DIDATTICA PER IL CORSO DI INFERENZA STATISTICA BAYESIANA

UN' INTRODUZIONE ALL'INFERENZA SU PROCESSI STOCASTICI (prof. A. Wedlin)

Il piu' semplice problema (statico) di inferenza statistica parametrica e' quello in cui la distribuzione di probabilita' comune a tutti i numeri aleatori (n.a.) osservabili X(t), t ≥ 1 , dipende da un parametro non noto Θ di cui si cerca una valutazione numerica approssimata (stima) sulla base dei valori di un numero prefissato n di osservazioni (o misurazioni) X(t) = x(t), t = 1,2,..... .....n. In altri termini, per il parametro incognito Θ si cerca uno stimatore S dipendente dai n.a. X(1),....., X(n) che goda di buone proprieta'. Ovviamente queste ultime devono garantire che la probabilita' di rilevanti errori di stima Θ - S sia opportunamente piccola. E' noto che tale obiettivo viene perseguito in vari modi ai quali corrispondono vari procedimenti o metodi di stima puntuale; i piu' usati sono il "metodo di massima verosimiglianza", il "metodo dei momenti", il "metodo di minima varianza", ...

Quando il parametro incognito Θ varia con t, cioe' quando esso costituisce un processo stocastico {Θ(t), t ≥ 1}, il problema inferenziale assume **un carattere dinamico**: per ogni t, si tratta di trovare uno stimatore S(t) = ξ[X(1),X(2),.....,X(t)] per il n.a. Θ(t) che abbia buone proprieta' nel senso anzidetto. La forma funzionale ξ[.....] dello stimatore dipende, ovviamente, dal metodo di stima adottato. Nel problema dinamico di stima che abbiamo ora introdotto compaiono due processi stocastici: l'insieme di variabili osservabili, o **processo di osservazione,** {X(t) ; t ≥ 1} e l'insieme dei parametri incogniti, o **processo delle variabili di stato**, {Θ(t) ; t ≥ 1}. La maggiore o minore complessita' del problema di stima dipende naturalmente dalle caratteristiche dei due processi stocastici e dal tipo di relazione sussistente tra essi.

Per un semplice esempio di problema inferenziale dinamico si pensi che X(t) rappresenti il numero aleatorio di chiamate telefoniche su una data linea nel periodo t-esimo, con riferimento ad una sequenza di periodi della stessa durata (per esempio oraria). La distribuzione di probabilita' condizionata comune a tutte le variabili osservabili X(t) sia di tipo poissoniano

$$P\{X(t) = n \mid \Theta(t) = \theta(t)\} = \frac{[\theta(t)]^n}{n!}.e^{-\theta(t)}$$

ove Θ(t) e' l'intensita' non nota delle chiamate nel t-esimo periodo. Il problema di inferenza statistica in questo caso consiste nella stima delle intensita' Θ(t), per ogni t, sulla base delle osservazioni dei valori di X(1),.....,X(t) , fissate che siano le caratteristiche dei due processi {X(t)} e {Θ(t)}.

Da quanto abbiamo detto finora si intuisce la centralita' della nozione di processo stocastico: presenteremo quindi sommariamente tale nozione e descriveremo alcune principali tipologie di processi.

La nozione di processo stocastico

Dal punto di vista matematico, fissato uno spazio di probabilità (Ω,A,P), un processo stocastico a valori reali deve essere considerato una **funzione di due variabili**, $X(\omega,t),\ \ \omega \in \Omega,\ \ t \in T,$ tale che:

1) per ogni fissato valore t' di t, in T , $X(.,t')$ è un numero aleatorio, cioè una funzione reale, A-misurabile, definita in Ω, con funzione di ripartizione $F_{t'}(x) = P\{X_{t'}^{-1}(-\infty, x]\}$;

2) per ogni fissato valore ω' di ω, in Ω, $X(\omega',.)$ è una funzione deterministica definita in T e denominata "traiettoria" o "realizzazione" del processo X(t).

Da qui nascono le due interpretazioni del processo stocastico quale "famiglia di numeri aleatori", secondo la 1), e quale "funzione aleatoria", secondo la 2).

Come per un numero aleatorio, anche per un processo stocastico si possono individuare più (in teoria infiniti) **livelli di conoscenza o di specificazione**: quello massimale comporta la conoscenza della "legge temporale", cioè della totalità delle funzioni di ripartizione congiunte di dimensione finita $F_{t_1,t_2,.....,t_n}(x_1, x_2, ..., x_n)$, n $\geq$ 1, $t_j \in$ T per ogni j.

Per i processi a parametro discreto in cui T = Z^+ (cioè per le sequenze illimitate $\{X_t ; t \geq 1\}$ di numeri aleatori), tale famiglia di distribuzioni può essere sostituita dalla più semplice successione $F_1(x_1),\ \ F_{1,2}(x_1, x_2),\ \ F_{1,2,3}(x_1, x_2, x_3),\ \ .....$ che soddisfi la seguente condizione di coerenza:

Teorema di A.N.Kolmogorov: condizione necessaria e sufficiente affinchè tale successione di funzioni di ripartizione costituisca la legge temporale di un processo stocastico a parametro discreto è che ogni elemento della successione sia implicato dai successivi come loro distribuzione marginale e che implichi i precedenti come sue distribuzioni marginali.

Un livello di conoscenza inferiore al precedente e molto usato nelle applicazioni è quello che comporta la conoscenza dei soli momenti del primo e secondo ordine dei numeri aleatori X_t , e cioè delle speranze matematiche, delle varianze e delle covarianze. Si definiscono per il processo $\{X_t; t \geq 1\}$ le due funzioni: la **funzione valor medio** $\varphi(t) = E(X_t)$ e la **funzione di covarianza** $\psi(s,t) = Cov(X_s, X_t)$; quest'ultima soddisfa le seguenti proprietà generali:

1) $\psi(t,t) = Var(X_t) \geq 0$, $\forall t$;

2) $|\psi(s,t)| \leq [\psi(s,s) . \psi(t,t)]^{1/2}$;

3) $\psi(s,t) = \psi(t,s)$;

4) $\sum_{s,t=1}^{N} a_s . a_t . \psi(s,t) \geq 0$; $\forall N$, $\mathbf{a} \in R^N$.

Le principali categorie di processi stocastici sono i processi **markoviani**, quelli **stazionari** ed i processi **martingala**; i fondamenti teorici di tali processi sono stati posti negli anni venti e trenta del secolo ventesimo. I contributi principali sono dovuti a P.Lèvy, A.N.Kolmogorov, A.Y.Khintchine, W.Feller, B. de Finetti, H.Cramèr, H.Wold, J.L.Doob e altri.

La nostra esposizione riguardera' soprattutto i processi a parametro discreto $\{X_t; t \in Z^+\}$ anche se, occasionalmente, parleremo di alcuni processi a parametro continuo $\{X(t), t \in R\}$; a questo punto introduciamo sinteticamente le definizioni dei tre tipi di processi suddetti quando il parametro operativo t assume valori in Z^+:

α) il processo $\{X_t; t \in Z^+\}$ e' **markoviano** se per ogni scelta dell'intero n e della sequenza $t_1 < t_2 < ..... < t_n$ risulta $F\left(x_{t_n} / x_{t_1}, ..., x_{t_{n-1}}\right) = F\left(x_{t_n} / x_{t_{n-1}}\right)$;

β) il processo $\{X_t; t \in Z^+\}$ e' **stazionario** se per ogni scelta dell'intero n e della sequenza $\left(t_1, t_2, ..., t_n\right)$ risulta $F\left(x_{t_1}, x_{t_2}, ..., x_{t_n}\right) = F\left(x_{t_1+h}, x_{t_2+h}, ..., x_{t_n+h}\right)$, ove h e' un qualunque intero tale che ogni $t_j + h \in Z^+$;

γ) il processo $\{X_t; t \in Z^+\}$ e' una **martingala** se per ogni intero positivo $t \geq 1$ risulta $E[|X_t|] < \infty$ e $E(X_t / X_1, ..., X_{t-1}) = X_{t-1}$.

Processi stocastici del secondo ordine

Cominciando con i processi stocastici a parametro discreto $\{Y_t; t = 0,1,2,.....\}$ diremo che $\{Y_t\}$ è una **sequenza aleatoria del secondo ordine** se i numeri aleatori che la costituiscono hanno momenti secondi finiti; formalmente: $E(Y_t^2) < \infty$, $\forall t$.

La totalità dei n.a. di questo tipo definiti in uno spazio di probabilità (Ω, A, P) si indica con H = $L_2(\Omega, A, P)$ e tale insieme costituisce uno spazio di Hilbert con funzione prodotto – interno $\langle X, Y \rangle = E(X.Y)$ e norma $\|X\| = \sqrt{\langle X, X \rangle} = \sqrt{E(X^2)}$; si definiscono anche la distanza tra due elementi di H come $d(X,Y) = \|X - Y\|$ e l'ortogonalità , $X \perp Y$, tra due elementi di H che è verificata se e solo se $\langle X, Y \rangle = 0$. Spesso per semplicità si considera anziché H, lo spazio di Hilbert H_0 dei n.a. definiti in (Ω, A, P), dotati di momento secondo finito e aventi valor medio nullo: in questo caso è $\langle X, Y \rangle = Cov(X,Y)$, $\|X\| = \sqrt{Var(X)}$, $d(X,Y) = \sqrt{Var(X - Y)}$ e infine $X \perp Y \leftrightarrow Cov(X,Y) = 0$.

Supponiamo che i n.a. della sequenza $\{Y_t; t = 0,1,2,.....\}$ siano elementi di H_0: diremo che essi convergono in norma al n.a. $Y \in H_0$ se $d(Y_t, Y) = \|Y_t - Y\| \to 0$ per $t \to \infty$; poiché è $d(Y_t, Y) = \|Y_t - Y\| = \sqrt{E(Y_t - Y)^2}$ si ha che la convergenza in norma coincide con la convergenza in media quadratica $Y_t \xrightarrow{q.m.} Y$.

Le definizioni suddette si applicano allo stesso modo a processi stocastici a parametro continuo $\{Y(t); t \in T \subseteq R\}$ detti anche "funzioni aleatorie". Per questi processi esiste un "calcolo differenziale in media quadratica" basato sulle nozioni di convergenza, continuità, derivabilità e integrabilità in media quadratica:

- convergenza ad Y per t → τ : $\lim E[Y(t) - Y]^2 = 0$;
- continuità per t → τ : $\lim E[Y(t) - Y(\tau)]^2 = 0$;
- derivabilità in t : $\exists Z = Y'(t)$, $\lim E\left\{\frac{1}{h}\left[Y(t+h) - Y(t)\right] - Z\right\}^2 = 0$ per h → 0 ;

- integrabilità su (a,b) : $\exists Z = \int_a^b Y(t)dt, \quad \lim E\left\{\sum_{i=0}^{n-1} Y(t_i).\left[t_{i+1} - t_i\right] - Z\right\}^2 = 0$

per $\max_i \left[t_{i+1} - t_i\right] \to 0$ in corrispondenza ad una sequenza di partizioni P_n sempre più fini di (a,b) , ove $P_n = \left(a = t_0 < t_1 < ..... < t_n = b\right)$.

Ognuna delle suddette proprietà della funzione aleatoria $\{Y(t); t \in T \subseteq R\}$ è ricollegabile a opportune proprietà della funzione di covarianza $\psi_Y(s,t) = Cov(Y_s, Y_t)$ del processo; per esempio, si ha la convergenza ad Y per t → τ se e solo se $\exists \lim \psi_Y(s,t)$ per s, t → τ ; si ha la continuità per t → τ se e solo se $\psi_Y(s,t)$ è continua in (τ, τ) e così via. Rinviamo il lettore interessato ad approfondire l'argomento, per esempio, al testo di J. Lamperti – Stochastic Processes (Springer – Verlag, 1977) oppure al manuale di M. Loeve – Probability Theory (Springer-Verlag, 1977).

Per qualche semplice esempio in proposito si considerino i seguenti casi ove $A_1 \quad e \quad A_2$ sono due numeri aleatori: se $Y(t) = A_1.t$ è $Y'(t) = A_1$; se $X(t) = Y^2(t) = A_1^2.t^2$ si ha $X'(t) = 2.A_1^2.t$; se $Z(t) = X'(t) = 2.A_1^2.t$ si ha $\int_0^t Z(s)ds = A_1^2.t^2$; se $W(t) = \sin A_1.t + \cos A_2.t$ si ha $W'(t) = A_1.\cos A_1.t - A_2.\sin A_2.t$. Invitiamo il lettore ad essere cauto sull'evidente analogia dei suddetti risultati con le regole ben note concernenti la derivazione e l'integrazione di funzioni deterministiche; ognuno dei risultati enunciati andrebbe dimostrato sulla base delle definizioni del calcolo in media quadratica.

Si è già detto che i tipi principali di processi stocastici sono quelli stazionari, markoviani e martingale: esistono versioni analoghe di queste condizioni per i processi del secondo ordine:

- si parla di **stazionarietà del secondo ordine** quando la funzione valor medio $\varphi_Y(t)$ è costante e quando la funzione di covarianza $\psi_Y(s,t)$ dipende soltanto dalla differenza s – t degli argomenti;

- si parla di **markovianità del secondo ordine** quando per ogni n e ogni n.a. X dipendente dai n.a. $\{Y_t; t = n, n+1, ...\}$ si ha $E_2\left(X / Y_n, Y_{n-1}, ..., Y_1\right) = E_2\left(X / Y_n\right)$, ove $E_2(X/Y)$ è l'approssimatore ottimale dei minimi quadrati per X in termini di Y;

- si parla di **martingalità del secondo ordine** se sussistono per ogni n le condizioni $E_2(Y_n / Y_1, Y_2, ..., Y_{n-1}) = Y_{n-1}$. Ovviamente $E_2(Y_n / Y_1, ..., Y_{n-1})$ indica l'approssimatore ottimale dei minimi quadrati per Y_n in termini di $Y_1, ..., Y_{n-1}$.

I processi stocastici del secondo ordine sono spesso impiegati nelle applicazioni in quanto la loro specificazione riguarda soltanto i momenti del primo e secondo ordine, e cioè valori medi, varianze e covarianze, senza chiamare in causa i momenti di ordine superiore al secondo o addirittura le distribuzioni di probabilità congiunte finite-dimensionali. Ad esempio, l'Analisi delle serie temporali (o storiche), disciplina statistica molto usata sia nelle Scienze naturali che in quelle sociali, usa modelli lineari nei quali sia l'input che l'output sono costituiti da processi (a parametro discreto) del secondo ordine. Più precisamente, i modelli matematici usati sono equazioni alle differenze finite lineari e stocastiche $Y_t - \sum_{i=1}^{p} a_i . Y_{t-i} = Z_t + \sum_{j=1}^{q} b_j . Z_{t-j}$ ove il processo input $\{Z_t ; t \geq 0\}$ è specificato dalle funzioni valor medio $\varphi_Z(t)$ e di covarianza $\psi_Z(s,t)$. Ovviamente la soluzione dell'equazione, e cioè il processo $\{Y_t ; t \geq 0\}$, non può avere una specificazione più dettagliata di quella di $\{Z_t ; t \geq 0\}$: risulteranno quindi determinate le sole due funzioni $\varphi_Y(t)$ e $\psi_Y(s,t)$.

Alcuni esempi.

1) Processo di "rumore bianco"(white noise): si tratta di un processo a parametro discreto costituito da n.a. equi (con valor medio zero), aventi una medesima varianza σ_Y^2 e mutuamente non correlati; viene solitamente indicato con $Y_t \approx WN(0, \sigma_Y^2)$.
2) Processo di "passeggiata aleatoria" (random walk): si tratta di un processo definito dalla equazione alle differenze $Y_t = Y_{t-1} + \varepsilon_t ; \quad \varepsilon_t \approx WN(0, \sigma_\varepsilon^2); \quad Y_0 = 0.$
3) Processo "autoregressivo del primo ordine" o AR(1): $Y_t = a . Y_{t-1} + \varepsilon_t ; \quad \varepsilon_t \approx WN(0, \sigma_\varepsilon^2);$ $Cov(Y_0, \varepsilon_t) \equiv 0; \quad Y_0 \approx (m_0, \sigma_0^2)$.

Cenni sui processi markoviani

La condizione di **dipendenza markoviana** è stata introdotta da A. Markov nel 1906, ma furono A.N.Kolmogorov nel 1931 e, più tardi, W.Feller ad impostare rigorosamente la teoria dei processi stocastici markoviani. Nel seguito distingueremo il caso di numeri aleatori con un insieme **discreto** di valori possibili (detti anche "stati") da quello di n.a. con un insieme **continuo** di valori e i processi a tempo **discreto** da quelli a tempo **continuo**.

Fissati arbitrariamente n valori $t_1 < t_2 < .... < t_n$ del parametro operativo la condizione di dipendenza markoviana è espressa dalla $F\left(x_{t_n} / x_{t_1}, ..., x_{t_{n-1}}\right) = F\left(x_{t_n} / x_{t_{n-1}}\right)$ o dalla corrispondente uguaglianza tra le densità di probabilità se queste esistono. Questa condizione ha la seguente notevole implicazione per le distribuzioni congiunte che supporremo dotate di densità:

$$f\left(x_{t_1}, ..., x_{t_n}\right) = f\left(x_{t_n} / x_{t_1}, ..., x_{t_{n-1}}\right) . f\left(x_{t_1}, ..., x_{t_{n-1}}\right) = f\left(x_{t_n} / x_{t_{n-1}}\right) . f\left(x_{t_1}, ..., x_{t_{n-1}}\right) =$$

$$= .... =$$

$$= f\left(x_{t_1}\right) . \prod_{j=2}^{n} f\left(x_{t_j} / x_{t_{j-1}}\right),$$

il chè significa che le distribuzioni congiunte del processo markoviano sono determinate dalle distribuzioni univariate marginali e dalle distribuzioni univariate condizionali. Una seconda notevole implicazione è la seguente: se si suppone che sia $s < \tau < t$ si ha

$$f\left(x_t / x_s\right) = \int_R f\left(x_t, x_\tau / x_s\right) dx_\tau = \int_R f\left(x_t / x_\tau, x_s\right) . f\left(x_\tau / x_s\right) dx_\tau = \int_R f\left(x_t / x_\tau\right) . f\left(x_\tau / x_s\right) dx_\tau ,$$

e tale relazione costituisce una condizione di coerenza per le distribuzioni univariate condizionali, nota in letteratura come condizione di Chapman – Kolmogorov. Le densità condizionate $f\left(x_t / x_s\right)$ sono dette "densità di transizione dallo stato x_s allo stato x_t".

a) Catene di Markov omogenee con un numero finito di stati.

Il tipo più semplice di processo markoviano è denominato "catena markoviana omogenea" ed è costituito da una successione di n.a. $\{X_n ; n \geq 0\}$, aventi ciascuno lo stesso insieme finito di N valori possibili, caratterizzata per ogni $n \geq 0$ dalla condizione di dipendenza markoviana omogenea:

$$\text{Prob}\{X_{n+1} = j / (X_0 = a) \wedge (X_1 = b) \wedge ....\wedge (X_n = i)\} = \text{Prob}\{X_{n+1} = j / X_n = i\} = p_{ij}(n) ,$$

Se le probabilità subordinate p_{ij} (n), dette "di transizione dallo stato i allo stato j", non dipendono dall'indice n allora si parla di "dipendenza markoviana omogenea". Si verifica facilmente che in tal caso la struttura probabilistica del processo è univocamente determinata dalla matrice $N \times N$ delle probabilità subordinate P = [p_{ij}] e dalla distribuzione del n.a. X_0 , detta "distribuzione iniziale". Ovviamente, la somma degli elementi di ogni riga di P è uguale a 1(matrice "stocastica").

