Appunti del corso di
Affidabilitá e Diagnostica dei Sistemi Elettrici

Andrea Cavallini, Gian Carlo Montanari
DIE-Universitá di Bologna
viale Risorgimento 2, 40136 Bologna
andrea.cavallini@mail.ing.unibo.it
http://limat.ing.unibo.it

A.A 1999/2000

Indice

Capitolo 8

Metodi empirici

I metodi empirici consentono una stima agevole ed intuitiva della distribuzione di probabilitá (popolazione) di una VA e dei suoi momenti sulla base dei risultati sperimentali (campione). Normalmente, i metodi empirici richiedono un modello probabilistico estremamente ampio. Infatti, il primo metodo presentato, quello delle carte di probabilitá, richiede solamente che sia specificata la legge di probabilitá dei dati. La stima empirica dei momenti non richiede alcuna ipotesi sul modello probabilistico dei dati.

8.1 Stima empirica delle leggi di probabilitá

Sia

$$\mathcal{X} = \{x_k, k = 1\ldots, n\} \tag{8.1}$$

l'insieme dei risultati sperimentali (campione) della variabile aleatoria $\boldsymbol{x}$. La stima empirica della distribuzione di probabilitá si ottiene ordinando per valori crescenti i dati misurati:

$$x_1 \le x_2 \le \ldots \le x_i \le \ldots \le x_n$$

Poiché la migliore stima della probabilitá di un evento é la frequenza relativa di esso, la migliore stima della probabilitá $\Pr(\boldsymbol{x} \le x_i)$ é data dal rapporto i/n:

$$\hat{F}(x_i) = \frac{i}{n} \tag{8.2}$$

Il "cappello" ($\hat{}$) indica che la grandezza calcolata é una stima, ottenuta in base ai dati del campione, della distribuzione di probabilitá della popolazione. La procedura seguita é mostrata, schematicamente, in figura 8.1. Tale procedura é, in realtá, comune a tutte le stime di tipo statistico e prevede: (1) la realizzazione di esperimenti necessaria per formare un campione, (2) l'applicazione di una procedura (stimatore) per (3) ottenere la stima del parametro o distribuzione cercata. Si noti che il risultato di questa operazione non é noto a priori e, pertanto, deve essere considerato a tutti gli effetti come la realizzazione di una variabile aleatoria. Quest'ultimo punto, particolarmente importante, sará approfondito in seguito.

Il grafico dei punti $(x_i, i/n)$ fornisce una rappresentazione visiva dell'andamento della funzione di distribuzione della VA $\boldsymbol{x}$. Nella figura 8.2 si mostrano le distribuzioni empirica e teorica di 20 dati estratti da una popolazione normale di media nulla e varianza unitaria.

Un modo alternativo di procedere sia il seguente. Siano $x_{min} = x_1$, $x_{max} = x_n$ (si ipotizza che i dati siano ordinati per valori crescenti). L'intervallo $[x_{min}, x_{max}]$ puó essere suddiviso in C sottointervalli:

$$U_i = \left[x_{C_i} - \frac{h}{2}, x_{C_i} + \frac{h}{2}\right] \tag{8.3}$$

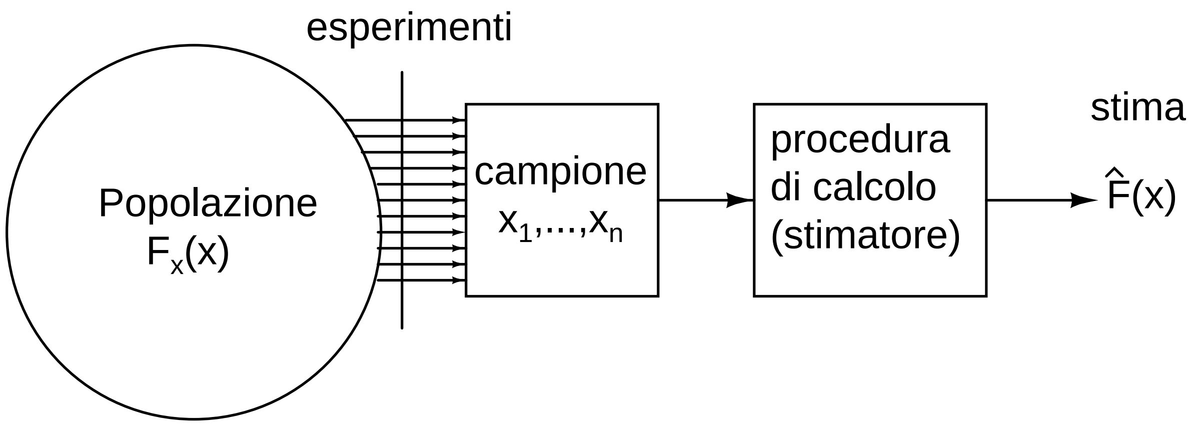

Figura 8.1: Procedura di stima della distribuzione di una variabile aleatoria.

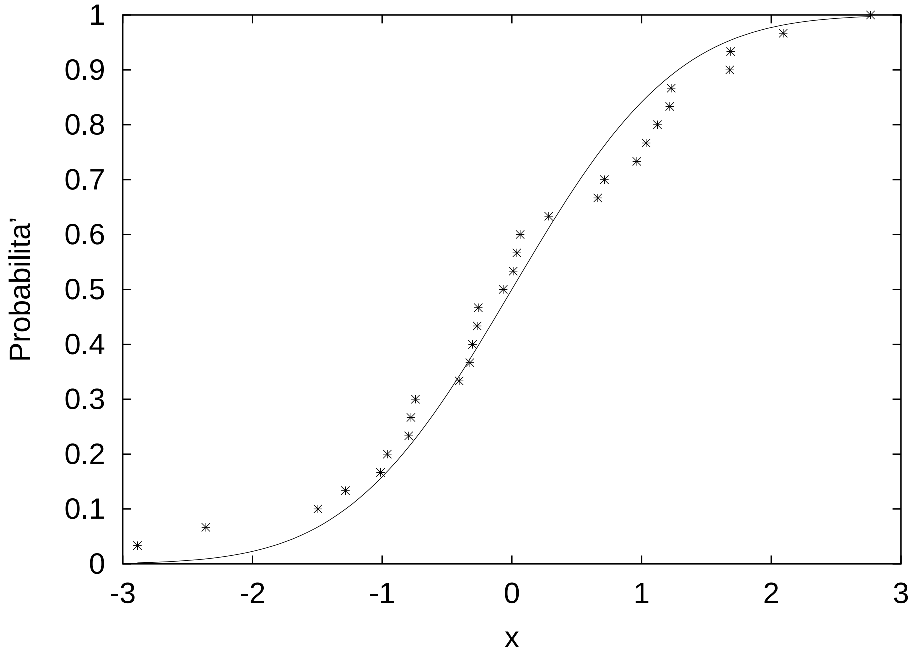

Figura 8.2: Distribuzione empirica (punti) e teorica (linea continua) di un campione di numerositá 20 estratto da una popolazione normale di media nulla e varianza unitaria.

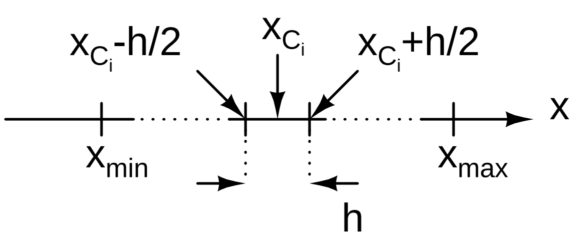

Figura 8.3: Partizione dell'asse reale.

