APPUNTI DI STATISTICA MATEMATICA

Domenico Marinucci
Dipartimento di Matematica, Università di Roma Tor Vergata

December 12, 2024

Abstract

Note di un corso di Statistica Matematica, con particolare enfasi sulla Teoria Asintotica

1 Disuguaglianze Fondamentali

In questo capitolo richiamiamo le disuguaglianze fondamentali che costituiscono gli strumenti principali per lo sviluppo di tutta la teoria asintotica.

Proposition 1 *(Disuguaglianza di Markov) Sia $X : \Omega \to \mathbb{R}$ una variabile aleatoria a valori non-negativi (cioè $X(\omega) \geq 0$ con probabilità 1) e con valor medio finito (cioè $\mathbb{E}[|X|] = \mathbb{E}[X] < \infty$). Allora $\forall t > 0$ abbiamo*

$$\mathbb{P}(X \geq t) \leq \frac{\mathbb{E}[X]}{t} \ .$$

Proof. E' sufficiente osservare che

$$\begin{aligned} \mathbb{P}(X \geq t) &= \mathbb{E}[\mathbb{I}_{[t,\infty)}(X)] = \int_t^\infty dF_X(x) \\ &\leq \int_t^\infty \frac{x}{t} dF_X(x) \leq \frac{1}{t}\int_0^\infty x dF_X(x) \\ &= \frac{\mathbb{E}[X]}{t} \ . \end{aligned}$$

■

Proposition 2 *(Disuguaglianza di Markov generalizzata) Sia $g : \mathbb{R} \to \mathbb{R}^+$ crescente e $X : \Omega \to \mathbb{R}$ una variabile aleatoria tale che $\mathbb{E}[g(X)] < \infty$. Allora*

$$\Pr(X \geq t) \leq \frac{\mathbb{E}[g(X)]}{g(t)} \ .$$

Proof. La dimostrazione è praticamente identica a quella già vista della disuguaglianza di Markov. In particolare, chiamando $Y = g(X)$ otteniamo

$$\begin{aligned}
\mathbb{P}\left(X \geq t\right) &= \mathbb{P}\left(g(X) \geq g(t)\right) = \mathbb{E}[\mathbb{I}_{[g(t),\infty)}(Y)] \\
&= \int_{g(t)}^{\infty} dF_Y(y) \leq \int_{g(t)}^{\infty} \frac{Y}{g(t)} dF_Y(y) \\
&= \int_{t}^{\infty} \frac{g(x)}{g(t)} dF_X(x) \leq \frac{1}{g(t)} \int_{0}^{\infty} g(x) dF_X(x) \\
&= \frac{\mathbb{E}[g(X)]}{g(t)} \ .
\end{aligned}$$

■

Remark 3 *La disuguaglianza di Markov e la sua versione generalizzata sono il primo esempio di disuguaglianze di concentrazione; in pratica, implicano che tanto più una variabile aleatoria ammette momenti finite, tanto più le code della sua distribuzione vanno a zero velocemente. Ad esempio, consideriamo una variabile aleatoria che ammetta funzione generatrice dei momenti finita in un intorno dell'origine, cioè tale per cui*

$$m_X(u) := \mathbb{E}[\exp(uX)] < \infty \ \textit{per} \ u \in [0, T], \ T > 1 \ .$$

Segue immediatamente che

$$\Pr(X \geq t) \leq \frac{\mathbb{E}[\exp(X)]}{\exp(t)} = \frac{m_X(1)}{\exp(t)} \ ,$$

cioè le code della distribuzione di X devono decadere almeno esponenzialmente.

Proposition 4 *(Disuguaglianza di Chebyshev) Sia $X : \Omega \to \mathbb{R}$ una variabile aleatoria con momento secondo finito (cioè $\mathbb{E}[X^2] < \infty$). Allora $\forall t > 0$ abbiamo*

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq t) \leq \frac{Var[X]}{t^2} \ .$$

Proof. E' sufficiente osservare che, utilizzando la disuguaglianza di Markov per $Y = |X - \mathbb{E}[X]|^2$

$$\begin{aligned}
\mathbb{P}(|X - \mathbb{E}[X]| &\geq t) = \mathbb{P}(|X - \mathbb{E}[X]|^2 \geq t^2) \\
&= \frac{\mathbb{E}[|X - \mathbb{E}[X]|^2]}{t^2} \ .
\end{aligned}$$

■

Remark 5 *Come discusso nei cenni storici più avanti, la disuguaglianza di Chebyshev è precedente a quella di Markov; in effetti Markov è stato un alunno di Chebyshev a San Pietroburgo.*

Example 6 *(Legge debole dei grandi numeri) Anticipando un poco la discussione nei prossimi capitoli, possiamo illustrare immediatamente l'applicazione più importante della disuguaglianza di Chebyshev. Siano infatti* $X_i : \Omega \to \mathbb{R}$ *variabili aleatorie indipendenti con momento secondo finito, e definiamo come al solito valor medio e varianza* $\mu := \mathbb{E}[X]$, $\sigma^2 := \mathbb{E}[(X-\mu)^2]$. *Introduciamo altresì la media aritmetica* $\overline{X}_n := \frac{1}{n}\sum_{i=1}^n X_i$; *dalla disuguaglianza di Chebyshev abbiamo immediatamente che, per ogni* $\varepsilon > 0$

$$\Pr\left\{|\overline{X}_n - \mu| > \varepsilon\right\} \leq \frac{Var\left\{\overline{X}_n\right\}}{\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \to 0 \text{ per } n \to \infty \ .$$

E' immediato verificare che l'ipotesi di indipendenza può essere generalizzata a quella di incorrelazione; anche le ipotesi di identica distribuzione può essere facilmente abbandonata, come vedremo nei prossimi capitoli.

Nelle precedenti disuguaglianze abbiamo mostrato che per variabili aleatorie che ammettono un numero di momenti maggiore le code della funzione di distribuzione decadono più velocemente. E' pertanto naturale chiedersi se si possano ottenere risultati più stringenti focalizzandosi su classi di variabili aleatorie che abbiano supporto compatto, risultando quindi uniformemente limitate. La risposta è affermativa, come mostrato nel prossimo risultato.

Proposition 7 *(Disuguaglianza di Hoeffding) Sia* $X_1, ..., X_n$ *una successione di variabili aleatorie indipendenti e tali che esistano due sequenze di numeri reali* a_i, b_i, $i = 1, 2, .., n$ *per cui valga* $a_i \leq X_i \leq b_i$ *per ogni* i; *assumiamo inoltre che* $\mathbb{E}[Y_i] = 0$ *per* $i = 1, ..., n$. *Per ogni* $t > 0$ *vale la disuguaglianza*

$$\Pr\left\{\sum_{i=1}^n X_i \geq t\right\} \leq \inf_{u \geq 0}\left\{e^{-ut}\prod_{i=1}^n e^{u^2(b_i-a_i)^2/8}\right\} \ .$$

Proof. Per la disuguaglianza di Markov abbiamo, quale che sia $u > 0$

$$\begin{aligned}
\Pr\left\{\sum_{i=1}^n X_i \geq t\right\} &= \Pr\left\{\sum_{i=1}^n uX_i \geq ut\right\} \\
&= \Pr\left\{e^{u\sum_{i=1}^n X_i} \geq e^{ut}\right\} \\
&\leq e^{-ut}\mathbb{E}\left[e^{u\sum_{i=1}^n X_i}\right] \\
&= e^{-ut}\prod_{i=1}^n \mathbb{E}\left[e^{uX_i}\right] \ .
\end{aligned}$$

Scriviamo ora

$$Y_i = \alpha b_i + (1-\alpha)a_i \ , \ \alpha = \frac{Y_i - a_i}{b_i - a_i} \ ;$$

per la convessità della funzione esponenziale, si ha che

$$\begin{aligned}
e^{uX_i} &\leq \frac{Y_i - a_i}{b_i - a_i}e^{ub_i} + (1 - \frac{Y_i - a_i}{b_i - a_i})e^{ua_i} \\
&= \frac{Y_i - a_i}{b_i - a_i}e^{ub_i} + \frac{b_i - Y_i}{b_i - a_i}e^{ua_i} \ ,
\end{aligned}$$

da cui, ricordando che $\mathbb{E}[Y_i] = 0$

$$\mathbb{E}\left[e^{uX_i}\right] \leq -\frac{a_i}{b_i - a_i} e^{ub_i} + \frac{b_i}{b_i - a_i} e^{ua_i} \ .$$

Scriviamo ora

$$-\frac{a_i}{b_i - a_i} e^{ub_i} + \frac{b_i}{b_i - a_i} e^{ua_i} = e^{g_i(v_i)}$$

dove

$$g_i(v_i) := -\gamma_i v_i + \log(1 - \gamma_i + \gamma_i e^{v_i}) \ , \ \gamma_i = -\frac{a_i}{b_i - a_i} \ , \ v_i := u(b_i - a_i) \ ;$$

notiamo infatti che

$$\begin{aligned} e^{g_i(v_i)} &= e^{-\gamma_i v_i}(1 + \frac{a_i}{b_i - a_i} - \frac{a_i}{b_i - a_i} e^{u(b_i - a_i)}) \\ &= e^{ua_i}(1 + \frac{a_i}{b_i - a_i} - \frac{a_i}{b_i - a_i} e^{u(b_i - a_i)}) \\ &= (\frac{b_i - a_i}{b_i - a_i} e^{ua_i} + \frac{a_i}{b_i - a_i} e^{ua_i} - \frac{a_i}{b_i - a_i} e^{ub_i}) \\ &= \frac{b_i}{b_i - a_i} e^{ua_i} - \frac{a_i}{b_i - a_i} e^{ub_i}. \end{aligned}$$

Notiamo che $g_i(0) = 0$ ed inoltre

$$g_i'(0) = -\gamma_i + \left. \frac{\gamma_i e^{v_i}}{(1 - \gamma_i + \gamma_i e^{v_i})} \right|_{v_i = 0} = 0 \ ,$$

$$g_i''(v) = \frac{\gamma_i e^v (1 - \gamma_i + \gamma_i e^{v_i}) - \gamma^2 e^{2v}}{(1 - \gamma_i + \gamma_i e^{v_i})^2} = \frac{\gamma_i e^v (1 - \gamma_i)}{(1 - \gamma_i + \gamma_i e^v)^2} \leq \frac{1}{4} \text{ per ogni } \gamma_i, v \ .$$

Infatti, poiché $\gamma_i, v > 0$

$$\begin{aligned} \frac{\gamma_i e^v (1 - \gamma_i)}{(1 - \gamma_i + \gamma_i e^v)^2} &= \frac{\gamma_i (1 - \gamma_i)}{e^{v/2}} \frac{1}{(1 - \gamma_i + \gamma_i e^{v/2})^2} \\ &\leq \frac{1}{4} \frac{1}{(1 + \gamma_i (e^{v/2} - 1))^2} \leq \frac{1}{4} \ . \end{aligned}$$

Per il teorema del valor medio di Lagrange, esiste $\xi \in (0, v_i)$ tale per cui

$$\begin{aligned} g_i(v_i) &= g_i(0) + g_i'(0) v_i + g_i''(\xi) \frac{v_i^2}{2} \\ &= g_i''(\xi) \frac{v_i^2}{2} \leq \frac{u^2 (b_i - a_i)^2}{8}. \end{aligned}$$

Ne segue che

$$\mathbb{E}\left[e^{tX_i}\right] \leq e^{g_i(v_i)} \leq e^{u^2 (b_i - a_i)^2 / 8},$$

ed la Proposizione è dimostrata. ■

Corollary 8 *(Media aritmetica di variabili aleatorie di Bernoulli) Siano $Y_1, ..., Y_n$ variabili aleatorie indipendenti identicamente distribuite con legge di Bernoulli $Ber(p)$. Per ogni $\varepsilon > 0$, abbiamo che*

$$\Pr\left\{|\overline{Y}_n - p| > \varepsilon\right\} \leq 2e^{-2n\varepsilon^2}.$$

Proof. Si prenda $X_i := \frac{1}{n}(Y_i - p)$. E' immediato verificato che $\mathbb{E}[X_i] = 0$ e $a_i \leq X_i \leq b_i$ per ogni $i = 1, ..., n$, con $a_i = -\frac{p}{n}$ e $b_i = \frac{1-p}{n}$, da cui $(b_i - a_i)^2 = \frac{1}{n^2}$. Applicando la disuguaglianza di Hoeffding si ha

$$\Pr\left\{\overline{Y}_n - p > \varepsilon\right\} \leq \inf_u \left\{e^{-u\varepsilon} \prod_{i=1}^{n} e^{u^2/8n^2}\right\} = \inf_u \left\{e^{-u\varepsilon} e^{u^2/8n}\right\}$$

e prendendo $u = 4n\varepsilon$ otteniamo

$$\inf_u \left\{e^{-u\varepsilon} e^{u^2/8n}\right\} \leq e^{-4n\varepsilon\times\varepsilon} e^{16n^2\varepsilon^2/8n} = e^{-2n\varepsilon^2}.$$

La dimostrazione è completata ripetendo lo stesso ragionamento per $\Pr\left\{\overline{Y}_n - p < -\varepsilon\right\}$. ■

Remark 9 *(Un confronto tra la disuguaglianza di Chebyshev e quella di Hoeffding). Nel caso di variabili aleatorie limitate, la disuguaglianza di Hoeffding è enormemente più efficiente della disuguaglianza di Chebyshev. Ad esempio, nel caso di variabili Bernoulliane di parametro $p = \frac{1}{2}$ con $n = 100$ ed $\varepsilon = 0.2$ la disuguaglianza di Chebyshev ci dà il seguente limite limite superiore:*

$$\Pr\left\{|\overline{Y}_n - p| > 0.2\right\} \leq \frac{Var\left\{\overline{Y}_n\right\}}{\varepsilon^2} = \frac{1}{4n\varepsilon^2} \simeq 0.0625$$

mentre con la disuguaglianza di Hoeffding si ottiene

$$\Pr\left\{|\overline{Y}_n - p| > 0.2\right\} \leq 2e^{-2\times 100\times(0.2)^2} \simeq 0.00067 \ .$$

Remark 10 *(Concentrazione del volume di un ipercubo). La disuguaglianza di Hoeffding ammette una interessante intepretazione geometrica quando viene applicata al comportmanto de volume di un ipercubo $[0,1]^n$, nel limite in cui $n \to \infty$. Consideriamo infatti l'iperpiano $\Gamma_n := \{x_i : x_1 + x_2 + ...x_n = n/2\}$; definitamo un "Tubo" di raggio ε intorno a Γ_n come*

$$Tub_\varepsilon(\Gamma_n) := \{\mathbf{x} \in \mathbb{R}^n : dist\{\mathbf{x}, \Gamma_n\} \leq \varepsilon\} \ .$$

Per ogni $\varepsilon > 0$, il volume dell'intersezione tra $Tub_\varepsilon(\Gamma_n)$ e $[0,1]^n$ converge esponenzialmente a 1 quando n diverge all'infinito; in altre parole, la massa del cubo si concentra esponenzialmente in un intorno arbitrariamente piccolo della diagonale principale.

Le precedenti disuguaglianze ci permettono di avere dei limiti superiori per il comportamento delle code di variabili aleatorie generiche. E' interessante confrontare il loro comportamento con quello delle code di una variabile Gaussiana standard; a questo scopo introduciamo la prossima proposizione.

Proposition 11 *(Disuguaglianza di Mill-Gordon) Sia $Z \sim N(0,1)$ una variabile Gaussiana standard. Si ha che*

$$\sqrt{\frac{1}{2\pi}}\frac{t}{1+t^2}\exp(-\frac{t^2}{2}) \leq \Pr\{Z > t\} \leq \sqrt{\frac{1}{2\pi}}\frac{1}{t}\exp(-\frac{t^2}{2}) \ .$$

Proof. La disuguaglianza di destra può essere stabilita con un semplice cambio di variabile; abbiamo:

$$\begin{aligned}
\Pr\{Z > t\} &= \int_t^\infty \frac{1}{\sqrt{2\pi}}\exp(-\frac{x^2}{2})dx \\
&\leq \int_t^\infty \frac{x}{t}\exp(-\frac{x^2}{2})dx \\
&= \frac{1}{t}\frac{1}{\sqrt{2\pi}}\int_{t^2/2}^\infty \exp(-y)dy \\
&= \frac{1}{t}\frac{1}{\sqrt{2\pi}}\exp(-\frac{t^2}{2}) \ .
\end{aligned}$$

dopo il cambio di variabile $y = \frac{x^2}{2}$. Per il limite inferiore, possiamo ragionare come segue:

$$\begin{aligned}
\Pr\{Z > t\} &= \int_t^\infty \frac{1}{\sqrt{2\pi}}\exp(-\frac{x^2}{2})dx \\
&\geq \int_t^\infty \frac{x^4+2x^2-1}{x^4+2x^2+1}\frac{1}{\sqrt{2\pi}}\exp(-\frac{x^2}{2})dx \\
&= \int_t^\infty \frac{x^2(x^2+1)+x^2-1}{(x^2+1)^2}\frac{1}{\sqrt{2\pi}}\exp(-\frac{x^2}{2})dx \\
&= \int_t^\infty \left\{\frac{x^2}{(x^2+1)^2}+\frac{x^2-1}{(x^2+1)^2}\right\}\frac{1}{\sqrt{2\pi}}\exp(-\frac{x^2}{2})dx \ .
\end{aligned}$$

Chiamiamo ora

$$\psi(x) = \frac{x}{1+x^2}\exp(-\frac{x^2}{2})$$

e notiamo che

$$\begin{aligned}
\psi'(x) &= \frac{d\psi(x)}{dx} = \frac{1+x^2-2x^2}{(1+x^2)^2}\exp(-\frac{x^2}{2}) - \frac{x^2}{1+x^2}\exp(-\frac{x^2}{2}) \\
&= -\left\{\frac{x^2}{(x^2+1)^2}+\frac{x^2-1}{(x^2+1)^2}\right\}\exp(-\frac{x^2}{2}) \ .
\end{aligned}$$

La dimostrazione è dunque completata dal Teorema Fondamentale del Calcolo:

$$-\frac{1}{\sqrt{2\pi}}\int_t^\infty \psi'(x)dx = -\frac{1}{\sqrt{2\pi}}\left[\psi(x)\right]_t^\infty = \sqrt{\frac{1}{2\pi}}\frac{t}{1+t^2}\exp(-\frac{t^2}{2}) \ .$$

■

Remark 12 *Queste due disuguglianze sono piuttosto efficienti; ad esempio, per* $t = 2$ *abbiamo*

$$\Pr\{Z > 2\} = \int_2^\infty \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}) dx = 0.02275$$

mentre le stime di Mill danno

$$\frac{1}{\sqrt{2\pi}} \frac{t}{1+t^2} \exp(-\frac{t^2}{2})\Big|_{t=2} = 0.021596 \leq \Pr\{Z > 2\} \leq \frac{1}{\sqrt{2\pi}} \frac{1}{t} \exp(-\frac{t^2}{2})\Big|_{t=2} = 0.027 \ .$$

Per $t = 3$ *il risulta diventa ancora più preciso; otteniamo*

$$\Pr\{Z > 3\} = \int_3^\infty \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2}) dx = 0.001349$$

mentre le stime di Mill danno

$$\begin{aligned} \frac{1}{\sqrt{2\pi}} \frac{t}{1+t^2} \exp(-\frac{t^2}{2})\Big|_{t=3} &= 0.001329 \\ &\leq \Pr\{Z > 3\} \\ &\leq \frac{1}{\sqrt{2\pi}} \frac{1}{t} \exp(-\frac{t^2}{2})\Big|_{t=3} = 0.0014773 \ . \end{aligned}$$

Remark 13 *Vediamo quale risultato otterremo per la somma di variabili Bernoulliane considerando l'approssimazione asintotica che segue dal teorema del Limte Centrale; prendendo per semplicità* $p = \frac{1}{2}$ *abbiamo:*

$$\overline{Y}_n - \mathbb{E}[\overline{Y}_n] \simeq \mathcal{N}(0, \frac{p(1-p)}{n}) = \mathcal{N}(0, \frac{1}{4n})$$

ed utilizzando la disuguaglianza di Mill abbiamo dunque

$$\begin{aligned} \Pr\left\{|\overline{Y}_n - \frac{1}{2}| > 0.2\right\} &= \Pr\left\{|\mathcal{N}(0, \frac{1}{4n})| > 0.2\right\} + o_{n\to\infty}(1) \\ &= \Pr\left\{|\mathcal{N}(0,1)| > 0.2 \times 2\sqrt{n}\right\} \\ &\leq \frac{2}{\sqrt{2\pi}} \frac{1}{0.2 \times 2\sqrt{n}} e^{-4\times n\times (0.2)^2/2}. \end{aligned}$$

Per $n = 100$ *otteniamo*

$$\frac{2}{\sqrt{2\pi}} \frac{1}{0.2 \times 20} e^{-2\times 100\times (0.2)^2} = 6.691\,5 \times 10^{-5},$$

un valore circa 10 volte inferiore a quello ottenuto dalla disuguaglianza di Hoeffding. Il risultato però non deve ingannare: la disuguaglianza di Hoeffding vale in senso stretto, mentre qui stiamo trovando un limite superiore alla probabilità Gaussiana, tralasciando il fatto che il termine di approssimazione $o_{n\to\infty}(1)$ *è molto più grande della maggiorazione ottenuta: si può dimostrare solo l'ordine* $O(n^{-1/2})$*. In altre parole, il Teorema del Limite Centrale NON può essere sfruttato per ottenere un limite così efficiente per la probabilità sulle code della variabile* $\overline{Y}_n$*.*

La prossima disuguaglianza dovrebbe essere ben nota dai corsi di Geometria ed Analisi.

Proposition 14 *(Disuguaglianza di Cauchy-Schwartz) Siano $X, Y : \Omega \to \mathbb{R}$ variabili aleatorie con momento secondo finito. Abbiamo che*

$$|\mathbb{E}[XY]|^2 \leq \mathbb{E}[X^2]\mathbb{E}[Y^2] \ .$$

La disuguaglianza è stretta, a meno che le variabili aleatorie siano tra loro in relazione lineare con probabilità 1, cioè esista un numero reale t^* tale per cui

$$\mathbb{E}[(Y - t^*X)^2] = 0 \ .$$

Proof. Consideriamo la funzione quadratica (in t)

$$g(t) : \mathbb{E}[(Y - tX)^2] = \mathbb{E}[Y^2] - 2t\mathbb{E}[XY] + t^2\mathbb{E}[X^2] \geq 0 \ .$$

L'equazione $g(t) = 0$ ovviamente ammette al più una radice reale di molteplicità due; quindi il discriminante deve avere valore non-positivo, in particolare

$$4(\mathbb{E}[XY])^2 - 4\mathbb{E}[X^2]\mathbb{E}[Y^2] \leq 0$$

da cui segue la disuguaglianza, che diviene una ugualianza se e solo se esiste t^* tale per cui $g(t^*) = 0$. ■

Remark 15 *Il richiamo alla geometria ci anticipa una idea che avrà grande importanza nei corsi di probabilità più avanzati: lo spazio delle variabili aleatorie di momento secondo finito può essere visto come uno spazio vettoriale dotato di un proodtto interno (prodotto scalare) definito da*

$$\langle X, Y \rangle := \mathbb{E}[XY] \ .$$

Nel caso particolare di variabili con valor medio nullo questo prodotto interno è la covarianza.