Indicata con il vettore riga **a**(0) la distribuzione di X_0, quella **a**(1) del n.a. X_1 è data dal prodotto **a**(1) = **a**(0) . P e, in generale, quella di X_n da **a**(n) = **a**(0) . P^n = **a**(n-1) . P . L'elemento generico della matrice P^n, $p_{ij}^{(n)}$, è detto "probabilità subordinata di transizione da i a j in n passi" e soddisfa, come si prova facilmente, la condizione di Chapman – Kolmogorov:

$$p_{ij}^{(n)} = \sum_{h=1}^{N} p_{ih}^{(n-m)} . p_{hj}^{(m)} ,$$

per ogni intero positivo $m < n$.

Una catena markoviana è detta "regolare" se esiste un intero n_0 tale che per ogni $n > n_0$ tutti gli elementi di P^n risultano positivi; per le catene regolari sussiste il seguente importante risultato:

Teorema di Markov: la matrice P^n converge, al divergere di n, ad una matrice U con elementi positivi, cioè risulta $p_{ij}^{(n)} \to u_j > 0$ per ogni i; le righe di U quindi sono tutte uguali e la somma degli elementi di riga è unitaria. Indicata con **u** la generica riga di U, è **u** . P = **u ,** per cui **u** è detta distribuzione stazionaria.

Alcuni esempi.

Si consideri innanzitutto il seguente schema generale di estrazioni ripetute da un'urna contenente inizialmente b palline bianche ed r palline rosse: si sceglie a caso una pallina dall'urna e, assieme ad essa, si immettono nell'urna c palline dello stesso colore di quella estratta e d palline di colore opposto; dopo la prima estrazione nell'urna ci sono quindi b+r+c+d palline. Si effettua una seconda estrazione casuale con la stessa procedura e così via: in generale la composizione dell'urna varia colpo per colpo. Alcuni modelli particolari sono i seguenti:

a) c = d = 0 (estrazioni ripetute con reimbussolamento della pallina estratta);

b) c = -1, d = 0 (estrazioni ripetute senza reimbussolamento);
c) c > 0, d = 0 (urna di Pòlya – modelli di contagio positivo);
d) c = -1, d = 1 (urna di Ehrenfest).

L'ultimo modello realizza una catena di Markov nel senso che le variabili aleatorie X_n che contano il numero di palline bianche nell'urna dopo le prime n estrazioni costituiscono una catena di Markov caratterizzata dalle probabilità subordinate di transizione $p_{i,i+1} = (b+r-i) / (b+r)$, $p_{i,i-1} = i / (b+r)$ e $p_{i,j} = 0$ se j è diverso da i+1 e i-1. E' facile costruire la corrispondente matrice P se si tiene presente che il numero degli stati è N = b+r+1.

Si verifica facilmente che le variabili X_n nel modello dell'urna di Pòlya costituiscono ancora una catena di Markov, mentre invece le variabili Y_n (indicatori di eventi) che assumono i valori 1 e 0 a seconda che dall'urna di Pòlya sia estratta, al colpo n-mo, una pallina bianca o rossa non verificano la condizione di dipendenza markoviana, ma costituiscono un processo stazionario scambiabile.

Infine, le variabili Z_n che denotano le proporzioni di palline bianche dopo n estrazioni dall'urna di Pòlya costituiscono un processo martingala.

Un esempio di catena di Markov con un'infinità numerabile di stati (e precisamente l'insieme degli interi non negativi) è costituito dal processo di passeggiata aleatoria $Y_t = Y_{t-1} + |E_t|, \;\; Y_0 = 0,$ ove gli eventi $E_t, \;\; t \geq 1,$ sono assunti indipendenti e ugualmente probabili con $P(E_t) \equiv p$.

b) Processi a tempo continuo con un insieme discreto di stati $\{s_i\}$.

In questi processi le probabilità di transizione in un intervallo di tempo di ampiezza t dallo stato s_i allo stato s_j , $P\left[X(t+\tau) = s_j / X(\tau) = s_i \right]$, sono indicate con $p_{ij}(t)$,per ogni τ, e soddisfano le seguenti condizioni:

$$p_{ij}(t) \geq 0; \;\; \sum_j p_{ij}(t) = 1, \;\; \forall t; \;\; p_{ij}(t+\tau) = \sum_h p_{ih}(t).p_{hj}(\tau) .$$

Se il numero di stati è finito, l'ultima condizione (le relazioni di Chapman – Kolmogorov) può essere espressa dalla $P(t+\tau) = P(t).P(\tau)$, ove $P(t) = \left[p_{ij}(t) \right]$, assumendo che sia $P(0) = I$. Il teorema di Markov stabilisce ora che l'ipotesi $p_{ij}(t) > 0$, per ogni i, j, t , implica l'esistenza di una matrice U tale che $U = \lim_{t \to \infty} P(t) \;\; e \;\; U.P(t) = U$. Come nel caso precedente e sempre nell'ipotesi che gli stati siano finiti, indicata con il vettore riga $a(0)$ la distribuzione di X(0) è:

$$a(t) = a(0).P(t) = a(0).P(\tau).P(t-\tau) = a(\tau).P(t-\tau)\,.$$

Esempio: **il processo di Poisson**.

L'insieme degli stati è ora l'insieme degli interi non negativi e le variabili del processo, N(t), contano il numero di eventi (di un tipo fissato) che si verificano nell'intervallo di tempo [0, t]. Si assume N(0) = 0 e che gli incrementi del processo N(t) – N(s) siano stocasticamente indipendenti e omogenei (o stazionari): ciò significa che se è $t_1 < t_2 \le t_3 < t_4$ gli incrementi $N(t_4) - N(t_3) \quad e$ $N(t_2) - N(t_1)$ sono indipendenti e che $N(t) - N(s) \quad e \quad N(t+\tau) - N(s+\tau)$ sono ugualmente distribuiti qualunque sia τ . Sotto ipotesi non molto restrittive risulta che

$$P\{N(t) = n\} = e^{-\lambda t}.(\lambda.t)^n / n!$$

ove λ > 0 è l'unico parametro della distribuzione; esso è detto "intensità del processo degli arrivi" e rappresenta il numero medio di eventi nell'intervallo di tempo unitario. Si calcola facilmente che è $E[N(t)] = Var[N(t)] = \lambda.t$.

Indicato con T_i il tempo di attesa tra l' (i-1)-esimo e l' i-esimo evento, si prova che tali n.a. sono indipendenti e ugualmente distribuiti con densità di probabilità esponenziale negativa $f(t) = \lambda.e^{-\lambda t}$. E' evidente che $\{N(t) = n\} = \{T_n \le t\} \wedge \{T_{n+1} > t\}$.

Semplici generalizzazioni del processo di Poisson introducono ai **processi di puro ingresso** (facendo dipendere l'intensità λ dallo stato j raggiunto) e ai **processi di ingresso e uscita**, o di nascita e morte, (introducendo oltre alle intensità di ingresso λ_j anche intensità di uscita μ_j). Si veda in proposito, per esempio, il capitolo XVII del primo volume di "An introduction to probability theory and its applications" di W.Feller.

Se si generalizzano ulteriormente i processi markoviani, eliminando l'ipotesi di omogeneità nel tempo, le probabilità di transizione da i a j in un intervallo di tempo (s,t) non dipendono più dalla sola ampiezza dell'intervallo t-s, ma da entrambi gli estremi s e t e quindi dovranno essere indicate col simbolo $P_{ij}(s,t)$. Esse verificano le condizioni $P_{ij}(s,t) \ge 0$, $\sum_j P_{ij}(s,t) = 1$ e le relazioni di

Chapman – Kolmogorov: $P_{ij}(s,t) = \sum_h P_{ih}(s,\tau) \,.\, P_{hj}(\tau,t)$, $s < \tau < t$.

c) Processi a tempo discreto e spazio degli stati continuo.

Assumeremo che lo spazio degli stati sia R, ma gli sviluppi formali che seguono non varierebbero se esso fosse un arbitrario spazio cartesiano, per esempio R^n.

Nel caso di un processo markoviano a parametro *discreto* le probabilità di transizione sono espresse da una funzione (stochastic kernel) K(x,S) = Prob $\{X_{n+1} \in S / X_n = x\}$ di due argomenti, $x \in R$ ed $S \subset R$; fissato x, K(x, .) è una probabilità sui sottoinsiemi boreliani di R, mentre fissato S, K(. ,S) è una funzione di Baire in R.

Frequentemente nelle applicazioni è $K(x,S) = \int_S k(x,y)\,dy$ e la funzione k(x,y) (stochastic density kernel) è definita in R^2. Le relazioni di Chapman – Kolmogorov per le funzioni K e k sono rispettivamente: $K^{(m+n)}(x,S) = \int_R K^{(m)}(x,dy).\,K^{(n)}(y,S)$ e $k^{(m+n)}(x,y) = \int_R k^{(m)}(x,z).\,k^{(n)}(z,y)\,dz$. Ponendo m = 1 e facendo crescere n si ottiene una definizione ricorsiva di $K^{(n)}(x,S)$ e $k^{(n)}(x,y)$; per esempio si hanno le $k^{(2)}(x,y) = \int_R k(x,z).k(z,y)dz$, $k^{(3)}(x,y) = \int_R k(x,z).k^{(2)}(z,y)dz$ e così via.

Un esempio: sia $k(x,y) = \lambda\,.\,y^{\lambda-1} / x^{\lambda}$, per $0 < y < x$; si ha facilmente

$$k^{(n)}(x,y) = \frac{\lambda^n .y^{\lambda-1}\left[\log(x/y)\right]^{n-1}}{\Gamma(n).x^{\lambda}}$$

ed inoltre, se S = [a , b] è

$$K(x\,,[a,b]) = \int_a^b k(x,y)\,dy = (b^{\lambda} - a^{\lambda}) / x^{\lambda}\ .$$

d)Processi a tempo continuo e spazio degli stati continuo.

Per un processo markoviano omogeneo a parametro *continuo* denoteremo con $Q_t(x,S)$ le probabilità di transizione

$$Q_t(x,S) = \text{Prob}\{ X(t+\tau) \in S / X(\tau) = x\} , \ \forall \tau .$$

Esse soddisfano le condizioni $Q_t(x,S) \geq 0$, $Q_t(x,R) = 1$ e le relazioni di Chapman – Kolmogorov :

$$Q_t(x,S) = \int_R Q_{t-\tau}(x,dy) \, . \, Q_\tau(y,S) , \quad 0 < \tau < t \ .$$

Un primo esempio molto importante è costituito dal **processo di Poisson composto.** Si tratta di un processo markoviano a parametro continuo molto usato nelle applicazioni la cui interpretazione è tipicamente la seguente: si pensi ad un n.a. N(t) di eventi (o arrivi) osservabili nel periodo di tempo [0, t] a ciascuno dei quali sia associato un "effetto" aleatorio Y_j e interessi conoscere la distribuzione dell'effetto complessivo S(t) = $\sum_{j=0}^{N(t)} Y_j$. Se indichiamo con $\{p_{nt} ; n \geq 0\}$ la distribuzione di probabilità di N(t) e con $F_Y(y)$ la funzione di ripartizione (f.d.r.) comune ai n.a. Y_j che assumiamo essere anche mutuamente indipendenti si ha che la distribuzione di $\sum_{j=0}^{N(t)} Y_j$ ha f.d.r. data alla:

$$\text{Prob}\{S(t) \leq x\} = \sum_{n\geq 0} P[N(t) = n].P[S(t) \leq x / N = n] = \sum_{n\geq 0} p_{nt}.F_Y^{n*}(x) \ ,$$

ove il simbolo $F_Y^{n*}(x)$ indica la f.d.r. della distribuzione della somma $\sum_{j=0}^{n} Y_j$ di n addendi (poiché per ipotesi è $Y_0 = 0$) i.i.d . Il nome di "processo di Poisson composto" è appropriato se la distribuzione $\{p_{nt} ; n \geq 0\}$ è poissoniana con parametro λt .

Il carattere markoviano di {S(t)} si dimostra facilmente osservando che se $0 < t_1 < \ldots\ldots < t_n < t$ e posto $N(t+\tau) - N(t) = N_\tau$ si ha:

$$\text{Prob}\{S(t+\tau) = k / [\bigcap_{i=1}^{n} [S(t_i) = j_i]] \cap [S(t) = j]\} = \text{Prob}\{\sum_{h=1}^{N_\tau} Y_h = k - j \} = \text{Prob}\{S(t+\tau) = k / S(t) = j\}.$$

Può essere utile riportare l'espressione della funzione caratteristica del processo S(t); si dimostra che è: $\Phi_{S(t)}(\xi) = \exp\{\lambda.t[\Phi_Y(\xi) - 1]\}$, ove $\Phi_Y(\xi)$ indica la funzione caratteristica della distribuzione di probabilità dei n.a. Y_j .

Nel seguito, del suddetto processo a parametro continuo {S(t)} considereremo gli incrementi relativi ad intervalli unitari di tempo $X_t = S(t) - S(t-1)$, t = 1, 2, …….. , definiti dalle $X_t = \sum_{j=0}^{N_t} Y_{tj}$, ove N_t = N(t) – N(t-1) è l'incremento del processo poissoniano di arrivi {N(t) ; t≥0} e ove Y_{tj} , che indica l'effetto associato al j-mo arrivo nell'intervallo unitario t-esimo di tempo, è l'elemento generico del "processo degli effetti" {Y_j ; j≥1}. Solitamente si assume che i n.a. N_t siano indipendenti dagli Y_j e che la comune distribuzione degli Y_j sia di tipo Gamma; per semplicità noi assumeremo $Y_j \sim$ Gamma (1 , β), cioè supporremo che ogni Y_j abbia una densità di probabilità di tipo esponenziale negativo, f(y) = β.exp(-β.y), per cui la convoluzione n-ma è una densità di tipo Gamma (n , β). Si ha allora:

$$F_{X_t}(x) = \sum_{n\geq 0} p_n . F_Y^{n*}(x) = \sum_{n\geq 0}\left(\frac{e^{-\lambda}.\lambda^n}{n!}\right).\int_0^x \frac{\beta^n}{\Gamma(n)} z^{n-1}.e^{-\beta.z}dz \text{ , ed anche } f_{X_t}(x) = \sum_{n\geq 0}\left(\frac{e^{-\lambda}.\lambda^n}{n!}\right).\left(\frac{\beta^n}{\Gamma(n)} x^{n-1}.e^{-\beta.x}\right).$$

Si ricavano, per i primi due momenti di X_t, i risultati :

$$E(X_t) = E(N_t).E(Y_j) = \lambda / \beta \text{ , } \quad E(X_t^2) = E(N_t).Var(Y_j) + E(N_t^2).\ E^2(Y_j) = \lambda.(\lambda+2) / \beta^2 \text{ ,}$$

dai quali si ottiene immediatamente Var $(X_t) = E(N_t).E(Y^2) = 2\lambda / \beta^2$.

Ci limitiamo ad accennare che il processo di Poisson composto ha anche una notevole importanza teorica, oltre che applicativa. Si dimostra, per esempio, che la più generale distribuzione "infinitamente divisibile" è rappresentabile come limite di un'appropriata sequenza di distribuzioni di Poisson composto, $Q_t(x) = \sum_{n\geq 0} \frac{(\alpha.t)^n.\exp(-\alpha.t)}{n!} F^{n*}(x)$. Ricordiamo che una distribuzione di probabilità F(x) è infinitamente divisibile se, indicata con φ(ξ) la corrispondente funzione caratteristica, questa può essere rappresentata, per ogni n ≥ 1, come potenza n-ma di una funzione caratteristica $\phi_n(\xi)$, cioè se $\varphi(\xi) = [\phi_n(\xi)]^n$.

Alcuni esempi di distribuzioni infinitamente divisibili sono costituiti dalle distribuzioni

- di Poisson con $\varphi(\xi) = \exp[\lambda(e^{i\xi} - 1)]$ e $\phi_n(\xi) = \exp[\lambda(e^{i\xi} - 1)/n]$,
- dalla distribuzione normale con $\varphi(\xi) = \exp(i\xi\mu - \sigma^2\xi^2/2)$ e $\phi_n(\xi) = \exp[(i\xi\mu - \sigma^2\xi^2/2)/n]$,

- dalla distribuzione Gamma con $\varphi(\xi) = \left(1-\frac{i\xi}{\beta}\right)^{-\alpha}$ e $\phi_n(\xi) = \left(1-\frac{i\xi}{\beta}\right)^{-\alpha/n}$,
- dalla distribuzione di Poisson composto con $\varphi_X(\xi) = \exp\{\lambda[\varphi_Y(\xi) - 1]\}$ e $\phi_n(\xi) = \exp\{\lambda[\varphi_Y(\xi) - 1]/n\}$.

Un altro esempio importante è costituito dal processo **"pseudo – poissoniano"** $\{X(t);\ t \geq 0\}$ caratterizzato dalle funzioni di transizione

$$Q_t(x,S) = \sum_{n=0}^{\infty}\left[\frac{e^{-\alpha t}.(\alpha.t)^n}{n!}\right].K^{(n)}(x,S)\ ,$$

ove K(x,S) è una fissata probabilità di transizione di una catena markoviana $\{Z_n\ ;\ n \geq 0\}$ ed {N(t)} è un processo di Poisson con n.a. N(t) indipendenti dai n.a. Z_n ; è allora $X(t) = Z_{N(t)}$.

Se $K(x,S) = \int_S \lambda.y^{\lambda-1}.x^{-\lambda}dy$ si prova che la corrispondente probabilità di transizione $Q_t(x,S)$ ha densità espressa dalla $q_t(x,y) = \frac{e^{-\alpha t}.\sqrt{\alpha\lambda t}.y^{\lambda-1}}{x^{\lambda}.\sqrt{\log(x/y)}}.I_1(2\sqrt{\alpha\lambda t.\log(x/y)})$, ove $I_1(z) = \sum_{k=0}^{\infty}\frac{1}{k!\Gamma(k+2)}.(z/2)^{2k+1}$ è la funzione di Bessel di ordine 1, per $0 < y < x$ ed una massa concentrata pari a $e^{-\lambda t}$ in y = 0 .

Chiaramente, un caso particolare del processo pseudo-poissoniano è costituito dal processo di Poisson composto se le variabili Z_n della catena markoviana sono somme di n.a. Y_j indipendenti e identicamente distribuiti.

Cenni sui processi con incrementi stazionari e indipendenti

I più semplici processi stocastici di questo tipo, a parametro **discreto**, sono quelli denominati "processi di passeggiata aleatoria" e definiti dalle $S_n = \sum_{k=1}^{n} X_k$, $n \geq 1$, ove i n.a. X_k sono indipendenti e ugualmente distribuiti (brevemente i.i.d.); per completezza si pone $S_0 = 0$. Equivalentemente, il processo $\{S_n\}$ può essere definito dalle $S_0 = 0$, $S_n = S_{n-1} + X_n$, $n \geq 1$. Si tratta, ovviamente, di processi markoviani e, più precisamente, di catene markoviane.