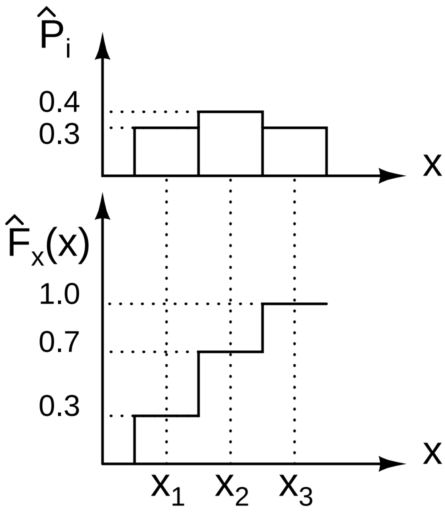

Figura 8.4: Stima della distribuzione di probabilitá a partire dai valori calcolati $\hat{P}_i$.

dove

$$h = \frac{x_{max} - x_{min}}{C} \tag{8.4}$$

e

$$x_{C_1} = x_{min} + \frac{h}{2} \tag{8.5}$$

E' immediato verificare che tali intervalli ricoprono completamente $[x_{min}, x_{max}]$ e non sono sovrapposti, cioé formano una partizione. Conseguentemente, i numeri in $[x_{min}, x_{max}]$ sono suddivisi in classi. Ad esempio, la classe i-esima comprende tutti i numeri appartenenti ad U_i.
Sia ora n_i il numero di volte in cui la variabile aleatoria assume valori in U_i. Poiché U_i definisce un evento della variabile aleatoria $\boldsymbol{x}$, se $P_i = \Pr(\boldsymbol{x} \in U_i)$, allora, per il teorema di Bernoulli é:

$$\hat{P}_i = \frac{n_i}{n} \tag{8.6}$$

Da $\hat{P}_i$ si puó ottenere una stima della distribuzione di probabilitá. Infatti, la probabilitá di osservare un valore della VA inferiore a $x_{C_i} + h/2$ é la somma della probabilitá di osservare la VA in uno qualsiasi (unione, OR logico) degli intervalli di indice inferiore o uguale ad i. Dunque:

$$\hat{F}(x_{C_i} + h/2) = \sum_{j=1}^{i} \hat{P}_j \tag{8.7}$$

Tale procedimento é mostrato, sinteticamente, in figura 8.4
Inoltre, poiché:

$$\Pr(\boldsymbol{x} \in [a, b]) = F_x(b) - F_x(a)$$

e

$$f_x(x) = \lim_{h \to 0} \frac{\Pr(\boldsymbol{x} \in [x, x+h])}{h}$$

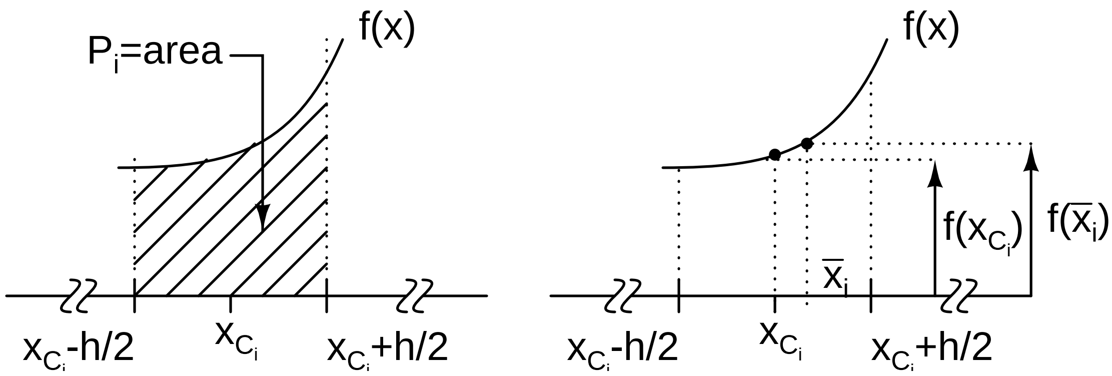

Figura 8.5: Per stimare correttamente f_{C_i} é necessario che l'ampiezza, h, dell'intervallo U_i tenda a zero. In caso contrario, ció che si stima é in valore medio di f in U_i, che non coincide, in generale, con f_{C_i}.

La stima della densitá di probabilitá é data da:

$$\hat{f}(x_{C_i}) = \frac{\hat{P}_i}{h} \tag{8.8}$$

Per il teorema di Bernoulli é:

$$P_i = \lim_{n\to\infty} \hat{P}_i \tag{8.9}$$

pertanto, i metodi empirici descritti forniscono stime della distribuzione di probabilitá e della densitá di probabilitá la cui accuratezza cresce con la dimensione del valore considerato (ció é vero, in generale, per ogni stima statistica). Per quanto riguarda la densitá é, inoltre, richiesto che l'ampiezza dei sottointervalli, h, tenda a zero. In caso contrario, ció che si stima é in valore medio di f in U_i, che non coincide, in generale, con f_{C_i}, come mostrato, schematicamente, in figura 8.5.