Per la prossima disuguaglianza ricordiamo innanzitutto che una funzione $g(.) : \mathbb{R} \to \mathbb{R}$ si dice convessa se per ogni $\alpha \in (0, 1)$ e per ogni x_1 e x_2 nel suo dominio (connesso), si ha che

$$g(\alpha x_1 + (1 - \alpha)x_2) \leq \alpha g(x_1) + (1 - \alpha)g(x_2) \ .$$

Data una funzione convessa per ogni punto x_0 esiste una costante $L = L_{x_0}$ tale per cui

$$g(x) \geq g(x_0) + L(x - x_0) \ ;$$

la costante non è necessariamente unica (se la funzione è derivabile, coincide con il valore della derivata in quel punto).

Proposition 16 *(Disuguaglianza di Jensen) Sia X una variabile aleatoria con valor medio finito e $g(.)$ una funzione convessa. Abbiamo che*

$$\mathbb{E}[g(X)] \geq g(\mathbb{E}[(X)]) \ ,$$

dove la grandezza a sinistra della disuguaglianza può essere infinita.

Proof. E' sufficiente osservare che, per la monotonia del valor medio e prendendo $x_0 = \mathbb{E}[(X)]$

$$\begin{aligned} \mathbb{E}[g(X)] &\geq \mathbb{E}[g(x_0) + L(X - x_0)] = \mathbb{E}[g(x_0)] + \mathbb{E}[L(X - x_0)] \\ &= g(\mathbb{E}[(X)]) + L\mathbb{E}[(X - \mathbb{E}[(X)])] = g(\mathbb{E}[(X)]) \ . \end{aligned}$$

■

2 Modalità di convergenza

Gli strumenti fondamentali per studiare procedure e metodi statistici da un punto di vista matematico sono forniti dai teoremi di convergenza asintotica. Prima di poterli enunciare, è molto importante capire quali siano le principali modalità di convergenza per sequenze di varianili aleatorie, e studiare le relazioni che intercorrono tra loro.

Definition 17 *((1): Convergenza in Legge, o Convergenza in Distribuzione) Sia* $\{X_n\}_{n\in\mathbb{N}}$ *una successione di variabili aleatorie; diremo che la sequenza converge in legge alla variabile aleatoria* X *(scritto* $X_n \to_d X$*) se e solo se si ha*

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x) \text{ per ogni } x \text{ punto di continuità di } F_X(.) \ .$$

Remark 18 *Nella definizione precedente non abbiamo specificato né lo spazio di probabilità su cui è definita la sequenza* $\{X_n\}_{n\in\mathbb{N}}$*, né quello su cui è definita* X*. Non è stata una dimenticanza: non è infatti necessario che tali spazi siano gli stessi, anzi lo spazio di probabilità può persino essere diverso al variare di* n*, cioè potremmo considerare v.a. definite su sequenze* $\{\Omega_n, \Im_n, \mathbb{P}_n\}$*.*

Remark 19 *Si potrebbe pensare di introdurre una definizione diversa di convergenza in probabilità, lasciando cadere il requisito che la convergenza avvenga solo nei punti di continuità della funzione di distribuzione* $F_X(.)$*. Si vede facilmnente che questo darebbe adito però ad alcuni gravi paradossi. Si consideri ad esempio la sequenza deterministica* $X_n = \frac{1}{n}$*; considerate come variabili aleatorie, le* X_n *hanno funzione di ripartizione* $F_{X_n}(x) = \mathbb{I}_{[\frac{1}{n},\infty)}(x)$*. Prendiamo adesso la variabile aleatoria identicamente uguale a zero* $X = 0$*, la cui funzione di distribuzione è* $F_X(x) = \mathbb{I}_{[0,\infty)}(x)$ *abbiamo che*

$$\lim_{n\to\infty} F_{X_n}(0) = \lim_{n\to\infty} 0 = 0 \neq F_X(0) = 1 \ .$$

Pertanto se richiedessimo la convergenza delle funzioni di continuità in ogni punto dovremmo concludere che la sequenza $\frac{1}{n}$ *non converge in distribuzione a* 0*, quando* $n \to \infty$*.*

Example 20 *Sia* X_n *una sequenza di variabilie binomiali con funzione di probabilità discreta*

$$\Pr\{X_n = k\} = \binom{n}{k} p_n^k (1-p_n)^{n-k}, \ k = 0, 1, 2, ..., n \ (0 \text{ altrimenti}),$$

dove $p_n := \frac{\lambda}{n}$*,* $\lambda > 0$*. Abbiamo che* $X_n \to_d X$*,* $X \sim Pois(\lambda)$ *. Infatti si verifica*

facilmente che

$$\begin{aligned}
\lim_{n\to\infty} \Pr\{X_n = k\} &= \lim_{n\to\infty} \binom{n}{k} p_n^k (1-p_n)^{n-k} \\
&= \lim_{n\to\infty} \frac{n!}{k!(n-k)!} \frac{\lambda^k}{n^k} (1-\frac{\lambda}{n})^n (1-\frac{\lambda}{n})^{-k} \\
&= \frac{\lambda^k}{k!} (\lim_{n\to\infty} \frac{n!}{(n-k)!} \frac{1}{n^k}) (\lim_{n\to\infty} (1-\frac{\lambda}{n})^n) (\lim_{n\to\infty} (1-\frac{\lambda}{n})^{-k}) \\
&= \frac{\lambda^k}{k!} e^{-\lambda},
\end{aligned}$$

usando limiti notevoli conosciuti dai primi corsi di Analisi.

Definition 21 *((2): Convergenza in Probabilità o Convergenza Debole) Sia* $\{X_n\}_{n\in\mathbb{N}}$ *una successione di variabili aleatorie definita sullo spazio di probabilità* $\{\Omega, \Im, \mathbb{P}\}$*; sia inoltre* X, $X : \Omega \to \mathbb{R}$ *una variabile aleatoria definita sullo stesso spazio di probabilità. Diremo che la sequenza converge in probabilità alla variabile aleatoria* X *(scritto* $X_n \to_p X$*) se e solo se si ha*

$$\lim_{n\to\infty} \mathbb{P}\{|X_n - X| > \varepsilon\} = 0 \text{ per ogni } \varepsilon > 0 \ .$$

Lemma 22 *((2)* $\Rightarrow$ *(1)) La convergenza in probabilità implica la convergenza in distribuzione, cioè* $\{X_n \to_p X\} \Rightarrow \{X_n \to_d X\}$.

Proof. Abbiamo che

$$\begin{aligned}
F_{X_n}(x) &= \mathbb{P}\{X_n \le x\} \\
&= \mathbb{P}\{X_n \le x, |X_n - X| > \varepsilon\} + \mathbb{P}\{X_n \le x, |X_n - X| \le \varepsilon\} \\
&\le \mathbb{P}\{|X_n - X| > \varepsilon\} + F_X\{X \le x + \varepsilon\} \ .
\end{aligned}$$

Similmente

$$\begin{aligned}
F_X(x-\varepsilon) &= \mathbb{P}\{X \le x - \varepsilon\} \\
&= \mathbb{P}\{X \le x - \varepsilon, |X_n - X| > \varepsilon\} + \mathbb{P}\{X \le x - \varepsilon, |X_n - X| \le \varepsilon\} \\
&\le \mathbb{P}\{|X_n - X| > \varepsilon\} + F_{X_n}\{X \le x\}
\end{aligned}$$

e dunque

$$F_X(x-\varepsilon) - \mathbb{P}\{|X_n - X| > \varepsilon\} \le F_{X_n}\{X \le x\} \ .$$

Di conseguenza

$$\begin{aligned}
F_X(x-\varepsilon) &= \lim\inf_{n\to\infty} [F_X(x-\varepsilon) - \mathbb{P}\{|X_n - X| > \varepsilon\}] \\
&\le \lim\inf_{n\to\infty} F_{X_n}\{X \le x\} \\
&\le \lim\sup_{n\to\infty} F_{X_n}\{X \le x\} \\
&\le \lim\sup_{n\to\infty} [\mathbb{P}\{|X_n - X| > \varepsilon\} + F_X\{X \le x + \varepsilon\}] \\
&= F_X(x+\varepsilon) \ .
\end{aligned}$$

Dato che ε è arbitrario, la convergenza segue immediatamente per tutti i punti di continuità di $F_X(.)$. ■

Remark 23 *E' immediato verificare che l'implicazione inversa non vale in generale; strettamente parlando, non avrebbe nemmeno senso porre la domanda, visto che (2) richiede che le variabili siano definite tutte sullo stesso spazio di probabilità, mentre (1) no. Comunque a fini esplicativi si possono prendere banali controesempi: sia $\{X_n\}_{n\in\mathbb{N}}$ una successione di variabili aleatorie Gaussiane standard indipendenti, e sia X una altra Gaussiana standard indipendente da questa sequenza: è ovvio che $X_n \stackrel{d}{=} X$ e pertanto $X_n \rightarrow_d X$, mentre invece*

$$\Pr\{|X_n - X| > \varepsilon\} = \Pr\left\{|X| > \frac{\varepsilon}{\sqrt{2}}\right\} \nrightarrow 0 \ ,$$

dove abbiamo usato $X_n - X \stackrel{d}{=} N(0,2)$. Una parziale implicazione inversa esiste però nel caso in cui sia abbia convergenza in distribuzione ad una costante.

Lemma 24 *Sia X_n una successione di variabili aleatorie tali per cui $X_n \rightarrow_d c$, dove $c \in \mathbb{R}$ è una qualsiasi costante. Allora esiste uno spazio di probabilità $\{\Omega, \Im, \mathbb{P}\}$ ed una successione di variabili aleatorie $\{X'_n\}$ definite su questo spazio tali per cui $X'_n \stackrel{d}{=} X_n$ per ogni n e $X'_n \rightarrow_p c$.*

Proof. Notiamo innanzitutto che $F_c(x) = \mathbb{I}_{[c,\infty)}(x)$, da cui

$$\begin{aligned}
\Pr\{|X'_n - c| > \varepsilon\} &= \Pr\{X'_n > c + \varepsilon\} + \Pr\{X'_n < c - \varepsilon\} \\
&= 1 - F_{X_n}(c+\varepsilon) + F_{X_n}(c-\varepsilon) \\
&\rightarrow 1 - \mathbb{I}_{[c,\infty)}(c+\varepsilon) + \mathbb{I}_{[c,\infty)}(c-\varepsilon) = 0 \ .
\end{aligned}$$

■

Remark 25 *Intuitivamente, la (1) sta richiedendo che X_n e X "tendano a dare gli stessi numeri con la stessa probablità", mentre la (2) sta richiedendo che "in una singola estrazione, il valore di X_n e X tende ad essere vicino". E' quindi naturale che la (2) sia una richiesta più forte della (1).*

Definition 26 *((3): Convergenza in Media r-esima) Sia $\{X_n\}_{n\in\mathbb{N}}$ una successione di variabili aleatorie definita sullo spazio di probabilità $\{\Omega, \Im, \mathbb{P}\}$; sia inoltre X, $X : \Omega \rightarrow \mathbb{R}$ una variabile aleatoria definita sullo stesso spazio di probabilità. Diremo che la sequenza converge in media $r-$esima alla variabile aleatoria X (scritto $X_n \rightarrow_p X$) se e solo se si ha*

$$\lim_{n\rightarrow\infty} \mathbb{E}|X_n - X|^r = 0 \ .$$

Remark 27 *Affinchè la definizione abbia senso stiamo implicitamente richiedendo che tutte le variabili aleatorie coinvolte abbiano momenti $r-$esimi finiti.*

Remark 28 *I casi più importanti sono quelli che corrispondono a $r = 2$ (Convergenza in Media Quadratica) e $r = 1$.*

Lemma 29 *La convergenza in media r-esima implica la convergenza in probabilità, cioè $\{X_n \to_r X\} \Rightarrow \{X_n \to_p X\}$.*

Proof. La dimostrazione è una immediata conseguenza della disuguaglianza di Markov:

$$\lim_{n\to\infty} \Pr\{|X_n - X| > \varepsilon\} \leq \lim_{n\to\infty} \frac{\mathbb{E}\left[|X_n - X|^r\right]}{\varepsilon^r} .$$

■

Remark 30 *L'implicazione opposta è falsa in generale - in effetti, la convergenza in probabilità non implica nemmeno l'esistenza di momenti di un ordine qualsiasi $r > 0$. Un controesempio più interessante è fornito dalla seguente sequenza:*

$$X_n = \begin{cases} 0 \text{ con probabilità } 1 - \frac{1}{n} \\ n \text{ con probabilità } \frac{1}{n} \end{cases} .$$

Questa sequenza ha valor medio finito e costante ($\mathbb{E}[X_n] = 1$), ma converge in probabilità alla variabile aleatoria identicamente pari a zero, che ha ovviamente valor medio nullo.

Remark 31 *C'è una parziale freccia inversa tra la convergenza in probabilità e quella in media r-esima. Intutivamente la convergenza in probabilità non implica la convergenza in media r-esima perché può capitare che con probabilità sempre più piccola vengano assunti valori sempre più grandi, come nel contro-esempio precedente; se però imponiamo che le variabili siano limitate, la possibilità di questi contro-esempi cade. In aprticolare, abbiamo il Lemma che segue.*

Lemma 32 *Sia $\{X_n\}$ una successione di variabili aleatorie limitate $|X_n| \leq M \in \mathbb{R}$ definite sullo spazio di probabilità $\{\Omega, \Im, P\}$ e tale per cui $X_n \to_p X$ (X è evidentemente definita sullo stesso spazio di probabilità; si noti che $|X| \leq M$ con probabilità 1). Allora $X_n \to_r X$ per ogni $r \in \mathbb{R}$.*

Proof. Abbiamo che

$$\begin{aligned} E|X_n - X|^r &= E|X_n - X|^r I_{\{|X_n - X| > \varepsilon\}} + E|X_n - X|^r I_{\{|X_n - X| \leq \varepsilon\}} \\ &\leq (2M)^r E[I_{\{|X_n - X| > \varepsilon\}}] + \varepsilon^r = (2M)^r \Pr\{|X_n - X| > \varepsilon\} + \varepsilon^r, \end{aligned}$$

e la dimostrazione è conclusa per l'arbitrarietà di ε e l'ipotesi sulla convergenza in probabilità. ■

Example 33 *(Legge debole dei grandi numeri) Sia $\{X_n\}_{n\in\mathbb{N}}$ una successione di variabili aleatorie indipendenti ed identicamente distribuite definita sullo spazio di probabilità $\{\Omega, \Im, \mathbb{P}\}$, con $\mathbb{E}[X] = \mu$ e $\mathbb{E}\left[(X - \mu)^2\right] = \sigma^2 < \infty$. Allora*

$$\overline{X}_n := \frac{1}{n} \sum_{i=1}^{n} X_i \to_2 \mu .$$

Infatti

$$\begin{aligned}\mathbb{E}\left[(\overline{X}_n - \mu)^2\right] &= Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\ &= \frac{1}{n^2}\sum_{i=1}^{n} Var\left[X_i\right] \\ &= \frac{\sigma^2}{n}\ .\end{aligned}$$

Remark 34 *Le ipotesi della legge debole dei grandi numeri possono essere grandemente generalizzate. Ad esempio, si può sostituire l'indipendenza con l'incorrelazione, e l'identica distribuzione con l'ipotesi che il valor medio sia costante e la varianza uniformemente limitata.*

E' importante ricordare la relazione che esiste tra la convergenza in media di ordini diversi.

Lemma 35 *Sia $\{X_n\}_{n\in\mathbb{N}}$ una successione di variabili aleatorie definita sullo spazio di probabilità $\{\Omega, \Im, \mathbb{P}\}$; sia inoltre X, $X : \Omega \to \mathbb{R}$ una variabile aleatoria definita sullo stesso spazio di probabilità. Allora $\{X_n \to_{r_2} X\} \Rightarrow \{X_n \to_{r_1} X\}$ per ogni $0 < r_1 < r_2$.*

Proof. Per la dimostrazione, basta considerare la funzione convessa $f(x) := |x|^{r_2/r_1}$ e notare che, per la disuguaglianza di Jensen$\cap\cap$

$$f(\mathbb{E}|X_n - X|^{r_1}) = (\mathbb{E}|X_n - X|^{r_1})^{r_2/r_1} \leq \mathbb{E}\left[f(|X_n - X|^{r_1})\right] = (\mathbb{E}|X_n - X|^{r_2})$$

e quindi

$$(\mathbb{E}|X_n - X|^{r_1})^{1/r_1} \leq (\mathbb{E}|X_n - X|^{r_2})^{1/r_2},$$

$$\left\{\lim_{n\to\infty} \mathbb{E}|X_n - X|^{r_2} = 0\right\} \Rightarrow \left\{\lim_{n\to\infty} \mathbb{E}|X_n - X|^{r_1} = 0\right\}\ .$$

■

Definition 36 *((4), Convergenza quasi certa) Sia $\{X_n\}_{n\in\mathbb{N}}$ una successione di variabili aleatorie definita sullo spazio di probabilità $\{\Omega, \Im, \mathbb{P}\}$; sia inoltre X, $X : \Omega \to \mathbb{R}$ una variabile aleatoria definita sullo stesso spazio di probabilità. Diciamo che X_n converge quasi certamente a X, scritto $X_n \to_{q.c.} X$, se e solo se*

$$\mathbb{P}\left\{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\} = \mathbb{P}\left\{\lim_{n\to\infty} X_n = X\right\} = 1\ .$$

Remark 37 *Notiamo che l'evento $\{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}$ può essere scritto come*

$$\left\{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\} = \cap_{k=1}^{\infty} \cup_{n=1}^{\infty} \cap_{m=n}^{\infty} \left\{\omega : |X_m(\omega) - X(\omega)| < \frac{1}{k}\right\},$$

cioè l'insieme di quegli ω tali per cui, per scelto $\frac{1}{k}$ piccolo quanto si vuole, $|X_m(\omega) - X(\omega)| < \frac{1}{k}$ definitivamente. Possiamo quindi definire la convergenza quasi certa imponendo che il complementare di questo evento abbia probabilità nulla, cioè

$$\begin{aligned} 0 &= \mathbb{P}\left(\left[\cap_{k=1}^{\infty}\cup_{n=1}^{\infty}\cap_{m=n}^{\infty}\left\{\omega : |X_m(\omega) - X(\omega)| < \frac{1}{k}\right\}\right]^c\right) \\ &= \mathbb{P}\left(\left[\cup_{k=1}^{\infty}\cap_{n=1}^{\infty}\cup_{m=n}^{\infty}\left\{\omega : |X_m(\omega) - X(\omega)| \geq \frac{1}{k}\right\}\right]\right) \\ &= \lim_{n\to\infty}\mathbb{P}\left(\cup_{m=n}^{\infty}\left\{\omega : |X_m(\omega) - X(\omega)| \geq \frac{1}{k}\right\}\right) \text{ , per ogni } k = 1, 2, ... \end{aligned}$$

Da quest'ultima riga è immediato vedere che la convergenza quasi certa implica la convergenza debole, $(4) \Rightarrow (2)$.

Remark 38 *La convergenza debole al contrario non implica la convergenza quasi certa. Un controesempio può essere costruito come segue: sia U una variabile aleatoria uniforme in $[0, 1]$, definita su uno spazio di probabilità adeguato. Consideriamo la sequenza $\{X_n\}_{n\in\mathbb{N}}$ definita come*

$$\begin{aligned} X_1 &= \mathbb{I}_{[0,\frac{1}{2})}(U) \text{ , } X_2 = \mathbb{I}_{[\frac{1}{2},1]}(U) \\ X_3 &= \mathbb{I}_{[0,\frac{1}{4})}(U) \text{ , } X_4 = \mathbb{I}_{[\frac{1}{4},\frac{1}{2})}(U) \text{ , } X_5 = \mathbb{I}_{[\frac{2}{4},\frac{3}{4})}(U) \text{ , } X_6 = \mathbb{I}_{[\frac{3}{4},1]}(U) \text{ , ...} \\ X_7 &= \mathbb{I}_{[0,\frac{1}{8})}(U) \text{ , } X_8 = \mathbb{I}_{[\frac{1}{8},\frac{2}{8})}(U) \text{ , ..., } X_{14} = \mathbb{I}_{[\frac{7}{8},1]}(U) \text{ , } \\ X_{15} &= \mathbb{I}_{[0,\frac{1}{16})}(U) \text{ , } X_8 = \mathbb{I}_{[\frac{1}{16},\frac{2}{16})}(U) \text{ , ..., } X_{30} = \mathbb{I}_{[\frac{15}{16},1]}(U) \text{ , } \\ &\ldots \\ X_{2^q-1} &= \mathbb{I}_{[0,\frac{1}{2^q})}(U) \text{ , } X_{2^q} = \mathbb{I}_{[\frac{1}{2^q},\frac{2}{2^q})}(U) \text{ , ...} \end{aligned}$$

Si ha che

$$\lim_{n\to\infty}\Pr\{|X_n| > 0\} = 0 \text{ , } \lim_{n\to\infty}\Pr\{\cup_{m=n}^{\infty}|X_m| > 0\} = 1 \text{ ,}$$

quindi la sequenza converge in probabilità a zero, anche se assume infinitamente spesso il valore 1 e pertanto non può convergervi quasi certamente.