Per un primo esempio si assuma $X_k = |\ E_k\ |$, con eventi E_k indipendenti e ugualmente probabili (processo bernoulliano); in tal caso S_n è la frequenza di successo su n eventi associata a $\{X_k\}$ e per il processo $\{S_k\}$ si hanno i ben noti risultati

a) Teorema di Bernoulli : p - lim $[S_n/n] = P(E_1)$;

b) Teorema di de Moivre – Laplace : $\frac{S_n - n.P(E_1)}{\sqrt{n.P(E_1).[1-P(E_1)]}} \xrightarrow{d} N(0;1)$.

Se, più in generale, i n.a. X_k, non necessariamente indicatori di eventi, hanno valor medio $E(X_k) = \mu$ e varianza $Var(X_k) = \sigma^2$ finiti sussistono i seguenti teoremi che generalizzano i risultati precedenti:

c) Legge debole dei grandi numeri: p – lim $[S_n/n] = \mu$;

d) Teorema centrale limite: $\frac{S_n - n.\mu}{\sqrt{n.\sigma^2}} \xrightarrow{d} N(0;1)$.

Per quanto concerne i momenti fino al secondo ordine del processo $\{S_k\}$ si ha che la funzione valor medio $\varphi_S(n) = E(S_n) = n.\mu$ è crescente con n e che la funzione di covarianza è data dalla $\psi_S(m,n) = Cov(S_m, S_n) = \min(m.\sigma^2, n.\sigma^2)$.

Gli incrementi del processo $\{S_k\}$ sono dati dai n.a. $S_n - S_{n-1} = X_n$ per $n > 1$ e $S_1 - S_0 = S_1 = X_1$: come si è già assunto, essi sono i.i.d.

Il processo in cui $X_k = |E_k|$ con punto iniziale l'origine e $P(E_k) = p$, $\forall k$, può descriversi come una catena di Markov con un'infinità numerabile di stati (i numeri interi, positivi nulli e negativi) e matrice di transizione con elementi tutti nulli salvo per $p_{j,j+1} = p$ e $p_{j,j-1} = 1-p$, $\forall j$.

Un processo a parametro **continuo** $\{S(t); t \in R\}$ ha incrementi indipendenti e stazionari se, considerata un'arbitraria partizione dell'intervallo $[s, s + t]$, $s = t_0 < t_1 < \ldots\ldots\ldots < t_n = s + t$, gli incrementi $S(t_k) - S(t_{k-1})$, $k = 1,\ldots..,n$, sono mutuamente indipendenti e la loro distribuzione dipende soltanto dalle differenze $t_k - t_{k-1}$.

Si supponga ora che gli intervalli $[t_k - t_{k-1}]$ abbiano la medesima lunghezza di modo che gli incrementi $\Delta_k = S(t_k) - S(t_{k-1})$ siano ugualmente distribuiti: dalla $S(s + t) - S(s) = \sum_{k=1}^{n} \Delta_k$ discende che la distribuzione di $S(s + t) - S(s)$ è uguale alla convoluzione n-ma della comune distribuzione dei Δ_k . Poiché il numero n degli intervalli $[t_k - t_{k-1}]$ è arbitrario, l'incremento $S(s + t) - S(s)$ ha necessariamente una distribuzione "infinitamente divisibile".

Riprendendo quanto già detto, un n.a. Z, la sua funzione di ripartizione $F_Z(z)$ e la sua funzione caratteristica $\Phi(u) = E[\exp(iuZ)] = \int_R e^{iuz} dF_Z(z)$ sono **infinitamente divisibili** se per ogni $n \geq 1$ esistono n.a. $\eta_1, \eta_2, \ldots\ldots, \eta_n$ i.i.d. tali che $F_Z(z) = F_{\eta_1}^{n*}(z)$, oppure $\Phi_Z(u) = [\Phi_\eta(u)]^n$. Alcune distribuzioni infinitamente divisibili (inf. div.) sono la normale, la gamma e la distribuzione di Cauchy tra quelle **continue** e la binomiale negativa (e quindi la geometrica), la distribuzione di Poisson e quella di Poisson composto tra quelle **discrete**.

La caratterizzazione delle distribuzioni inf. div. è stata ottenuta nei primi anni 30 attraverso il lavoro di B. de Finetti, A.N. Kolmogorov, P. Lèvy e, più tardi, di W. Feller e A.Y. Khintchine.

Di fondamentale importanza è il seguente risultato

Teorema: sia $\{\Phi_n(u); n \geq 1\}$ una successione di funzioni caratteristiche; condizione necessaria e sufficiente affinchè esista una funzione limite continua $\Phi(u) = \lim [\Phi_n(u)]^n$ è che esista, continua, la funzione limite $\lim n.[\Phi_n(u) - 1] = \chi(u)$ e in tal caso si ha $\Phi(u) = \exp\{-\chi(u)\}$.

Ricordando che $\exp\{ n.[\Phi_n(u) - 1] \}$ è la funzione caratteristica della distribuzione di Poisson composto con intensità n e funzione caratteristica della distribuzione degli effetti $\Phi_n(u)$, il teorema afferma che ogni distribuzione inf. div. è rappresentabile come limite di una opportuna successione di distribuzioni di Poisson composto. Inoltre esso afferma che ogni limite continuo di una successione di funzioni caratteristiche inf. div. è esso stesso inf. div..

Tra le distribuzioni inf. div. sono importanti le **distribuzioni stabili** ; tra quelle sopra indicate sono stabili soltanto la normale e quella di Cauchy. Esse sono definite al modo seguente: una distribuzione F e la sua funzione caratteristica Φ sono stabili se, per ogni $n \geq 1$ esistono costanti $a_n > 0$ e b_n tali che $[\Phi_F(u)]^n = \Phi_{an.F+ bn}(u) = \exp\{i.b_n.u\} . \Phi_F(a_n.u)$ cioè, a parole, se la distribuzione della somma di n numeri aleatori X_i indipendenti con distribuzione F (la cui funzione caratteristica è $[\Phi_F(u)]^n$) coincide con quella del n.a. $a_n.X_1 + b_n$ (con funzione caratteristica $\exp\{i.b_n.u\} . \Phi_F(a_n.u)$).

Si dimostrano le due seguenti proposizioni concernenti le distribuzioni stabili:

- le costanti a_n hanno la forma $a_n = n^{1/\alpha}$, con $0 < \alpha \leq 2$: per la distribuzione normale è $\alpha = 2$ mentre per quella di Cauchy è $\alpha = 1$;

- le distribuzioni stabili **simmetriche** hanno funzione caratteristica avente la forma Φ(u) = $\exp\{-\gamma_\alpha . |u|^\alpha\}$ con γ_α reale positivo : per la distribuzione N(0 ; 1) è Φ(u) = $\exp\{-\frac{1}{2}u^2\}$ mentre per quella di Cauchy è Φ(u) = $\exp\{-\theta . |u|\}$, ove è θ > 0 .

Presentiamo ora alcuni esempi di processi {S(t)} con incrementi indipendenti e stazionari:

- processi di Poisson composto, ove Prob{S(t) ≤ s} = $\sum_k \frac{e^{-\lambda t}(\lambda t)^k}{k!} F^{k*}(s)$;
- processi di Poisson: si ricavano dai precedenti specificando che la distribuzione F degli effetti è concentrata in 1; in tal caso è $\Phi_{S(t)}(u) = \exp\{\lambda t[e^{iu} - 1]\}$ in quanto e^{iua} è la funzione caratteristica della distribuzione concentrata in a ;
- processi di Wiener – Lèvy (o di moto browniano) caratterizzati dall'ipotesi S(0) = 0 e dal fatto che gli incrementi S(s+t) – S(s) = S(t) – S(0) = S(t) hanno distribuzione N(0 ; $\sigma^2 . t$) .

Cenni sui processi stocastici martingala

Un processo stocastico $\{X_t\}$ è detto essere una martingala se per ogni intero positivo $t \geq 1$ si ha $E[|X_t|] < \infty$ ed è verificata (con probabilità 1) la condizione:

$$E(X_t / X_1, \ldots\ldots\ldots\ldots, X_{t-1}) = X_{t-1} .$$

E' facile verificare che sussiste, per ogni t, l'uguaglianza $E[X_t] = E[X_{t-1}]$ e che, per s intero positivo, è $Cov(X_t, X_{t-s}) = Var(X_{t-s})$. Si ha inoltre $E(X_{t+h} / X_1, \ldots\ldots\ldots\ldots, X_{t-1}) = X_{t-1}$ per ogni h intero non negativo, cosicchè ogni sottosequenza $X_{t_1}, X_{t_2}, \ldots\ldots\ldots\ldots, X_{t_n}, \ldots\ldots\ldots\ldots$ di un processo martingala è una martingala.

Una definizione più flessibile di martingala è la seguente**: il processo {Xt} è una martingala rispetto al processo {Y_t }** se $E[|X_t|] < \infty$ e se $E[X_{t+1} / Y_1, \ldots\ldots\ldots, Y_t] = X_t$ per ogni t. Una definizione ancora più generale di martingala è quella rispetto ad una sequenza crescente di σ-algebre di eventi anziché rispetto a sequenze di vettori ($Y_1, \ldots\ldots\ldots, Y_t$). Noi ci atterremo peraltro alla prima, più semplice, definizione.

Teorema di rappresentazione: il processo $\{X_t\}$ è una martingala se e solo se $X_t = Y_1 + Y_2 + \ldots\ldots\ldots + Y_t + c$ ove c è una costante e i n.a. Y_t sono **assolutamente equi** $[E(Y_1) = 0$ e $E(Y_t/Y_1,\ldots\ldots, Y_{t-1}) = 0, \forall\ t]$. In forza di questo risultato, n.a. Y_t assolutamente equi sono detti "differenze, o incrementi, di martingala". Si dimostra che le differenze di martingala sono n.a. equi e mutuamente non correlati.

Teorema di convergenza: se i momenti secondi dei n.a. X_t del processo martingala sono equilimitati, cioè se esiste un numero reale m per cui è $E[X_t^{\,2}] \leq m < \infty$, allora esiste un n.a. X verso cui la martingala converge, con probabilità 1, al divergere di t. Inoltre è $E(X) = E(X_t)$ per ogni t.

Alcuni esempi :

1) Le somme successive di n.a. Z_t equi e stocasticamente indipendenti costituiscono una martingala (è lo schema di un gioco equo): infatti i n.a. Z_t sono anche assolutamente equi per cui, ponendo $c = 0$ e $X_t = Z_1 + Z_2 + \ldots\ldots\ldots + Z_t$ si ha che X_t è una martingala per il teorema di rappresentazione.
2) Le proporzioni X_t di palline bianche dopo ogni estrazione casuale da un'urna di Polya (contenente inizialmente b palline bianche ed r palline rosse, essendo c le nuove palline immesse nell'urna, dello stesso colore di quella appena estratta) costituiscono una martingala.
3) La successione delle funzioni di regressione $X_t = E[Z / Y_1,\ldots\ldots\ldots,Y_t]$ di un n.a. Z rispetto ai n.a. $Y_1,\ldots\ldots\ldots,Y_t,\ldots\ldots$ è un processo martingala . Occorre, in questo esempio, adottare per i n.a. X_t la definizione di martingala rispetto alla sequenza $Y_1, Y_2,\ldots\ldots,Y_t,\ldots\ldots$.

Ritornando all'esempio 1), e considerando che X_t sia una martingala rispetto a $\{Z_t\}$, è

opportuno tener conto della possibilità che un giocatore decida colpo per colpo se giocare o no all'epoca n: si può far questo introducendo una "funzione di decisione" $\{\varepsilon_n\}$, ove $\varepsilon_n = 0$ oppure $\varepsilon_n = 1$ a seconda che il giocatore decida di non giocare, o giocare, la partita n-ma. Indicando con W_n il guadagno totale nelle prime n partite si ha $W_n = W_{n-1} + \varepsilon_n \,.\, Z_n = W_{n-1} + \varepsilon_n \,.(X_n - X_{n-1})$.

Osserviamo che in generale $W_{n-1} \neq X_{n-1} = \sum_{t=1}^{n-1} Z_t$ perché il giocatore può decidere di non giocare qualcuna delle prime n-1 partite; se, per esempio, è uguale a 1 solo la prima variabile decisionale ε_1 mentre tutte le successive n-2 sono nulle si ha $W_{n-1} = Z_1 = X_1$.

Poiché, a partire dall'equazione sopra scritta, è:

$$E(W_n / Z_1,\ldots\ldots\ldots,Z_{n-1}) = W_{n-1} + \varepsilon_n \,.\, [E(X_n / Z_1,\ldots\ldots\ldots,Z_{n-1}) - X_{n-1}] \,,$$

e ricordando che $\{X_n\}$ è una martingala rispetto a $\{Z_n\}$ risulta $E(W_n/Z_1,....,Z_{n-1}) = W_{n-1}$ cosicchè anche $\{W_n\}$ è una martingala rispetto a $\{Z_n\}$. Tale risultato è noto come

Teorema di impossibilità (di modificare la struttura di martingala): qualunque funzione di decisione $\{\varepsilon_n\}$ trasforma una martingala $\{X_n\}$ in una nuova martingala $\{W_n\}$.

Una particolare funzione di decisione $\{\varepsilon_t\}$ è quella in cui esiste un intero n tale che $\varepsilon_t = 1$ per ogni $t \leq n$ e $\varepsilon_t = 0$ se $t > n$. Tipicamente, l'intero n è considerato aleatorio (e denotato N) di modo che si ha $\varepsilon_t = 1$ se $N > t-1$ mentre $\varepsilon_t = 0$ se $N \leq t-1$.

Ovviamente, neanche l'introduzione di un **tempo aleatorio di arresto** (stopping time) modifica la proprietà di martingala. Aggiungiamo alcune altre nozioni collegate a quella di martingala.

Un processo $\{X_t\}$ è detto costituire una **sub-martingala** (super-martingala) se soddisfa le condizioni $E(X_t / X_1,.....,X_{t-1}) \geq X_{t-1}$ ($\leq X_{t-1}$). Sussiste l'importante teorema:

"se u(.) è una funzione convessa ed $\{X_t\}$ una martingala allora il processo $\{u(X_t)\}$ è una sub-martingala a condizione che esista finita la $E[|u(X_t)|]$ per ogni t ".

Un processo $\{X_t\}$ è una **martingala in senso lato** se $E_2[X_t / X_1,.....,X_{t-1}] = X_{t-1}$ ove il primo membro indica l'approssimazione lineare dei minimi quadrati di X_t rispetto ai n.a. $X_1,.....,X_{t-1}$. Si dimostra che un processo $\{X_t\}$ è una martingala in senso lato se e solo se $X_t = Y_1+.....+Y_t + c$ ove i n.a. Y_t sono equi e mutuamente ortogonali.

Il processo martingala presentato nel terzo esempio è costituito da **approssimatori dei minimi quadrati** $X_t = E[Z / Y_1,.....,Y_t]$ del n.a. Z in termini di funzioni arbitrarie dei n.a. Y_t supposti osservabili. La condizione $E(Z^2) < \infty$ implica la $E(X_t^2) \leq m < \infty$ per ogni t e dunque la convergenza in media quadratica di $\{X_t\}$ ad un n.a. X: quest'ultimo dovrebbe potersi interpretare come approssimatore ottimale in quanto utilizza la totalità dei n.a. osservabili Y_t . Inoltre, è naturale pensare che al crescere della numerosità della sequenza $Y_1,.....,Y_t$ migliori progressivamente l'approssimazione di Z perché si utilizza una quantità di informazione crescente: è infatti quanto viene stabilito dal seguente teorema.

Teorema: posto $D_t = [E(Z - X_t)^2]^{1/2}$, al crescere di t la sequenza $\{D_t\}$ decresce e tende ad un limite D, mentre la sequenza $\{E(X_t^2)\}$ cresce tendendo al limite $E(Z^2) - D^2$. La dimostrazione è semplice se si ricorda che $E[Z / Y_1,\ldots\ldots,Y_t]$ è interpretabile come proiezione ortogonale di Z sul sottospazio lineare chiuso dell'insieme dei n.a. con momento secondo finito costituito dalle funzioni $\psi(Y_1,\ldots\ldots,Y_t)$ tali che sia $E[\psi(Y_1,\ldots\ldots,Y_t)]^2 < \infty$. Intanto risulta $0 \leq D_t \leq [E(Z^2)]^{1/2}$ e poiché la sequenza {Dt} decresce al crescere di t esiste un limite D. Inoltre, per la proprietà di ortogonalità di $Z - E[Z / Y_1,\ldots\ldots,Y_t]$ nei confronti di tutte le funzioni $\psi(Y_1,\ldots\ldots,Y_t)$ si ha $Z - Xt \perp Xt$ e dunque $E(Z^2) = D_t^2 + E(X_t^2)$. Risulta quindi che $E(X_t^2) \uparrow E(Z^2) - D^2$.

Cenni sui processi stazionari

Un processo stocastico a parametro discreto $\{X_t; t \geq 1\}$ è **stazionario in senso stretto** se le sue distribuzioni congiunte finite-dimensionali soddisfano la seguente condizione di invarianza

$$F\left(x_{t_1}, x_{t_2}, \ldots\ldots\ldots, x_{t_n}\right) = F\left(x_{t_1+h}, x_{t_2+h}, \ldots\ldots\ldots, x_{t_n+h}\right)$$

per ogni intero n, ogni sequenza di interi distinti $\left(t_1, t_2, \ldots\ldots\ldots, t_n\right)$ e ogni vettore di numeri reali $\left(x_{t_1}, x_{t_2}, \ldots\ldots\ldots, x_{t_n}\right)$. Ne discende che i numeri aleatori della sequenza $\{X_t; t \geq 1\}$ sono ugualmente distribuiti, le coppie di n.a. $\left(X_s, X_t\right)$ aventi uguale differenza degli indici s – t sono ugualmente distribuite, ……. .

Un processo stocastico a parametro discreto $\{X_t; t \geq 1\}$ è **stazionario in senso lato** se i n.a. X_t hanno momenti secondi finiti e se

1) la funzione valor medio è costante, cioè se $\varphi_X(t) = E\left(X_t\right) \equiv E\left(X_1\right)$,

2) la funzione di covarianza $\Psi_X(s,t) = Cov\left(X_s, X_t\right)$ dipende da s e t solo attraverso s – t .

Da quanto detto si ricava che la stazionarietà in senso stretto implica quella in senso lato soltanto se i n.a. X_t hanno tutti momenti secondi finiti; l'implicazione inversa sussiste solo quando le distribuzioni congiunte dipendono dai soli momenti del primo e secondo ordine, come accade per esempio alle distribuzioni di tipo Gaussiano e Student – t .