Teoricamente, dunque, il numero di classi deve tendere all'infinito. In generale, il numero di classi puó essere legato alla dimensione n del campione attraverso la regola empirica:

$$C = 1 + 3.3 \log n \tag{8.10}$$

Una scelta di C sbagliata puó influire significativamente sul risultato della stima. In generale, bassi valori di C non permettono di apprezzare i dettagli delle funzioni. Alti valori di C, al contrario, generano una stima "rumorosa". Per chiarire queste affermazioni si osservino i grafici di figura 8.6. Tali grafici mostrano la stima empirica della densitá di probabilitá di una popolazione la cui distribuzione é la somma di due normali di varianza unitaria ed aventi media 0 e 5, rispettivamente:

$$f(x) = \frac{1}{\sqrt{2\pi}} \frac{\exp(-0.5x^2) + \exp(-0.5(x-5)^2)}{2}$$

Si vede come per $C = 5$ la stima della densitá non fornisce i dettagli della distribuzione, per $C = 100$ la stima fornisce dettagli ma l'errore rispetto alla densitá vera puó essere, localmente, molto forte (rumore di stima). Per $C = 1+3.3 \log 200 \approx 18$ la stima é sufficientemente aderente alla densitá della popolazione.

Si noti, infine, che la partizione puó essere anche realizzata con intervalli di ampiezza disuguale, non uniforme. Ció é particolarmente utile per distribuzioni aventi "code" fortemente allungate come, ad esempio, la distribuzione esponenziale, lognormale, di Gumbel, ecc.