Definition 39 *((5), Convergenza completa) Sia $\{X_n\}_{n\in\mathbb{N}}$ una successione di variabili aleatorie definita sullo spazio di probabilità $\{\Omega, \Im, \mathbb{P}\}$; sia inoltre X, $X : \Omega \to \mathbb{R}$ una variabile aleatoria definita sullo stesso spazio di probabilità. Diciamo che X_n converge completamente a X, scritto $X_n \to_{c.c.} X$, se e solo se*

$$\sum_{n=1}^{\infty}\mathbb{P}\{|X_n - X| > \varepsilon\} < \infty \text{ , per ogni } \varepsilon > 0 \text{ .}$$

Remark 40 *La convergenza completa implica quella quasi certa ($(5) \Rightarrow (4)$); infatti, per la subadditività della misura di probabilità abbiamo che*

$$\begin{aligned} &\lim_{n\to\infty}\mathbb{P}\left(\cup_{m=n}^{\infty}\left\{\omega : |X_m(\omega) - X(\omega)| \geq \frac{1}{k}\right\}\right) \\ \leq\ &\lim_{n\to\infty}\sum_{m=n}^{\infty}\mathbb{P}\left(\left\{\omega : |X_m(\omega) - X(\omega)| \geq \frac{1}{k}\right\}\right) = 0 \text{ ,} \end{aligned}$$

perchè il resto n−esimo di una serie convergente va a zero. Il vice versa non è vero: consideriamo ad esempio una variabile uniforme in $[0,1]$ U*, ed introduciamo la sequenza*

$$X_n := \mathbb{I}_{[0,\frac{1}{n}]}(U) \ .$$

Chiaramente

$$\sum_{n=1}^{N} \Pr\{X_n > 0\} = \sum_{n=1}^{N} \frac{1}{n} \to \infty \ \textit{per}\ N \to \infty \ ;$$

d'altra parte però per ogni ω *tale che* $U(\omega) \neq 0$ *abbiamo* $\lim_{n\to\infty} \mathbb{I}_{[0,\frac{1}{n}]}(U) = 0$ *.*

Come abbiamo visto, in generale la convergenza in probabilità è molto più debole della convergenza quasi certa, e quindi a maggior ragione di quela completa. E' comunque possibile trovare una parziale controimplicazione, come nel Lemma che segue.

Lemma 41 *Sia* $\{X_n\}_{n\in\mathbb{N}}$ *una successione di variabili aleatorie definita sullo spazio di probabilità* $\{\Omega, \Im, \mathbb{P}\}$*; sia inoltre* X*,* $X : \Omega \to \mathbb{R}$ *una variabile aleatoria definita sullo stesso spazio di probabilità, e si abbia la convergenza in probabilità* $X_n \to_p X$*. Allora esiste una sottosuccessione* X_{n_k} *tale che* $X_{n_k} \to_{c.c.} X$*.*

Proof. Per la dimostrazione, è sufficiente scegliere la sottosuccessione X_{n_k} tale per cui

$$\Pr\left\{|X_{n_k} - X| > \frac{1}{k}\right\} \leq \frac{1}{2^k} \ ,$$

in modo che si abbia, per ogni $\varepsilon > 0$

$$\begin{aligned} & \sum_{n_k=1}^{\infty} \Pr\{|X_{n_k} - X| > \varepsilon\} \\ \leq & \sum_{n_k=1}^{\lceil \frac{1}{\varepsilon} \rceil} \Pr\{|X_{n_k} - X| > \varepsilon\} + \sum_{n_k=\lceil \frac{1}{\varepsilon} \rceil + 1}^{\infty} \Pr\left\{|X_{n_k} - X| > \frac{1}{k}\right\} \\ \leq & \left\lceil \frac{1}{\varepsilon} \right\rceil + \sum_{n_k=1}^{\infty} \frac{1}{2^k} < \infty \ . \end{aligned}$$

■

Remark 42 *Si può verificare che non esiste implicazione, né in un senso né nell'altro, tra la convergenza completa e la convergenza in media r-esima. Si consideri infatti la sequenza:*

$$X_n = \begin{cases} 0 \ \textit{con probabilità}\ 1 - \frac{1}{n^2} \\ n^2 \ \textit{con probabilità}\ \frac{1}{n^2} \end{cases} \ .$$

Le X_n *convergono completamente (e quindi quasi certamente) alla variabile aleatoria che vale identicamente zero, ma non convergono nemmeno in media prima; infatti* $\mathbb{E}[X_n] = 1$ *per ogni* n*.*

Abbiamo quindi stabilito le implicazioni

$$\begin{aligned} (5) \quad &\Rightarrow \quad (4) \Rightarrow (2) \Rightarrow (1) \ , \\ (3) \quad &\Rightarrow \quad (2) \Rightarrow (1) \ ; \end{aligned}$$

concludiamo mostrando una implicazione tra (1) e (4).

Proposition 43 *(Skorohod) Sia* $\{X_n\}_{n\in\mathbb{N}}$ *una successione di variabili aleatorie definita sullo spazio di probabilità tale per cui* $X_n \to_d X$. *Allora esiste uno spazio di probabilità* $\{\Omega, \Im, \mathbb{P}\}$ *su cui sono definite variabili aleatorie* X'_n, X' *tali per cui* $X_n \stackrel{d}{=} X'_n$, $X \stackrel{d}{=} X'$, *e*

$$\mathbb{P}\left\{\lim_{n\to\infty} X'_n = X\right\} = 1 \ .$$

Proof. Per semplicità consideriamo il caso in cui le funzioni di distribuzione $F_{X_n}(.), F_X(.)$ siano crescenti e continue. Prendiamo $\{\Omega, \Im, \mathbb{P}\} = \{[0,1], \mathbb{B}[0,1], Leb\}$ con la variabile aleatoria identità $U(\omega) = \omega$, cioè l'uniforme in $[0,1]$. Definiamo

$$X'_n = F^{-1}_{X_n}(U) \ , \ X' = F^{-1}_X(U) \ ;$$

queste inverse sono ben poste per le ipotesi sulla funzione di ripartizione ed il risultato segue immediatamente perchè

$$\Pr(X'_n \le x) = \Pr(F^{-1}_{X_n}(U) \le x) = \Pr(U \le F_{X_n}(x)) = F_{X_n}(x) \ ,$$

e similmente per X'. La convergenza quasi certa è una conseguenza della convergenza (quasi) ovunque delle funzioni di distribuzione e delle loro inverse. ■

3 Stochastic Orders of Magnitude

In molti casi può essere utile avere alcuni strumenti per studiare il comportamento di variabili aleatorie che non convergono in nessuno dei sensi discusso nel capitolo precedente, ma nonstante questo rimagono limitate, dopo essere state opportunamente normalizzate. A questo fine, introduciamo la seguente

Definition 44 *(Stochastic Orders of Magnitude: $O_p(.)$ e $o_p(.)$) Sia $\{X_n\}_{n\in\mathbb{N}}$ una successione di variabili aleatorie e sia $\{f_n\}_{n\in\mathbb{N}}$ una successione deterministica, $f_n > 0$ per ogni n. Diremo che $X_n = O_p(f_n)$ se e solo se*

$$\forall \varepsilon > 0 \ , \ \exists K \in \mathbb{R} \ , \ \exists n_0 \in \mathbb{N} \ : \ \forall n > n_0 \ \Pr\left\{ \frac{|X_n|}{f_n} \leq K \right\} \geq 1 - \varepsilon \ .$$

Diremo che $X_n = o_p(f_n)$ se e solo se

$$\forall \varepsilon > 0 \ , \ \forall \delta > 0 \ , \ \exists n_0 \in \mathbb{N} \ : \ \forall n > n_0 \ \Pr\left\{ \frac{|X_n|}{f_n} \leq \delta \right\} \geq 1 - \varepsilon \ .$$

Remark 45 *Dalla definizione, segue immediatamente che $\{X_n = o_p(f_n)\} \Longleftrightarrow \left\{ \frac{X_n}{f_n} \rightarrow_p 0 \right\}$.*

Remark 46 *Nella definizione di $X_n = O_p(f_n)$ avremmo potuto rimpiazzare la condizione $\exists n_0 \in \mathbb{N} \ : \ \forall n > n_0$ con la condizione (apparentemente più generale) $\forall n \in \mathbb{N}$. Infatti se la sequenza $\frac{|X_n|}{f_n}$ è limitata per n sufficientemente grande è sempre possibile scegliere una costante K' (magari più grande di K) per la quale l'affermazione $\Pr\left\{ \frac{|X_n|}{f_n} \leq K' \right\} \geq 1 - \varepsilon$ sia soddisfatta con $n = 1, ..., n_0$.*

Remark 47 *E' evidente che $\{X_n = o_p(f_n)\} \Rightarrow \{X_n = O_p(f_n)\}$. Sul piano discorsivo, $O_p(.)$ implica che la sequenza normalizzata X_n/f_n appartiene ad un compatto; se la sequenza X_n/f_n converge è zero, appartiene ad un compatto arbitrariamente piccolo intorno all'origine, per n grande abbastanza.*

Proposition 48 *(Algebra degli stochastic orders) Siano $\{f_n, g_n\}$ due sequenze deterministiche strettamente positive.*

1) Abbiamo che

$$\{X_n = O_p(f_n) \ , \ Y_n = O_p(g_n)\} \Rightarrow \{X_n + Y_n = O_p(\max(f_n, g_n))\}$$

$$\{X_n = O_p(f_n) \ , \ Y_n = O_p(g_n)\} \Rightarrow \{X_n \times Y_n = O_p(f_n \times g_n)\}$$

2) Inoltre

$$\{X_n = o_p(f_n) \ , \ Y_n = o_p(g_n)\} \Rightarrow \{X_n + Y_n = o_p(\max(f_n, g_n))\}$$

$$\{X_n = o_p(f_n) \ , \ Y_n = o_p(g_n)\} \Rightarrow \{X_n \times Y_n = o_p(f_n \times g_n)\}$$

3) Infine

$$\{X_n = O_p(f_n) \ , \ Y_n = o_p(g_n)\} \Rightarrow \{X_n \times Y_n = o_p(f_n \times g_n)\}$$

Proof. Per la dimostrazione del punto 1), fissato $\varepsilon > 0$ prendiamo due costanti C_1, C_2 sufficientemente grandi tali per cui

$$\Pr\left\{\frac{|X_n|}{f_n} > C_1\right\} < \frac{\varepsilon}{2}\ ,\ \Pr\left\{\frac{|Y_n|}{g_n} > C_1\right\} < \frac{\varepsilon}{2}\ ;$$

l'esistenza di tali costanti è garantita dall'ipotesi $X_n = O_p(f_n)$, $Y_n = O_p(g_n)$. Abbiamo evidentemente

$$\begin{aligned} & \Pr\{|X_n + Y_n| > (C_1 \vee C_2)(f_n \vee g_n)\} \\ \leq\ & \Pr\{|X_n| > C_1 f_n\} + \Pr\{|Y_n| > C_2 g_n\} = \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon\ , \end{aligned}$$

dove la prima disuguaglianza segue dal fatto che l'evento $\{|X_n + Y_n| > (C_1 \vee C_2)(f_n \vee g_n)\}$ implica necessariamente il verificarsi di almeno uno dei due eventi $\{|X_n| > C_1 f_n\}$ o $\{|Y_n| > C_2 g_n\}$. In maniera analoga possiamo notare che

$$\begin{aligned} & \Pr\{|X_n \times Y_n| > (C_1 \times C_2)(f_n \times g_n)\} \\ \leq\ & \Pr\{|X_n| > C_1 f_n\} + \Pr\{|Y_n| > C_2 g_n\} = \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon\ , \end{aligned}$$

di nuovo perché l'evento $\{|X_n \times Y_n| > (C_1 \times C_2)(f_n \times g_n)\}$ implica necessariamente il verificarsi di almeno uno dei due eventi $\{|X_n| > C_1 f_n\}$ o $\{|Y_n| > C_2 g_n\}$.

La dimostrazione del punto 2) è del tutto analoga.

Per il punto 3), per ipotesi abbiamo $X_n = O_p(f_n)$, $Y_n = o_p(g_n)$; notiamo che si ha, per qualsiasi costante $C_1 > 0$

$$\begin{aligned} & \Pr\{|X_n \times Y_n| > C_1(f_n \times g_n)\} \\ =\ & \Pr\left\{|X_n \times Y_n| > (\frac{C_1}{C_2} \times C_2)(f_n \times g_n)\right\} \\ \leq\ & \Pr\{|X_{nn}| > C_2 f_n\} + \Pr\left\{|Y_n| > \frac{C_1}{C_2} g_n\right\}. \end{aligned}$$

Per ipotesi, possiamo scegliere C_2 tale per cui $\Pr\{|X_{nn}| > C_2 f_n\} \leq \frac{\varepsilon}{2}$; sempre per ipotesi, per $\frac{C_1}{C_2}$ fissato esiste senz'altro n_0 tale che

$$\Pr\left\{|Y_n| > \frac{C_1}{C_2} g_n\right\} \leq \frac{\varepsilon}{2} \text{ per ogni } n > n_0\ ,$$

ed il risultato è dimostrato.

E' naturale domandarsi come si inseriscano gli stochastic orders of magnitude nella gerarchia di modalità di convergenza che abbiamo discusso nella sezione precedente. Informalmente, possiamo considerare una sequenza $X_n = O_p(1)$ come se fosse nella categoria "0", nel senso che tale proprietà è implicata da tutte le modalità di convergenza precedentemente introdotte:

$$\begin{aligned} (5) \quad &\Rightarrow \quad (4) \Rightarrow (2) \Rightarrow (1) \Rightarrow O_p(1)\ , \\ (3) \quad &\Rightarrow \quad (2) \Rightarrow (1) \Rightarrow O_p(1)\ ; \end{aligned}$$

più rigorosamente, possiamo dimostrare il risultato qui di seguito. ■

Lemma 49 *Sia X_n una sequenza di variabili aleatorie che converge in distribuzione ad una variabile limite X. Allora $X_n = O_p(1)$.*

Proof. Notiamo prima che per ogni $\varepsilon >$ esiste sicuramente un compatto $K = K_\varepsilon \subset \mathbb{R}$ tale per cui $\Pr\{X \in K_\varepsilon\} > 1 - \varepsilon$. E' sufficiente infatti prendere $K_\varepsilon = [-M_\varepsilon, M_\varepsilon]$, dove $M_\varepsilon > 0 \in \mathbb{R}$ è tale per cui $F_X(-M_\varepsilon) < \frac{\varepsilon}{2}$ e $F_X(M_\varepsilon) > 1 - \frac{\varepsilon}{2}$; questo M_ε esiste senz'altro perché tutte le funzioni di distribuzione devono soddisfare $\lim_{x\to\infty} F_x(-x) = 0$, $\lim_{x\to-\infty} F_x(x) = 1$ (questa proprietà è talvolta chiamata la regolarità delle misure di probabilità). Preso ora $\varepsilon > 0$, scegliamo $M_{\varepsilon/2}$ punto di continuità di F_X tale per cui

$$\Pr\left\{X \in [-M_{\varepsilon/2}, M_{\varepsilon/2}]\right\} > 1 - \frac{\varepsilon}{2} \ ;$$

si scelga ora n_0 tale per cui $|F_{X_n}(-M_{\varepsilon/2}) - F_X(-M_{\varepsilon/2})|, |F_{X_n}(M_{\varepsilon/2}) - F_X(M_{\varepsilon/2})| \leq \frac{\varepsilon}{4}$ per tutti gli $n > n_0$; allora per un tale n abbiamo che

$$\begin{aligned} & \Pr\left\{X_n \in [-M_{\varepsilon/2}, M_{\varepsilon/2}]\right\} \\ = \ & F_{X_n}(-M_{\varepsilon/2}) + 1 - F_{X_n}(M_{\varepsilon/2}) \\ \leq \ & |F_{X_n}(-M_{\varepsilon/2}) - F_X(-M_{\varepsilon/2})| + F_X(-M_{\varepsilon/2}) \\ & +1 - F_X(M_{\varepsilon/2}) + |F_X(M_{\varepsilon/2}) - F_{X_n}(M_{\varepsilon/2})| \\ \leq \ & \frac{\varepsilon}{2} + \frac{\varepsilon}{4} + \frac{\varepsilon}{4} = \varepsilon \ , \end{aligned}$$

da cui segue immediatamente che $X_n = O_p(1)$, come richiesto. ■

Example 50 *E' interessante produrre alcuni esempi in cui la condizione $X_n = O_p(1)$; tre particolarmente semplici sono le sequenze*

$$X_n = n \ , \ X_n \stackrel{d}{=} N(0, n) \ , \ X_n \stackrel{d}{=} U(-n, n) \ .$$

Nel corso di Analisi 1 si è visto, come importante corollario del Teorema di Bolzano-Weierstrass, che ogni successione limitata ammette sempre una sottosuccessione (estratta) convergente. La nozione di sequenza $O_p(1)$ si può interpretare come una forma di equilimitatezza stocastica (detta tipicamente *tightness*) nella letteratura internazionale; è naturale domandarsi se possa esistere una proprietà simile per successioni di variabili aleatorie. La risposta (affermativa) è fornita dal seguente Teorema.

Theorem 51 *(Prohorov 1956) Ogni successione $X_n = O_p(1)$ ammette una estratta che converge in distribuzione.*

Remark 52 *Il teorema di Prohorov è in realtà molto più generale di questo, perché vale per qualsiasi successione tight di elementi aleatori a valori su uno spazio completo, metrico e separabile.*

Proof. Dal Teorema di Helly-Bray (si veda ad esempio....) sappiamo che ogni successione di funzioni di distribuzione ammette una estratta convergente ad

una funzione di distribuzione possibilmente difettiva, cioè una funzione F^* che sia non decrescente e continua da destra, ma per la quale non valga necessariamente $\lim_{x\to\infty} F^*(-x) = 0$ e $\lim_{x\to\infty} F^*(x) = 1$. Sia F_{n_k} questa estratta, e dimostriamo che se la corrispondente sequenza di variabile aleatorie è tight necessariamente la funzione limite F^* deve indurre una misura di massa 1 (cioè deve valere $\lim_{x\to\infty}(F^*(x) - F^*(-x)) = 1$). Poiché F^* è non decrescente, il limite $\lim_{x\to\infty} F^*(x) = \eta$ sicuramente esiste; supponiamo per assurdo che sia strettamente minore di 1. Allora per ogni insieme compatto $K \subset \mathbb{R}$ si avrebbe necessariamente $\Pr\{X_n \in K\} \leq \Pr\{X_n \leq \max(K)\} \leq \eta < 1$, e la sequenza non potrebbe essere $O_p(1)$. ■

Il teorema di Levy ci garantisce che una successione di funzioni di ripartizione individua una variabile aleatoria limite (in distribuzione) se e solo se le corrispondenti funzioni caratteristiche convergono ad una funzione limite continua all'origine. Questa condizione può essere interpretata come una condizione di tightness; ad esempio, per la sequenza di Gaussiane $N(0, n)$ le corrispondenti funzioni caratteristiche sono date da $\phi_n(t) = \exp(-n\frac{t^2}{2})$ che convergono ad una funzione limite $\phi_\infty(t) = I_{\{t=0\}}(t)$, evidentemente discontinua all'origine.

4 Uniforme Integrabilità

Abbiamo visto in precedenza che la convergenza in probablità può implicare la convergenza in media r-esima per sequenze di variabili aleatorie uniformemente limitate, perché la limitatezza impedisce che con probabilità anche sempre più piccola si assumano valori sempre più grandi. E' naturale domandarsi se questa richiesta di uniforme limitatezza non sia eccessiva; dovrebbe essere possibile "compensare" il comportamento sulle code in modo da far sì che la massa corrispondente decada comunque a zero. Questa è proprio l'intuizione dietro alla nozione di Uniforme Integrabilità.

Definition 53 *La sequenza di variabili aleatorie $\{X_n\}$ è detta Uniformemente Integrabile se*

$$\lim_{M\to\infty} \sup_n E[|X_n| I_{\{|X_n|>M\}} = 0 \ .$$

Remark 54 *Notiamo innanzitutto che il fatto che i momenti siano uniformememente limitati non è sufficiente a garantire l'uniforme integrabilità: basta infatti considerare l'esempio visto precedentemente, con :*

$$X_n = \begin{cases} 0 \text{ con probabilità } 1 - \frac{1}{n^2} \\ n^2 \text{ con probabilità } \frac{1}{n^2} \end{cases} \ .$$

Qui $E[X_n] = 1$ per ogni n, quindi il valor medio è unifomemente limitato, però

$$\lim_{M\to\infty} \sup_n E[|X_n| I_{\{|X_n|>M\}}] = 1 \ ,$$

quindi la sequenza non è uniformemente integrabile. D'altra parte si vede subito che la limitatezza dei momenti è una condizione necessaria.

Lemma 55 *(Condizione necessaria per uniforme integrabilità). Se $\{X_n\}$ è uniformemente integrabile, $E[X_n]$ è uniformemente limitato.*

Proof. Si scelga $M = M_\varepsilon$ tale per cui

$$\sup_n E[|X_n| I_{\{|X_n|>M\}}] < \varepsilon \ ;$$

allora $E[|X_n|] = E[|X_n| I_{\{|X_n|\le M_\varepsilon\}}] + E[|X_n| I_{\{|X_n|>M_\varepsilon\}}] \le M_\varepsilon + \varepsilon$, da cui segue che il valor medio è limitato uniformemente in n. ■

E' interessante enunciare due condizioni sufficienti per l'Uniforme Integrabilità.

Lemma 56 *(Condizioni sufficienti per l'uniforme integrabilità)*

a) Se la sequenza $\{X_n\}$ è costituita da variabili aleatorie identicamente distribuite con valor medio finito, essa è uniformemememente integrabile

b) Se esiste $\eta > 0$ tale per cui $E[|X_n|^{1+\eta}] < M_\eta < \infty$ per ogni n, allora la sequenza è uniformemente integrabile

Proof. Per a), è sufficiente notare che

$$\lim_{M\to\infty} \sup_n E[|X_n| I_{\{|X_n|>M\}}] = \lim_{M\to\infty} E[|X_1| I_{\{|X_1|>M\}}] = 0 \ ,$$

per l'ipotesi del momento primo finito. Per la b), abbiamo che

$$\begin{aligned} \lim_{M\to\infty} \sup_n E[|X_n| I_{\{|X_n|>M\}}] &\leq \lim_{M\to\infty} \sup_n E[\frac{|X_n|^\eta}{M^\eta} |X_n| I_{\{|X_n|>M\}}] \\ &= \lim_{M\to\infty} \frac{1}{M^\eta} \sup_n E[|X_n|^{1+\eta} I_{\{|X_n|>M\}}] \\ &\leq \lim_{M\to\infty} \frac{M_\eta}{M^\eta} = 0 \ . \end{aligned}$$

■

Lemma 57 *Sia* $\{X_n\}$ *una sequenza uniformemente integrabile. Allora* $X_n \to_p X$ *implica* $X_n \to_1 X$.

Proof. Basta scrivere

$$E[|X_n - X|] \leq E[|X_n - X_n \wedge M|] + E[|X - X \wedge M|] + E[|X_n \wedge M - X \wedge M|] \ .$$

Il primo elemento può essere reso arbitrariamente piccolo per l'uniforme integrabilità, il secondo perché X ha necessariamente valor medio finito (dimostrarlo), il terzo perché abbiamo a che fare con variabili limitate che convergono in probabilità, quindi possiamos fruttare il risultato dimostrato precedentemente.