Un primo esempio di processo stazionario in entrambi i sensi è costituito da una sequenza di n.a. indipendenti e ugualmente distribuiti con varianza finita σ_X^2 . La funzione di covarianza $\Psi_X(s,t)$ ha due valori, $\Psi_X(s,t) = 0$ se $s \neq t$ e $\Psi_X(s,t) \equiv \sigma_X^2$ se $s = t$. I casi più significativi di stazionarietà sono però quelli in cui i n.a. X_t sono mutuamente dipendenti: l'esempio più prossimo al precedente è dato da una sequenza di n.a. ugualmente distribuiti con una funzione di covarianza con due valori, $\Psi_X(s,t) = \gamma \neq 0$ se $s \neq t$ e $\Psi_X(s,t) \equiv \sigma_X^2$ se $s = t$ (nel seguito useremo il termine di "scambiabilità del secondo ordine" per indicare un modello di questo tipo).

a) Funzione di covarianza e funzione spettrale

Lo studio dei processi stazionari (del secondo ordine) può avvenire da due diversi punti di vista: l'approccio **temporale** che utilizza la funzione di covarianza quale strumento principale e l'approccio **frequenziale** (o spettrale) che utilizza quale strumento principale la funzione spettrale, definita come la trasformata di Fourier della funzione di covarianza. Si ha in proposito il seguente

Teorema: se $\{X_t; t \geq 1\}$ è un processo stazionario in senso lato, a valori reali, con funzione valor medio identicamente nulla e funzione di covarianza $\Psi_X(h)$, h $\in$ Z, si ha :

a) $\Psi_X(h)$ è semi-definita positiva, cioè soddisfa, per ogni intero positivo N ed ogni sequenza di numeri reali ($a_1, \ldots\ldots\ldots, a_N$), la condizione $\sum_{i,j=1}^{N} a_i . a_j . \Psi_X(i-j) \geq 0$ e, viceversa, ogni funzione definita in Z e semi-definita positiva è la funzione di covarianza di un processo stazionario;

b) esiste una funzione (spettrale) F(λ), a valori reali, monotona non decrescente e limitata, tale che $\Psi_X(h)$ ha la rappresentazione $\Psi_X(h) = \int_{-\pi}^{\pi} \cos(\lambda h)$ d F(λ) ; la corrispondenza tra l'insieme delle $\Psi_X(h)$ e quello delle funzioni spettrali F(λ) può essere resa biunivoca imponendo, per esempio, l'ulteriore condizione F(-π) = 0 e assumendo F(λ) continua a destra in ogni eventuale punto di discontinuità;

c) se $\Psi_X(h)$ è assolutamente sommabile, cioè se verifica la condizione $\sum_{-\infty}^{+\infty} |\Psi_X(h)| < +\infty$, allora esiste f(λ) = F'(λ), detta densità spettrale, continua in (-π , π), simmetrica rispetto a λ = 0, tale che $\Psi_X(h) = \int_{-\pi}^{\pi} \cos(\lambda h)$. f(λ) dλ e che f(λ) = $(2\pi)^{-1}$. $\sum_{h=-\infty}^{+\infty} \cos(\lambda h) . \Psi_X(h)$.

Per la dimostrazione di questo teorema e degli altri che seguiranno rinviamo il lettore per esempio al testo di P.J.Brockwell e R.A.Davis "Time Series: Theory and Methods" della Springer. Vediamo ora alcuni esempi di processi stazionari e di funzioni spettrali:

1. **Processi con numeri aleatori i.i.d**. e dotati di momento secondo finito: la funzione di covarianza $\Psi_X(h)$ assume il valore σ_X^2 se h = 0 ed è nulla per gli altri valori di h; la corrispondente densità spettrale è identicamente uguale a $\sigma_X^2 / 2\pi$. Si noti che anche il processo WN (0 ; σ^2) ha la medesima densità spettrale.

2. **Processo armonico** $X_t = Y.\cos(\lambda t) + Z.\sin(\lambda t)$, $0 < \lambda < \pi$, ove le speranze matematiche di Y e Z sono assunte nulle, le varianze sono assunte uguali e denotate con σ^2 e la loro covarianza è assunta nulla; il parametro λ, denominato "frequenza angolare", è assunto positivo e non maggiore di π. E' immediato provare che è $\varphi(t) = 0$ e si trova facilmente la

$\psi(s,t) = \sigma^2 . \cos \lambda(t-s)$. Dal momento che questa funzione di covarianza **non** è assolutamente sommabile, non esiste una corrispondente densità spettrale. La funzione spettrale $F(\lambda)$ corrispondente alla $\psi(s,t)$ è una funzione a gradini con discontinuità in $-\lambda$ e λ e l'ampiezza dei salti è pari a $\sigma_X^2 /2$. Questo stesso processo stocastico può avere altre due **rappresentazioni equivalenti**; la prima è $X_t = V. \sin(\lambda t + \eta)$, ove il nuovo parametro η, detto "spostamento di fase", è definito ponendo $Y = V.\sin\eta$ e $Z = V.\cos\eta$, con $V = (Y^2 + Z^2)^{1/2}$, di modo che è banalmente $\sin\eta = Y/V$ e $\cos\eta = Z/V$.

La seconda rappresentazione alternativa è $X_t = A^*.e^{-i\lambda t} + A .e^{i\lambda t}$, ove si è posto A = (Y– iZ)/2 e ove A* denota il numero complesso coniugato di A. Questa seconda rappresentazione introduce formalmente la frequenza angolare negativa $-\lambda$ e numeri aleatori a valori complessi, ma si vedrà nel seguito che tale complicazione formale consente un'effettiva semplificazione negli sviluppi matematici.

3. Processi ottenuti da **somme finite di processi armonici**.

Sia $X_t = \sum_{j=1}^{N} (Y_j.\cos \lambda_j t + Z_j.\sin \lambda_j t)$, ove $0 < \lambda_1 < \ldots\ldots < \lambda_N < \pi$; i numeri aleatori Y_j e Z_j abbiano speranza matematica nulla, varianza pari a σ^2_j e siano tutti mutuamente non correlati. La funzione valor medio di $\{X_t\}$ è identicamente nulla e la sua funzione di covarianza è data da $\psi(s,t) = \sum_{j=1}^{N} \sigma^2_j . \cos \lambda_j(t-s)$. La corrispondente funzione spettrale $F(\lambda)$ ha 2N discontinuità, simmetriche rispetto a $\lambda = 0$, di ampiezza $\sigma_j^2 / 2$ e risulta $F(\pi) = \sum_{j=1}^{N} \sigma_j^2$.

Come per il precedente processo armonico, sussistono altre due rappresentazioni possibili per l'attuale processo : $X_t = \sum_{j=1}^{N} V_j . \sin (\lambda_j t + \eta_j)$ e $X_t = \sum_{j=-N}^{N} A_j . \exp (i\lambda_j t)$. Nell'ultima rappresentazione, le condizioni affinchè $\{X_t\}$ sia un processo a valori reali sono $A_j = (Y_j – iZ_j)/2$ se $j \geq 0$, $A_{-j} = A_j^*$ e $\lambda_{-j} = - \lambda_j$.

Tale processo può essere esteso ad un numero infinito di addendi se si suppone finita la somma della serie delle varianze σ^2_j ; questa estensione ha una rilevante importanza dal punto di vista teorico per la rappresentazione di processi stazionari "con spettro discreto".

Sussiste in proposito il seguente **teorema di E.E. Slutskij** (1938): ogni processo stazionario in senso lato con funzione valor medio nulla e funzione spettrale costante a tratti (spettro discreto) può essere rappresentato secondo la

$$X_t = \sum_{j=1}^{\infty} b_j .(Y_j .\cos \lambda_j t + Z_j .\sin \lambda_j t) ,$$

con $\sum_{j=1}^{\infty} b_j^{\,2} < \infty$, ove i n.a. Y_j e Z_j sono tutti equi, mutuamente non correlati con $V(Y_j) = V(Z_j) = \sigma_j^2$. La generalizzazione di questo risultato a tutti i processi stazionari in senso lato con funzione valor medio nulla e funzione spettrale arbitraria è dovuta a Kolmogorov, Cramér e Loève.

4) **Processo MA(1)** : Xt = ε_t +b.ε_{t-1} con $\varepsilon_t \sim$ WN (0 ; σ_ε^2) . Si ha $\Psi_X(0) = (1+b^2).\sigma_\varepsilon^2$,

$\Psi_X(1) = b.\sigma_\varepsilon^2$, mentre $\Psi_X(h) = 0$ per h > 1. Evidentemente esiste una densità spettrale, poiché $\sum_{-\infty}^{+\infty}|\Psi_X(h)| = (1+b^2).\sigma_\varepsilon^2 + 2\,b.\sigma_\varepsilon^2$, e la sua espressione è :

$$f(\lambda) = \sigma_\varepsilon^2\ (2\pi)^{-1}.(1+2b.\cos\lambda+b^2)\ .$$

b) Filtri lineari invarianti

Si definisce "filtro lineare invariante" applicato ad un processo stocastico $\{Y_t; t \geq 1\}$ una trasformazione lineare $X_t = \sum_{j=-\infty}^{+\infty} c_j.Y_{t-j}$ del processo {Yt} con coefficienti c_j indipendenti da t e tali da soddisfare una qualche condizione del tipo $\sum_{j=-\infty}^{+\infty} c_j^2 < \infty$ o $\sum_{j=-\infty}^{+\infty}|c_j| < \infty$.

Sussiste il seguente:

Teorema: se la successione di coefficienti { c_j } è assolutamente sommabile e il processo {Yt} è stazionario in senso lato allora {Xt} è stazionario in senso lato con funzione di autocovarianza $\Psi_X(h) = \sum_{j,k=-\infty}^{+\infty} c_j.c_k.\Psi_Y(h-j+k)$ e funzione spettrale verificante la $dF_X(\lambda) = |H(\lambda)|^2.dF_Y(\lambda)$, ove H(λ) è la funzione di trasferimento del filtro definita dalla $H(\lambda) = \sum_{j=-\infty}^{+\infty} c_j.e^{-ij\lambda}$. Se esiste la densità spettrale $f_Y(\lambda)$ allora esiste anche $f_X(\lambda)$ che verifica la relazione $f_X(\lambda) = |H(\lambda)|^2 . f_Y(\lambda)$.

Nell'esempio 4) è $H(\lambda) = 1 + b.\exp(-i\lambda)$ e $|H(\lambda)|^2 = 1 + 2b\cos(\lambda) + b^2$ per cui si ha $f_X(\lambda) = \sigma_\varepsilon^2.(1+2b.\cos\lambda+b^2)/2\pi$, essendo $\sigma_\varepsilon^2/2\pi$ la densità spettrale del processo $\{\varepsilon_t\}$.

5. **Processo AR(1)** : $X_t - a.X_{t-1} = \varepsilon_t$, con $\{\varepsilon_t\} \sim$ WN (0 ; σ_ε^2) e $|a| < 1$.

Sostituendo in $X_t = a.X_{t-1} + \varepsilon_t$ a X_{t-1} l'espressione $a.X_{t-2} + \varepsilon_{t-1}$ si ottiene $X_t = a^2.X_{t-2} + \varepsilon_t + a.\varepsilon_{t-1}$; continuando allo stesso modo le sostituzioni si perviene alla $X_t = a^n.X_{t-n} + \varepsilon_t + \sum_{k=1}^{n-1} a^k.\varepsilon_{t-k}$.

Passando al limite per n tendente all'infinito si ha $X_t = \varepsilon_t + \sum_{k=1}^{\infty} a^k.\varepsilon_{t-k}$, dalla quale il processo{X_t}

appare ottenuto con l'applicazione al processo $\{\varepsilon_t\}$ di un filtro lineare invariante. La funzione di trasferimento di questo filtro è la $H(\lambda) = 1 + \sum_{k=1}^{+\infty} a^k . e^{-ik\lambda} = (1 - a.e^{-i\lambda})^{-1}$ e risulta $|H(\lambda)|^2 =$

$(1 - 2a.\cos\lambda + a^2)^{-1}$. Si ha dunque $f_X(\lambda) = |H(\lambda)|^2 . f_Y(\lambda) = \sigma_\varepsilon^2 . (1 - 2a.\cos\lambda + a^2)^{-1} / 2\pi$.

6. **Processo ARMA (1 , 1)** : $X_t - a.X_{t-1} = \varepsilon_t + b.\varepsilon_{t-1}$, con $\{\varepsilon_t\} \sim WN(0 ; \sigma_\varepsilon^2)$ e $|a| < 1$.

E' opportuno riguardare il modello suddetto come un modello AR(1), $X_t - a.X_{t-1} = Y_t$, con $Y_t = \varepsilon_t + b.\varepsilon_{t-1}$. Per quanto già visto, {Yt} è un processo stazionario con densità spettrale $f_Y(\lambda) = \sigma_\varepsilon^2 (1 + 2b\cos\lambda + b^2) / 2\pi$ per cui è:

$$f_X(\lambda) = |H(\lambda)|^2 . f_Y(\lambda) = \frac{\sigma_\varepsilon^2}{2\pi} . \frac{(1 + 2b.\cos\lambda + b^2)}{(1 - 2a.\cos\lambda + a^2)} .$$

c) Cenni sui processi scambiabili in senso lato

Vedremo che i processi scambiabili sono una sottoclasse di processi stazionari; se si suppongono finiti i momenti secondi dei n.a. X_t di un processo scambiabile, ha senso considerare la condizione di scambiabilità in senso lato: essa è caratterizzata dal fatto che la funzione valor medio $\varphi_X(t)$ è costante e la funzione di covarianza $\psi_X(h)$ ha soltanto i due valori $\psi_X(0) = \sigma_X^2$ e $\psi_X(h) = \gamma > 0$ se h è diverso da 0. Evidentemente la $\psi_X(h)$ non è assolutamente sommabile e si prova che la funzione spettrale $F(\lambda)$ ha una discontinuità nell'origine e che la sua espressione è $F(\lambda) = \gamma . F_1(\lambda) + (\sigma_X^2 - \gamma) . F_2(\lambda)$, ove $F_1(\lambda)$ è una funzione di ripartizione che concentra la massa unitaria nell'origine ed $F_2(\lambda)$ una funzione di ripartizione con densità uniforme in $[-\pi , \pi]$.

7. **Sia $X_t = Y + \varepsilon_t$** ove $\{\varepsilon_t\}$ è un processo WN $(0 ; \sigma_\varepsilon^2)$. Il n.a. Y abbia valor medio nullo,

varianza σ_Y^2 e sia non correlato con ciascun n.a. ε_t. E' facile verificare che $E(X_t) \equiv 0$ e che la funzione di covarianza $\Psi_X(h)$ ha due valori: $\Psi_X(0) = \sigma_Y^2 + \sigma_\varepsilon^2$ e $\Psi_X(h) = \sigma_Y^2$ se $h \neq 0$. Siamo in presenza di una funzione di covarianza a due valori, per cui, ponendo $\sigma_Y^2 + \sigma_\varepsilon^2 = \sigma^2$ e $\sigma_Y^2 = \gamma$, la funzione spettrale è la $F(\lambda) = \gamma . F_1(\lambda) + (\sigma^2 - \gamma) . F_2(\lambda)$. In altri termini, il processo $\{X_t\}$ è scambiabile in senso lato.

Si può provare il seguente **teorema di rappresentazione** : tutti e soli i processi stocastici scambiabili in senso lato sono rappresentabili come $X_t = Y + \varepsilon_t$ ove i n.a. a secondo membro hanno le caratteristiche suddette con Var (Y) = $\Psi_X(h)$, con $h \neq 0$, e $\sigma_\varepsilon^2 = \Psi_X(0) - \Psi_X(h)$.

d) Processi stazionari a valori complessi

Esempio: $X_t = \sum_{j=1}^{N} A_j . \exp(i\lambda_j t)$, ove i n.a. complessi A_j sono equi, con varianza σ_j^2 e non correlati, mentre i numeri certi λ_j appartengono all'intervallo $(-\pi , \pi]$. Si trova facilmente che i n.a. X_t sono equi e che la loro funzione di covarianza ha l'espressione $\psi_X(h) = \sum_{j=1}^{N} \sigma_j^2 . \exp(i\lambda_j h)$. La corrispondente funzione spettrale è costante a tratti con discontinuità di ampiezza σ_j^2 nei punti $\lambda = \lambda_j$; naturalmente è $F_X(\lambda_N) = F_X(\pi) = \sum_{j=1}^{N} \sigma_j^2$.

Forniremo ora due espressioni formali equivalenti per $\psi_X(h)$ e per il processo $\{X_t\}$: poiché è $F_X(\lambda) = \sum_{j:\lambda_j \le \lambda} \sigma_j^2$ possiamo scrivere $\psi_X(h) = \int_{-\pi}^{\pi} \exp(i\lambda h)\, \mathrm{d}\, F_X(\lambda)$ e rappresentare quindi $\psi_X(h)$ come un integrale di Stieltjes con funzione peso $F_X(\lambda)$. Ponendo invece $Z(\lambda) = \sum_{j:\lambda_j \le \lambda} A_j$ si può scrivere $X_t = \int_{-\pi}^{\pi} \exp(i\lambda h)\, \mathrm{d}\, Z(\lambda)$: si tratta di un integrale stocastico, del tipo detto "integrale di Wiener", rispetto al processo{Z(λ)}. Quest'ultimo è un processo a parametro continuo definito in $(-\pi , \pi]$ con funzione valor medio identicamente nulla e con incrementi non correlati; inoltre è $E|Z(\lambda)|^2 = F_X(\lambda)$. Viene denominato "processo spettrale" associato a {Xt}.

Il fatto notevole è che le due suddette rappresentazioni sussistono, con diverse funzioni peso $F(\lambda)$ e Z(λ), per tutti i processi stazionari in covarianza con valori complessi e con momenti secondi assoluti finiti. Sono entrambe rappresentazioni nel dominio frequenziale o spettrale e costituiscono decomposizioni della funzione di covarianza e del processo stazionario in componenti cicliche di frequenze angolari $\lambda \in (-\pi , \pi]$.

Diamo ora gli enunciati dei corrispondenti teoremi:

Teorema di Herglotz: la funzione di covarianza di un processo $\{Y_t\}$ stazionario in senso lato con $E|Y_t|^2 < +\infty$, in quanto funzione semidefinita positiva, è rappresentabile come

$$\psi_Y(h) = \int_{-\pi}^{\pi} \exp(i\lambda h)\ \mathrm{d}\, F_Y(\lambda) \ ,$$

ove $F_Y(\lambda)$ è a valori reali, monotona non decrescente e limitata. Solitamente si assume anche che essa sia continua a destra negli eventuali punti di discontinuità e che $F_Y(-\pi) = 0$.

Teorema di Kolmogorov, Cramér, Loève: ad ogni processo $\{Y_t\}$ stazionario in senso lato con $E|Y_t|^2 < +\infty$ e funzione valor medio nulla può essere associato un processo stocastico{$Z_Y(\lambda)$}, $\lambda \in$ $(-\pi , \pi]$, con incrementi non correlati tale che

$$Y_t = \int_{-\pi}^{\pi} \exp(i\lambda t)\, d\, Z_Y(\lambda) \ ,$$

ove $E\,[Z_Y(\lambda)] = 0$, $E\,|Z_Y(\lambda)|^2 = F_Y(\lambda)$, $E\,|dZ_Y(\lambda)|^2 = d\,F_Y(\lambda)$

Infine enunceremo un terzo fondamentale teorema concernente i processi stazionari in covarianza con funzione valor medio identicamente nulla:

Teorema di H. Wold: ogni processo $\{Y_t\}$ stazionario in senso lato con $E|Y_t|^2 < +\infty$ e funzione valor medio nulla può essere rappresentato al modo seguente

$$Y_t = V_t + \sum_{j=0}^{\infty} \gamma_j . \varepsilon_{t-j}$$

ove $\{\varepsilon_t\} \sim$ WN $(0\,; \sigma_\varepsilon^2)$, $\sum_{j=0}^{\infty} \gamma_j^2 < \infty$, $\gamma_0 = 1$, $Cov(V_t, \varepsilon_s) \equiv 0$ e ove il processo stocastico $\{V_t\}$ è "deterministico" nel senso che esso è completamente determinato dal suo passato.

