non dare l'illusione che la convergenza a normale sia sempre rapida, facciamo un controesempio notevole. Per la proprietà riproduttiva della poissoniana, la somma di n poissoniane indipendenti danno ancora luogo ad una poissoniana. Quindi, sotto certe condizioni, essa tenderà a normale (vedi paragrafo successivo). Se però consideriamo un miliardo di variabili, ciascuna avente $\lambda = 10^{-9}$ la distribuzione finale sarà ancora poissoniana con $\lambda = 1$, ben diversa da una normale! (In questo caso particolare la convergenza a normale si ha per $n \gg 10^9$.)

10.10.1 Distribuzione della media aritmetica

Diamo qui una prima applicazione del teorema alla distribuzione della media aritmetica. La variabile casuale media aritmetica è pari, a meno di un fattore di proporzionalità, ad una sommatoria e quindi, al crescere del numero di osservazioni sulle quali essa è effettuata, sarà descritta sempre meglio da normale:

$$\overline{X} \sim \mathcal{N}(\mu, \sigma/\sqrt{n}) \,. \tag{10.58}$$

10.10.2 Convergenza in distribuzione della binomiale e della poissoniana

Sia la binomiale che la poissoniana godono della proprietà riproduttiva. Questa proprietà permette di pensare una distribuzione caratterizzata da un valore medio elevato come una somma di tante variabili casuali provenienti dallo stesso tipo di distribuzione ma caratterizzate da valori medi più piccoli. Siccome per il teorema del limite centrale una somma di variabili casuali tende ad una distribuzione normale, la distribuzione binomiale e quella di Poisson tendono alla distribuzione normale al crescere, rispettivamente, di np e di λ:

- distribuzione binomiale: per valori di np e di nq "abbastanza grandi" la distribuzione binomiale tende ad una distribuzione normale di $\mu = np$ e $\sigma = \sqrt{pqn}$;
- distribuzione di Poisson: per valori di λ "abbastanza grandi" la distribuzione di Poisson tende ad una distribuzione normale di $\mu = \lambda$ e $\sigma = \sqrt{\lambda}$.

La condizione di valore "abbastanza grande" dipende dal grado di accuratezza con cui si vuole calcolare la probabilità. Per la maggior parte delle applicazioni che si incontrano nella trattazione degli errori e per la determinazione degli intervalli di fiducia la condizione è soddisfatta per valori di np, nq e λ maggiori di 10-15.

Bisogna precisare cosa si intende per limite di una distribuzione discreta a una distribuzione continua. Infatti in un caso la funzione di distribuzione ha il significato di una probabilità e nell'altro di una densità di probabilità. Il limite va allora inteso per la funzione di ripartizione:

$$\lim_{n\to\infty} F(x_i) = \int_{-\infty}^{x_i+1/2} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$