■

5 Teoremi Limite

Theorem 58 *(Legge dei grandi numeri per sequenze uniformemente integrabili). Sia $\{X_n\}$ una sequenza di variabili aleatorie con valor medio nullo, indipendenti e uniformemente integrabili. Allora $\overline{X}_n = n^{-1}\sum_{i=1}^{n} X_i \to_1 0$.*

Proof. Notiamo innanzitutto che non abbiamo ipotizzato che queste variabili abbiano momento secondo finito e limitato, altrimenti il risultato seguirebbe banalmente dalla legge dei grandi numeri di Chebyshev. Le ipotesi del Lemma sono verificate, ad esempio, se le variabili sono identicamente distribuite. Possiamo scrivere

$$\begin{aligned} X_n &= X_n I_{\{|X_N|\leq M\}} + X_n I_{\{|X_N|>M\}} \\ &= X'_n + X''_n = X'_n - E[X'_n] + X''_n - E[X''_n] \ , \end{aligned}$$

perché $E[X'_n] + E[X''_n] = 0$. Abbiamo

$$\frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\sum_{i=1}^{n}\{X'_i - E[X'_i]\} + \frac{1}{n}\sum_{i=1}^{n}\{X''_i - E[X''_i]\} \ .$$

La prima sequenza aleatoria è la media aritmetica di variabili aleatore con valor medio nullo e varianza uniformemente limitata da M^2, quindi va a zero in media quadratica per la legge dei grandi numeri di Chebyshev. Per la seconda parte possiamo scrivere

$$\begin{aligned} E[|\frac{1}{n}\sum_{i=1}^{n}\{X''_i - E[X''_i]\}|] &\leq 2\sup_{i=1,2,3,\ldots} E[|X_n I_{\{|X_N|>M\}}|] \\ &\leq \varepsilon \ , \end{aligned}$$

per l'uniforme integrabilità, scegliendo M in modo appropriato. ■

Theorem 59 *(Legge forte dei grandi numeri - Borel) Sia $\{X_n\}_{n\in\mathbb{N}}$ una successione di variabili aleatorie indipendenti ed identicamente distribuite con valor medio $\mathbb{E}[X] = \mu$ e momento quarto finito $\mathbb{E}\left[X^4\right] = \mu_4$. Allora*

$$\overline{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i \to_{c.c.} \mu \ .$$

Proof. Senza perdita di generalità consideriamo $\mu = 0$; infatti

$$\left\{\frac{1}{n}\sum_{i=1}^{n} X_i \to_{c.c.} \mu\right\} \Longleftrightarrow \left\{\frac{1}{n}\sum_{i=1}^{n} X'_i \to_{c.c.} 0 \ , \ X'_i := X_i - \mu\right\} \ .$$

Notiamo innanzitutto che

$$\mathbb{E}\left[\overline{X}_n^4\right] = O(\frac{1}{n^2}) \ ;$$

infatti

$$\begin{aligned}
\mathbb{E}\left[\overline{X}_n^4\right] &= \frac{1}{n^4} \sum_{i_1,i_2,i_3,i_4=1}^{n} \mathbb{E}\left[X_{i_1} X_{i_2} X_{i_3} X_{i_4}\right] \\
&= \frac{1}{n^4} \sum_{i=1}^{n} \mathbb{E}\left[X_i^4\right] + \frac{3}{n^4} \sum_{\substack{i_1,i_2=1 \\ i_i \neq i_2}}^{n} \mathbb{E}\left[X_{i_1}^2\right] \mathbb{E}\left[X_{i_2}^2\right] \\
&= \frac{n}{n^4} \mathbb{E}\left[X_1^4\right] + \frac{3n(n-1)}{n^4} \mathbb{E}\left[X_1^2\right] \mathbb{E}\left[X_2^2\right] \\
&= O(n^2) \ .
\end{aligned}$$

La prima uguaglianza è dovuta al fatto che $\mathbb{E}\left[X_{i_1} X_{i_2} X_{i_3} X_{i_4}\right] = 0$ se c'è almeno un indice diverso da tutti gli altri; basta quindi focalizzarsi sul caso in cui gli indici siano tutti uguali, oppure uguali a coppie (diverse tra loro). Ci sono n termini del primo tipo, $3n(n-1)$ del secondo. A questo punto, utilizzando la disuguaglianza di Markov, per ogni $\varepsilon > 0$ si ha

$$\sum_{n=1}^{\infty} \Pr\left\{\overline{X}_n > \varepsilon\right\} \leq Const \times \sum_{n=1}^{\infty} \frac{1}{n^2 \varepsilon^4} < \infty \ ,$$

e la disuguaglianza è dimostrata. ■

Remark 60 *E' interessante notare come questa legge forte dei grandi numeri sia stata enunciata per la prima volta all'inizio del novecento da Emile Borel all'interno di problemi di Teoria dei Numeri. In particolare, Borel la utilizzo per dimostrare che i numeri tra* $[0, 1]$*, espressi in forma binaria, hanno quasi sempre una proporzione uguale di* 0 *e di* 1*. Questo esclude automaticamente tutti i numeri con espansione finita o periodica, cioè i razionali; ma in realtà dimostra che hanno misura nulla anche numeri di classi più ampie. Paradossalmente, per pochissimi numeri conosciuti è stato effettivamente dimostrato che hanno questa struttura "normale".*

6 Il Teorema del Limite Centrale

6.1 Cenni storici

La storia del teorema del limite centrale è forse una tra le più interessanti in ambito matematico, o almeno in ambito probabilistico. Come noto, il calcolo delle probabilità nasce essenzialmente nel 1654, con il carteggio tra Pascal e Fermat in merito ad alcuni semplici problemi di gioco d'azzardo e calcolo combinatorio. Dal loro lavoro e da quelli successivi, in aprticolare ad opera dei Bernoulli, si arriva alla formulazione della legge binomiale: ripetendo in modo indipendente n esperimenti in ciascuno dei quali si campioni una Bernoulliana con probabilità di successo pari a p, la probabilità di avere k successi (con $0 \leq k \leq n$) è data da $\Pr\{X_n = k\} = \binom{n}{k} p^k (1-p)^k$. Mentre questo risolve completamente il problema dal punto di vista teorico, dal punto di vista pratico la formula ottenuta è del tutto inapplicabile quando il valore di n diventa nell'ordine di 100 o superiore, anche utilizzando gli attuali supercomputer - nel 1700 il calcolo avrebbe dovuto essere effettuato a mano e sarebbe stato verosimilmente infattibile anche per valori di n molto bassi.

Il primo a cercare di ottenere una approssimazione numerica per queste probabilità fi Abraham De Moivre, che tra il 1733 ed il 1738 si rese conto che i valori della legge binomiale si potevano approssimare con la funzione

$$\Pr\{X_n = k\} = const \times \exp(-\frac{1}{2(np(1-p))}(k-np)^2) \ ;$$

il calcolo delle constante di normalizzazione fu ottenuto circa 50 anni da Laplce, che fu quindi il primo a scrivere la densità nella funzione Gaussiana nella forma che oggi conosciamo,

$$\phi_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}(\frac{x-\mu}{\sigma})\right\} \ , \ \mu = np \ , \ \sigma^2 = np(1-p) \ .$$

L'accuratezza di questa approssimazione può essere verificata con alcuni semplici giochi da tavoli (la cosiddetta Galton Board).

A questo punto la densità Gaussiana sembrava poco più di un utile strumento di approssimazione numerica; iniziò a occuparsene però Gauss, nell'ambito del suo lavoro come astronomo reale all'Osservatorio di Berlino introno al 1810. In particolare, Gauss si trovò di fronte al problema di dover confrontare le traiettorie dei pianeti previste dalla meccanica Newtoniana con le osservazioni da lui raccolte; esistevano naturalmente alcune discrepanze dovute principalmente ad errori di misura, e Gauss pensò di diminuirne l'effetto prendendo semplicemente la media aritmetica di un maggior numero di osservazioni successive. Si pose quindi una domanda - quale legge dovrebbero seguire gli errori di misurazione affinché prendere la media aritmetica sia la scelta ottimale? Per rispondere a questa domanda, bisogna prima specificare cosa si intenda per "ottimale" - Gauss scelse un criterio che fu officialmente codificato più di 100 anni dopo di lui da Ronald Fisher, la cosiddetta Massima Verosimiglianza, tuttora alla base di quasi tutta la teoria degli stimatori. L'idea della massima verosimiglianza

è semplice: si prende come valore "stimato" del parametro quello che rende massimilmente probabile osserva quello che ho effettivamente osservato. Ad esempio, supponiamo di avere di fronte una scatola che contiene palline bianche e nere; non sappiamo in quale proporzione però sappiamo che se la scatola è di tipo "1" ci saranno 90 palline bianche ed 10 nere, se la scatola è di tipo "2" ci saranno 10 bianche e 90 nere. Effettuiamo una estrazione, che produce una pallina bianca - come stima del parametro della scatola ovviamente prendiamo il valore 1, perché rende molto più probabile osservare quello che abbiamo effettivamente osservato.

Proseguendo con questa ragionamento, Gauss suppone che gli errori seguano una certa legge ignota (oggi diremmo densità di probabilità) che possiamo chiamare $f(x)$; avendo effetuato n misurazioni indipendenti, conclude che quella che oggi chiameremmo la loro legge congiunta deve avere forma

$$f(X_1)f(X_2)...f(X_n) \ .$$

E' immediato verificare che la media aritmetica minimizza in c la funzione $-\sum_{i=1}^{n}(X_i - c)^2$; quindi la media aritmetica costituisce uno stimatore ottimale quando si ha

$$f(X_1)f(X_2)...f(X_n) = const \times \exp\left\{-const \times \sum_{i=1}^{n}(X_i - c)^2\right\},$$

cioè esattamente la densità congiunta di n Gaussiane indipendenti, a meno di costanti. Non sappiamo se Gauss abbia anche verificato empiricamente il fatto che gli errori seguissero questa legge, nel qual caso sarà rimasto senz'altro sbalordito: come se la Natura stesse cercando segretamente di aiutarlo, gli errori di osservazione seguono proprio quella legge che rende ottimale prendere come stimatore la media aritmetica, nonostante avesse scelto questa procedura senza sapere a priori che fosse ottimale. In ogni caso, il fatto che la stessa legge introdotta da De Moivre esclusivamente per motivazioni numeriche avesse anche questa proprietà importantissima di ottimalità per la media numerica, oltre ad essere quella effettivamente seguita empiricamente dagli errori di osservazione, deve aver avuto sicuramente un effetto notevole su Gauss ed i suoi successori (Galton diceva che se l'avessero scoperto i Greci avrebbero aggiunto la legge Gaussiana all'Olimpo dei loro dei).

Il passo successivo si ha con Maxwell nel 1860, quando si scopre che la Gaussiana determina la legge delle velocità delle molecole in gas perfetti. Da quel momento la Gaussiana comincia ad apparire in una infinità di campi diversi, dalla biologia alla termodinamica, dalla finanza alla medicina. Ad esempio, Einstein nel 1905 lega l'equazione termodinamica della diffusione del calore ad un modello probabilistico di diffusione di particelle; nello stesso anno Bachelier crea di fatto la finanza matematica modellando con una Gaussiana le fluttuazione dei rendimenti dei titoli di borsa. Le applicazioni della Gaussiana nel corso del ventesimo e ventunesimo secolo sono troppe per poter essere elencate: sul sito arxiv.org risultano circa 450 articoli nel 2022 che hanno il termine "Central

Limit theorem" nell'abstract e più di 6500 che hanno il termine "Gaussian". Queste applicazioni sono centrali a tutte le aree della matematica; qui sotto alcuni esempi un po' inaspettati.

Example 61 *(Teoria dei Numeri) E' ben noto che ogni numero intero $n \in \mathbb{N}$ si fattorizza in modo univoco nella somma dei suoi fattori primi, $n = p_1 \times p_2 \times ... \times p_k$. Quanti sono questi fattori per un numero scelto "a caso", nel limite in cui $n \to \infty$? si può dimostrare che vale un Teorema del Limite Centrale, ed in particolare scrivendo $\omega(J_n)$ per il numero di fattori primi di J_n, con J_n uniformemente distribuita tra 1 e n, si ha*

$$\frac{\omega(J_n) - \log\log n}{\sqrt{\log\log n}} \to_d N(0,1) \ ,$$

si veda ad esempio P. Erd¨os and M. Kac. The Gaussian law of errors in the theory of additive number theoretic functions., Amer. J. Math., 62:738–742, 1940, Chen, Louis H. Y.; Jaramillo, Arturo; Yang, Xiaochuan A generalized Kubilius-Barban-Vinogradov bound for prime multiplicities. ALEA Lat. Am. J. Probab. Math. Stat. 20 (2023), no. 1, 713–730.

Example 62 *(Geometria in alta dimensione) Scegliamo un punto a caso $(x_1, ..., x_d)$ su una sfera unitaria di dimensione d,*

S^d; qui per "a caso" intendiamo in maniera uniforme, cioè tutte le regioni con la stessa area hanno la stessa probabilità di essere estratte. Fissiamo un intero k, e consideriamo le primi k coordinate di queste punto; allora per una opportuna sequenza di normalizzazione $\sigma(d)$ si ha

$$\frac{1}{\sigma(d)}(x_1, ..., x_k) \to_d (Z_1, ..., Z_k) \ ,$$

dove le Z_i sono variabili aleatorie Gaussiane standard indipendenti tra loro. In particolare, qualsiasi coordinate tende (dopo una rinormalizzazione per l'opportuno fattore di scala) ad una Gaussiana standard. (Freedman e Diaconis, Poincaré...)

Example 63 *(Polinomi trigonometrici) Fissiamo un insieme di pesi a_k, b_k $k = 1, 2, ...n$; per fissare le idee potremmo anche prendere tutti questi pesi uguali a 1. Consideriamo ora il polinomio trigonometrico (deterministico!)*

$$p_n(x) = \sum_{k=1}^{n} a_k \cos(kx) + b_k \sin(kx) \ , \ x \in [-\pi, \pi] \ .$$

Dopo una opportuna normalizzazione $\sigma(n)$, per quasi tutte (nel senso di con probabilità 1, se si vedono questi pesi come variabili aleatorie) le scelte dei pesi (a_k, b_k) si ha

$$\frac{1}{2\pi} Misura \left\{ x : c_1 \leq \frac{p_n(x)}{\sigma(n)} \leq c_2 \right\} \to_{n\to\infty} \int_{c_1}^{c_2} \frac{\exp(-x^2/2)}{\sqrt{2\pi}} dx = \Pr\{Z \in [c_1, c_2]\} \ ,$$

cioè in altre parole i valori (deterministici) presi dal polinomio trigonometrico si distribuiscono come una Gaussiana standard: facendo un plot di questi valori si otterrebbe la classica curva a campana. (Per alcune referenze relative a questo risultato, si veda ad esempio il classico articolo R. Salem and A. Zygmund. "Some properties of trigonometric series whose terms have random signs". In: Acta Math. 91 (1954), pp. 245–301, oppure i più recenti Jürgen Angst and Guillaume Poly "Variations on Salem–Zygmund results for random trigonometric polynomials: application to almost sure nodal asymptotics". In: Electronic Journal of Probability 26 (2021), pp. 1–36, e Louis Gass, Almost-sure asymptotics for Riemannian random waves. In Bernoulli 29, (2023), no.1, 625–651.) Un risultato analogo, ma forse più semplice da enunciare, è il seguente: sia n_k una sequenza di numeri interi tale per cui $1 < n_{k+1} - n_k = O(1)$; definiamo il polinomio trigonometrico deterministico

$$q_n(x) = \frac{\sqrt{2}}{\sqrt{n}} \sum_{k=1}^{n} \cos(n_k x) \ .$$

Allora quando $n \to \infty$ si ha

$$\frac{1}{2\pi} Misura \left\{ x : c_1 \leq q_n(x) \leq c_2 \right\} \to_{n\to\infty} \int_{c_1}^{c_2} \frac{\exp(-x^2/2)}{\sqrt{2\pi}} dx = \Pr\left\{ Z \in [c_1, c_2] \right\} \ .$$

Quindi i valori deterministici del polinomio $q_n(x)$ si distribuiscono seguendo una curva Gaussiana standard. Per questo risultato si veda ad esempio Katusi Fukuyama, A central limit theorem to trigonometric series with bounded gaps, Probab.Theory Related Fields, 149, no. 1-2, 139–148, (2011).

7 Il teorema del limite centrale

La formulazione più semplice del teorema limite centrale che si incontra in un primo corso di probabilità è la seguente.

Theorem 64 *Sia $X_1, ..., X_n$ una seuenza di variabili aleatorie indipendenti ed identicamente distribuite con valor medio μ e varianza (finita) σ^2. Allora abbiamo la convergenza in distribuzione*

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n\sigma^2}} \to_d Z \sim N(0,1) \ , \ per \ n \to \infty \ .$$

Proof. Ricordiamo innanzitutto la definizione di funzione caratteristica di una variabile aleatoria X :

$$\psi_X(t) := E[\exp(iXt)] \ .$$

Ricordiamo altri due fatti dai corsi elementari di probabilità: Il Teorema di continuità di Lévy ci garantisce che se la sequenza di funzioni di caratteristiche $\psi_{X_n}(t) = E[\exp(iX_n t)]$ converge a $\psi_X(t) = E[\exp(iXt)]$, necessariamente si ha

la convergenza in distribuzione $X_n \rightarrow_d X$ (e viceversa). Ovviamente nel nostro caso è sufficiente dimostrare che

$$\frac{\sum_{i=1}^n \widetilde{X}_i}{\sqrt{n}} \rightarrow_d Z \sim N(0,1) \text{ , per } n \rightarrow \infty \text{ , con } \widetilde{X}_i := \frac{X_i - \mu}{\sigma} \text{ .}$$

Abbiamo dunque, usando le proprietà standard della funzione caratteristica

$$\begin{aligned} \psi_{\frac{\sum_{i=1}^n \widetilde{X}_i}{\sqrt{n}}}(t) &= E[\exp(i \sum_{i=1}^n \widetilde{X}_i \frac{t}{\sqrt{n}})] \\ &= \psi_{\sum_{i=1}^n \widetilde{X}_i}(\frac{t}{\sqrt{n}}) \\ &= \psi_{\widetilde{X}_1}(\frac{t}{\sqrt{n}}) \times ... \times \psi_{\widetilde{X}_n}(\frac{t}{\sqrt{n}}) \\ &= \left\{ \psi_{\widetilde{X}_1}(\frac{t}{\sqrt{n}}) \right\}^n \text{ ,} \end{aligned}$$

dove nel secondo passaggio abbiamo usato il fatto che le variabili sono indipendenti e nel terzo il fatto che sono identicamente distribuite. Ricordiamo ora che

$$\psi_{\widetilde{X}_1}(t) = 1 + itE[\widetilde{X}_1] + \frac{(it)^2}{2} E[\widetilde{X}_1^2] + o((it)^2) \text{ ,}$$

da cui

$$\begin{aligned} \left\{ \psi_{\widetilde{X}_1}(\frac{t}{\sqrt{n}}) \right\}^n &= \left\{ 1 + i\frac{t}{\sqrt{n}} E[\widetilde{X}_1] + \frac{(it)^2}{2n} E[\widetilde{X}_1^2] + o((\frac{it}{\sqrt{n}})^2) \right\}^n \\ &= \left\{ 1 - \frac{t^2}{2n} + o(\frac{t^2}{n}) \right\}^n \rightarrow \exp(-\frac{t^2}{2}) \text{ .} \end{aligned}$$

La dimostrazione è conclusa ricordando che la Gaussiana standard ha proprio funzione caratteristica

$$\psi_Z(t) = \exp(-\frac{t^2}{2}) \text{ .}$$

■

Nell'ambito di questo corso, avremo bisogno di un Teorema del Limite Centrale di portata ben più generale. A tale fine, introduciamo la cosiddetta condizione di Lindeberg.

Condition 65 *(Lindeberg) Sia $\{X_i\}$ una successione di variabili inidpendenti (non necessariamente identicamente distribuite) definite su uno spazio di probabilità $\{\Omega, \Im, P\}$ e con momento secondo finito; scriviamo $\mu_i = E[X_i]$ e $\sigma_i^2 = Var[X_i]$. Questa sequenza soddisfa la condizione di Lindeberg se per ogni $\varepsilon > 0$ si ha*

$$\lim_{n \rightarrow \infty} \frac{1}{\sum_{i=1} \sigma_i^2} \sum_{i=1}^n E[X_i^2 I_{\left\{ X_i^2 > \varepsilon \sqrt{\sum_{i=1} \sigma_i^2} \right\}}] = 0 \text{ .}$$

Theorem 66 *Sia $\{X_i\}$ una successione di variabili aleatorie che soddisfa la condizione di Lindeberg. Allora*

$$\frac{\sum_{i=1}^n X_i - \sum_{i=1}^n \mu_i}{\sqrt{\sum_{i=1}^n \sigma_i^2}} \rightarrow_d Z \sim N(0,1) \ .$$

Remark 67 *La dimostrazione di questo risultato si basa su approssimazioni successive della funzione caratteristica e non è riportata per brevità. E' però interessante capire il significato della condizione di Lindeberg, che intuitivamente tende ad escludere due casi:*

a) quello in cui la somma delle varianza NON diverga, cioè $\sum_{i=1} \sigma_i^2$ vada ad una costante. In questo caso stiamo in pratica sommando solo un numero finito di variabili aleatorie e pertanto il TLC non può valere, a meno che le variabili non siano già in partenza Gaussiane

b) quello in cui la varianza nella coda delle ultime variabili aleatorie sommate sia dello stesso ordine di grandezza di tutte le altre; anche in questo caso il TLC non può valere perché è come se sommassimo due singole variabili, una identificata dalla somma delle prime $n-1$ e l'altra costituita dall'ultima.