Si può affermare che il teorema di H. Wold costituisce il fondamento probabilistico della tecnica statistica nota come Analisi delle serie temporali (o delle serie storiche).

e) Processi scambiabili e parzialmente scambiabili

La nozione di **processo scambiabile** è stata introdotta da B. de Finetti nel 1928 con una comunicazione al Congresso Internazionale dei Matematici di Bologna dal titolo "Funzione caratteristica di un fenomeno aleatorio" con la quale:

- introdusse la nozione di processo stocastico scambiabile,

- dimostrò il corrispondente teorema di rappresentazione e

- risolse il problema dell'inferenza statistica in ipotesi di scambiabilità delle variabili osservabili fornendo una giustificazione razionale del principio di induzione .

Mentre la modesta frase iniziale della suddetta comunicazione – "Scopo di questa comunicazione è di mostrare come il metodo della funzione caratteristica, già così vantaggiosamente introdotto nella teoria delle variabili casuali, si presti pure assai utilmente allo studio dei fenomeni aleatori" – non lascia trasparire i notevoli contributi innovativi in essa contenuti, la frase finale - "Queste conclusioni e questi esempi possono chiarire l'influenza che sulla valutazione di una probabilità esercitano i dati dell'esperienza." – fa intravedere che la ragione principale della introduzione della nozione di scambiabilità era stata l'intenzione di chiarire le condizioni in cui l'osservazione di una frequenza su una sequenza di prove fornisce coerentemente la base per una valutazione di probabilità o per una previsione della frequenza su una sequenza di prove ancora da eseguire. In altri termini l'autore chiarisce l'insieme di ipotesi che giustifica il "principio di induzione", cioè la propensione a valutare la probabilità di un evento E_{n+1} in termini della frequenza relativa di successo osservata su n eventi **analoghi** ad E_{n+1}; formalmente:

$$P(E_{n+1}/K) \cong \frac{m}{n} \text{ se } K = \left\{ \sum_{j=1}^{n} \left| E_j \right| = m \right\} .$$

Dal punto di vista formale, i processi scambiabili sono caratterizzati dalla condizione di invarianza delle distribuzioni congiunte rispetto a permutazioni arbitrarie degli indici dei n.a., cioè dall'invarianza delle distribuzioni rispetto all'ordine dei n.a. . Si richiede cioè che sia

$$F_{1,\ldots\ldots\ldots,n}(x_1,\ldots\ldots\ldots,x_n) = F_{j_1,\ldots\ldots\ldots,j_n}(x_1,\ldots\ldots\ldots,x_n)$$

per ogni intero positivo n $\geq$ 1, per ogni permutazione ($j_1,\ldots\ldots\ldots\ldots,j_n$) della sequenza (1,……,n) e per ogni sequenza di argomenti ($x_1,\ldots\ldots\ldots,x_n$). Per quanto concerne i momenti fino al secondo ordine si ha evidentemente $\varphi_X(t) = E(X_t) \equiv E(X_1)$ e $\psi_X(s,t) = Cov(X_s, X_t) = \psi_X(s-t) = \sigma^2$ o $\psi_X(s-t) = \gamma > 0$ a seconda che sia s – t = 0 o, rispettivamente, s – t $\neq$ 0 . La suddetta definizione implica che ogni insieme finito di n n.a. distinti X_t abbia la stessa distribuzione di probabilità congiunta, cioè

$$F_{t_1,\ldots\ldots\ldots,t_n}(x_1,\ldots\ldots\ldots,x_n) = F_n(x_1,\ldots\ldots\ldots,x_n),$$

qualunque sia l'intero positivo n, la sequenza di interi positivi distinti ($t_1,\ldots\ldots\ldots,t_n$) e la sequenza di argomenti ($x_1,\ldots\ldots\ldots,x_n$) e ove F_n indica la funzione di ripartizione congiunta per un qualunque insieme di n n.a. distinti.

Dal punto di vista interpretativo la condizione di scambiabilita' traduce l'idea di **analogia o equivalenza** tra i n.a. osservabili X_t in modo ben piu' efficace della condizione di indipendenza stocastica e uguale distribuzione degli stessi: infatti, se gli X_t sono assunti mutuamente indipendenti l'apprendimento dall'esperienza non puo' avvenire. Per il caso $X_t = |E_t|$ l'assunzione di indipendenza tra gli eventi equivale ad assumere che sia $P(E_{n+1}/K) \equiv P(E_{n+1})$ qualunque sia l'evento osservabile K riguardante gli eventi $E_1,.....,E_n$. Per contro, l'assunzione di scambiabilita' tra gli eventi introduce tipicamente una dipendenza stocastica tra essi e questa implica la $P(E_{n+1}/K) \cong \frac{m}{n}$, come facilmente si puo' provare.

Sussiste un importante teorema di rappresentazione per i processi scambiabili illimitati, cioè costituiti da una infinità numerabile di n.a. X_t .

Teorema di B. de Finetti: tutti e soli i processi scambiabili illimitati $\{X_t; t \geq 1\}$ sono combinazioni lineari (o misture) di processi stocastici $\{X_t^{(\omega)}; t \geq 1\}$, $\omega \in \Omega$, i cui n.a. sono indipendenti e dotati di una comune funzione di ripartizione $F^{(\omega)}(x)$ nel senso che esiste una funzione di ripartizione G(ω) tale che per ogni intero positivo n ed ogni sequenza finita di indici ($t_1,.....,t_n$) sussiste la

$$F_{t_1,.....,t_n}(x_1,.....,x_n) = \int_\Omega F^{(\omega)}(x_1,.....,x_n)\ dG(\omega) = \int_\Omega \left[\prod_{j=1}^{n} F^{(\omega)}(x_j)\right]\ dG(\omega).$$

In questa sede ci limitiamo a fornire due semplici esemplificazioni del suddetto teorema di rappresentazione: nel primo esempio le funzioni di ripartizione univariate $F^{(\omega)}$ siano tutte di tipo normale o Gaussiano con valor medio ω e varianza unitaria; supponiamo ancora che G(ω) sia la funzione di ripartizione di una distribuzione normale con parametri $\mu \quad e \quad \sigma^2$. Si ha allora:

$$F_{t_1,.....,t_n}(x_1,.....,x_n) = \int_R \left[\prod_{j=1}^{n} (2\pi)^{-1/2}.\exp\{-(x_j-\omega)^2/2\}\right]\ dG(\omega) =$$

$$= (2\pi)^{-n/2} \int_R \left[\exp\{-\frac{1}{2}\sum_{j=1}^{n}(x_j-\omega)^2\}\right]\left[(2\pi\sigma^2)^{-1/2}.\exp\{-\frac{\sigma^2}{2}(\omega-\mu)^2\}\right] d\omega$$

dalla quale si ricava, con facili calcoli, che $F_{t_1,.....,t_n}(x_1,....., x_n)$ corrisponde ad una distribuzione normale n – dimensionale caratterizzata da valori medi tutti uguali a μ, varianze tutte uguali a $1+\sigma^2$ e covarianze tutte uguali a σ^2.

Nel secondo esempio i n.a. X_t siano indicatori di eventi $|E_t|$ scambiabili per i quali si considera il prodotto logico $A=\{E_1^{'} \wedge E_2^{'} \wedge ..... \wedge E_n^{'}\}$ implicante l'evento $\{\sum_{j=1}^{n}|E_j| = m\}$; supponiamo anche che sia noto il valore logico (vero o falso) di ogni evento del prodotto logico A. Le funzioni di ripartizione $F^{(\omega)}(x)$ sono identicamente nulle per x < 0, uguali a ω per $0 \le x < 1$ e identicamente uguali a 1 per x > 1. Se assumiamo che $G(\omega)$ sia la funzione di ripartizione di una distribuzione Beta con parametri numerici $\alpha \quad e \quad \beta$ l'applicazione del teorema di rappresentazione fornisce la

$$P(A)=\int_0^1 \left[\omega^m.(1-\omega)^{n-m}\right].\left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha).\Gamma(\beta)}\omega^{\alpha-1}.(1-\omega)^{\beta-1}\right]d\omega$$

dalla quale si ricava, con facili calcoli, che è $P(A)=\frac{(\alpha)_m.(\beta)_{n-m}}{(\alpha+\beta)_n}$, ove $(\alpha)_m=\prod_{j=0}^{m-1}(\alpha+j)$.

La condizione di **scambiabilità parziale** è stata introdotta sempre da B. de Finetti nel 1937 ed è più flessibile rispetto a quella di scambiabilità e quindi più adatta a modellizzare schemi nei quali l'analogia tra le unità campionarie non è considerata perfetta. Si pensi a misurazioni ripetute di una stessa grandezza effettuate con strumenti di misura aventi precisioni diverse, a rischi assicurativi (per esempio contratti R.C.A.) giudicati non omogenei (per esempio per la diversa cilindrata degli autoveicoli), ai risultati aleatori di lanci ripetuti di monete diverse e così via.

Il modello più semplice di processo parzialmente scambiabile consiste di un **insieme finito** (per esempio una coppia, o una terna,.....) **di processi stocastici scambiabili, tra loro correlati**. Si considerino, per semplicità, due processi stocastici distinti $\{X_t; t \ge 1\}$ e $\{Y_t; t \ge 1\}$ ciascuno dei quali è supposto essere scambiabile. Circa i legami di dipendenza stocastica tra i due processi si possono ipotizzare tante situazioni diverse: i casi limite sono quello di indipendenza tra essi e, all'opposto, quello di scambiabilità semplice generalizzata (in cui non c'è bisogno di considerare distinti i n.a. del primo da quelli del secondo processo). Tra queste situazioni limite c'è tutta la gamma di situazioni di scambiabilità parziale non banale: ciò che le accomuna è la condizione di uguale covarianza tra ogni n.a. X_t del primo processo e tutti i n.a. Y_s del secondo, cioè $\mathrm{Cov}(X_t, Y_S) = \gamma$ per ogni coppia di valori (t,s) degli indici. Ovviamente dovrà essere $|\gamma| \le \sqrt{\sigma_X^2.\sigma_Y^2}$.

Dal punto di vista formale, i processi parzialmente scambiabili (costituiti da due processi scambiabili) sono caratterizzati dalla seguente proprieta' di invarianza delle distribuzioni congiunte:

$$F(x_{t_1},.....,x_{t_n};y_{s_1},.....,y_{s_m}) = F(x_1,.....,x_n;y_1,.....,y_m)$$

per ogni coppia di interi non negativi n, m e ogni coppia di sequenze di interi positivi distinti $t_1,\ldots\ldots\ldots,t_n \quad e \quad s_1\ldots\ldots\ldots\ldots\ldots,s_m$.

Sussiste il seguente risultato dovuto a B. De Finetti:

Teorema di rappresentazione: tutti e soli i processi parzialmente scambiabili illimitati (costituiti da due processi scambiabili componenti) sono combinazioni lineari (o misture) di processi stocastici i cui n.a. sono indipendenti, con una funzione di ripartizione $F_1^{(\omega)}(x)$ comune a tutti i n.a. X_t e una funzione di ripartizione $F_2^{(\eta)}(y)$ comune a tutti i n.a. Y_s nel senso che esiste una distribuzione congiunta $G(\omega,\eta)$ tale che sia

$$F\left(x_{t_1},\ldots\ldots\ldots\ldots,x_{t_n};y_{s_1},\ldots\ldots\ldots\ldots,y_{s_m}\right)=\iint_{\{\omega,\eta\}}\left[\prod_{i=1}^{n}F_1^{(\omega)}(x_{t_i})\right].\left[\prod_{j=1}^{m}F_2^{(\eta)}(y_{s_j})\right]dG(\omega,\eta)$$

per ogni coppia di interi non negativi (n,m) e ogni coppia di sequenze di interi positivi $t_1,\ldots\ldots\ldots\ldots,$ $t_n \quad e \quad s_1\ldots\ldots\ldots\ldots\ldots,s_m$.

In un primo esempio di applicazione del teorema precedente in cui $X_t \quad e \quad Y_s$ sono indicatori di eventi sia $g(\omega,\eta)$, la densità congiunta corrispondente a $G(\omega,\eta)$, una densita' di probabilita' di tipo Beta bivariata con parametri reali positivi $\nu_1,\nu_2 \quad e \quad \nu_3$, cioè

$$g(\theta_1,\theta_2)=\frac{\Gamma(\nu_1+\nu_2+\nu_3)}{\Gamma(\nu_1).\Gamma(\nu_2).\Gamma(\nu_3)}\theta_1^{\nu_1-1}.\theta_2^{\nu_2-1}.(1-\nu_1-\nu_2)^{\nu_3-1}.$$

Si prova allora che i processi $\{X_t\} \quad e \quad \{Y_s\}$ riescono entrambi scambiabili con $P\left\{\sum_{t=1}^{n}X_t=n\right\}=\frac{(\nu_1)_n}{(\nu_1+\nu_2+\nu_3)_n}$ e, rispettivamente, $P\left\{\sum_{t=1}^{n}Y_t=n\right\}=\frac{(\nu_2)_n}{(\nu_1+\nu_2+\nu_3)_n}$; si ha infine

$$P\left\{[\sum_{t=1}^{n}X_t=n]\wedge[\sum_{s=1}^{m}Y_s=m]\right\}=\frac{(\nu_1)_n.(\nu_2)_m}{(\nu_1+\nu_2+\nu_3)_{n+m}}.$$

Quest'ultima probabilità è quella che n eventi del primo tipo (con indicatori X_t) ed m eventi del secondo tipo (con indicatori Y_s) siano tutti veri, cioè che la frequenza di successo su n eventi del primo tipo ed m eventi del secondo tipo sia pari a n + m.

In un secondo esempio di applicazione le funzioni di ripartizione univariate $F_1^{(\omega)}(x)$, comune a tutti i n.a. X_t, ed $F_2^{(\eta)}(y)$, comune a tutti i n.a. Y_s, siano entrambe di tipo Gaussiano con valori medi ω ed η e varianze unitarie; supponiamo ancora che $G(\omega,\eta)$ sia la funzione di ripartizione di una distribuzione Gaussiana bivariata con vettore medio $\mu=(\mu_1,\mu_2)^T$ e matrice di varianze e covarianze $\Gamma=\begin{bmatrix}\gamma_1 & \gamma\\ \gamma & \gamma_2\end{bmatrix}$. Si verifica allora facilmente che $F\left(x_{t_1},\ldots\ldots\ldots\ldots,x_{t_n};y_{s_1},\ldots\ldots\ldots\ldots,y_{s_m}\right)$ e' la frunzione di ripartizione di una distribuzione Gaussiana (n+m) – dimensionale caratterizzata dai

momenti $E(X_t) \equiv \mu_1$, $E(Y_s) \equiv \mu_2$, $Var(X_t) \equiv 1+\gamma_1$, $Var(Y_s) \equiv 1+\gamma_2$, $Cov(X_t, Y_s) \equiv \gamma$.

Per ulteriori approfondimenti sui processi scambiabili e parzialmente scambiabili si possono consultare, per esempio:

1) B. de Finetti – "Teoria delle probabilita'" (G. Einaudi, 1970 oppure, la corrispondente edizione in lingua inglese edita da J. Wiley, 1974).

2) L. Daboni e A. Wedlin – "Statistica: un'introduzione all'impostazione neo-bayesiana" (UTET, 1982).

3)J.M. Bernardo and A.F.M. Smith – "Bayesian Theory" (J. Wiley, 1994).

Quale testo di riferimento per la teoria dei processi stocastici a parametro discreto indichiamo:

A.N. Shiryayev – "Probability" (Springer-Verlag, 1984).

Cenni sui modelli lineari dinamici stocastici a tempo discreto

1) Rappresentazioni per modelli lineari dinamici deterministici

I modelli lineari sono la categoria più semplice di modelli dinamici e ciò spiega la loro ampia utilizzazione nelle applicazioni. Esiste per essi una teoria sufficientemente generale della quale tenteremo di fornire gli elementi introduttivi. Bisogna subito distinguere i modelli a tempo discreto da quelli a tempo continuo: i primi sono tipicamente impiegati nella rappresentazione dei fenomeni economici; i secondi in quella dei fenomeni fisici. I tipi principali di modelli matematici lineari per l'analisi dinamica sono i seguenti:

1) equazioni (e sistemi di equazioni) differenziali e alle differenze finite;
2) modelli nello spazio degli stati (o markoviani);
3) modelli input – output (definiti in termini della funzione di trasferimento).

Per prima cosa tratteremo dei modelli lineari dinamici **a tempo discreto** per i quali, con riferimento all'elenco precedente, il primo tipo di rappresentazione è costituito da equazioni alle differenze finite, lineari, con coefficienti costanti:

$$y_t - \sum_{j=1}^{p} a_j . y_{t-j} = u_t, \quad t = 1, 2, \ldots\ldots\ldots ,$$

ove $\{u_t\}$ è una successione nota ed i coefficienti a_j sono fissati, mentre $\{ y_t \}$ è la successione incognita che va determinata risolvendo l'equazione dopo aver fissato una "condizione iniziale", per esempio la sequenza iniziale ($y_1, \ldots\ldots\ldots\ldots, y_p$) della $\{ y_t \}$.

La seconda rappresentazione, corrispondente alla precedente, prevede l'introduzione delle "variabili di stato" $x_t^{(h)}$, h = 1,…..,p , per il cui vettore $\mathbf{x}_t = \left[x_t^{(1)}, \ldots\ldots\ldots\ldots, x_t^{(p)} \right]^T$ sussistono le equazioni seguenti:

$$\mathbf{x}_t = F \,.\, \mathbf{x}_{t-1} + \mathbf{b} \,.\, u_t$$

$$y_t = \mathbf{h}^T . \mathbf{x}_t$$

ove F è una matrice quadrata di dimensione (p,p), **b** ed **h** sono vettori colonna p-dimensionali. La scelta delle variabili di stato, per uno stesso modello, non è unica; nel nostro caso una scelta possibile è la seguente: $x_t^{(1)} = y_t$, ……, $x_t^{(p)} = y_p$. In corrispondenza, si ha:

$$F = \begin{bmatrix} a_1 & a_2 & a_3 & & a_p \\ 1 & 0 & 0 & & 0 \\ 0 & 1 & 0 & & 0 \\ & & & & \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix}, \quad \mathbf{h} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} .$$

In questa seconda rappresentazione, l'equazione alle differenze lineare di ordine p della prima è trasformata in una equazione vettoriale alle differenze, lineare, del primo ordine.