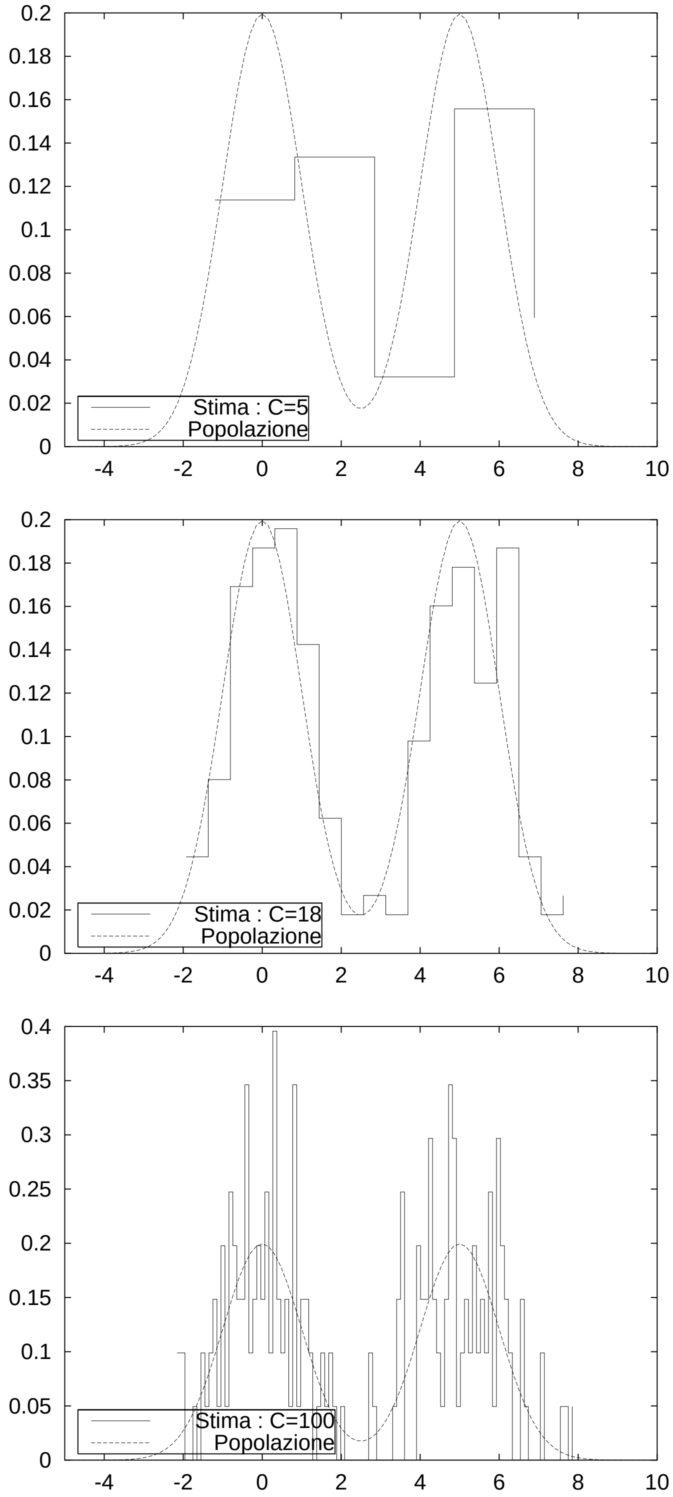

Figura 8.6: Calcolo della distribuzione empirica di una popolazione effettuata a partire da un campione di numerositá $n = 200$.

8.2 Percentili

Il procedimento per la stima empirica dei percentili é analogo a quello fatto per stimare la distribuzione di probabilitá. Infatti, il primo passo é quello di ordinare i dati nel campione per ordine crescente. Se $x_1, \ldots, x_i \ldots, x_n$ sono i dati ordinati, ad x_i si associa la probabilità i/n, stima della distribuzione in x_i. Il percentile a probabilitá P é il numero x_i per cui $i/n \leq P$. Eventualmente, la stima puó essere migliorata interpolando la distribuzione di probabilitá empirica.

8.3 Carte probabilistiche

Il metodo delle carte probabilistiche consente di verificare rapidamente una ipotesi sulla distribuzione di probabilitá della popolazione. Tale ipotesi, detta ipotesi nulla, si indica sinteticamente come:

$$H_0 : F_x(x) \in \mathcal{F}_0 \tag{8.11}$$

cioé si ipotizza che la VA $\boldsymbol{x}$ segua una distribuzione di probabilitá appartenente ad una famiglia parametrica prefissata, $\mathcal{F}_0$. Per essere concreti, si puó ipotizzare che la popolazione segua una distribuzione normale. Oppure, che la distribuzione segua una legge di Weibull.

Il metodo delle carte probabilistiche si basa sull'esistenza di due opportune funzioni, che indicheremo come Φ e Ψ. La funzione Ψ, se esiste, é tale che i punti $(\Phi(x), \Psi(F(x)))$ si dispongono secondo una retta. Ad esempio, per una distribuzione esponenziale:

$$F(x) = 1 - \exp(-\lambda x) \tag{8.12}$$

la funzione Φ é la funzione identitá ($\Phi(x) = x$), la funzione $\Psi(F)$ (si omette la dipendenza da x per semplicitá) é la seguente:

$$\Psi(F) = -\log(1 - F(x)) \tag{8.13}$$

I punti $(x, -\log(1 - F(x)))$ si dispongono infatti secondo rette di pendenza pari a λ.

Generalizzando, si supponga di avere un campione $\mathcal{X}$ consistente di n misure di una VA. Come si applica il metodo delle carte probabilistiche? La procedura é la seguente:

- scegliere una famiglia parametrica $\mathcal{F}_0$ e, se esistono, le appropriate funzioni Φ e Ψ;
- ordinare il campione in modo che $x_1 \leq x_2 \ldots \leq x_n$;
- stimare $F(x_i)$ mediante, ad esempio, $\hat{F}(x_i) = (i - 0.5)/n$;
- calcolare $\Phi(x)$ e $\Psi(\hat{F})$;
- verificare se le coppie $(\Phi(x), \Psi(\hat{F}))$ si dispongono, approssimativamente, in linea retta. Se ció é vero, allora si accetta l'ipotesi H_0;

L'ultimo punto, cioé la verifica se i punti si dispongono secondo una linea retta o no, puó essere fatta, come suggerito sopra, in modo soggettivo. Alternativamente, dopo avere verificato che i punti non siano disposti secondo una legge fortemente non lineare, é possibile calcolare il coefficiente di correlazione campionario $\hat{R}_{xy}$ (il procedimento di stima sará discusso in seguito). Se tale coefficiente é prossimo a 1, allora si accetta l'ipotesi H_0.