Remark 68 *La condizione di Lindeberg è difficile da verificare in pratica e si usa più spesso una meno generale, ma di più facile verifica, la cosiddetta condizione di Lyapunov.*

Condition 69 *(Lyapunov) Sia $\{X_i\}$ una successione di variabili aleatori indipendenti ed a media nulla e definiamo*

$$\widetilde{X}_{i,n} = \frac{X_i}{\sqrt{\sum_{i=1}^n E[X_i^2]}} \ ,$$

in modo tale che $\sum_{i=1}^n E[\widetilde{X}_{i,n}^2] = 1$ per ogni n. La sequenza soddisfa la condizione di Lyapunov se

$$\lim_{n\rightarrow\infty} \max_{i=1,\ldots,n} E[|\widetilde{X}_{i,n}|^{2+\delta}] = 0 \text{ per qualche } \delta > 0 \ .$$

Remark 70 *Anche la condizione di Lyapunov può essere vista come la richiesta che non ci sia nessuna variabile aleatoria con varianza non trascurabile rispetto alla somma di tutte le altre.*

Remark 71 *E' importante discutere la relazione esistente tra teorema del limite centrale e legge dei garndi numeri. Consideriamo il caso di variabili indipendenti X_i, e normalizziamole in modo che abbiano valor medio nullo e varianza 1. Sappiamo che la somma di questa variabili aleatorie ha varianza $Var[\sum_{i=1}^n X_i] = n$, e pertanto $\sum_{i=1}^n X_i = O_p(\sqrt{n})$; in altre parole questa sequenza di somme se divisa per $\sqrt{n}$ vivrà su un compatto con probabilità che può essere resa arbitrariamente vicina ad uno. la legge dei grandi numeri ci dice che dividendo per n abbiamo*

$$\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i \rightarrow_p 0 \ ;$$

in un certo senso, il Teorema del Limite Centrale si può pensare come uno zoom sulle fluttuazioni che diventano infinitesimali della media intorno allo zero: riscalando per un fattore $\sqrt{n}$ otteniamo

$$\sqrt{n}\overline{X}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i \to_d N(0,1) \ .$$

Quindi le fluttuazioni nella somma di n variabili aleatorie indipendenti a valor medio nullo sono di ordine $\sqrt{n}$; le fluttuazioni della media aritmetica sono di ordine $\frac{1}{\sqrt{n}}$; le fluttuazioni nel teorema del limite centrale sono di ordine $O(1)$.

Example 72 *(L'uso del Teorema del limite centrale nei sondaggi di opinione.)...*

8 Il Metodo Delta

Il Continuous Mapping Theorem ci garantice che se $a_n(X_n - E[X_n])$ è una sequenza di variabili aleatorie che converge in distribuzione ad un limite Z e g è una funzione continua, allora $g(a_n(X_n - E[X_n]))$ converge in distribuzione a $G(Z)$; ad esempio, se $a_n(X_n - E[X_n])$ converge ad una Gaussiana standard e $g(x) = x^2$, allora $g(a_n(X_n - E[X_n]))$ converge in distribuzione ad una Gaussiana con un grado di libertà. Ci poniamo ora una domanda diversa; se applichiamo la trasformazione direttamente a X_n invece che alla sua versione standardizzata, cosa possiamo concludere sulla convergenza? In altre parole, cosa possiamo dire sulla convergenza in distribuzione di $a_n(g(X_n) - E[g(X_n)]$? A questa domanda risponde il cosiddetto metodo delta.

Lemma 73 *Sia G una funzione definita in un intorno dell'origine (da $\mathbb{R}^n$ in $\mathbb{R}^n$) e ivi continua e tale che $G(h) = o(\|h\|)$. Allora*

$$X_n \to_p 0 \Longrightarrow G(X_n) = o_p(\|X_n\|) \ , \ cioè \ \frac{G(X_n)}{\|X_n\|} = o_p(1) \ .$$

Proof. E' sufficiente definire la funzione continua

$$g(h) = \begin{cases} \frac{G(h)}{\|h\|} \ , \text{ se } \|h\| \neq 0 \\ 0 \ , \text{ se } \|h\| = 0 \end{cases} \ .$$

Se $X_n \to_p 0$, $g(X_n) \to_p 0$ per il Lemma di Slutzky. ■

Theorem 74 *(Metodo Delta) Sia $g : \mathbb{R}^k \to \mathbb{R}^m$ differenziabile, con matrice Jacobiana limitata in $\mu \in \mathbb{R}^n$. Supponiamo che la sequenza di vettori aleatori X_n sia tale per cui*

$$a_n(X_n - \mu) \to_d X \ ,$$

per a_n sequenza deterministica che diverge all'infinito. Allora

$$a_n(g(X_n) - g(\mu)) \to_d (J(g(\mu))X \ ,$$

dove il vettore X ha dimensioni $k \times 1$ e la matrice Jacobiana $(J(g(\mu))$ ha dimensioni $m \times k$,

$$J(g(\mu) = \begin{pmatrix} \frac{\partial g_1}{\partial x_1} & ... & \frac{\partial g_1}{\partial x_k} \\ ... & ... & ... \\ \frac{\partial g_m}{\partial x_1} & ... & \frac{\partial g_m}{\partial x_k} \end{pmatrix} \ .$$

Proof. Per il teorema di Taylor multivariato abbiamo che

$$g(\mu + y) - g(\mu) = (J(g(\mu))^T y + G(y) \ ,$$

con $G(y) = o(\|y\|)$. Prendendo $y = X_n - \mu$ abbiamo

$$g(X_n) - g(\mu) = (J(g(\mu))(X_n - \mu) + G(X_n - \mu) \ \ ,$$

e per il precedente Lemma sappiamo che

$$\begin{aligned} a_n G(X_n - \mu) &= a_n \|X_n - \mu\| \frac{G(X_n - \mu))}{\|X_n - \mu\|} \\ &= \|a_n(X_n - \mu)\| \frac{G(X_n - \mu))}{\|X_n - \mu\|} \\ &= O_p(1)o_p(1) = o_p(1) \ . \end{aligned}$$

Segue che

$$\begin{aligned} a_n(g(X_n) - g(\mu)) &= a_n(J(g(\mu))(X_n - \mu) + a_n G(X_n - \mu) \\ &= (J(g(\mu))a_n(X_n - \mu) + o_p(1) \rightarrow_d (J(g(\mu))X \ . \end{aligned}$$

■

Example 75 *Sia $X_1, ...X_n$ una sequenza di variabili IID, per le quali si ha*

$$\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \rightarrow_d Z \sim N(0,1) \ .$$

Allora

$$\sqrt{n}(e^{\overline{X}_n} - e^{\mu}) \rightarrow_d N(0, e^{2\mu}\sigma^2) \ .$$

Example 76 *Sia $X_1, ...X_n$ una sequenza di variabili IID, per le quali si ha*

$$\sqrt{n}\frac{\overline{X}_n - \mu}{\sigma} \rightarrow_d Z \sim N(0,1) \ .$$

Allora

$$\sqrt{n}(\overline{X}_n^2 - \mu^2) \rightarrow_d 2\mu Z \ .$$

Example 77 *Sia $X_1, ...X_n$ una sequenza di vettori aleatori IID, per i quali si ha*

$$\sqrt{n}(\overline{X}_n - \mu) = \sqrt{n}\left\{ \begin{pmatrix} \overline{X}_{1n} - \mu_1 \\ \overline{X}_{2n} - \mu_2 \end{pmatrix} \right\} \rightarrow_d \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} \sim N(0, \begin{pmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{pmatrix}) \ .$$

Allora

$$\sqrt{n}(\overline{X}_{1n}\overline{X}_{2n} - \mu_1\mu_2) \rightarrow_d \begin{pmatrix} \mu_2 & \mu_1 \end{pmatrix} \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = \mu_2 Z_1 + \mu_1 Z_2 \ .$$

9 Stimatori e loro proprietà

Definition 78 *Dediniamo campione aleatorio (random sample) un insieme di variabili aleatorie indipendenti ed identicamente distribuite $X_1, ..., X_n$ definite su uno spazio di probabilità $\{\Omega, \Im, P\}$.*

Remark 79 *In realtà né l'ipotesi di indipendenza né quella di identica distribuzione sono necessarie e saranno entrambe abbandonate più avanti; partiamo da queste condizioni super-semplificate per necessità.*

Supponiamo ora che la misura di probabilità P sia nota solo a meno del valore di un certo parametro θ. In genere, si definisce *Statistica Parametrica* la disciplina che studia le tecniche per risalire al valore di θ nel caso in cui esso appartenga ad uno spazio di dimensione finita, $\theta \in \mathbb{R}^p$, $p \in \mathbb{N}$; si parla invece di *Statistica NonParametrica* la disciplina che studia il caso in cui θ ha dimensione infinita. Ad esempio, se sappiamo che P è una legge Gaussiana ma ne ignoriamo valor medio e varianza siamo nell'ambito della Statistica Parametrica; se sappiamo che P ammette una densità e cerchiamo di ricavarla senza sapere che tipo di legge sia, siamo nell'ambito della Statistica Nonparametrica. Per distinguere un valore generico del parametro dal valore "vero" che caratterizza la legge da cui è stato estratto un certo campione aleatorio, puòà essere conveniente scrivere θ_0 in questo ultimo caso.

Definition 80 *(Stimatore) Uno stimatore di un certo parametro $\theta \in \mathbb{R}^p$ è una funzione (o meglio una successione di funzioni) misurabile $T_n : \mathbb{R}^n \to \mathbb{R}^p$ di un campione aleatorio; scriveremo equivalentemente*

$$T_n = \widetilde{\theta}_n = T_n(X_1, ...X_n) \ .$$

Remark 81 *La definizione è volutamente molto generica: qualsiasi funzione misurabile può essere considerata uno stimatore, quindi il punto cruciale sarà determinare quali siano le proprietà che consentono di valutare la bontà della scelta fatta.*

Example 82 *L'esempio più ovvio di stimatore (per il valore medio $\mu = E[X_i]$) è la media aritmetica , cioè*

$$\widetilde{X}_n = \widetilde{\mu}_n := \frac{1}{n}\sum_{i=1}^{n} X_i \ .$$

Altro stimatore naturale è la cosiddetta varianza empirica (che stima $\sigma^2 = Var[X_i]$), cioè

$$S_n^2 = \widetilde{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \widetilde{\mu}_n)^2.$$

9.1 Criteri per Valutare gli stimatori

- Non-distorsione/asintotica non distorsione
- Consistenza (in probabilità, in media r-esima, quasi certa, completa)
- Efficienza (assoluta e relativa)
- Asintotica Gaussianità

Example 83 *Media aritmetica - banale mostrare non-distorsione, consistenza ed asintotica Gaussianità; si può mostrare con Cramer-Rao (vedi dopo) che efficiente in senso assoluto nel caso Gaussiano. nel caso non-Gaussiano, si consideri un altro stimatore lineare*

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

(dove la condizione che la somma dei pesi sia 1 garantisce la non-distorsione). La varianza è evidentemente $\sum_{i=1}^{n} c_i^2 Var[X_i]$*, e pertanto nel caso di variabili incorrelate ed identicamente distribuite possiamo porcil problema di massimo vincolato*

$$\min \sum_{i=1}^{n} c_i^2 \ \ s.t. \ \ \sum_{i=1}^{n} c_i = 1 \ .$$

Con i moltiplicatori di Lagrange troviamo

$$\psi(c_1, ..., c_n; \lambda) = \sum_{i=1}^{n} c_i^2 + \lambda \left\{ \sum_{i=1}^{n} c_i - 1 \right\}$$

con derivate prime

$$\frac{\partial \psi(c_1, ..., c_n; \lambda)}{\partial c_i} = 2c_i + \lambda \ .$$

Ne segue che il problema di minimo sarà risolto imponendo che i coefficienti c_i *siano tutti uguali e pertanto pari a* n^{-1}*; questo rende la media aritmetica uno stimatore BLUE (Best Linear Unbiased Estimator).*

Example 84 *(Varianza Campionaria) In questo caso possiamo scrivere*

$$\begin{aligned} S_n^2 &= \frac{1}{n} \sum_{i=1}^{n} (X_i - \widetilde{\mu}_n)^2 = \frac{1}{n} \sum_{i=1}^{n} X_i^2 - \widetilde{\mu}_n^2 \ , \\ E[S_n^2] &= \frac{1}{n} \sum_{i=1}^{n} E[X_i^2] - (E\widetilde{\mu}_n)^2 - Var[\widetilde{\mu}_n] \\ &= \frac{1}{n} \left(\sum_{i=1}^{n} \left\{ E[X_i^2] - \mu^2 \right\} - \sigma^2 \right) = \frac{n-1}{n} \sigma^2. \end{aligned}$$

La varianza campionaria risulta quindi distorta, anche se asintoticamente non-distorta; in un certo senso è come se perdessimo un grado di liberta centrando le osservazioni sulla media empirica invece di quella teorica (per $n = 1$ la varianza campionaria è identicamente nulla). Per quello che riguarda la consistenza debole, è una semplice conseguenza del Lemma di Slutzky; se le X_i hanno momenti quarti finiti, si può dimostrare la convergenza in media quadratica usando la disuguaglianza di Chebyshev.

Exercise 85 *Determinare opportune condizioni per dimostrare la convergenza quasi certa della varianza campionaria al valore della varianza σ^2; sotto quali condizioni vale il teorema del limte centrale?*

10 Il metodo dei momenti

I due stimatori più naturali sono media e varianza empirica; in entrambi i casi possono essere interpretati come lo stimatore che si ottiene sostituendo la distribuzione empirica a quella teorica nel calcolo di parametri che possono esser visti come valori medi di momenti del campione aleatorio. Il metodo dei momenti può essere visto come la generalizzazione di questa idea.

Supponiamo ora di avere un campione aleatorio $X_1, ..., X_n$ estratto da una legge di probabilità P_θ, con $\theta = (\theta_1, ..., \theta_k)$. Supponiamo anche che questa legge ammetta k momenti finiti, cioè $\mu_j = E[X_1^j]$ sia ben definito per $j = 1, ..., k$. Chiamiamo m_j i momenti empirici $m_j := n^{-1} \sum_{i=1} X_i^j$; l'idea di fondo è scegliere il valore dei parametri $\theta_1, ..., \theta_k$ in modo che valga l'uguaglianza vettoriale:

$$\begin{aligned} m_1 &= \mu_1(\widetilde{\theta}_1, \widetilde{\theta}_2, ..., \widetilde{\theta}_k) \\ m_2 &= \mu_2(\widetilde{\theta}_1, \widetilde{\theta}_2, ..., \widetilde{\theta}_k) \\ &... \\ m_k &= \mu_k(\widetilde{\theta}_1, \widetilde{\theta}_2, ..., \widetilde{\theta}_k) \end{aligned}$$

Più precisamente

Definition 86 *Assumiamo che esista un diffeomorfismo $g : \mathbb{R}^k \to \mathbb{R}^k$ tale che $g(\mu_1, ..., \mu_k) = (\theta_1, ..., \theta_k)$, per ogni valore di $(\theta_1, ..., \theta_k)$. Lo stimatore del metodo dei momenti è allora definito da*

$$(\widetilde{\theta}_1, ..., \widetilde{\theta}_k) := g(m_1, ..., m_k) \ .$$

Per determinare le proprietà asintotiche del nostro stimatore dobbiamo rinforzare leggermente le nostre ipotesi.

Condition 87 *Le variabili aleatorie $X_1, ...X_n$ sono indipendenti ed identicamente distribuite con momenti di ordine $2k$ finiti.*

Proposition 88 *Sotto la precedente condizione, vale il teorema del limite centrale multivariato*

$$\sqrt{n} \begin{pmatrix} m_1 - \mu_1 \\ ... \\ m_k - \mu_k \end{pmatrix} \to_d N \left(\begin{pmatrix} 0 \\ ... \\ 0 \end{pmatrix}, \Omega \right)$$

dove

$$\Omega := \begin{pmatrix} E[(X_1 - \mu_1)^2] & E[(X_1 - \mu_1)(X_1^2 - \mu_2)] & ... & E[(X_1 - \mu_1)(X_1^k - \mu_k)] \\ & ... & ... & ... \\ & & & E[(X_1^k - \mu_k)^2] \end{pmatrix} \ .$$

Proposition 89 *Sotto la precedente condizione, lo stimatore del metodo dei momenti converge in probabilità al valore vero dei parametri e vale il teorema del limte centrale:*

$$\sqrt{n} \begin{pmatrix} \widetilde{\theta}_1 - \theta_1 \\ ... \\ \widetilde{\theta}_k - \theta_k \end{pmatrix} \to_d N \left(\begin{pmatrix} 0 \\ ... \\ 0 \end{pmatrix}, Jg(\mu_1, ..., \mu_k) \Omega Jg(\mu_1, ..., \mu_k)^T \right) \ .$$

Proof. Delta Method multivariato. ■

Example 90 *(Legge Gamma) Supponiamo di avere un campioni di variabili IID con legge Gamma di parametri α e β, cioè con densità*

$$\begin{aligned} f_X(x) &= \frac{1}{\Gamma(\alpha)\beta^\alpha} x^{\alpha-1} \exp(-\frac{y}{\beta}) I_{[0,\infty)}(x) \ , \\ \Gamma(\alpha) &= \int_0^\infty x^{\alpha-1} \exp(-x) dx \ . \end{aligned}$$

Come noto (si veda l'appendice) abbiamo

$$\begin{aligned} E[X_1] &= \alpha\beta \ , \\ E[X_1^2] &= (\alpha\beta)^2 + \alpha\beta^2 = \alpha\beta^2(\alpha+1) \\ Var[X_1] &= \alpha\beta^2. \end{aligned}$$

Possiamo pertanto scrivere

$$\begin{aligned} \frac{\mu_2 - \mu_1^2}{\mu_1} &= \frac{(\alpha\beta)^2 + \alpha\beta^2 - (\alpha\beta)^2}{\alpha\beta} = \beta \ , \\ \frac{\mu_1^2}{\mu_2 - \mu_1^2} &= \frac{\alpha^2\beta^2}{\alpha\beta^2} = \alpha \ . \end{aligned}$$

Lo stimatore del metodo dei momenti è pertanto

$$\begin{pmatrix} \widetilde{\alpha} \\ \widetilde{\beta} \end{pmatrix} = \begin{pmatrix} \frac{m_1^2}{m_2 - m_1^2} \\ \frac{m_2 - m_1^2}{m_1} \end{pmatrix}$$

con matrice Jacobiana

$$\begin{aligned} Jg(\mu_1(\alpha,\beta), \mu_2(\alpha,\beta)) &= \begin{pmatrix} \frac{2\mu_1\mu_2}{(\mu_2 - \mu_1^2)^2} & -\frac{\mu_2}{\mu_1^2} - 1 \\ -\frac{\mu_1^2}{(\mu_2 - \mu_1^2)^2} & \frac{1}{\mu_1} \end{pmatrix} \\ &= \begin{pmatrix} \frac{2\alpha\beta(\alpha^2\beta^2 + \alpha\beta^2)}{(\alpha\beta^2)^2} & -\frac{(\alpha^2\beta^2 + \alpha\beta^2)}{\alpha^2\beta^2} - 1 \\ -\frac{\alpha^2\beta^2}{(\alpha\beta^2)^2} & \frac{1}{\alpha\beta} \end{pmatrix} \\ &= \begin{pmatrix} \frac{2(\alpha+1)}{\alpha\beta} & -2 - \frac{1}{\alpha} \\ -\frac{1}{\beta^2} & \frac{1}{\alpha\beta} \end{pmatrix} \ . \end{aligned}$$

Example 91 *Suipponiamo che in una certa città avvengano una serie di reati; poiché però non tutti i reati vengono denunciati, la legge del numero dei crimini osservati può essere vista come una binomiale con k e p incogniti, cioè il nostro campione aleatorio $X_1, ..., X_n$ ha legge*

$$\binom{k}{x} p^x (1-p)^{k-x},$$

con k, p parametri incogniti da stimare. Abbiamo

$$\mu_1 = kp \ , \ \mu_2 = k^2p^2 + kp(1-p) \ ,$$

e lo stimatore col metodo dei momenti si ottiene invertendo le relazioni

$$\begin{aligned} \overline{X}_n &= kp \ , \\ \frac{1}{n}\sum_{i=1}^{n} X_i^2 &= kp(1-p) + k^2p^2. \end{aligned}$$

Si può ottenere

$$\begin{aligned} \widetilde{p}_n &= \frac{\overline{X}_n}{\widetilde{k}_n} \ , \\ \widetilde{k}_n &= \frac{\overline{X}_n^2}{\overline{X}_n - \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2} \ . \end{aligned}$$

Il metodo dei momenti fu introdotto da Pearson (1857-1936), matematico ed avvocato, che cambiò il nome da Carl in Karl per la sua ammirazione nei riguardi di Karl Marx. E' da notare che non richiede la consocenza completa della legge di probabilità delle X_i, a differenza del metodo della massima verosimiglianza che discutiamo nella prossima sezione.

11 Il metodo della massima verosimiglianza

Supponiamo di avere di fronte a noi due urne, una con 90 palline bianche e 10 nere, e l'altra con 10 nere e 90 bianche; identifichiamo le urne con la proporzione di palline bianche che contengono (denotata con p), in modo tale che la prima urna corrisponda a $p = .9$ e la seconda a $p = .1$. Viene estratta una singola pallina e non sappiamo da quale urne venga; osserviamo però che la pallina è bianca. Dovendo cercare di risalire all'urna di provenienza, sembra naturale scegliere quella che rende più probabile osservare quello che abbiamo effettivamente osservato, cioè l'urna che contiene la maggiore proporzione di palline bianche; in altre parole, dall'osservazione "è stata estratta una pallina bianca" sembra naturale far discendere lo stimatore $\widehat{p} = 0.9$.

Supponiamo ora che invece di una sola pallina ne siano estratte 5, e che risultino essere 3 bianche e 2 nere; immaginiamo altresì che l'alternativa non sia solo tra due urne, ma tra un continuo di urne con tutte le possibili proporzioni di palline bianche tra 0 ed 1, identificate sempre con p. In questo caso, la probabilità di osservare 3 bianche e 2 nere è data da una legge binomiale:

$$\Pr\{\text{3 bianche e 2 nere}\} = \binom{5}{3} p^3 (1-p)^2,$$

ed è facilmente verificabile che questa probabilità è massimizzata prendendo $\widehat{p} = \frac{3}{5}$.

Questo esempio molto semplice dovrebbe aiutare a capire l'idea che è alla base degli stimatori di massima verosimiglianza - si tratta di costruire la funzione di probabilità (o di densità, nel caso continuo) relativa ad un certo campione aletaorio, e vederla quindi non più come funzione del campione, ma come funzione (aleatoria) dei parametri, prendendo come date le osservazioni. In altre parole, si ha la seguente definizione:

Definition 92 *(Funzione di verosimiglianza) Sia $X_1, ..., X_n$ un campione aleatorio con legge (=funzione di probabilità o densità, a seconda del caso discreto o continuo) f_θ, $\theta \in \mathbb{R}^p$. La funzione di verosimiglianza $L : \mathbb{R}^p \to \mathbb{R}$ è definita da*

$$L(\theta; X_1, ..., X_n) := \prod_{i=1}^{n} f(X_i; \theta) = f(X_1, ..., X_n; \theta) \ .$$

Remark 93 *Con un leggero abuso di notazione, a sinistra della prima uguaglianza abbiamo scritto f sia per la legge univariata che per qualla congiunta; inoltre usiamo $f(.)$ sia per il caso discreto che per quello continuo. Il punto più importante è però un altro: la verosimiglianza è funzione del parametro prendendo i valori di $X_1, ...X_n$ come dati, mentre la densità congiunta prende il parametro θ come dato ed è funzione dei possibili valori delle osservazioni $x_1, ..., x_n$.*

Remark 94 *Ovviamente si ha (nel caso continuo)*

$$\int_{\mathbb{R}^n} f(x_1, ..., x_n; \theta) dx_1 ... dx_n = 1 \ ;$$

non c'è però nessun motivo di aspettarsi l'uguaglianza

$$\int_{\mathbb{R}^p} L(\theta; X_1, ..., X_n) d\theta \stackrel{???}{=} 1 \ ,$$

anzi in generale questo integrale può divergere. La funzione di verosimiglianza pertanto NON identifica una densità di probabilità.