Il terzo tipo di rappresentazione, molto usato nelle scienze ingegneristiche, si ottiene applicando la "trasformazione Z" ad entrambi i membri della prima equazione alle differenze ottenendo l'espressione

$$Y(z) = H(z).U(z)$$

ove H(z) rappresenta la "funzione di trasferimento" del sistema; essa ha la seguente espressione in termini dei coefficienti dell'equazione alle differenze:

$$H(z) = (1 - \sum_{j=1}^{p} a_j.z^{-j})^{-1} \ .$$

La denominazione di rappresentazione "input – output" corrisponde all'interpretazione di U(z) come input di un "trasformatore", descritto dalla funzione di trasferimento H(z), e a quella di Y(z) come output dello stesso.

Dal momento che nel seguito non impiegheremo questa rappresentazione ci limitiamo a dire che la "trasformazione Z" sostituisce alle successioni numeriche $\{y_t; t \geq 0\}$ e $\{u_t; t \geq 0\}$ le funzioni di variabile complessa $Y(z) = \sum_{j=0}^{\infty} y_j.z^{-j}$ e $U(z) = \sum_{j=0}^{\infty} u_j.z^{-j}$

Esempio: l'equazione della produzione aggregata nel modello di P.A. Samuelson.

Prendiamo in considerazione il modello economico di P.A. Samuelson costituito dalle seguenti equazioni:

$$C_t = \beta.Y_{t-1}, \quad I_t = \gamma.(C_t - C_{t-1}), \quad Y_t = C_t + I_t + G_t,$$

ove $Y_t, C_t \ \ e \ \ I_t$ (produzione, consumo e investimento aggregati) sono variabili endogene, mentre G_t (spesa della pubblica amministrazione) è l'unica variabile esogena del modello. La terza equazione, con le sostituzioni indicate dalle

$$Y_t = C_t + I_t + G_t = \beta.Y_{t-1} + \{\gamma[(\beta.Y_{t-1}) - (\beta.Y_{t-2})]\} + G_t = \beta.(1+\gamma)Y_{t-1} - \beta.\gamma.Y_{t-2} + G_t \ ,$$

costituisce un'equazione alle differenze finite del secondo ordine, lineare, a coefficienti costanti nella successione incognita dei livelli della produzione aggregata $\{Y_t\}$.

A questa prima rappresentazione corrisponde la seguente rappresentazione in termini delle variabili di stato $X_t^{(1)} = Y_t \quad e \quad X_t^{(2)} = Y_{t-1}$, componenti del vettore $\mathbf{X}_t$:

$$\mathbf{X}_t = \begin{bmatrix} \beta.(1+\gamma) & -\beta.\gamma \\ 1 & 0 \end{bmatrix}. \mathbf{X}_{t-1} + \begin{bmatrix} 1 \\ 0 \end{bmatrix}.G_t \ ,$$

$$Y_t = (1 \quad 0). \mathbf{X}_t .$$

Affermiamo che la risoluzione del suddetto sistema di equazioni fornisce, se le condizioni iniziali poste nei due casi sono coerenti, la stessa soluzione della precedente equazione alle differenze del secondo ordine.

2) Rappresentazioni per modelli lineari dinamici stocastici

Quando qualche elemento del modello dinamico, uno o piu' coefficienti oppure l'input $\{u_t\}$ oppure la condizione iniziale (o anche piu' elementi tra questi) non e' noto con certezza allora si parla di **modello dinamico stocastico**. L'elemento non completamente noto dev'essere allora caratterizzato in senso probabilistico e cio' implica che anche la soluzione $\{y_t\}$ possa essere caratterizzata solo probabilisticamente.

Se, per esempio, l'input $\{u_t\}$ del modello non e' noto con certezza esso va considerato un processo stocastico $\{U_t\}$ e definito probabilisticamente sulla base delle informazioni disponibili: se queste consentono di specificare soltanto la funzione valor medio $\varphi_U(t) = \{E(U_t); t \geq 1\}$ e la funzione di covarianza $\psi_U(s,t) = \{Cov[U_s, U_t]; s,t \geq 1\}$ di $\{U_t\}$ anche la specificazione del processo stocastico output $\{Y_t\}$ sara' limitata alle funzioni $\varphi_Y(t)$ e $\psi_Y(s,t)$. Ovviamente, il livello di specificazione probabilistica dell'output non puo' essere piu' elevato di quello dell'input.

Le rappresentazioni gia' menzionate nel paragrafo precedente diventano ora, se l'input e' un processo stocastico $\{U_t\}$:

1) un'equazione alle differenze finite, lineare, stocastica

$$Y_t - \sum_{j=1}^{p} a_j . Y_{t-j} = U_t, \quad t = 1,2,.....$$

e, rispettivamente,

2) un sistema di due equazioni lineari stocastiche

$$\mathbf{X}_t = F \,.\, \mathbf{X}_{t-1} + \mathbf{b} \,.\, U_t \ ,$$

$$Y_t = \mathbf{h}^T . \mathbf{X}_t \ .$$

La risoluzione del modello in ciascuna delle due forme precedenti, fissata che sia la condizione iniziale, non fornisce la corrispondente successione dei livelli per l'output, ma la corrispondente successione dei livelli aleatori $\{Y_t ; t \geq 1\}$ con un livello di specificazione probabilistica non superiore a quello del processo stocastico input $\{U_t ; t \geq 1\}$. Un semplice esempio si ottiene supponendo che $\{U_t ; t \geq 1\}$ sia un processo White Noise (0; σ_U^2) : l'equazione 1) o il corrispondente sistema 2) caratterizzano il processo $\{Y_t ; t \geq 1\}$ come un processo AR(p) la cui eventuale stazionarieta' asintotica dipende dalla ben nota condizione che le radici dell'equazione caratteristica $\lambda^t - \sum_{j=1}^{p} a_j . \lambda^{t-j} = 0$ siano tutte minori di 1 in modulo.

Esempio: la produzione aggregata nella versione stocastica del modello di P.A. Samuelson

Le tre note equazioni ora diventano

$$C_t = \beta . Y_{t-1} + \varepsilon_t, \quad I_t = \gamma . (C_t - C_{t-1}) + \eta_t, \quad Y_t = C_t + I_t + G_t \ ,$$

ove le nuove variabili $\varepsilon_t \quad e \quad \eta_t$ sono "perturbazioni aleatorie" che riassumono l'influenza su C_t e, rispettivamente, su I_t di tutti quei fattori economici e di altro tipo (sociologici, politici,.....) che nel modello sono trascurati. Molto spesso si assume che le perturbazioni aleatorie siano processi stocastici di tipo White Noise, $\varepsilon_t \approx WN(0, \sigma_\varepsilon^2) \quad e \quad \eta_t \approx WN(0, \sigma_\eta^2)$, eventualmente correlati tra loro. L'equazione per la produzione aggregata diventa

$$Y_t - \beta.(1+\gamma).Y_{t-1} + \beta.\gamma.Y_{t-2} = G_t + U_t \ ,$$

ove $U_t = (1+\gamma).\varepsilon_t - \gamma.\varepsilon_{t-1} + \eta_t$, cioe' un'equazione alle differenze finite, lineare, stocastica la cui soluzione fornisce il processo stocastico $\{Y_t ; t \geq 1\}$. Evidentemente e' $E(U_t) \equiv 0$ per cui, dalla precedente equazione, si ottiene

$$E(Y_t) - \beta.(1+\gamma).E(Y_{t-1}) + \beta.\gamma.E(.Y_{t-2}) = E(G_t)$$

che e' un'equazione alle differenze deterministica nella successione incognita $\varphi_Y(t)$ dei valori medi dei livelli della produzione aggregata. E' anche facile ottenere la funzione di covarianza, $\psi_Y(s,t)$, del processo $\{Y_t ; t \geq 1\}$ il che fornisce per quest'ultimo una specificazione del secondo ordine.

Inferenza statistica sulle variabili di stato

Siamo infine arrivati a trattare l'argomento centrale della nostra esposizione. Il modello nello spazio degli stati e' stato introdotto nel paragrafo precedente nella forma

$$\mathbf{X}_t = F \,.\, \mathbf{X}_{t-1} + \mathbf{b} \,.\, U_t \ ,$$

$$Y_t = \mathbf{h}^T .\, \mathbf{X}_t \ ,$$

che costituiva una rappresentazione equivalente dell'equazione alle differenze stocastica

$$Y_t - \sum_{j=1}^{p} a_j .Y_{t-j} = U_t \ .$$

Nel seguito considereremo la seguente generalizzazione dello schema suddetto:

$$\mathbf{X}_t = F \,.\, \mathbf{X}_{t-1} + \mathbf{U}_t \ ,$$

$$\mathbf{Y}_t = H \cdot \mathbf{X}_t + \mathbf{V}_t ,$$

ove la prima equazione, detta “equazione di evoluzione o di stato”, e’ un’equazione alle differenze lineare e stocastica nella successione dei vettori di stato $\{\mathbf{X}_t\}$, mentre la seconda equazione, detta “di osservazione”, mette in relazione il vettore di stato con il vettore delle variabili osservabili $\mathbf{Y}_t$.

I vettori $\mathbf{U}_t$ e $\mathbf{V}_t$ sono perturbazioni aleatorie vettoriali di tipo White Noise non correlate tra loro, cioe’si assume che sia $E(\mathbf{U}_t) = \mathbf{0}$, $E(\mathbf{V}_t) = \mathbf{0}$ e $Cov(\mathbf{U}_t) \equiv \Sigma_U$, $Cov(\mathbf{V}_t) \equiv \Sigma_V$, $Cov(\mathbf{U}_t, \mathbf{V}_s) \equiv [0]$. Nella piu’ semplice di tali generalizzazioni tutti i vettori hanno le stesse dimensioni e le matrici quadrate F ed H, costanti nel tempo, sono note, come anche le matrici di dispersione Σ_U e Σ_V .

Il problema inferenziale che considereremo, detto anche problema di “filtraggio”, consiste nella stima puntuale lineare del vettore di stato non osservabile $\mathbf{X}_t$ sulla base della conoscenza dei vettori $\mathbf{Y}_1$ ....., $\mathbf{Y}_t$. Il metodo di stima, o di approssimazione, che adotteremo e’ quello dei minimi quadrati; ricordiamo al lettore che tale metodo richiede soltanto la specificazione del secondo ordine dei processi stocastici coinvolti nel problema.

1) Approssimazioni lineari dei minimi quadrati per numeri aleatori

Si consideri lo spazio lineare H_0 (di dimensione infinita) dei numeri aleatori X, Y, Z, …… che supponiamo dotati di speranza matematica nulla e momento secondo finito. Sia Y il n.a. di interesse per il quale si voglia costruire una stima (o previsione o approssimazione) in termini di una qualche funzione $\varphi(\mathbf{X})$ dei n.a. X_1,....., X_n costituenti il vettore aleatorio **X**.

Se non si introducono vincoli particolari per la funzione stimatore $\varphi(\mathbf{X})$, eccetto quello $E[\varphi(\mathbf{X})]^2 < \infty$, si prova che la funzione ottimale, nel senso dei minimi quadrati, è la funzione di regressione $E(Y/\mathbf{X})$; formalmente, per ogni ammissibile funzione $\varphi(.)$ si ha :

$$E\left[Y - E(Y/X)\right]^2 \le E\left[Y - \varphi(X)\right]^2 .$$

Se per φ(**X**) si impone il vincolo di **linearità** , cioè se si assume che φ(**X**) sia una funzione lineare, $\sum_{j=1}^{n} \alpha_j . X_j$, dei n.a. $X_1,\ldots\ldots\ldots,X_n$, si devono trovare gli n coefficienti $\hat{\alpha}_j$, j = 1,……,n , per i quali risulta:

$$E\left[Y-\sum_j \hat{\alpha}_j . X_j\right]^2 \le E\left[Y-\sum_j \alpha_j . X_j\right]^2$$

in corrispondenza ad ogni n-pla $\alpha_1,\ldots\ldots\ldots,\alpha_n$ di numeri reali. Si dimostra facilmente che il vettore dei coefficienti ottimali $\hat{\underline{\alpha}}$ è dato da $\hat{\underline{\alpha}}$ = [Cov(**X**)] $^{-1}$. E(Y.**X**) sotto la condizione che la matrice di dispersione di **X** sia invertibile (il che accade se i n.a. Xj sono linearmente indipendenti).

Per dimostrarlo si tratta di porre uguali a 0 le derivate parziali rispetto ad ogni α_j di $E\left[Y-\sum_j \alpha_j . X_j\right]^2$: il sistema lineare che si ottiene ha l'espressione Cov (**X**) . $\hat{\underline{\alpha}}$ = E(Y.**X**) e la sua soluzione è unica se Cov (**X**) è invertibile . Il previsore (o stimatore o approssimatore) lineare ottimale per Y è allora $\hat{Y} = (X_1,\ldots\ldots\ldots,X_n).\left[Cov(X)\right]^{-1}.E(Y.X)$.

Tale numero aleatorio ha un'interessante interpretazione geometrica: $\hat{Y}$ coincide con la **proiezione ortogonale** $P(Y/L)$ di Y sul sottospazio lineare L di H_0 generato dai n.a. $X_1,\ldots\ldots\ldots,X_n$

Per attribuire un significato preciso a tale proposizione è necessario introdurre le seguenti definizioni: in H_0 la **lunghezza** (o norma) del vettore geometrico associato ad un n.a. Z è definita dalla $\|Z\| = \left[Var(Z)\right]^{1/2}$; la **distanza** tra due n.a. Z e V è definita dalla d(Z,V) = $\left[Var(Z-V)\right]^{1/2}$; la **condizione di ortogonalità** tra Z e V è espressa dalla E(Z.V) = 0 (poiché E(Z) = E(V) = 0, Z e V sono ortogonali se Cov(Z,V) = 0).

La suddetta interpretazione geometrica di $\hat{Y}$ si consegue applicando il "principio di ortogonalità" il cui enunciato in questo caso stabilisce che **considerato un qualunque n.a. Y di H_0, il n.a. di L a minima distanza da esso coincide con la proiezione ortogonale di Y su L** quando si osservi che $E\left[Y-\sum_j \alpha_j . X_j\right]^2$ esprime il quadrato della distanza tra Y ed il generico elemento del sottospazio L .

Importa osservare che mentre l'utilizzazione dell'approssimatore ottimale di Y , costituito dalla funzione di regressione E(Y/**X**), richiede la conoscenza della distribuzione congiunta F(

$y, x_1, ..., x_n$) dei n.a. considerati, o almeno della distribuzione subordinata F($y / x_1, ..., x_n$) , la costruzione dell'approssimatore lineare ottimale di Y, costituito dal n.a. $\hat{Y} = P(Y / L)$, richiede la conoscenza (o meglio la specificazione) dei soli momenti del primo e secondo ordine dei n.a. $Y, X_1, ..., X_n$. Una seconda osservazione rilevante è che per le distribuzioni implicanti una funzione di regressione E(Y/**X**) lineare negli elementi di **X** accade che gli approssimatori E(Y/**X**) e $\hat{Y} = P(Y / L)$ coincidono ; le distribuzioni più note aventi questa caratteristica sono quella normale e quella Student – t multivariate.

2) Proprieta' iterativa dell'approssimazione lineare dei minimi quadrati

Si considerino tre spazi lineari tali che sia $L_{n-1} \subset L_n \subset H_0$, ove l'ultimo spazio (di dimensione infinita) contiene tutti i numeri aleatori (n.a.) equi e dotati di varianza finita, mentre $L_{n-1} = L(X_1, ..., X_{n-1})$ ed $L_n = L(X_1, ..., X_n)$, essendo i n.a. X_j linearmente indipendenti ed osservabili. Sia $Y \in H_0$ il n.a. non osservabile per il quale si vogliono determinare le approssimazioni lineari dei minimi quadrati basate sul processo osservabile $\{X_j ; j = 1,2,.....\}$.

Proveremo la seguente **proprietà iterativa**:

$$\hat{Y}_n = \mathrm{P}(Y / L_n) = \mathrm{P}(Y / L_{n-1}) + \mathrm{P}[Y / X_n - \mathrm{P}(X_n / L_{n-1})] ,$$

ove la differenza $X_n - \mathrm{P}(X_n / L_{n-1})$ è il n.a. "innovazione di X_n" che risulta ortogonale a L_{n-1} .

La dimostrazione è fornita dalla seguente semplice sequenza di uguaglianze:

$$\hat{Y}_n = \mathrm{P}(Y / X_1, ..., X_{n-1}, X_n) = \mathrm{P}[Y / X_1, ..., X_{n-1}, X_n - \mathrm{P}(X_n / X_1, ..., X_{n-1})] =$$

$$= \mathrm{P}(Y / X_1, ..., X_{n-1}) + \mathrm{P}(Y / X_n - \hat{X}_{n/n-1}) = \hat{Y}_{n-1} + \mathrm{P}(Y / X_n - \hat{X}_{n/n-1}) .$$

Il significato geometrico del suddetto ragionamento formale si ricava dalle seguenti considerazioni: indicando per brevità con Y_n^* la differenza $Y - \hat{Y}_n$, sussistono evidentemente le due decomposizioni ortogonali $Y = \hat{Y}_{n-1} + Y_{n-1}^*$ e $Y = \hat{Y}_n + Y_n^*$ ed inoltre si ha $\hat{Y}_n = Z_1 + Z_2$, ove $Z_1 = \mathrm{P}(\hat{Y}_n / L_{n-1}) = \hat{Y}_{n-1}$ perché $\mathrm{P}(\hat{Y}_n / L_{n-1}) = \mathrm{P}[\mathrm{P}(Y / L_n) / L_{n-1}] = \mathrm{P}(Y / L_{n-1})$ per la proprietà (δ) del proiettore ortogonale citata nel prossimo numero. Si ha dunque:

$$Y = \hat{Y}_n + Y_n^* = (Z_1 + Z_2) + Y_n^* = \hat{Y}_{n-1} + (Z_2 + Y_n^*)$$

cosicché dev'essere $Z_2 + Y_n^* = Y_{n-1}^*$ la quale implica che sia $\hat{Y}_n = \hat{Y}_{n-1} + Z_2$. Rimane da provare che $Z_2 = \mathrm{P}[Y / X_n - \mathrm{P}(X_n / L_{n-1})] = \mathrm{P}(Y / X_n - \hat{X}_{n/n-1})$ e ciò è già stato ottenuto se si tiene conto della unicità delle decomposizioni ortogonali.

3) Altre proprietà del proiettore ortogonale

α) $P\left[P(X / L_n) / L_n\right] = P(X / L_n)$

β) $\mathrm{P}(a_1.X_1 + a_2.X_2 / L_n) = a_1.P(X_1) + a_2.P(X_2)$,

γ) $\left\langle P(X / L_n), Y \right\rangle = \left\langle X, P(Y / L_n) \right\rangle$, ove $\left\langle X, Y \right\rangle$ = E(X . Y) indica il prodotto interno tra X e Y ,

δ) $P\left[P(X / L_{n-1}) / L_n\right] = P\left[P(X / L_n) / L_{n-1}\right] = P(X / L_{n-1})$,

ε) $P(X / Y_1, Y_2, ..., Y_n) = \sum_{j=1}^{n} P(X / Y_j)$ se i n.a. Y_j sono due a due ortogonali.

4) **Filtro di R. E. Kalman a tempo discreto**

Assumiamo di essere interessati ai n.a. di un processo stocastico {Y_n ; n = 1,2,……}, non osservabile, generato da un modello stocastico lineare espresso dalle equazioni alle differenze finite

(1) $Y_{n+1} = a.Y_n + V_n$, ove $V_n \approx WN(O;\sigma_V^2)$ e ove $V_n \perp Y_m$, $\forall m \le n$.