Per determinare $\hat{F}(x_i)$ é possibile utilizzare, alternativamente, le seguenti espressioni:

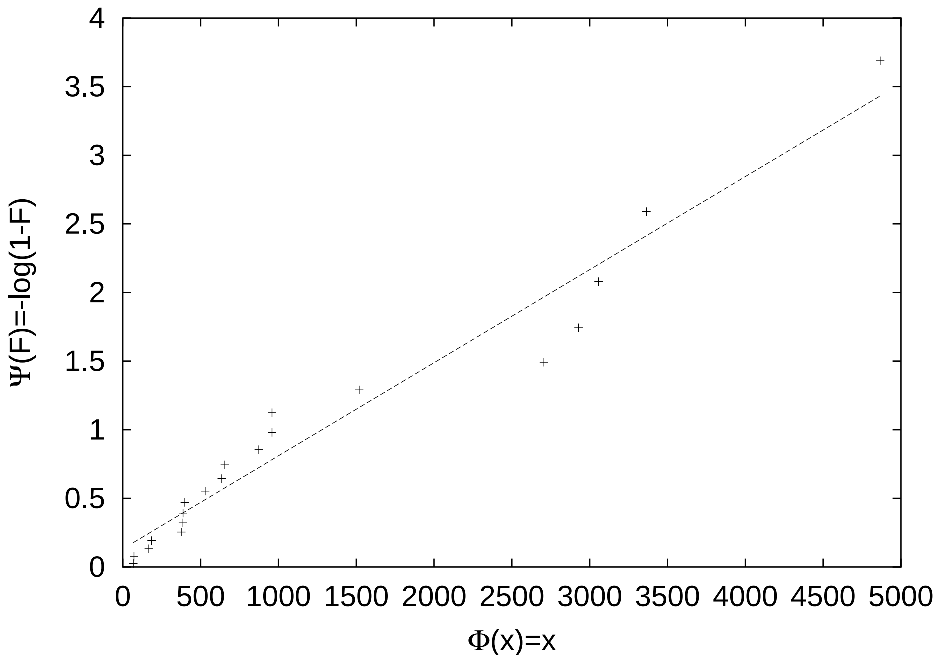

Figura 8.7: Rappresentazione mediante carta probabilistica di dati distribuiti secondo una legge esponenziale di parametro $\lambda = 0.001$.

$$\hat{F}(x_i) = \frac{i - 0.5}{n} \tag{8.14}$$

$$\hat{F}(x_i) = \frac{i}{n + 1} \tag{8.15}$$

$$\hat{F}(x_i) = \frac{i - 0.5}{n + 0.5} \tag{8.16}$$

Tali espressioni sono approssimate e sono state derivate in base all'ipotesi di equiprobabilitá degli eventi elementari. Una stima esatta (ma piuttosto complessa) richiede l'utilizzo della funzione Beta.

Un esempio di applicazione della procedura discussa (con $\hat{F}(x_i) = (i - 0.5)/n$ é mostrato in figura 8.7 per dati distributi secondo una legge esponenziale di parametro $\lambda = 0.001$.