Example 95 *(Valor medio nel caso Gaussiano) Sia $X_1, ..., X_n$ un campione aleatorio estratto da una legge Gaussiana $N(\mu, 1)$; è immediato verificare che la funzione di verosmiglianza prende la forma*

$$L(\mu; X_1, ..., X_n) = \frac{1}{(2\pi)^{n/2}} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} (X_i - \mu)^2 \right\} \ .$$

In questo caso (e in quelli a seguire) è conveniente lavorare con il logaritmo della funzione di verosimiglianza; trattandosi di una trasformazione monotona crescente, il valore che massimizza la funzione (cioè l'argmax) non cambia, e pertanto nemmeno lo stimatore. Si ha in particolare

$$\begin{aligned} \ell(\mu; X_1, ..., X_n) &: = \log L(\mu; X_1, ..., X_n) \\ &= -\frac{n}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^{n} (X_i - \mu)^2, \end{aligned}$$

funzione che è evidentemente massimizzata in

$$\widehat{\mu}_{MLE} = \frac{1}{n} \sum_{i=1}^{n} X_i \ .$$

Example 96 *(Variabili Bernoulliane) Sia $X_1, ..., X_n$ un campione aleatorio di variabili $Ber(p)$; la probabilità di k successi per $k = 0, ..., n$ è evidentemente data da $p^k (1-p)^{n-k}$, e la verosimiglianza è pertanto*

$$L(\mu; X_1, ..., X_n) = p^{\sum_{i=1}^n X_i} (1-p)^{n - \sum_{i=1}^n X_i}.$$

La massimizzazione della log-verosimiglianza ci fa ottenere facilmente

$$\begin{aligned} \frac{\partial \log L(\mu; X_1, ..., X_n)}{\partial p} &= (\sum_{i=1}^{n} X_i) \frac{\partial}{\partial p} \log p + (n - \sum_{i=1}^{n} X_i) \frac{\partial}{\partial p} \log(1-p) \\ &= \frac{\sum_{i=1}^n X_i}{p} - \frac{(n - \sum_{i=1} X_i)}{1-p} \ ; \end{aligned}$$

imponendo che la derivata sia pari a zero, si vede facilmente che

$$\widehat{p}_{MLE} = \frac{1}{n} \sum_{i=1} X_i \ .$$

Remark 97 *Si sarebbe potuto scrivere la verosimiglianza come*

$$L(\mu; X_1, ..., X_n) = \binom{n}{\sum_{i=1}^n X_i} p^{\sum_{i=1}^n X_i} (1-p)^{n-\sum_{i=1}^n X_i}.$$

Quale è la versione "giusta" della verosimiglianza? Come sarebbero cambiate le stime utilizzando questa seconda versione? Si tratta di un risultato generale o c'è un principio generale sottostante?

11.0.1 Consistency and Asymptotic Gaussianity

To study consistency, we need first to introduce the Kullback-Leibler distance between two probility densities/probability mass function. This is defined by

Definition 98 *Siano f, g funzioni di densità o funzioni di probabilità. La distanza di Kullback-Leibler tra f è definita da*

$$D(f,g) = \int \log(\frac{f(x,\theta)}{g(x,\theta)}) f(x;\theta) dx$$

se f è assolutamente continua rispetto a g; altrimenti la distanza viene posta pari a $+\infty$. Analogamente, nel caso discreto

$$D(p,q) = \sum_{x_i} \log(\frac{p(x_i,\theta)}{q(x_i,\theta)}) p(x_i;\theta) \ .$$

Remark 99 *E' facilmente verificabile che la distanza di Kullback-Leibler è sempre non-negativa; infatti, utilizzando la disuguaglianza di Jensen*

$$\begin{aligned} D(f,g) &= E_f[\log \frac{f(X)}{g(X)}] = -E_f[\log \frac{g(X)}{f(X)}] \\ &\geq \log\left(\int \frac{g(x,\theta)}{f(x,\theta)}) f(x;\theta) dx\right) = \log\left(\int g(x,\theta) dx\right) = 0 \ . \end{aligned}$$

D'altra parte la distanza di Kullback-Leibler non identifica una metrica, ad esempio perché non è simmetrica negli argomenti (e non vale la disuguaglianza triangolare).

Quale è la relazione tra funzione di verosimiglianza e distanza di Kullback-Leibler? Focalizziamoci come al solito sulla log-verosimiglianza; considerando che la stimatore è invariante se la funzione è trasformata linearmente con coefficienti che non dipendono dal parametro, possiamo passare da $\log L = \sum_{i=1}^n \log f(X_i;\theta)$ a

$$M_n(\theta) := \frac{1}{n}\sum_{i=1}^n \log f(X_i;\theta) - \frac{1}{n}\sum_{i=1}^n \log f(X_i;\theta_0) = \frac{1}{n}\sum_{i=1}^n \log \frac{f(X_i;\theta)}{f(X_i;\theta_0)} \ ,$$

dove θ_0 denota il "vero" valore del parametro. Sotto alcune delle condizioni di regolarità che abbiamo visto precedentemente affinchè valga la legge dei grandi numeri, abbiamo che, *ad un valore di θ fissato:*

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\log\frac{f(X_i;\theta)}{f(X_i;\theta_0)} &\rightarrow {}_pE[\log\frac{f(X_1;\theta)}{f(X_1;\theta_0)}] \\ &= \int\log\frac{f(x;\theta)}{f(x;\theta_0)}f(x;\theta_0)dx \\ &= -D(f_{\theta_0},f_\theta)\ .\end{aligned}$$

L'idea euristica dietro alla dimostrazione della consistenza è dunque la seguente: $M_n(\theta)$ converge a $-D(f_{\theta_0},f_\theta)$, che è sempre negativa tranne quando $\theta=\theta_0$, dove vale zero. Poiché lo stimatore di massima verosimiglianza è definito come il valore di θ che massimizza $M_n(\theta)$, a sua volta esso convergerà al valore che massimizza $-D(f_{\theta_0},f_\theta)$, cioè $\theta=\theta_0$. Nel seguito, scriviamo per brevità $M(\theta)=-D(f_{\theta_0},f_\theta)$; abbiamo la seguente proposizione:

Proposition 100 *(Consistenza) Assumiamo che valgano le seguenti condizioni di regolarità:*

a) (Legge dei grandi numeri uniforme) Si ha che

$$\sup|M_n(\theta)-M(\theta)|=o_p(1)\ \textit{per}\ n\rightarrow\infty\ ;$$

b) (Convessità) Si ha che per ogni $\varepsilon>0$ esiste $\eta_\varepsilon>0$ tale che

$$|\theta-\theta_0|>\varepsilon\Longrightarrow M(\theta_0)-M(\theta)>\eta_\varepsilon\ .$$

Allora $\widehat{\theta}_{MLE;n}\rightarrow_p\theta_0$.

Remark 101 *L'ipotesi a) è più forte delle leggi dei grandi numeri che abbiamo dimostrato in precedenza, perché richiede che la convergenza sia uniforme in θ, il che non è garantito anche se la convergenza avviene puntualmente per ogni valore fisso di θ. Le leggi dei grandi numeri uniformi sono ampiamente trattate nei corsi della magistrale (Teoremi di Glivenko-Cantelli) ma non fanno parte del programma di questo corso. L'ipotesi b) è essenzialmente una condizione di identificabilità: se esistesse un insieme di valori diversi di θ a distanza di Kullback-Leibler 0 da θ_0 il problema della consistenza sarebbe evidentemente irrisolvibile (e lo stimatore potrebbe non essere nemmeno ben definito).*

Proof. L'idea è di mostrare che per ogni $\varepsilon>0$ fissato si ha $\Pr\left\{M(\theta_0)-M(\widehat{\theta}_n)>\eta_\varepsilon\right\}\rightarrow 0$; grazie alla condizione di convessità abbiamo

$$\Pr\left\{|\theta_0-\widehat{\theta}_n|>\varepsilon\right\}\leq\Pr\left\{M(\theta_0)-M(\widehat{\theta}_n)>\eta_\varepsilon\right\}$$

e conseguentemente la proposizione sarà dimostrata. A questo proposito notiamo che, per definizione di stimatore di massima verosimiglianza

$$M_n(\widehat{\theta}_n)-M_n(\theta_0)>0$$

e quindi

$$\begin{aligned} M(\widehat{\theta}_n) - M(\theta_0) &= M_n(\theta_0) - M(\widehat{\theta}_n) + M(\theta_0) - M_n(\theta_0) \\ &\leq M_n(\widehat{\theta}_n) - M(\widehat{\theta}_n) + M(\theta_0) - M_n(\theta_0) \ . \end{aligned}$$

Segue quindi che

$$\begin{aligned} \Pr\left\{ M(\theta_0) - M(\widehat{\theta}_n) > \eta_\varepsilon \right\} &\leq \Pr\left\{ |M(\theta_0) - M_n(\theta_0)| > \frac{\eta_\varepsilon}{2} \right\} + \Pr\left\{ |M_n(\widehat{\theta}_n) - M(\widehat{\theta}_n)| > \frac{\eta_\varepsilon}{2} \right\} \\ &\to 0 \ , \text{ per } n \to \infty \ , \end{aligned}$$

usando per il primo termine la legge dei grandi numeri standard e per il secondo quella uniforme (che implica ovviamente quella standard). ■

11.0.2 La funzione punteggio

Definition 102 *(Funzione Punteggio) La funzione punteggio (score function in inglese) è definita da*

$$s_n(\theta; X_1, ..., X_n) = \frac{\partial}{\partial \theta} \log L(\theta; X_1, ..., X_n) \ ,$$

assumendo ovviamente che la derivata esista. In genere, si tratta di una funzione da $\mathbb{R}^p$ in $\mathbb{R}^p$.

Nella discussione che segue inizieremo dal caso $p = 1$. Notiamo innanzitutto che la funzione punteggio è essa stessa una variabile aleatoria, per ogni valore di θ fissato. Ha quindi senso domandarsi quale sia il suo valor medio e la sua varianza, ammesso che esistano. Iniziamo quindi a imporre alcune condizioni di regolarità.

Definition 103 *(Condizioni di regolarità) Valgono le seguenti ipotesi:*

- *Le condizioni di consistenza*
- *Le osservazioni $X_1, ...X_n$ sono indipendenti ed identicamente distribuite*
- *La log-verosimiglianza ammette due derivate continue, $\frac{\partial^2}{\partial \theta^2} \log L_n \in C^2$*
- *La derivata seconda ha momento secondo finito, $E[|\frac{\partial^2}{\partial \theta^2} \log L_n|] < \infty$*
- *Il vero valore del parametro θ_0 appartiene all'interno dell'insieme dei valori ammissibili, $\theta_0 \in Int(\Theta)$*
- *E' possibile scambiare due volte la derivata con l'integrale nel valor medio della log-verosimiglianza*

Remark 104 *Queste condizioni sono più forti del necessario per qullo che riguarda i risultati sulla funzione punteggio, ma preferiamo enunciarle una volta per tutte in funzione del risultato sulla asinotitca Gaussianità degli stimatori ML.*

Lemma 105 *Sotto le condizioni di regolarità, abbiamo che*

$$E_{\theta_0}[s_n(\theta_0; X_1, ..., X_n)] = 0 \ .$$

Proof. Per definizione abbiamo che

$$\begin{aligned} E_{\theta_0}[s_n(\theta_0; X_1, ..., X_n)] &= \int \left\{\log' L(\theta_0; x_1, ..., x_n)\right\} f(x_1, ..., x_n; \theta_0) dx_1 ... dx_n \\ &= \int \frac{\frac{\partial f(x_1,...,x_n;\theta_0)}{\partial \theta}\Big|_{\theta=\theta_0}}{f(x_1, ..., x_n; \theta_0)} f(x_1, ..., x_n; \theta_0) dx_1 ... dx_n \\ &= \int \left.\frac{\partial f(x_1, ..., x_n; \theta_0)}{\partial \theta}\right|_{\theta=\theta_0} dx_1 ... dx_n \ . \end{aligned}$$

Poiché per ipotesi possiamo scambiare derivate ed integrali, abbiamo che

$$\begin{aligned} \int \left.\frac{\partial f(x_1, ..., x_n; \theta_0)}{\partial \theta}\right|_{\theta=\theta_0} dx_1 ... dx_n &= \left.\frac{\partial}{\partial \theta} \int f(x_1, ..., x_n; \theta) dx_1 ... dx_n\right|_{\theta=\theta_0} \\ &= \left.\frac{\partial}{\partial \theta} 1\right|_{\theta=\theta_0} = 0 \ . \end{aligned}$$

■

Example 106 *Se consideriamo il caso della verosimiglianza Gaussiana, abbiamo facilmente (assumendo varianza nota e pari a 1, X_i iid $N(\mu_0, 1)$)*

$$s_n(\mu; X_1, ..., X_n) = \sum_{i=1}^{n} (X_i - \mu) \ ,$$

ed è evidentemente che il valor medio si annulla prendendo $\mu = \mu_0$:

$$E[s_n(\mu; X_1, ..., X_n)] = n(\mu - \mu_0) \ .$$

Lemma 107 *Sotto le condizioni di regolarità, abbiamo che*

$$\begin{aligned} & E\left[\left.\frac{\partial^2 \log L(\theta; x_1, ..., x_n)}{\partial \theta^2}\right|_{\theta=\theta_0}\right] \\ = & \int \left\{\log'' L(\theta_0; x_1, ..., x_n)\right\} f(x_1, ..., x_n; \theta_0) dx_1 .... dx_n \\ = & -\int \left\{\log' L(\theta_0; x_1, ..., x_n)\right\}^2 f(x_1, ..., x_n; \theta_0) dx_1 .... dx_n \\ = & -E\left[\left\{\left.\frac{\partial \log L(\theta; x_1, ..., x_n)}{\partial \theta}\right|_{\theta=\theta_0}\right\}^2\right] \ . \end{aligned}$$

Proof. Si noti innanzitutto che, per qualsiasi valore di θ

$$H(\theta) := \int \left\{\log' L(\theta; x_1, ..., x_n)\right\} f(x_1, ..., x_n; \theta) dx_1 ....dx_n \equiv 0 \ .$$

Di nuovo per le condizioni di regolarità possiamo allora considerare

$$\begin{aligned}
0 &= \frac{\partial H(\theta)}{\partial \theta} = H'(\theta) \\
&= \frac{\partial}{\partial \theta} \int \left\{\log' L(\theta; x_1, ..., x_n)\right\} f(x_1, ..., x_n; \theta) dx_1 ....dx_n \\
&= \int \left\{\log'' L(\theta; x_1, ..., x_n)\right\} f(x_1, ..., x_n; \theta) dx_1 ....dx_n \\
&\quad + \int \left\{\log' L(\theta; x_1, ..., x_n)\right\} f'(x_1, ..., x_n; \theta) dx_1 ....dx_n \\
&= \int \left\{\log'' L(\theta; x_1, ..., x_n)\right\} f(x_1, ..., x_n; \theta) dx_1 ....dx_n \\
&\quad + \int \left\{\log' L(\theta; x_1, ..., x_n)\right\} \frac{f'(x_1, ..., x_n; \theta)}{f(x_1, ..., x_n; \theta)} f(x_1, ..., x_n; \theta) dx_1 ....dx_n \\
&= \int \left\{\log'' L(\theta; x_1, ..., x_n)\right\} f(x_1, ..., x_n; \theta) dx_1 ....dx_n \\
&\quad + \int \left\{\log' L(\theta; x_1, ..., x_n)\right\}^2 f(x_1, ..., x_n; \theta) dx_1 ....dx_n \ ,
\end{aligned}$$

ed il risultato è dimostrato. ■

Remark 108 *E' importante notare che in $\theta = \theta_0$*

$$\begin{aligned}
&\int \left\{\log' L(\theta_0; x_1, ..., x_n)\right\}^2 f(x_1, ..., x_n; \theta_0) dx_1 ....dx_n \\
= \ & E[(s_n(\theta_0))^2] = Var(s_n(\theta_0)) \ .
\end{aligned}$$

La varianza della funzione punteggio calcolata nel valore vero del parametro è pertanto uguale al reciproco della derivata seconda della log-verosimiglianza nello stesso punto.

Definition 109 *(Informazione di Fisher) La varianza della funzione punteggio è nota come informazione di Fisher e si indica con $I_n(\theta_0)$. In generale, si tratta di una matrice di dimensione $p \times p$.*

11.0.3 L'asintotica Gaussianità degli stimatori MLE

Theorem 110 *Sotto le condizioni di regolarità di cui sopra, si ha che*

$$\sqrt{n}\left\{\widehat{\theta}_n - \theta_0\right\} \rightarrow_d N(0, I_1^{-1}(\theta_0)) \ .$$

Equivalentemente, abbiamo che

$$I_n^{1/2}(\theta_0)\left\{\widehat{\theta}_n - \theta_0\right\} \rightarrow_d N(0, Id_p) \ ,$$

dove Id_p indica la matrice di identità di ordine p.

Proof. Consideriamo per semplicità il caso $p = 1$. Per il teorema del valor medio di Lagrange, esiste $\overline{\theta}_n$ intermedio tra $\widehat{\theta}_n$ e θ_0 tale per cui vale l'uguaglianza

$$0 = \log' L(\widehat{\theta}_n) = \log L'(\theta_0) + \log L''(\overline{\theta}_n)(\widehat{\theta}_n - \theta_0) \ ,$$

da cui

$$\sqrt{n}(\widehat{\theta}_n - \theta_0) = -\frac{\log L'(\theta_0)/\sqrt{n}}{\log L''(\overline{\theta}_n)/n} \ .$$

Per il numeratore abbiamo una somma di variabili aleatorie IID con valor medio nullo e varianza finita; siamo quindi nel dominio di applicabilità del teorema del limite centrale ed otteniamo

$$\log L'(\theta_0)/\sqrt{n} = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \left. \frac{\partial \log f(X_i;\theta)}{\partial \theta} \right|_{\theta=\theta_0} \rightarrow_d N(0, I_1(\theta_0)) \ .$$

Per il denominatore abbiamo una somma di variabili IID con valor medio finito, quindi per la legge dei grandi numeri su variabili uniformemente integrabili ed il teorema di Slutzky otteniamo

$$\log L''(\overline{\theta}_n)/n = \frac{1}{n} \sum_{i=1}^{n} \left. \frac{\partial^2 \log f(X_i;\theta)}{\partial \theta^2} \right|_{\theta=\overline{\theta}_n} \rightarrow_p I_1(\theta_0) \ .$$

Combinando i due risultati ed usando di nuovo Slutzky si arriva all'enunciato del Teorema. ■

Remark 111 *Questa dimostrazione rende evidente come la consistenza degli stimatori sia un prerequisito necessario perché abbia senso la domanda sulla loro asintotica Gaussianità.*

12 Il limite inferiore di Cramér-Rao

Un risultato notevole riguarda la determinazione della varianza minima degli stimatori non-distorti, sotto condizioni di regolarità; in particolare, emerge che tale varianza minima coincide con quella degli stimatori di massima verosimiglianza, sotto condizioni di regolarità. Per tale motivo, gli stimatori di massima verosimiglianza sotto opportune ipotesi (che coprono gran parte delle distribuzioni di uso comune, almeno nei casi più semplici) risultano essere non solo consistenti ed asinotitcamente Gaussiani, ma anche efficienti in senso assoluto.

Remark 112 *E' evidente che porsi la questione sulla varianza minima ha senso solo per stimatori che siano non-distorti; altrimenti qualsiasi stimatore con valore identicamente costante non potrebbe essere migliorato, avendo varianza identicamente pari a zero.*

Consideriamo ora uno stimatore generico di un parametro $\theta \in \mathbb{R}$,

$$W_n = W_n(X_1, ..., X_n)$$

e supponiamo che la funzione $\psi(\theta) := E_{f_\theta}[W(X_1, ..., X_n)]$ sia di classe C^1 per ogni n; assumiamo inoltre che la densità congiunta f_θ sia tale per cui, per ogni $h \in C^1$, $h : \mathbb{R}^n \to \mathbb{R}$ con $E_\theta[|h(X_1, ..., X_n)|] < \infty$ valga la scambiabilità

$$\frac{d}{d\theta} E_\theta[h(X_1, ..., X_n)] = \int_{\mathbb{R}^n} h(x_1, ..., x_n) \frac{\partial}{\partial \theta} f(x_1, ..., x_n; \theta) dx_1 ... dx_n \ .$$

Chiamiamo ora $b(\theta) = E[W_n] - \theta$ il bias della nostra statistica, e $b'(.)$ la sua derivata rispetto a θ. Abbiamo allora il seguente

Theorem 113 *(Cramér-Rao) Per $\theta \in \mathbb{R}$*

$$Var[W_n] \geq \frac{\{b'(\theta_0)\}^2}{I_n(\theta_0)} \ .$$

Remark 114 *Per semplicità, abbiamo enunciato e dimostreremo il teorema solo nel caso $p = 1$. La generalizzazione a p generico è comunque abbastanza semplice; consideriamo lo stimatore W_n come un vettore colonna $p \times 1$, con valor medio $\Psi(\theta)$ che ha matrice Jacobiana $p \times p$ denotata con $J\Psi()$. Allora la matrice di varianza e covarianza di W_n soddisfa la disuguaglianza*

$$Var[W_n] \geq J\Psi(\theta_0) I_n^{-1}(\theta_0) J\Psi(\theta_0)^T .$$

La disuguaglianza va intesa none solito senso di disuguaglianza tra matrici simmetriche e non-negative definite: la loro differenza deve essere non-negativa definita.