Si supponga di poter osservare i n.a. X_n definiti dalla trasformazione lineare affine e stocastica

(2) $X_n = b.Y_n + W_n$, ove $W_n \approx WN(0;\sigma_W^2)$ e ove $W_n \perp V_m$, $\forall n, m$.

Il modello stocastico espresso dalle (1) e (2) è denominato "modello lineare dinamico" o anche "modello nello spazio degli stati" in quanto le variabili Y_n sono indicate come "variabili di stato".

Ci porremo il problema di determinare le **approssimazioni lineari dei minimi quadrati** $\hat{Y}_{n/n}$ per le variabili Y_n basate sull'osservazione delle ($X_1,\ldots\ldots\ldots,X_n$); in letteratura questo problema è noto come "problema di filtraggio". Il n.a. $\hat{Y}_{n/n}$ è definito dalle condizioni $E(Y_n - \hat{Y}_{n/n})^2 \le$ $\le E(Y_n - \sum_{j=1}^{n} \alpha_j.X_j)^2$ per ogni n-pla di coefficienti reali $(\alpha_1,\ldots\ldots\ldots,\alpha_n)$ ed è anche indicato con il simbolo $\mathrm{P}(Y_n / L_n)$ a causa della sua interpretazione geometrica quale "proiezione ortogonale di Y_n sullo spazio lineare L_n generato dai n.a. osservabili $X_1,\ldots\ldots\ldots,X_n$ ". Ricordiamo infine al lettore che se si assume che tutti i n.a. considerati abbiano valor medio nullo e varianze finite e si pone $\underline{X}$ = ($X_1,\ldots\ldots\ldots,X_n$)$^{\mathsf{T}}$ allora $\hat{Y}_{n/n}$ è dato dalla

(3) $\hat{Y}_{n/n} = \mathrm{P}(Y_n / L_n) = E(Y_n.\underline{X}^{\mathsf{T}})$. (Cov $\underline{X}$)$^{-1}$. $\underline{X}$.

Ci proponiamo di presentare il procedimento ricorsivo di approssimazione lineare noto come "filtro di Kalman", apparso in letteratura nel 1960; esso è costituito dalle seguenti equazioni, ove si

è indicato in generale con $P_{n/m}$ la varianza $E(Y_n - \hat{Y}_{n/m})^2$ dell'errore di approssimazione $Y_n - \hat{Y}_{n/m}$:

(4) $$\hat{Y}_{n+1/n} = a \,.\, \hat{Y}_{n/n} \ ,$$

(5) $$P_{n+1/n} = E(Y_{n+1} - \hat{Y}_{n+1/n})^2 = a^2.P_{n/n} + \sigma_V^2 \ ,$$

(6) $$\hat{Y}_{n+1/n+1} = \hat{Y}_{n+1/n} + \frac{b.P_{n+1/n}}{b^2.P_{n+1/n} + \sigma_W^2}.(X_{n+1} - b.\hat{Y}_{n+1/n}) \ ,$$

(7) $$P_{n+1/n+1} = \frac{\sigma_W^2.P_{n+1/n}}{b^2.P_{n+1/n} + \sigma_W^2} \ .$$

Le precedenti equazioni presuppongono che si siano già determinati ricorsivamente $\hat{Y}_{n/n}$ e $P_{n/n}$ a partire dalle fissate posizioni iniziali $\hat{Y}_{0/0}$ e $P_{0/0}$; le (4) e (5) sono dette "equazioni di previsione" mentre le (6) e (7) sono dette "equazioni di aggiornamento" cosicché ogni stadio del procedimento ricorsivo consiste di una fase di previsione e una di aggiornamento.

Forniremo ora una dimostrazione delle precedenti equazioni:

- la (4) si prova in base alla proprietà di linearità del proiettore ortogonale perché risulta
  $$\hat{Y}_{n+1/n} = \mathrm{P}(Y_{n+1} / L_n) = \mathrm{P}(a.Y_n + V_n / L_n) = a.\mathrm{P}(Y_n / L_n) + \mathrm{P}(V_n / L_n) = a \,.\, \hat{Y}_{n/n}$$
  in quanto $\mathrm{P}(V_n / L_n) = 0$ giacché si è supposto $V_n \perp Y_m$, $\forall m \le n$, e $W_n \perp V_m$, $\forall n, m$;

- la (5) si prova utilizzando il risultato precedente in quanto è:
  $$P_{n+1/n} = E(Y_{n+1} - \hat{Y}_{n+1/n})^2 = E[a.(Y_n - \hat{Y}_{n/n}) + V_n]^2 = a^2.P_{n/n} + \sigma_V^2 \ ;$$

- la (6) si dimostra utilizzando la proprietà iterativa del proiettore ortogonale che scriveremo indicando con X_{n+1}^* l'innovazione $X_{n+1} - \hat{X}_{n+1/n}$:

$$\hat{Y}_{n+1/n+1} = \hat{Y}_{n+1/n} + \mathrm{P}(Y_{n+1} / X^{*}_{n+1}) = \hat{Y}_{n+1/n} + \frac{Cov(Y_{n+1}, X^{*}_{n+1})}{Var(X^{*}_{n+1})}.X^{*}_{n+1} = \hat{Y}_{n+1/n} + \frac{b.P_{n+1/n}}{b^2.P_{n+1/n} + \sigma^2_W}.X^{*}_{n+1}$$

e l'ultima espressione coincide con la (6) se si tiene presente che è

$$X^{*}_{n+1} = X_{n+1} - \mathrm{P}(X_{n+1} / L_n) = X_{n+1} - \mathrm{P}(b.Y_{n+1} + W_{n+1} / L_n) = X_{n+1} - b.\mathrm{P}(Y_{n+1} / L_n) = X_{n+1} - b.\hat{Y}_{n+1/n} ;$$

- la (7) si prova in base alla definizione di $P_{n+1/n+1}$ e ricorrendo a semplici, anche se noiose sostituzioni:

$$P_{n+1/n+1} = E(Y_{n+1} - \hat{Y}_{n+1/n+1})^2 = \frac{\sigma^2_W.P_{n+1/n}}{b^2.P_{n+1/n} + \sigma^2_W} .$$

Per completezza, riporteremo ora le equazioni della versione vettoriale del filtro di Kalman che riguarda la stima lineare ricorsiva di un processo stocastico vettoriale {$\mathbf{Y}_n$; n = 1,2……..} basata sull'osservabilità del processo vettoriale {$\mathbf{X}_n$; n = 1,2……..}. Le due equazioni del modello lineare dinamico sono ora espresse dalle:

$$\mathbf{Y}_{n+1} = \mathrm{A} \,.\, \mathbf{Y}_n + \mathbf{V}_n \quad \text{e} \quad \mathbf{X}_n = \mathrm{B} \,.\, \mathbf{Y}_n + \mathbf{W}_n \ ;$$

in esse i due processi vettoriali di rumore sono ancora mutuamente non correlati e del tipo White Noise: $\mathbf{V}_n \sim \mathrm{WN}(\mathbf{0}\,;\mathrm{Q})$ e $\mathbf{W}_n \sim \mathrm{WN}(\mathbf{0}\,;\mathrm{R})$.

Supponendo noti $\hat{Y}_{n/n}$ e la matrice di momenti secondi $P_{n/n} = E[(Y_n - \hat{Y}_{n/n}).(Y_n - \hat{Y}_{n/n})^T]$, le equazioni del filtro di Kalman vettoriale sono espresse dalle:

(4') $\hat{Y}_{n+1/n} = A\,.\,\hat{Y}_{n/n}$,

(5') $P_{n+1/n} = A.P_{n/n}.A^T + Q$,

(6') $\hat{Y}_{n+1/n+1} = \hat{Y}_{n+1/n} + P_{n+1/n}.B^T(B.P_{n+1/n}.B^T + R)^{-1}.(X_{n+1} - B.\hat{Y}_{n+1/n})$,

(7′) $P_{n+1/n+1} = P_{n+1/n} - P_{n+1/n}.B^T (B.P_{n+1/n}.B^T + R)^{-1}.B.P_{n+1/n}$.

Per chi volesse approfondire queste prime nozioni suggeriamo, in ordine di completezza crescente della trattazione:

Bittanti S. – “Teoria della Predizione e del Filtraggio” (Pitagora Editrice Bologna, 2000) ,

Catlin D.E. – “Estimation, Control and the Discrete Kalman Filter” (Sprinter-Verlag, 1989) ,

Anderson B.D.O. – Moore J.B. – “Optimal Filtering” (PRENTICE-HALL, INC.,1979) .

Cenni sui modelli lineari dinamici stocastici a tempo continuo

1) Rappresentazioni per modelli lineari dinamici a tempo continuo

Dedicheremo ora qualche cenno ai modelli lineari dinamici a parametro continuo, usati sistematicamente nella Fisica, ma impiegati raramente in Economia matematica. Da qualche tempo, pero’, si assiste ad un loro uso massiccio nella Finanza matematica. I tipi principali di modelli lineari per l’analisi dinamica a tempo continuo sono i seguenti:

a) equazioni (e sistemi di equazioni) differenziali,
b) modelli input – output (definiti in termini della funzione di trasferimento);
c) modelli nello spazio degli stati (o markoviani).

Il primo tipo di rappresentazione è costituito da equazioni differenziali ordinarie lineari con coefficienti costanti:

$$y^{(n)}(t) + \sum_{k=1}^{n} a_k . y^{(n-k)}(t) = u(t) \ ,$$

ove u(t) è una funzione nota (o un processo stocastico) ed i coefficienti a_k sono fissati, mentre $y(t)$ è la funzione deterministica (o stocastica) incognita.

Il secondo tipo di rappresentazione si ottiene applicando la "trasformazione di Laplace" ad entrambi i membri della precedente equazione differenziale ottenendo l'espressione

$$y(s) = H(s).u(s) \ ,$$

ove H(s) rappresenta la "funzione di trasferimento" del sistema; essa ha la seguente espressione in termini dei coefficienti dell'equazione differenziale:

$$H(s) = \frac{1}{s^n + \sum_{k=1}^{n} a_k . s^{n-k}} \ .$$

La denominazione di rappresentazione "input – output" corrisponde all'interpretazione di u(s) come input del "trasformatore", descritto dalla funzione di trasferimento H(s), e a quella di y(s) come output dello stesso. Non faremo uso di questa rappresentazione.

La terza rappresentazione, corrispondente alle due precedenti, prevede l'introduzione delle "variabili di stato" $x_h(t)$, h = 1,.....,n , per il cui vettore **x**(t) sussistono le equazioni seguenti:

$$\frac{d}{dt} x(t) = F.x(t) + b.u(t) \, ,$$

$$y(t) = h^T .x(t) \, ,$$

ove F è una matrice quadrata di dimensione n, **b** ed **h** sono vettori colonna n-dimensionali. La scelta delle variabili di stato, per uno stesso modello, non è unica; nel nostro caso una scelta possibile è la seguente: $x_1(t) = y(t)$, $x_2(t) = y'(t)$, $x_3(t) = y''(t)$, ..., $x_n(t) = y^{(n-1)}(t)$. In corrispondenza, si ha:

$$\mathbf{F} = \begin{bmatrix} 0 & 1 & 0 & \vdots & 0 \\ 0 & 0 & 1 & \vdots & 0 \\ 0 & 0 & 0 & \vdots & 0 \\ 0 & 0 & 0 & \vdots & 1 \\ -a_n & -a_{n-1} & -a_{n-2} & \vdots & -a_1 \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \quad \mathbf{h} = \begin{bmatrix} 1 \\ 0 \\ 0 \\ 0 \\ 0 \end{bmatrix} .$$

In quest'ultima rappresentazione, l'equazione differenziale lineare di ordine n di partenza è trasformata in un sistema di equazioni differenziali lineari del primo ordine.

Esempio: oscillatore armonico.

Prendiamo in considerazione un punto materiale di massa m che si muove lungo l'asse x, attorno alla sua posizione di riposo, per effetto di una forza elastica di richiamo (esercitata per esempio da una molla) e che per effetto dell'attrito e/o della resistenza dell'aria tende progressivamente a fermarsi. Se indichiamo con $y(t)$ lo spostamento del punto dalla posizione di riposo, l'equazione del moto è data dalla

$$m \cdot y''(t) + a \cdot y'(t) + k \cdot y(t) = 0 ,$$

ove i coefficienti a e k dipendono dall'attrito e, rispettivamente, dalla forza di richiamo. Posto a/m = β e k/m = ω^2, l'equazione del moto viene espressa dalla seguente equazione differenziale lineare omogenea del secondo ordine:

$$y''(t) + \beta \cdot y'(t) + \omega^2 \cdot y(t) = 0 .$$

Ponendo y(t) = exp(λt) nella suddetta equazione differenziale e risolvendo l'equazione algebrica che ne risulta, e cioè $(\lambda^2 + \beta.\lambda + \omega^2).\exp(\lambda t) = 0$, detta equazione caratteristica, si trovano le due radici $\lambda_1 = (-\beta + \sqrt{\beta^2 - 4.\omega^2}) / 2$ e $\lambda_2 = (-\beta - \sqrt{\beta^2 - 4.\omega^2}) / 2$ e con ciò la soluzione generale della equazione differenziale omogenea :

1) $y(t) = c_1 . \exp(\lambda_1.t) + c_2 . \exp(\lambda_2.t)$, se le due radici sono reali e distinte ;
2) $y(t) = (c_1 + c_2.t) . \exp(-\beta.t / 2)$, se $\lambda_1 = \lambda_2 = -\beta / 2$;
3) $y(t) = \exp\{-\beta.t / 2\}. [(c_1 + c_2).\cos b.t + i.(c_1 - c_2).\sin b.t]$, se λ_1 e λ_2 sono complesse coniugate e ove $b = \frac{1}{2}\sqrt{4\omega^2 - \beta^2}$.

Chiaramente, i tre casi corrispondono al fatto che il discriminante $\beta^2 - 4\omega^2$ dell'equazione $\lambda^2 + \beta.\lambda + \omega^2 = 0$ sia positivo, nullo o negativo. Nei primi due casi il moto è aperiodico perché l'attrito è sufficientemente grande da impedire oscillazioni; nel terzo caso il moto è oscillatorio e smorzato attorno al punto di riposo.

La rappresentazione nello spazio degli stati corrispondente all'equazione differenziale omogenea, ponendo $x_1(t) = y(t)$ e $x_2(t) = y'(t)$, è data dal sistema:

$$\frac{d}{dt}x(t) = \begin{bmatrix} x_1'(t) \\ x_2'(t) \end{bmatrix} = F.x(t) = \begin{bmatrix} 0 & 1 \\ -\omega^2 & -\beta \end{bmatrix}.\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix},$$

$$y(t) = (1 \quad 0).x(t) \ .$$

Le radici λ_1 e λ_2 trovate precedentemente corrispondono agli autovalori della matrice F, cioè alle soluzioni dell'equazione det $(\lambda.I - F) = \lambda^2 + \beta.\lambda + \omega^2 = 0$.

Nel caso che esista una funzione input deterministica uguale a $u(t) = f_0 \cos \Omega t$ la rappresentazione nello spazio degli stati diventa:

$$\frac{d}{dt}x(t) = \begin{bmatrix} x_1'(t) \\ x_2'(t) \end{bmatrix} = F.x(t) = \begin{bmatrix} 0 & 1 \\ -\omega^2 & -\beta \end{bmatrix}.\begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} + \begin{bmatrix} 0 \\ 1 \end{bmatrix}.f_0 \cos \Omega t \ ,$$

$$y(t) = (1 \quad 0).x(t) \ .$$

Rinviamo il lettore agli innumerevoli manuali sull'argomento per la soluzione di questo sistema.

2) Modelli lineari dinamici stocastici a tempo continuo

Se la funzione a secondo membro, u(t), dell'equazione differenziale lineare ordinaria $y^{(n)}(t)+\sum_{k=1}^{n} a_k . y^{(n-k)}(t)=u(t)$ non è deterministica, ma stocastica, cioè **se u(t) rappresenta un processo stocastico a tempo continuo**, allora l'equazione descrive una trasformazione lineare di tale processo il cui risultato è un altro processo stocastico y(t), soluzione dell'equazione differenziale.

Un primo semplice esempio di equazione differenziale lineare stocastica è fornito dall'equazione di Langevin, $v'(t)+b.v(t)=u(t)$, $b>0$, riguardante il fenomeno di "moto Browniano" di una qualunque microscopica particella sospesa in un fluido e soggetta ad innumerevoli urti da parte delle molecole del fluido stesso, a loro volta soggette all'agitazione termica . In tale equazione v(t) indica la velocità aleatoria della particella, u(t) il processo stocastico input del sistema definito sulla base di ragionevoli ipotesi sulle caratteristiche del fenomeno fisico, il coefficiente b il grado di viscosità del fluido. Imponendo la condizione iniziale $v(0) = V_0$ e risolvendo l'equazione si trova:

$$v(t) = V_0 . e^{-bt} + \int_0^t e^{-b.(t-s)} . u(s) ds$$

che rappresenta il processo stocastico output.

Ritornando all'equazione differenziale generale $y^{(n)}(t)+\sum_{k=1}^{n} a_k . y^{(n-k)}(t)=u(t)$, per il processo stocastico u(t) deve essere fornita una qualche specificazione probabilistica: ci si può limitare a specificare la funzione valor medio E[u(t)] e/o la funzione di covarianza Cov[u(t),u(τ)] oppure, all'estremo opposto, si può caratterizzare completamente u(t) specificando la famiglia delle sue distribuzioni di probabilità congiunte. In corrispondenza del livello di specificazione prescelto per il processo input, dalla soluzione dell'equazione differenziale si otterrà lo stesso livello di specificazione per il processo output.

Nell'esempio dell'equazione di Langevin, se ci limitassimo a specificare $E[u(t)] \equiv 0$ si otterrebbe $E[v(t)] = E(V_0) . e^{-bt} + \int_0^t e^{-b.(t-s)} . E[u(s)] ds = E(V_0) . e^{-bt}$. Se oltre alla $E[u(t)] \equiv 0$ si fosse specificata anche la funzione di covarianza Cov[u(t),u(τ)] risulterebbe determinabile anche la funzione di covarianza del processo v(t).