Alcune coppie di funzioni Φ e Ψ per distribuzioni di uso comune sono riportate in seguito:

$$\text{Esponenziale} \begin{cases} \Phi(x) = x \\ \Psi(F) = -\log(1 - F) \end{cases}$$

$$\text{Weibull} \begin{cases} \Phi(x) = \log(x) \\ \Psi(F) = \log(-\log(1 - F)) \end{cases}$$

$$\text{Normale} \begin{cases} \Phi(x) = x \\ \Psi(F) = \sqrt{2}\text{erf}^{-1}(2F - 1) \end{cases}$$

$$\text{Lognormale} \begin{cases} \Phi(x) = \log(x) \\ \Psi(F) = \sqrt{2}\text{erf}^{-1}(2F - 1) \end{cases}$$

8.4 Stima empirica di momenti e percentili

8.4.1 Valore atteso

Per il momento di ordine 1, cioé il valore atteso, lo stimatore empirico é il valore medio dei dati nel campione. Quindi, a seconda che i dati siano raggruppati o meno in classi, si possono utilizzare le seguenti espressioni:

$$\hat{\eta} = \overline{x} = \frac{1}{n}\sum_{k=1}^{n} x_k \tag{8.17}$$

$$\hat{\eta} = \overline{x} = \sum_{i=1}^{C} v_i \frac{n_i}{n} \tag{8.18}$$

dove n é la numerositá del campione, C il numero di classi, v_i il valore assunto dalla VA in ciascuna classe ed n_i il numero di osservazioni in ciascuna classe.

La naturale interpretazione dell'esperimento aleatorio multiplo appena compiuto é che sono stati ripetuti n esperimenti indipendenti, ognuno dei quali ha fornito un valore (realizzazione) della VA $\boldsymbol{x}$. Al fine di costruire modelli statistici, un esperimento come quello descritto puó essere essere considerato in modo differente. Infatti, é possibile considerare che i dati $x_1, \ldots, x_n$ siano ottenuti dall'osservazione di n variabili aleatorie $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ identicamente distribuite ed indipendenti (i.i.d.). Dunque, qualsiasi stimatore é in realtá una funzione di n variabili aleatorie e la stima é, pertanto, una variabile aleatoria anch'essa. Nel seguito saranno mostrate alcune proprietá degli stimatori. Qui, verifichiamo che il valore atteso della stima del valore atteso coincide con il valore atteso stesso. Infatti:

$$\boldsymbol{E}[\hat{\boldsymbol{\eta}}] = \boldsymbol{E}[\frac{1}{n}\sum_{k=1}^{n} \boldsymbol{x}_k] = \frac{1}{n}\sum_{k=1}^{n} \boldsymbol{E}[\boldsymbol{x}_k] = \boldsymbol{E}[\boldsymbol{x}] \tag{8.19}$$

dove la penultima eguaglianza é stata ottenuta sfruttando la linearitá del valore atteso, l'ultima sfruttando il fatto che la distribuzione (e quindi il valore atteso) di ogni variabile aletoria $\boldsymbol{x}_k$ coincide con quella di $\boldsymbol{x}$. Uno stimatore il cui valore atteso coincide con il valore del parametro che si intende stimare si dice non polarizzato.

Si ricorda che $\sigma_{\overline{x}}^2 = \sigma_x^2/n$ e che il lemma di Tchebycheff asserisce che, per una VA $\boldsymbol{z}$ di valore atteso η_z e varianza σ_z^2 é:

$$\Pr(\mid \boldsymbol{z} - \eta_z \mid < \epsilon) \geq 1 - \sigma_z^2/\epsilon^2$$

E' possibile quindi affermare che, per il valore medio di una VA $\boldsymbol{x}$ é:

$$\Pr(\mid \overline{\boldsymbol{x}} - \boldsymbol{E}[\boldsymbol{x}] \mid < \epsilon) \geq 1 - \delta \tag{8.20}$$

dove

$$\delta = \frac{\sigma_x^2}{n\epsilon^2} \tag{8.21}$$

Pertanto, lo scostamento fra media campionaria e valore atteso di $\boldsymbol{x}$ puó essere reso arbitrariamente piccolo se é possibile controllare la dimensione del campione.