Proof. (caso $p = 1$). L'idea di fondo è utilizzare la disuguaglianza di Cauchy-Schwartz, scrivendola come

$$Var[Y] \geq \frac{Cov^2(X, Y)}{Var[X]} \ .$$

Prendiamo come X la funzione punteggio $\frac{d}{d\theta} \log L(\theta; X_1, ..., X_n)\big|_{\theta=\theta_0}$; sappiamo che $E[X] = 0$, da cui $Cov(X, Y) = E[XY]$. Prendiamo quindi $Y = W_n$, e osserviamo che

$$\begin{aligned}
E[XY] &= \int_{\mathbb{R}^n} W(x_1, ..., x_n) \left. \frac{d}{d\theta} \log L(\theta; x_1, ..., x_n) \right|_{\theta=\theta_0} f(x_1, ..., x_n; \theta_0) dx_1 ... dx_n \\
&= \int_{\mathbb{R}^n} W(x_1, ..., x_n) \frac{\frac{d}{d\theta} f(\theta; x_1, ..., x_n)\big|_{\theta=\theta_0}}{f(x_1, ..., x_n; \theta_0)} f(x_1, ..., x_n; \theta_0) dx_1 ... dx_n \\
&= \int_{\mathbb{R}^n} W(x_1, ..., x_n) \left. \frac{d}{d\theta} f(\theta; x_1, ..., x_n) \right|_{\theta=\theta_0} dx_1 ... dx_n \\
&= \left. \frac{d}{d\theta} \int_{\mathbb{R}^n} W(x_1, ..., x_n) f(\theta; x_1, ..., x_n) dx_1 ... dx_n \right|_{\theta=\theta_0} \\
&= \left. \frac{d}{d\theta} E_\theta[W_n] \right|_{\theta=\theta_0} .
\end{aligned}$$

Questo conclude la dimostrazione nel caso con densità; il caso discreto è identico. ■

Remark 115 *E' interessante studiare cosa succede nel caso in cui le condizioni di regolarità, ed in particolare la possibilità di scambiare la derivata con l'integrale nel valor medio, non siano soddisfatte. Prendiamo ad esempio un campione di variabili i.i.d., uniformi in* $[0, \theta]$; *si verifica facilmente che la funzione di verosimiglianza prende la forma*

$$L(\theta; X_1, ..., X_n) = \prod_{i=1}^{n} \frac{1}{\theta} I_{[0,\theta]}(X_i) = \frac{1}{\theta^n} I_{[0,\theta]}(X_{(n)}) \ ,$$

dove $X_{(n)}$ *indica la più grande delle n osservazioni; abbiamo inoltre*

$$\widehat{\theta}_{ML;n} = X_{(n)} \ .$$

Ora, si vede anche facilmente che, per ogni $\varepsilon > 0$

$$\begin{aligned} \Pr\left\{\theta_0 - X_{(n)} > \varepsilon\right\} &= \prod_{i=1}^{n} \Pr\left\{\theta_0 - X_i > \varepsilon\right\} \\ &= \left\{\Pr\left\{\theta_0 - \varepsilon > X_1\right\}\right\}^n \\ &= \left\{\frac{1}{\theta_0}(\theta_0 - \varepsilon)\right\}^n = (1 - \frac{\varepsilon}{\theta_0})^n. \end{aligned}$$

Poichè queste probabilità sono sommabili, abbiamo che lo stimatore è completamente convergente (e pertanto converge quasi certamente). Inoltre abbiamo che

$$\begin{aligned} \Pr\left\{n(\theta_0 - X_{(n)}) > \varepsilon\right\} &= \left\{\Pr\left\{\theta_0 - \frac{\varepsilon}{n} > X_1\right\}\right\}^n \\ &= (1 - \frac{\varepsilon}{n\theta_0})^n \to \exp(-\frac{\varepsilon}{\theta_0}) \end{aligned}$$

per $n \to \infty$: *abbiamo quindi una forma di superconsistenza, perché la convergenza avviene a velocità* n^{-1} *invece che* $n^{-1/2}$ *come nel teorema precedente. D'altra parte la distribuzione limite è esponenziale invece che Gaussiana. Questo esempio mostra come, quando le condizioni di regolarità vengono a mancare, non necessariamente le proprietà degli stimatori di massima verosimiglianza debbano peggiorare: possono addirittura essere superiori, sia come modalità che come velocità di convergenza.*

13 Statistiche sufficienti

Una domanda naturale che possiamo porci, specialmente in un momento in cui masse enormi di dati sono a disposizione, è la seguente: posso comprimere un campione di dati osservati senza perdere informazione sul parametro che mi interessa? Questa domanda ci porta alla nozione di statistiche sufficienti.

Definition 116 *(Statistiche sufficienti) Una statistica $T_n = T(X_1, ..., X_n)$ si dice sufficiente per il parametro θ se la distribuzione del campione condizionata a T_n non dipende da θ.*

In altre parole, la legge $p(X_1, ..., X_n|T_n)$ non dipende dal parametro. In questa sezione, supponiamo sempre per semplicità di avere a che fare con variabili aleatorie discrete; in questo ambito, è immediato verificare che conoscendo solo la statistica sufficiente possiamo generare un nuovo campione uguale in distribuzione all'originale. Abbiamo infatti, per qualsiasi scelta di valori $(x_1, x_2, ..., x_n)$

$$\Pr(X_1 = x_1, ..., X_n = x_n)$$

$$\begin{aligned}
&= \Pr(X_1 = x_1, ..., X_n = x_n \cap T(X_1, ..., X_n) = T(x_1, ..., x_n)) \\
&= \Pr(X_1 = x_1, ..., X_n = x_n | T(X_1, ..., X_n) = T(x_1, ..., x_n)) \Pr(T(X_1, ..., X_n) = T(x_1, ..., x_n)) \\
&= \Pr(Y_1 = x_1, ..., Y_n = x_n | T(Y_1, ..., Y_n) = T(x_1, ..., x_n)) \Pr(T(X_1, ..., X_n) = T(x_1, ..., x_n)) \\
&= \Pr(Y_1 = x_1, ..., Y_n = x_n)
\end{aligned}$$

dove le Y_i sono generate dalla distribuzione condizionata delle X_i a T; poichè l'uguaglianza vale per tutte le scelte di $(x_1, ..., x_n)$, abbiamo quindi dimostrato che

$$(Y_1, ..., Y_n) \stackrel{d}{=} (X_1, ..., X_n) \ .$$

Si noti che nella dimostrazione (primo passaggio) abbiamo anche utilizzato il fatto che l'evento $X_1 = x_1, ..., X_n = x_n$ è contenuto nell'evento $T(X_1, ..., X_n) = T(x_1, ..., x_n)$, e pertanto la probabilità della loro intersezione è uguale alla probabilità di $X_1 = x_1, ..., X_n = x_n$. Più in generale, denotiamo $q(T(X_1, ..., X_n); \theta)$ la funzione di probabilità della statistica sufficiente T; dalla definizione di probabilità condizionata è immediato verificare che T è sufficiente se e solo se $p(X_1, ..., X_n; \theta)/q(T(X_1, ..., X_n; \theta)$ non dipende da θ.

Example 117 *Si consideri un campione IID di variabili Bernoulliane $X_i \sim Ber(\theta)$. La loro legge congiunta è data da*

$$p(X_1, ..., X_n; \theta) = \theta^{\sum_{i=1}^n X_i} (1 - \theta)^{n - \sum_{i=1}^n X_i}.$$

Verifichiamo che $T_n := \sum_{i=1}^n X_i$ sia una statistica sufficiente; la sua funzione di probabilità è una binomiale, data quindi da

$$q(T_n(X_1, ..., X_n); \theta) = \binom{n}{\sum_{i=1}^n X_i} \theta^{\sum_{i=1}^n X_i} (1 - \theta)^{n - \sum_{i=1}^n X_i}$$

e quindi

$$\frac{p(X_1, ..., X_n; \theta)}{q(T_n(X_1, ..., X_n); \theta)} = \frac{1}{\binom{n}{\sum_{i=1}^n X_i}} \ ,$$

che è indipendente da θ.

Example 118 *Si consideri un campione IID di variabili Gaussiane $X_i \sim N(\theta, 1)$. La loro legge congiunta è data da*

$$p(X_1, ..., X_n; \mu) = \frac{1}{(2\pi)^{n/2}} \exp(-\frac{1}{2} \sum_{i=1}^{n} (X_i - \mu)^2) \ ;$$

consideriamo ora la statistica $\overline{X}_n = n^{-1} \sum_{i=1}^{n} X_i$, con legge $N(\mu, \frac{1}{n})$, cioè

$$q(\overline{X}_n; \mu) = \frac{\sqrt{n}}{(2\pi)^{1/2}} \exp(-\frac{n}{2} (\overline{X}_n - \mu)^2) \ .$$

E' facile verificare che

$$\sum_{i=1}^{n} (X_i - \mu)^2 = \sum_{i=1}^{n} (X_i - \overline{X}_n + \overline{X}_n - \mu)^2 = \sum_{i=1}^{n} (X_i - \overline{X}_n)^2 + n(\overline{X}_n - \mu)^2 \ ,$$

da cui

$$\begin{aligned} \frac{p(X_1, ..., X_n; \mu)}{q(\overline{X}_n; \mu)} &= Const \times \frac{\exp(-\frac{1}{2} \sum_{i=1}^{n} (X_i - \mu)^2)}{\exp(-\frac{n}{2} (\overline{X}_n - \mu)^2)} \\ &= Const \times \exp(-\frac{1}{2} \sum_{i=1}^{n} (X_i - \overline{X}_n)^2) \ , \end{aligned}$$

e la dimostrazione della sufficienza è completata.

A metà del ventesimo secolo Halmos e Savage hanno dato una caratterizzazione della suffiicenza leggermente più generale, come segue.

Theorem 119 *(Criterio di Fattorizzazione di Halmos e Savage) Una statistica è sufficiente se e solo se la funzione di probabilità (o di densità) di $X_1, ..., X_n$ si fattorizza come*

$$p(X_1, ..., X_n; \theta) = g(T(X_1, ..., X_n; \theta)) h(X_1, ..., X_n)$$

dove la funzione $h(.)$ non dipende da θ.

Proof. L'implicazione sufficienza $\Longrightarrow$ fattorizzazione è ovvia: basta prendere per $g(T; \theta) = q(T; \theta)$ (la legge di q) e h la legge delle X_i condizionata a T, che sappiamo essere indipendente da θ. L'interesse del Teorema è quindi nella seconda parte, cioè nel caso in cui la fattorizzazione non corrisponda necessariamente alla coppia legge di T/ legge di $X_1, ..., X_n$ condizionata a T. Come sempre in questa sezione ci limitiamo il caso discreto, e partizioniamo il dominio di $X_1, ..., X_n$ in classi di equivalenza corrispondenti ai valori della statistica sufficiente; definiamo cioè i sottoinsiemi di $\mathbb{R}^n$

$$A_{T(x_1, ..., x_n)} = \{(x'_1, ..., x'_n) : T(x'_1, ..., x'_n) = T(x_1, ..., x_n)\} \ .$$

Andiamo a calcolare, per $x_1, ..., x_n$ qualsiasi ma fissati

$$\begin{aligned}
\frac{p(x_1,...,x_n;\theta)}{q(T(x_1,...,x_n);\theta)} &= \frac{g(T(x_1,...,x_n;\theta))h(x_1,...,x_n)}{q(T(x_1,...,x_n);\theta)} \\
&= \frac{g(T(x_1,...,x_n;\theta))h(x_1,...,x_n)}{\sum_{(x'_1,...,x'_n)\in A_{T(x_1,...,x_n)}} p(x'_1,...,x'_n;\theta)} \\
&= \frac{g(T(x_1,...,x_n;\theta))h(x_1,...,x_n)}{\sum_{(x'_1,...,x'_n)\in A_{T(x_1,...,x_n)}} g(T(x_1,...,x_n;\theta))h(x'_1,...,x'_n)} \\
&= \frac{h(x_1,...,x_n)}{\sum_{(x'_1,...,x'_n)\in A_{T(x_1,...,x_n)}} h(x'_1,...,x'_n)} \ ;
\end{aligned}$$

l'ultima frazione non dipende da θ, il che è sufficiente a concludere la dimostrazione della sufficienza. ■

13.1 Il Teorema di Rao-Blackwell

L'idea intutiva dietro le statistiche sufficienti è quella di condensare l'informazione del campione aleatorio conservando quello che è utile per risalire al valore "vero" del parametro. Sembra quindi inuitivo che gli stimatori efficienti debbano essere costruiti solo a partire da statistiche sufficienti. Questa idea è resa rigorosa dal teorema di Rao-Blackwell; prima di enunciarlo, dobbiamo ricordare alcune nozioni sul valor medio condizionato, restringendoci solamente al caso discreto.

Siano X, Y valori aleatorie discrete con momento secondo finito e funzione di probabilità congiunta $P_{XY}(x,y)$; la densità di X condizionata a $Y = y$ è data da

$$p_{X|Y}(x,y) = \begin{cases} \frac{p_{XY}(x,y)}{p_X(x)} \text{ , se } p_X(x) > 0 \\ 0 \text{ altrimenti} \end{cases} ,$$

ed il valor medio condizionato a $Y = y$ è naturalmente

$$E[X|Y=y] = \sum_{x\in Dom(X)} x\frac{p_{XY}(x,y)}{p_Y(y)} \ .$$

Se evitiamo di fissare il valore della variabile aleatoria Y, avremo ora una nuova variabile aleatoria (funzione di Y) della forma $\psi(Y) = E[X|Y]$, che ad ogni valore di Y associa il valore del valor medio condizionato corrispondente a quel valore

$$E[X|Y] = \sum_{x\in Dom(X)} x\frac{p_{XY}(x,Y)}{p_Y(Y)} = \sum_{x\in Dom(X)} x p_{X|Y}(x|Y) \ ;$$

in altre parole, $\psi(Y)$ assume il valore $\psi(y) = E[X|Y=y]$ con probabilità $p_Y(y)$ (la legge marginale di Y. E' immediato verificare che (la cosiddetta legge del valor medio iterato)

$$E[E[X|Y]] = E[X] \ ;$$

infatti

$$\begin{aligned} E[E[X|Y]] &= \sum_y E[X|Y=y]p_Y(y) = \sum_y \sum_{x\in Dom(X)} x\frac{p_{XY}(x,y)}{p_Y(x)}p_Y(y) \\ &= \sum_{x\in Dom(X)} x\sum_y p_{XY}(x,y) \\ &= \sum_{x\in Dom(X)} xp_X(x) = E[X] \ . \end{aligned}$$

Allo stesso modo possiamo definire la variabile aleatoria (funzione di Y)

$$Var[X|Y] = \sum_{x\in Dom(X)} (x - E[X|Y])^2 p_{X|Y}(x|Y) \ ,$$

che corrisponde alla varianza condizionata di X, in funzione del valore aleatorio di Y. Vale il seguete risultato.

Lemma 120 *Siano X, Y variabili aleatorie di quadrato integrabile definite sullo stesso spazio di probabilità. Abbiamo*

$$Var[X] = E[Var[X|Y]] + Var[E[X|Y]] \ .$$

Proof. Abbiamo che

$$\begin{aligned} Var[X] &= E[(X - E[X])^2] \\ &= E[(X - E[X|Y] + E[X|Y] - E[X])^2] \\ &= E[(X - E[X|Y])^2] + E[(E[X|Y] - E[X])^2] \\ &\quad +2E[(X - E[X|Y])(E[X|Y] - E[X])] \ . \end{aligned}$$

Ora per definizione abbiamo

$$E[(E[X|Y] - E[X])^2] = E_Y[(E[X|Y] - E[X])^2] = Var[E[X|Y]] \ ;$$

inoltre

$$\begin{aligned} E[(X - E[X|Y])(E[X|Y] - E[X])] &= E_Y[E[(X - E[X|Y])(E[X|Y] - E[X])|Y]] \\ &= E_Y[(E[X|Y] - E[X])E[(X - E[X|Y])|Y]] \\ &= 0 \ , \end{aligned}$$

perché $E[(X - E[X|Y])|Y] = 0$. Infine

$$E[(X - E[X|Y])^2] = E[E[(X - E[X|Y])^2|Y]] = E[Var[X|Y]] \ ,$$

e il Lemma è dimostrato. ■

Possiamo finalmente arrivare al risultato principale, che ci garantisce che condizionando uno stimatore non distorto su una statistica sufficiente se ne ottiene un altro ancora non distorto e con varianza non maggiore.

Theorem 121 *(Rao-Blackwell) Sia $W_n = W_n(X_1, ..., X_n)$ uno stimatore non distorto e con varianza finita del parametro θ, e sia $T_n = T(X_1, ..., X_n)$ una statistica suffiicente per θ. Definiamo $\psi(T) = E[W_n|T]$; allora abbiamo*

$$E[\psi(T)] = \theta \text{ e } Var[\psi(T)] \leq Var[W_n] \ .$$

Proof. La dimostrazione è una applicazione immediata dei risultati precedenti sul valor medio condizionato. In particolare

$$\begin{aligned} E[\psi(T)] &= E[E[W_n|T]] = \theta \ , \\ Var[W_n] &= E[Var[W_n|T]] + Var[E[W_n|T]] \\ &\geq Var[E[W_n|T]] = Var[\psi(T)] \ . \end{aligned}$$

■

Remark 122 *Dove abbiamo usato nella dimostrazione precedente la sufficienza? Apparentemente non gioca alcun ruolo, ma in realtà ci garantisce che il valor medio condizionato non dipenda dal parametro, e quindi che $\psi(T)$ sia effettivamente uno stimatore. Ad esempio consideriamo un campione di sole due osservazioni X_1, X_2 Gaussiane indipendenti con valor medio μ e varianza 1, e consideriamo lo stimatore $\overline{X}_2 = \frac{1}{2}(X_1 + X_2)$. Consideriamo*

$$\psi(X_1) := E[\overline{X}_2|X_1] = \frac{X_1 + \mu}{2} \ .$$

Questa variabile aleatoria ha valor medio μ e varianza $\frac{1}{4} < Var(\overline{X}_2) = \frac{1}{2}$; non si tratta però di uno stimatore, perchè il suo valore non può essere determinato senza conoscere il valore del parametro μ.

14 Stimatori UMVUE, Completezza

.....

15 Stimatori Bayesiani

Ricordiamo innanzitutto la *Formula di Bayes:*

Theorem 123 *(Bayes) Siano $H_1, ...H_m$ eventi disgiunti ed esaustivi, cioè $H_i \cap H_j = \varnothing$ per ogni $i \neq j$ e $\cup_{i=1}^m = \Omega$. Sia inoltre $E \in \Im$ un evento con probabilità strettamente positiva; allora*

$$\Pr(H_i|E) = \frac{\Pr(E|H_i)\Pr(H_i)}{\sum_{j=1}^m \Pr(E|H_j)\Pr(H_j)} \ .$$

La derivazione della formula di Bayes è matematicamente banale (si riduce essenzialmente a ricordare che $\Pr(E) = \sum_{j=1}^m \Pr(E|H_j)\Pr(H_j)$; l'interpretazione però è molto importante, perché permette di combinare in modo matematicamente la probabilità a priori di m cause disgiunte H_j e l'evidenza empirica sul fatto che E si sia verificato.

Example 124 *...*

L'approccio Bayesiano all'inferenza statistica è profondamente diverso da quello che abbiamo seguito finora. L'idea di fondo è che non esista un "vero" parametro da stimare $\theta = \theta_0$, ma che il parametro sia esso stesso una variabile (o un vettore) aleatorio la cui distribuzione, che rappresenta il nostro stato di conoscenza, viene aggiornata tramite la formula di Bayes alla luce delle osservazioni. In altre parole, oltre alla legge delle osservazioni $f(X_1, ..., X_n|\theta)$ dobbiamo supporre di conoscere la legge $\pi(\theta)$ del parametro prima di aver effettuato osservazioni. Si noti come abbiamo scritto $f(X_1, ..., X_n|\theta)$ invece di $f(X_1, ..., X_n; \theta)$, perché ora ha senso parlare della legge di $X_1, ..., X_n$ condizionatamente al valore θ del parametro. L'oggetto centrale dell'inferenza diviene quindi la legge a posteriori, che attraverso la formula di Bayes è data da

$$\pi(\theta|X_1, ..., X_n) = \frac{f(X_1, ..., X_n|\theta)\pi(\theta)}{\int_\Theta f(X_1, ..., X_n|\theta)\pi(\theta)d\theta} \ ,$$

ed analogamente nel caso discreto.

Example 125 *Consideriamo $X_1, ..., X_n$ variabili Bernoulliane di parametro p; per quest'ultimo, assumiamo che abbia una distribuzione a priori di tipo Beta con parametri α, β, cioè*

$$\pi(p) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1}, \ p \in [0,1] \ .$$

Ricordiamo innanzitutto i valori di valor medio e varianza

$$E[p] = \frac{\alpha}{\alpha+\beta} \ , \ Var[p] = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \ ;$$

infatti

$$\begin{aligned} E[p] &= \int_0^1 p \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} dp \\ &= \frac{\Gamma(\alpha+1)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\alpha+\beta+1)} \int_0^1 \frac{\Gamma(\alpha+\beta+1)}{\Gamma(\alpha+1)\Gamma(\beta)} p^{\alpha}(1-p)^{\beta-1} dp \\ &= \frac{\alpha}{\alpha+\beta} \end{aligned}$$

e

$$\begin{aligned} E[p^2] &= \int_0^1 p^2 \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} dp \\ &= \frac{\Gamma(\alpha+2)\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\alpha+\beta+2)} \int_0^1 \frac{\Gamma(\alpha+\beta+2)}{\Gamma(\alpha+2)\Gamma(\beta)} p^{\alpha}(1-p)^{\beta-1} dp \\ &= \frac{\alpha(\alpha+1)}{(\alpha+\beta+1)(\alpha+\beta)} \ , \end{aligned}$$

$$
\begin{aligned}
Var[p] &= \frac{\alpha(\alpha+1)}{(\alpha+\beta+1)(\alpha+\beta)} - \frac{\alpha^2}{(\alpha+\beta)^2} \\
&= \frac{\alpha(\alpha+1)(\alpha+\beta) - \alpha^2(\alpha+\beta+1)}{(\alpha+\beta+1)(\alpha+\beta)^2} = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)} \ .
\end{aligned}
$$

Abbiamo inoltre, scrivendo $y = \sum_{i=1}^{n} X_i$

$$
\begin{aligned}
f(y|p)\pi(p) &= \binom{n}{y} p^y (1-p)^{n-y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{\alpha-1}(1-p)^{\beta-1} \\
&= \binom{n}{y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{y+\alpha-1}(1-p)^{n-y+\beta-1}.
\end{aligned}
$$

La marginale a denominatore si ottiene come segue:

$$
\begin{aligned}
f(y) &= \int_0^1 \binom{n}{y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{y+\alpha-1}(1-p)^{n-y+\beta-1} dp \\
&= \binom{n}{y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(y+\alpha)\Gamma(n-y+\beta)}{\Gamma(n+\alpha+\beta)} \ .
\end{aligned}
$$

Infine la distribuzione a posteriori è data da

$$
\begin{aligned}
\pi(p|y) &= \frac{\binom{n}{y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} p^{y+\alpha-1}(1-p)^{n-y+\beta-1}}{\binom{n}{y} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(y+\alpha)\Gamma(n-y+\beta)}{\Gamma(n+\alpha+\beta)}} \\
&= \frac{\Gamma(n+\alpha+\beta)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)} p^{y+\alpha-1}(1-p)^{n-y+\beta-1},
\end{aligned}
$$

cioè è ancora una beta, con parametri aggiornati. Quando la distribuzioni a posteriori assume la stessa forma dell'a priori si parla di ***leggi coniugate****.*

Remark 126 *Una interpretazione rigorosa dell'approccio Bayesiano dovrebbe concludersi con la derivazione della distribuzione a posteriori: il calcolo di uno stimatore "puntuale"non ha strettamente senso, visto che il paramrto non ha un singolo valore. In pratica però il calcolo degli stimatori Bayesiani si conclude molto spesso con la derivazione di un singolo valore di sintesi, come ad esempio il valore che massimizza la distribuzione a posteriori o ancora più spesso il valore medio a posteriori. Nel caso della binomiale con a priori beta otteniamo*

$$
\begin{aligned}
\widehat{p}_{Bayes} &= \int_0^1 p \frac{\Gamma(n+\alpha+\beta)}{\Gamma(y+\alpha)\Gamma(n-y+\beta)} p^{y+\alpha-1}(1-p)^{n-y+\beta-1} dp \\
&= \frac{y+\alpha}{n+\alpha+\beta} = \frac{y}{n}\frac{n}{n+\alpha+\beta} + \frac{\alpha}{\alpha+\beta}\frac{\alpha+\beta}{n+\alpha+\beta} \ .
\end{aligned}
$$

L'ultima espressione è illuminante, perchè rappresenta lo "stimatore Bayesiano" come una media ponderata di due elementi: lo stimatore classico di massima verosimiglianza (in questo caso, la semplice media aritmetica) ed il valore atteso a priori:

$$
\frac{y}{n} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}_n \ , \ \frac{\alpha}{\alpha+\beta} \ .
$$

Il peso dello stimatore di massima verosimiglianza converge ad 1 quando la dimensione del campione n diverge all'infinito; intuitvamente, le nostre opinioni a priori sono schiacciate dalla forza dell'evidenza empirica.

