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).