8.4.2 Varianza

La varianza é definita come:

$$\boldsymbol{E}[(\boldsymbol{x} - \boldsymbol{E}[\boldsymbol{x}])^2] = \boldsymbol{E}[\tilde{\boldsymbol{x}}^2] \tag{8.22}$$

L'ultima espressione, in cui $\tilde{\boldsymbol{x}} = \boldsymbol{x} - \boldsymbol{E}[\boldsymbol{x}]$ suggerisce che, conoscendo il valore atteso della variabile aleatoria, la stima della varianza potrebbe essere ottenuta calcolando $\tilde{x}_k = x_k - \boldsymbol{E}[\boldsymbol{x}]$ e quindi utilizzando la stima del valore atteso basata sul valore medio (eq. (8.17)). Tuttavia, salvo casi particolari, la stima della varianza implica anche una stima congiunta del valore medio. E' possibile dimostrare che:

$$\hat{\boldsymbol{\sigma}^2} = \frac{1}{n-1}\sum_{k=1}^{n}(\boldsymbol{x}_k - \overline{\boldsymbol{x}})^2 \tag{8.23}$$

fornisce una stima priva di errore sistematico della varianza, quindi $\boldsymbol{E}[\hat{\sigma^2}] = \sigma^2$. La dimostrazione é lasciata al lettore.

Espressioni equivalenti della varianza sono:

$$\hat{\sigma^2} = \frac{1}{n-1}\sum_{k=1}^{n} x_k^2 - n\overline{x}^2 \tag{8.24}$$

algebricamente equivalente a (8.23), Se i dati sono raggruppati in C classi di valore v_i e numero di ossevazioni n_i si utilizza:

$$\hat{\sigma^2} = \frac{1}{n-1}\sum_{i=1}^{C} v_i^2 n_i - n\overline{x}^2 \tag{8.25}$$

Si noti infine che $\sqrt{\hat{\sigma^2}}$ é una stima dello scarto quadratico medio, quindi:

$$\hat{\sigma}_x = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \overline{x})^2} \tag{8.26}$$

Legge di probabilitá della varianza campionaria

E' possibile dimostrare che se $\boldsymbol{x}$ segue una distribuzione normale di deviazione standard σ, allora:

$$\boldsymbol{t} = (n-1)\frac{\hat{\boldsymbol{\sigma}}}{\sigma^2} = \frac{\sum_{i=1}^{n}(\boldsymbol{x}_i - \overline{\boldsymbol{x}})^2}{\sigma^2} \tag{8.27}$$

segue la distribuzione chi-quadrato con $n-1$ gradi di libertá.

8.4.3 Covarianza e correlazione empiriche

La covarianza é definita come:

$$C_{xy} = \boldsymbol{E}[(\boldsymbol{x} - \boldsymbol{E}[\boldsymbol{x}])(\boldsymbol{y} - \boldsymbol{E}[\boldsymbol{y}])] = \boldsymbol{E}[\boldsymbol{x}\boldsymbol{y}] - \boldsymbol{E}[\boldsymbol{x}]\boldsymbol{E}[\boldsymbol{y}] \tag{8.28}$$

Si osservi che la varianza puó essere interpretata come la covarianza di una VA con se stessa, dunque, procedendo in modo formalmente analogo a quanto fatto in precedenza, si puó dimostrare che

$$\hat{C}_{xy} = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \overline{x})(y_k - \overline{y}) \tag{8.29}$$

e' una stima non polarizzata della covarianza. La dimostrazione, analoga alla precedente, é lasciata al lettore. La stima della correlazione é ottenibile da quella della covarianza come:

$$\hat{R}_{xy} = \frac{\hat{C}_{xy}}{\hat{\sigma}_x \hat{\sigma}_y} \tag{8.30}$$

con

$$\hat{\sigma}_x = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(x_k - \overline{x})^2} \tag{8.31}$$

$$\hat{\sigma}_y = \sqrt{\frac{1}{n-1}\sum_{k=1}^{n}(y_k - \overline{y})^2} \tag{8.32}$$

8.4.4 Momenti

Il q-esimo momento centrato della variabile aleatoria ($\boldsymbol{E}[(\boldsymbol{x} - \boldsymbol{E}[x])^q]$) puó essere stimato utilizzando la seguente espressione:

$$\hat{m}_q = \frac{1}{n-1}\sum_{k=1}^{n}(x_k - \overline{x})^q \tag{8.33}$$

Il corrispondente momento non centrato ($\boldsymbol{E}[\boldsymbol{x}^q]$) é dato da:

$$\hat{k}_q = \frac{1}{n}\sum_{k=1}^{n} x_k^q \tag{8.34}$$