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.