Exercise 127 *Sia X una Gaussiana con valor medio θ e varianza σ^2, con θ essa stessa Gaussiana di parametri μ, τ^2. Abbiamo che*

$$\pi(\theta|X) \sim N(\frac{\tau^2}{\tau^2+\sigma^2}X + \frac{\sigma^2}{\sigma^2+\tau^2}\mu, \frac{\sigma^2\tau^2}{\sigma^2+\tau^2}) \ .$$

16 Il Modello lineare multivariato

Nella pratica, è estremamente comune la situazione in cui si cerchi di stimare la relazione esistente tra una particolare variabile (che chiameremo genericamente Y) ed un gruppo di altre (che chiameremo genericamente $X_1, ..., X_k$), dove queste ultime sono considerate esplicative del comportamento della Y. Si tratta di un caso particolare di un ambito di ricerca al momento estremamente attivo, e che viene spesso indicato come Supervise Learning; in generale, lo studio di relazioni come $Y = f(X)$ va anche sotto il nome di Machine Learning o Statistical Learning. In questo corso, ci limiteremo a trattare il caso più comune, che è anche il più semplice; quello in cui assumiamo che $f(.)$ abbia una forma lineare in un vettore di parametri β. In particolare, supponiamo che valga la seguente relazione, per $i = 1, 2, ..., n$:

$$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + ... + x_{ki}\beta_k + \varepsilon_i \ ,$$

dove le variabili ε_i sono considerate dei "residui" o degli "errori" e catturano tutto quello che non è spiegato dalla relazione lineare tra le y e le x_i. In forma matriciale, possiamo scrivere

$$Y = X\beta + \varepsilon \ , \tag{1}$$

dove Y è un vettore $n \times 1$ della forma $Y = (y_1, ..., y_n)^T$, X è una matrice $n \times k$ le cui colonne sono costituite da $(x_{j1}, ..., x_{jn})^T$, e ε è un vettore di residui; iniziamo supponendo che si tratti di residui Gaussiani, con valor medio nullo e matrice di varianza/covarianza $E[\varepsilon\varepsilon^T] = \Omega$. Comunemente la prima colonna di X viene scelta come il vettore di costanti $(1, 1, ..., 1)$; così ad esempio per $k = 2$ abbiamo

$$y_i = \beta_1 + \beta_2 x_i + \varepsilon_i \ .$$

Abbiamo bisogno di una condizione ulteriore sulle variabili X :

Condition 128 *Le variabili X sono deterministiche ed il rango della matrice è esattamente pari a k.*

Ambedue le condizioni possono essere rimosse ed effettivamente lo sono nella ricerca più recente; si tratta però di ipotesi semplificatrici che costituiscono un punto di partenza naturale. Si noti che questa condizione implica immediatamente $k \leq n$; questa ipotesi in particolare viene abbandonata negli studi più recenti (high-dimensional statistics/big data).

Proposition 129 *Lo stimatore di massima verosimiglianza nel modello 1 ha la seguente forma:*

$$\begin{aligned}\widehat{\beta}_{MLE} &= (X^T X)^{-1} X^T Y \\ \widehat{\sigma}^2_{MLE} &= \frac{1}{n}(\widehat{\varepsilon}^T \widehat{\varepsilon}) \ , \ \widehat{\varepsilon} = Y - \widehat{Y} = Y - X\widehat{\beta} \ .\end{aligned}$$

Si ha inoltre che

$$\begin{aligned}&\widehat{\beta}_{MLE} \stackrel{d}{=} N(\beta, \sigma^2 (X^T X)^{-1}) \ , \\ &n \times \frac{\widehat{\sigma}^2_{MLE}}{\sigma^2} \stackrel{d}{=} \chi^2_{n-k} \ .\end{aligned}$$

Proof. Notiamo innanzitutto che la funzione di verosimiglianza prende la forma

$$\begin{aligned}L(\beta, \sigma^2; Y, X) &= \frac{1}{(2\pi)^{n/2}} \frac{1}{(\sigma^2)^{n/2}} \exp\left\{ -\frac{1}{2\sigma^2}(Y - X\beta)^T (Y - X\beta) \right\} \ , \\ \log L(\beta, \sigma^2; Y, X) &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}(Y - X\beta)^T (Y - X\beta) \ .\end{aligned}$$

Il calcolo del gradiente ci dà facilmente

$$\begin{aligned}\nabla_\beta \log L(\beta, \sigma^2; Y, X) \\ \frac{\partial \log L(\beta, \sigma^2; Y, X)}{\partial \sigma^2}\end{aligned} \quad \begin{aligned}&= \begin{array}{c} \nabla_\beta (Y^T Y - \beta^T X^T Y - Y^T X\beta + \beta^T X^T X\beta \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(Y - X\beta)^T (Y - X\beta) \end{array} \\ &= \begin{array}{c} -2X^T Y + 2X^T X\beta \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(Y - X\beta)^T (Y - X\beta) \end{array}\end{aligned}$$

and setting these quantities to zero leads to the previous expressions for $\widehat{\beta}_{MLE}, \widehat{\sigma}^2_{MLE}$. It is also immediate to see that

$$\begin{aligned}\widehat{\beta}_{MLE} &= (X^T X)^{-1} X^T Y \\ &= (X^T X)^{-1} X^T (X\beta + \varepsilon) \\ &= \beta + (X^T X)^{-1} X^T \varepsilon \ ,\end{aligned}$$

whence

$$\begin{aligned}E[(\widehat{\beta}_{MLE} - \beta)(\widehat{\beta}_{MLE} - \beta)^T] &= E[(X^T X)^{-1} X^T \varepsilon ((X^T X)^{-1} X^T \varepsilon)^T] \\ &= E[(X^T X)^{-1} X^T \varepsilon\varepsilon^T X (X^T X)^{-1}] \\ &= (X^T X)^{-1} X^T E[\varepsilon\varepsilon^T] X (X^T X)^{-1} \\ &= (X^T X)^{-1} X^T \sigma^2 I_n X (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} \\ &= \sigma^2 (X^T X)^{-1} \ .\end{aligned}$$

La dimostrazione del risultato sullo stimatore della varianza richiede alcune ulteriori discussioni e sarà riportata più avanti. ■

Remark 130 *Lo stimatore $\widehat{\beta}_{MLE}$ è anche noto come OLS (Ordinary Least Squares).*

16.1 Il rapporto con i teoremi di proiezione

Lo stimatore di massima verosimiglianza nel modello lineare multivariato si presta ad una importante interpretazione usando gli strumenti dell'algebra lineare. In particolare, notiamo che il valor medio di Y dato X è dato da $E[Y] = X\beta$; il valore di β è ignoto, ma è naturale definire il valore $\widehat{Y} := X\widehat{\beta}$ come il valore "previsto" per il vettore Y sulla base delle stime $\widehat{\beta}$ e del valore dei regressori X. Analogamente, il vettore degli "errori" $\varepsilon = Y - X\beta$ si può stimare come $\widehat{\varepsilon} = Y - X\widehat{\beta}$. Notiamo ora che

$$\begin{aligned} \widehat{Y} &: = X\widehat{\beta} = X(X^T X)^{-1} X^T Y = P_X Y \ , \\ P_X &: = X(X^T X)^{-1} X^T \ . \end{aligned}$$

E' immediato verificare che P_X è una matrice di proiezione, cioè è simmetrica ed idempotente; infatti

$$P_X^2 = X(X^T X)^{-1} X^T X (X^T X)^{-1} X^T = X(X^T X)^{-1} X^T \ .$$

In particolare, l'azione della matrice P_X (che ha dimensioni $n \times n$) corrisponde a proiettare il vettore Y sullo spazio vettoriale di dimensione k generato dalle colonne di X (ricordiamo che queste colonne sono linearmente indipendenti per ipotesi). Analogamente abbiamo

$$\begin{aligned} \widehat{\varepsilon} &= Y - X\widehat{\beta} = M_X Y \ , \\ M_X &= I - P_X = I - X(X^T X)^{-1} X^T \ . \end{aligned}$$

Anche M_X è simmetrica ed idempotente:

$$M_X^2 = (I - P_X)^2 = I - 2P_X + P_X^2 = I - 2P_X + P_X = M_X \ .$$

In particolare, l'azione di M_X consiste nel proiettare Y nello spazio ortogonale a quello generato dalle colonne di X. Questo ha alcune conseguenze importanti; si ha infatti

$$M_X P_X = P_X M_X = \mathbf{0} \ ,$$

dove intendiamo con $\mathbf{0}$ la matrice $n \times n$ costituita da tutti zeri. Come ulteriore conseguenza, notiamo che

$$X^T \widehat{\varepsilon} = 0 \ ,$$

ed in particolare il vettore dei residui stimati è ortogonale a qualsiasi vettore che giaccia nello spazio generato dalle colonne di X.

Remark 131 *Supponiamo che la matrice X contenga un termine costante, cioè una colonna della forma $e = (1, ..., 1)^T$. Allora abbiamo $e^T \widehat{\varepsilon} = 0$, cioè $\sum_{i=1}^n \widehat{\varepsilon}_i = 0$; in altre parole, la somma dei residui di regressione è sempre nulla se tra i regressori è incluso un termine costante.*

Remark 132 *Il fatto che lo stimatore di massima verosimiglianza nel modello lineare multivariato coincida con la soluzione di un problema puramente geometrico di proiezione su sottospazi vettoriali è assolutamente degno di nota. Si tratta dell'ennesima sorprendente proprietà della legge Gaussiana.*

Possiamo ora enunciare le seguenti ulteriori proprietà delle matrici P_X, M_X.

Lemma 133 *Le matrici P_X,M_X hanno tutti autovalori pari a zero o uno e rango $k, n-k$, rispettivamente.*

Proof. Poichè le matrici sono reali e simmetriche, possiamo diagonalizzarle come

$$\begin{aligned} P_X &= Q\Lambda Q^T \text{ , con } QQ^T = Q^TQ = I_n \text{ ,} \\ \Lambda &= \begin{pmatrix} \lambda_1 & 0 & ... & 0 \\ 0 & \lambda_2 & 0 & ... \\ ... & 0 & ... & 0 \\ 0 & ... & 0 & \lambda_n \end{pmatrix} . \end{aligned}$$

Si ha allora

$$P_X^2 = Q\Lambda Q^TQ\Lambda Q^T = Q\Lambda^2 Q^T = Q\Lambda Q^T \text{ ,}$$

da cui segue che necessariamente $\lambda_i = 0$ o 1, per $i = 1, 2, ..., n$; ragionamento identico si applica a M_X. Per quello che riguarda il rango, notiamo che esso eguaglia il numero di autovalori diversi da zero (con molteplicità), quindi nel caso di matrici di proiezioni la traccia (cioè la sommadi autovalori, uguale al numero di quelli che valgono 1). Ricordando che $Tr(AB) = Tr(BA)$, possiamo scrivere

$$Tr(P_X) = Tr(X(X^TX)^{-1}X^T) = Tr((X^TX)^{-1}X^TX) = Tr(I_k) = k = Rg(P_X) \text{ .}$$

La dimostrazione per M_X è identica. ■

Siamo ancora in grado di concludere la dimostrazionne sul comportamento dello stimatore di massima verosimiglianza della varianza. Ricordiamo innanzitutto la densità delle variabili aleatorie $\Gamma(\alpha, \beta)$:

$$f_\Gamma(t) = \frac{1}{\Gamma(\alpha)\beta^{\alpha-1}} t^{\alpha-1} \exp(-\frac{t}{\beta}) \mathbb{I}_{[0,\infty)}(t) \text{ ,}$$

a cui corrisponde un valor medio pari a $\alpha\beta$ ed una varianza pari a $\alpha\beta^2$. Ricordiamo che la variabile chi-quadro con p gradi di libertà corrisponde ad una Gamma di parametri $\alpha = \frac{p}{2}$, $\beta = 2$, e quindi con valor medio p e varianza $2p$. La proprietà fondamentale dellle chi-quadro e che esse sono uguali in distribuzioni alla somma di p Gaussiane standard indipendneti elevate al quadrato:

$$\chi_p^2 \stackrel{d}{=} Z_1^2 + ... + Z_p^2 \text{ , } Z_i \stackrel{d}{=} N(0,1) \text{ .}$$

Poniamo ora senza perdita di generalita $\sigma^2 = 1$, e notiamo che

$$\widehat{\varepsilon}^T\widehat{\varepsilon} = (M_X\varepsilon)^TM_X\varepsilon = \varepsilon^TM_X^2\varepsilon = \varepsilon^TM_X\varepsilon \text{ .}$$

Ricordiamo ora la diagonalizzazione $M_X = Q\widetilde{\Lambda}Q^T$, dove $\widetilde{\Lambda}$ ha $n-k$ elementi pari ad uno sulla diagonale principale ed i restanti tutti nulli. Otteniamo quindi

$$\begin{aligned}
\varepsilon^T M_X \varepsilon &= \varepsilon^T Q\widetilde{\Lambda}Q^T\varepsilon \\
&= u^T\widetilde{\Lambda}u \\
&= \sum_{i=1}^{n}\widetilde{\lambda}_i u_i^2 \\
&= \sum_{i=1}^{n-k}\widetilde{\lambda}_i u_i^2 \text{ , con } u := Q^T\varepsilon \stackrel{d}{=} N(0, I_n) \ ,
\end{aligned}$$

perchè la legge Gaussiana standard è invariante per rotazioni. L'ultimo termine ha una legge chi-quadro per quanto scritto sopra, il che completa la dimostrazione.

Example 134 *Si noti che anche il problema della stima del valor medio per variabili indipendenti $Y_i \sim N(\mu, \sigma^2)$ può essere riletta in termini di regressione. Possiamo infatti scrivere*

$$Y = \begin{pmatrix} 1 \\ 1 \\ ... \\ ... \\ 1 \end{pmatrix} \mu + \varepsilon \ , \ \varepsilon = Y - \begin{pmatrix} 1 \\ 1 \\ ... \\ ... \\ 1 \end{pmatrix} \mu \ ;$$

lo stimatore diventa

$$\widehat{\mu} = (e^T e)^{-1} e^T Y = (\sum_{i=1}^{n} 1)^{-1} \times \sum_{i=1}^{n}(1 \cdot Y_i) = \overline{Y}_n \ ,$$

con

$$\widehat{\mu} \sim N(\mu, \sigma^2(e^T e)^{-1}) = N(\mu, \frac{\sigma^2}{n}) \ .$$

Nel caso $k = 2$ (con costante) il modello diviene

$$Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$$

ed un poco di algebra mostra che

$$\begin{aligned}
\widehat{\beta}_1 &= \overline{Y}_n - \widehat{\beta}_2\overline{X}_n \ , \\
\widehat{\beta}_2 &= \frac{\sum_{i=1}^{n}(X_i - \overline{X}_n)(Y_i - \overline{Y}_n)}{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2} \ ,
\end{aligned}$$

con

$$\widehat{\beta}_2 \sim N(\beta_2, \frac{\sigma^2}{\sum_{i=1}^{n}(X_i - \overline{X}_n)^2}) \ .$$

16.2 Lo stimatore GLS (Generalized Least Squares)

Possiamo ora generalizzare il modello che abbiamo studiato sinora, immaginando che i residui ε abbiano una struttura di dipendenza molto più complessa di variabili indipendenti. In particolare, ipotizziamo che $E[\varepsilon\varepsilon^T] = \Omega$ con matrice positiva definita di rango (pieno) n. La funzione di verosimiglianza prende la forma

$$L(\beta; Y, X) = \frac{1}{(2\pi)^n} \frac{1}{\sqrt{\det(\Omega)}} \exp\left\{ -\frac{1}{2}(Y - X\beta)^T \Omega^{-1/2} (Y - X\beta)^T \right\} .$$

Potremmo procedere come nel Caso ordinario, ma c'è una strategia più semplice. La matrice Ω si può diagonalizzare come $\Omega = Q\Lambda_\Omega Q^T$, per qualche matrice ortonormale Q che non corrisponde a quelle che abbiamo introdotto prima. Possiamo definire quindi $\Omega^{-1/2} = Q\Lambda_\Omega^{-1/2} Q^T$, con l'ovvia proprietà che

$$\begin{aligned} \Omega^{-1/2}\Omega\Omega^{-1/2} &= Q\Lambda_\Omega^{-1/2} Q^T Q\Lambda_\Omega Q^T Q\Lambda_\Omega^{-1/2} Q^T \\ &= Q\Lambda_\Omega^{-1/2}\Lambda_\Omega\Lambda_\Omega^{-1/2} Q^T \\ &= QQ^T \\ &= I_n . \end{aligned}$$

Possiamo dunque definire il vettore $\widetilde{\varepsilon} := \Omega^{-1/2}\varepsilon$, che (si verifica facilmente) ha matrice di varianza/covarianza pari all'identità. Quindi

$$Y = X\beta + \varepsilon \mapsto \Omega^{-1/2} Y = \Omega^{-1/2} X\beta + \Omega^{-1/2}\varepsilon \mapsto \widetilde{Y} = \widetilde{X}\beta + \widetilde{\varepsilon} ,$$

con

$$\widetilde{Y} := \Omega^{-1/2} Y , \widetilde{X} := \Omega^{-1/2} X , \widetilde{\varepsilon} := \Omega^{-1/2}\varepsilon .$$

Lo stimatore diventa quindi

$$\widetilde{\beta}_{GLS} = (\widetilde{X}^T\widetilde{X})^{-1}\widetilde{X}\widetilde{Y} = (X^T\Omega^{-1}X)^{-1} X\Omega^{-1} Y ,$$

che è uno stimatore non distorto con legge Gaussiana e matrice di varianza e covarianza

$$E[(\widetilde{\beta}_{GLS} - \beta)(\widetilde{\beta}_{GLS} - \beta)^T] = \Omega^{-1}.$$

Remark 135 *Lo stimatore GLS può essere visto come i coefficienti di proiezione su un sottospazio generato dalle colonen di X quando la proiezione sia effettuata con una metrica Riemanniana indotta dalla matrice positive definita Ω^{-1}.*

Exercise 136 *Calcolare la forma dello stimatore GLS quando la matrice Ω sia nulla al di fuori della diagonale, nel caso $k = 1$.*

Exercise 137 *Calcolare la matrice di varianza e covarianza dello stimatore OLS (non GLS) quando la matrice di varianza di ε sia Ω non diagonale (cioè quando lo stimatore OLS sia applicato assumendo ipotesi in realtà non soddisfatte).*

Part I
Appendice: richiami di calcolo delle probabilità

17 Richiami di Calcolo delle Probabilità

Definition 138 *Uno spazio di probabilità è una terna* $\{\Omega, \Im, \mathbb{P}\}$ *dove* Ω *rappresenta lo spazio ambiente,* $\Im$ *una* σ*-algebra,* $\mathbb{P} : \Im \to [0,1]$ *una misura di probabilità, cioè una funzione di insieme che soddisfa le seguenti proprietà:*

- $\mathbb{P}(\Omega) = 1$
- $\mathbb{P}(A) \geq 0$, $\forall A \in \Im$
- $\mathbb{P}(\cup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$ *, per ogni sequenze* A_i *tale che* $A_i \in \Im$, $\forall i$ *e* $A_i \cap A_j = \varnothing$ *per ogni* $i \neq j$

Richiamare variabili aleatorie continue e discrete, valor medio e varianza, densità e funzione di ripartizione, funzione caratteristica, cambio di variabile e trasformazioni di densità

18 Esempi ed esercizi

Nel seguito, intendiamo con ε_i, $i = 1, 2, ...$, una sequenza di variabili aleatorie indipendenti ed identicamente distribuite con momento quarto finito.

1) Si determinino gli stochastic orders of magnitude di queste sequenze:

a) $X_n = \sum_{i=1}^n i^\alpha \varepsilon_i$

b) $Y_n = X_n^2 \times \varepsilon_n$

c) $W_n = (\sum_{i=1}^n \varepsilon_i^2) \times (\sum_{i=1}^n \varepsilon_i)$

d) $U_n =$

2) Si studi la convergenza delle seguenti sequenze, al variare di α e β :

a) $X_n = n^{-\beta} \sum_{i=1}^n i^\alpha \varepsilon_i$

b) $Y_n = X_n^2 \times \varepsilon_n$

c) $U_n = n^\alpha (\sum_{i=1}^n \varepsilon_i^2 - n^\beta)$

3) Si studi la convergenza delle seguenti sequenze, al variare di α:

a)

$$X_n = \frac{n^\alpha \sum_{i=1}^n \varepsilon_i}{\sum_{i=1}^n \varepsilon_i^2}$$

b)

$$X_n = \frac{n^\alpha \sum_{i=1}^n \varepsilon_i^2}{\sum_{i=1}^n \varepsilon_i^4}$$

c)

$$X_n = \frac{n^\alpha \sum_{i=1}^n \varepsilon_{3i+1}^2}{(\sum_{i=1}^n \varepsilon_{3i}^4) \times (\sum_{i=1}^n \varepsilon_{3i+2}^4)}$$