Particolarmente semplice risulta la specificazione probabilistica dei processi stocastici se si adotta, quando ciò è possibile, l'ipotesi di **stazionarietà** in quanto, com'è noto, tale ipotesi implica che la funzione valor medio del processo sia costante e che la sua funzione di covarianza dipenda da un solo argomento, l'ampiezza dell'intervallo temporale t - τ . Ricordiamo anche che in questo caso alla funzione di covarianza può essere associata mediante la trasformazione di Fourier, in modo unico, la **funzione spettrale** F(ω) al modo seguente:

$$\text{Cov [u(t),u(}\tau\text{)]} = \psi_u(t-\tau) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{i\omega(t-\tau)} dF_u(\omega) \ .$$

Quando $F_u(\omega)$ riesce derivabile, la sua derivata $F_u'(\omega) = f_u(\omega)$ è chiamata **densità spettrale** e per essa sussistono le:

$$\psi_u(t-\tau) = \frac{1}{2\pi}\int_{-\infty}^{+\infty} e^{i\omega(t-\tau)} . f_u(\omega) d\omega \ ,$$

$$f_u(\omega) = \int_{-\infty}^{+\infty} e^{-i\omega\tau} . \psi_u(\tau) d\tau \ .$$

Ricordiamo infine che se la specificazione probabilistica riguarda soltanto i momenti fino al secondo ordine, cioè se la specificazione concerne soltanto valori medi, varianze e covarianze dei numeri aleatori considerati, allora un processo stocastico stazionario u(t) viene caratterizzato dalla scelta del comune valor medio e della funzione di covarianza $\psi_u(\tau)$ oppure da quella del comune valor medio e della funzione spettrale $F_u(\omega)$ (o densità spettrale $f_u(\omega)$) ; se si assume anche che tutte le distribuzioni congiunte del processo siano **di tipo normale** le scelte alternative precedenti specificano completamente il processo stocastico.

Nella rappresentazione in termini del vettore **x**(t) delle variabili di stato $\frac{d}{dt}x(t) = F.x(t) + b.u(t)$, $y(t) = h^T.x(t)$, si ha che anche il vettore **x**(t) è stocastico, essendo esso la soluzione di un sistema di equazioni differenziali stocastiche.

Presenteremo a questo punto due esemplificazioni di modelli lineari dinamici stocastici: la prima riguarda un modello fisico analogo, dal punto di vista matematico, all'oscillatore armonico e cioè un circuito elettrico RCL con input stocastico; la seconda esemplificazione, di notevolissimo interesse storico, concerne la stima di un processo di segnale basata sull'osservabilità della somma del segnale e di un disturbo casuale, cioè aleatorio.

a) Circuito RCL.

Il circuito elettrico è costituito da un resistore, un condensatore e un solenoide connessi in serie e alimentati da un generatore di tensione. Sono note le caratteristiche dei tre primi componenti e cioè la resistenza R (misurata in ohm) del resistore, la capacità C del condensatore (misurata in farad) e l'induttanza L del solenoide (misurata in henry). Ci porremo il seguente problema: cosa si può affermare sull'intensità di corrente I(t) nel circuito se la tensione V(t) del generatore applicato al circuito è solo parzialmente nota? Assumeremo che la tensione V(t) sia un processo stocastico stazionario con funzione valor medio $\varphi_V(t)$ e funzione di covarianza $\psi_V(\tau)$ note.

Nel circuito scorre una corrente elettrica aleatoria I(t) determinata dalla seguente equazione integro-differenziale

$$L.I'(t) + R.I(t) + \frac{1}{C}.\int I(t)dt = V(t)$$

basata sulle leggi dell'Elettrotecnica.

Poiché la carica elettrica $Q(t)$ del condensatore è legata all'intensità di corrente I(t) dalla relazione I(t) = $Q'(t)$ la precedente equazione può essere trasformata nella seguente equazione differenziale del secondo ordine

$$Q''(t) + 2\xi\omega.Q'(t) + \omega^2.Q(t) = \frac{1}{L}.V(t)$$

nella funzione (aleatoria) incognita $Q(t)$. Le costanti ξ e ω sono legate alle caratteristiche note degli elementi passivi del circuito dalle relazioni $\xi = \frac{R}{2}.\sqrt{\frac{C}{L}}$ e $\omega = \frac{1}{\sqrt{C.L}}$.

Ponendo **Z**(t) = [$Q(t)$, $Q'(t)$] ^{T} la precedente equazione scalare si trasforma nell'equazione vettoriale

$$\mathbf{Z}'(t) = \begin{bmatrix} Q'(t) \\ Q''(t) \end{bmatrix} = \begin{bmatrix} 0 & 1 \\ -\omega^2 & -2\xi\omega \end{bmatrix}.\begin{bmatrix} Q(t) \\ Q'(t) \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{1}{L}.V(t) \end{bmatrix} .$$

Indicando con F la matrice del sistema, la soluzione dell'equazione differenziale vettoriale è data dalla

$$\mathbf{Z}(t) = e^{F.t}.Z(0) + \int_0^t e^{F.(t-s)}.\begin{bmatrix} 0 \\ \frac{1}{L}.V(s) \end{bmatrix} ds .$$

Gli autovalori λ_i, $(i=1,2)$, di F sono le radici dell'equazione $\det(F-\lambda.I)=0$ ed hanno le espressioni $\lambda_{1,2}=-\xi.\omega\pm\omega.\sqrt{\xi^2-1}$. Assumendo che sia $\xi^2<1$ i due autovalori sono complessi coniugati

$$\lambda_{1,2}=-\xi.\omega\pm i\omega.\sqrt{1-\xi^2}$$

e gli autovettori corrispondenti sono

$$v_1=\begin{pmatrix}1 & -\xi\omega+i\omega_1\end{pmatrix}^T \quad \text{e} \quad v_2=\begin{pmatrix}1 & -\xi\omega-i\omega_1\end{pmatrix}^T$$

ove si è posto $\omega_1=\omega.\sqrt{1-\xi^2}$; è facile constatare che $v_1 \;\; e \;\; v_2$ sono linearmente indipendenti.

Indicando con $V=(v_1 \;\; v_2)$ la matrice degli autovettori e ponendo $e^{\Lambda.t}=diag\left(e^{\lambda_1.t},e^{\lambda_2.t}\right)$ si ha la rappresentazione $e^{F.t}=V.e^{\Lambda.t}.V^{-1}$ ed anche la $e^{F.(t-s)}=V.e^{\Lambda.(t-s)}.V^{-1}$ dalle quali, con un noioso ma

non difficile calcolo, si ottengono le seguenti soluzioni corrispondenti alla condizione iniziale **Z**(0) = **0 :**

$$Q(t)=\frac{1}{\omega_1.L}\int_0^t e^{-\xi\omega(t-s)}.\sin\left[\omega_1.(t-s)\right].V(s)ds\ ,$$

$$Q'(t)=I(t)=\frac{1}{L}\int_0^t\{\ e^{-\xi\omega(t-s)}.\left[\cos\omega_1(t-s)-\xi.\frac{\omega}{\omega_1}\sin\omega_1(t-s)\right]\ \}.V(s)ds\ .$$

Gli integrali stocastici che compaiono nelle due espressioni sono numeri aleatori definiti come limiti in media quadratica di opportune successioni di somme integrali. Ci limitiamo a dire che dalle funzioni $\varphi_V(t)$ e $\psi_V(\tau)$ del processo {V(t)} si possono determinare le corrispondenti $\varphi_Q(t)$ e $\psi_Q(\tau)$

e $\varphi_I(t)$ e $\psi_I(\tau)$.

b) Stima lineare dei minimi quadrati di un processo stocastico a tempo continuo.

Indicheremo con S(t) il processo stocastico di segnale che vogliamo stimare e con U(t) un processo di disturbo (o di rumore) indesiderato ma non eliminabile che si sovrappone al primo; con Y(t) = S(t) + U(t) indicheremo il processo di osservazione o generatore di dati. Assumeremo, per semplicità, che i due processi S(t) e U(t) non siano correlati tra loro e faremo le seguenti ipotesi sulle loro strutture stocastiche:

$$S(t) \sim N(0\,;\ \sigma_S^2.\exp\{-\alpha.\tau\})\,, \qquad U(t) \sim NWN(0\,;\ \sigma_u^2)\,.$$

A parole, entrambi i processi sono assunti di tipo normale o Gaussiano: il primo, noto in letteratura come "processo di Ornstein-Uhlembeck", è un processo stazionario con funzione valor medio identicamente nulla e funzione di covarianza $\psi_S(\tau) = \sigma_S^2.\exp\{-\alpha.\tau\}$; il secondo, noto in letteratura come processo "normal-white noise", è anch'esso stazionario con variabili mutuamente indipendenti, aventi la comune varianza pari a σ_u^2 e funzione valor medio identicamente nulla.

Il modello introdotto non corrisponde apparentemente a nessuna delle tre rappresentazioni di modelli dinamici precedentemente introdotte; è però possibile fornire una equivalente rappresentazione nello spazio degli stati sostituendo alla suddetta specificazione probabilistica per il processo {S(t)} una equazione differenziale lineare stocastica del primo ordine la cui soluzione è costituita dal processo {S(t)}:

$$\frac{d}{dt}S(t) + \alpha.S(t) = \sigma.W(t)\ ,$$

ove W(t) è un processo normal - white noise con varianza unitaria. La rappresentazione completa nello spazio degli stati è allora la seguente:

$$\frac{d}{dt}S(t) + \alpha.S(t) = \sigma.W(t)\,,$$

$$Y(t) = S(t) + U(t)\,,$$

essendo i due processi W(t) e U(t) non correlati tra loro. Si tratta della versione a tempo continuo del modello lineare dinamico già presentato nel paragrafo precedente. Il problema di stima dei minimi quadrati delle variabili di stato S(t) sulla base delle osservazioni $\{Y(s); s \le t\}$ si risolve utilizzando la versione a tempo continuo del filtro di E. Kalman (nota come "filtro di Kalman –

Bucy") . Si dimostra che, indicando con $\hat{S}(t)$ lo stimatore lineare dei minimi quadrati per S(t), esso è determinato dalla seguente equazione differenziale

$$\frac{d}{dt}\hat{S}(t) = -\alpha.\hat{S}(t) + (\beta - \alpha).[Y(t) - \hat{S}(t)] = -\beta.\hat{S}(t) + (\beta - \alpha).Y(t)$$

con la condizione iniziale $\hat{S}(0) = 0$ e ove si è posto $\beta = \sqrt{\frac{2\alpha\sigma_S^2}{\sigma_U^2} + \alpha^2}$.

Uno studio sistematico di questi temi può basarsi su:

M.H.A. Davis – Linear estimation and Stochastic Control (Chapman and Hall, 1977),

A.H. Jazwinski – Stochastic Processes and Filtering Theory (Academic Press, 1970),

R.S.Liptser, A.N. Shiryaev – Statistics of Random Processes (Springer, 1978).

Appendici

APPENDICE n. 1 : Cenni sui processi di ramificazione (Branching processes)

E' un esempio di processo a **catena markoviana con un insieme numerabile di stati**. Si pensi ad un individuo (essere umano, animale, particella subatomica, …..) idoneo a "generare" un numero aleatorio di discendenti: esso costituisce la generazione zero mentre i suoi discendenti formano la prima generazione; i discendenti dei discendenti formano la seconda generazione e così via.

Il parametro operativo individua le successive generazioni cosicché si ha: $X_0 = 1$, X_1 denota il numero di discendenti costituenti la prima generazione, X_2 il numero di individui della seconda generazione, etc. Supporremo che i numeri di discendenti di differenti individui, appartenenti alla stessa o a differenti generazioni, siano stocasticamente indipendenti tra loro e dotati della medesima distribuzione di probabilità $\{p_j; j \geq 0\}$ della quale indicheremo con $G(s) = \sum_j p_j . s^j$ la funzione generatrice (p.g.f.). Per evitare casi banali assumeremo $p_0 > 0$ e $p_0 + p_1 < 1$.

Per quanto detto si ha:

$$X_{n+1} = \sum_{i=1}^{X_n} Z_i = Z_1 + Z_2 + ... + Z_{X_n} \ ,$$

ove Z_i indica il numero di discendenti generati dall' i-esimo individuo della generazione X_n ; gli addendi Z_i sono assunti i.i.d. con la comune distribuzione $\{p_j\}$ e p.g.f. G(s).

Il carattere markoviano del processo $\{X_n\}$ è chiaramente rivelato dalle:

$$\text{Prob } [X_{n+1} = k_{n+1} / \bigcap_{i=1}^{n} (X_i = k_i)] = \text{Prob } [\sum_{i=1}^{k_n} Z_i = k_{n+1}] = \text{ Prob } [X_{n+1} = k_{n+1} / X_n = k_n] .$$

Per quanto concerne le probabilità non condizionate si ha:

$$p_k(n+1) = \text{Prob } [X_{n+1} = k] = \sum_{j=0}^{\infty} \Pr ob(X_{n+1} = k / X_n = j) . \Pr ob(X_n = j) =$$

$$= \sum_j \Pr ob(Z_1 + Z_2 + ... + Z_j = k) . \Pr ob(X_n = j) = \sum_j p_{jk} \ . p_j(n) \ ,$$

avendo indicato con p_{jk} le probabilità subordinate Prob ($\sum_{i=1}^{j} Z_i = k$) = Prob ($X_{n+1} = k / X_n = j$) = $\{p_k\}^{j*}$ ove l'ultima espressione indica la convoluzione j-ma di $\{p_j\}$ con se stessa.

Indicando con $F_{n+1}(s)$ la p.g.f. di X_{n+1}, la precedente relazione equivale alla

$$F_{n+1}(s) = \sum_j G^j(s).p_j(n) = F_n\,[G(s)]\,,$$

relazione ricorrente per le funzioni generatrici delle probabilità dei n.a. X_n che può anche scriversi

$$F_{n+1}(s) = G\,[F_n(s)]$$

poiché, essendo $X_0 = 1$ e quindi $F_0(s) = s$, si ha $F_1(s) = F_0[G(s)] = G(s)$, $F_2(s) = F_1[G(s)] = G[G(s)] = G^{(2)}(s)$ (seconda iterata di G) e, in generale, $F_{n+1}(s) = G^{(n+1)}(s) = G[G^{(n)}(s)] = G[F_n(s)]$.

La relazione ricorrente $F_{n+1}(s) = F_n\,[G(s)] = G[F_n(s)]$ ha un'importanza fondamentale nella teoria dei processi di ramificazione anche se, da un punto di vista operativo, non può essere impiegata per la determinazione delle $F_n(s)$. Essa può però fornire informazioni importanti sui momenti dei n.a. X_n: indicati con μ e ν la speranza matematica e la varianza della distribuzione $\{p_j\}$, si ricavano facilmente le relazioni $E(X_{n+1}) = \mu\,.\,E(X_n)$ e $Var(X_{n+1}) = \mu^2\,Var(X_n) + \nu\,E(X_n)$.

Un problema importante concerne la possibilità di estinzione dell'insieme dei discendenti, cioè la possibilità che i n.a. X_n siano definitivamente nulli. In proposito sussiste il seguente risultato dovuto a Steffensen (1930):

Teorema : la probabilità ξ di estinzione in tempo finito è data dalla più piccola radice positiva dell'equazione G(x) = x. Se $\mu = G'(1) \leq 1$ allora è $\xi = 1$ e l'estinzione è certa; se $\mu > 1$ allora $\xi < 1$ e la probabilità che la popolazione cresca indefinitamente è data da $1 - \xi$.

Una traccia di dimostrazione si ricava dalle seguenti considerazioni:

1) innanzitutto la successione delle $p_0(n) = Prob\,(X_n = 0)$ è monotona e limitata cosicché esiste il relativo limite ξ per $n \to \infty$;
2) dalla relazione $F_{n+1}(s) = G\,[F_n(s)]$ si ricava la $p_0(n+1) = G[p_0(n)]$ e passando al limite si ottiene $\xi = G(\xi)$;
3) si verifica facilmente che G(s) è concava verso l'alto e quindi monotona crescente; poiché $G(0) = p_0 > 0$ per ipotesi e G(1) = 1, si ha che l'esistenza di una o due radici dell'equazione G(x) = x in [0 , 1] dipende dal fatto che $G'(1) = \mu$ sia $\leq$ o > di 1; quindi, se $\mu \leq 1$ è $\xi = 1$ mentre se $\mu > 1$ allora $\xi < 1$ e la probabilità che la popolazione cresca indefinitamente è $1-\xi$.

Un esempio: G(s) = (ps + q) / (1 + r – rs) , con p + q = 1 e p, q, r non negativi.

Si ricava $\mu = G'(s) = p + r$; inoltre è $p_0 = q / (1-r)$, $p_1 = (p+r) / (1+r)^2$, $p_2 = r / (1+r)^3$ e in generale $p_n = r^{n-1} / (1+r)^{n+1}$. L'equazione G(s) = s ha due radici pari a q/r e 1; essendo $\xi = \min (q/r , 1)$ è $\xi = (q/r)$ se p+r > 1.

In una ricerca riguardante l'estinzione dei cognomi familiari per i maschi bianchi degli USA nel 1920, Lotka trovò che la p.g.f. suddetta con q/(1+r) = 0.4981 e r/(1+r) = 0.5586 rappresentava bene il fenomeno concreto. Si ottiene, per tale funzione, $\mu = 1.14$ e $\xi = 0.89$: a parole, un cognome di un maschio bianco aveva negli USA del 1920 una probabilità di estinzione pari a 0.89, mentre quella di non estinzione era 0.11 .

APPENDICE N. 2 : Urna di Pòlya ed alcuni processi stocastici associati

Com'è noto, è detta **"urna di Pòlya"** un sistema di estrazioni casuali da un'urna contenente inizialmente b palline bianche ed r palline rosse; dopo un'estrazione casuale di una pallina, questa viene rimessa nell'urna assieme a c palline dello stesso colore di quella estratta: è chiaro che la composizione dell'urna varia colpo per colpo. Considereremo tre processi stocastici associati alla sequenza potenzialmente illimitata di estrazioni successive dall'urna:

a) il processo $\{Y_n ; n \geq 1\}$ dei risultati delle successive estrazioni, ove $Y_n = 1$ o 0 a seconda che l'ennesima estrazione dia pallina bianca oppure rossa;
b) il processo $\{X_n ; n \geq 1\}$ relativo al numero di palline bianche nell'urna dopo le estrazioni successive, ove X_n indica il numero di palline bianche presenti nell'urna dopo le prime n estrazioni;
c) il processo $\{Z_n ; n \geq 1\}$ delle frazioni o percentuali di palline bianche nell'urna dopo le successive estrazioni; è chiaramente $Z_n = X_n / (b + r + n.c)$, in quanto dopo n estrazioni le palline nell'urna sono b+ r + n.c .

Si vedrà che il processo sub a) è **stazionario**, anzi scambiabile, il processo sub b) è **markoviano** e che infine il processo sub c) è una **martingala**; ovviamente tutti i processi sono a parametro discreto. Sono quindi rappresentate, nello schema di estrazioni suddetto, le tre principali categorie di processi stocastici. Per motivi di semplicità, nel seguito supporremo c = 1 di modo che il numero complessivo di palline nell'urna, dopo ogni estrazione, aumenta di un'unità; dopo n estrazioni le palline nell'urna passano dalle iniziali b + r a b + r + n, quelle bianche da b a $b + S_n$ e quelle rosse da r a $r + n – S_n$, ove S_n indica il numero aleatorio $Y_1 + ……..+ Y_n$, cioè il numero di palline bianche uscito dall'urna nelle prime n estrazioni successive. Porremo inoltre $X_0 = b$ e $Z_0 = b / (b + r)$.

E' abbastanza evidente che lo schema dell'urna di Pòlya può costituire un modello semplificato per **fenomeni di contagio**: infatti dopo l'estrazione di una pallina bianca (o rossa) aumenta il numero e la

percentuale di bianche (o rosse) nell'urna per cui aumenta anche la probabilità di un'altra estrazione di pallina bianca (o rossa) al colpo successivo.

Cominceremo a fare alcune considerazioni di carattere generale: con riferimento ad un generico prodotto logico dei primi n eventi $E_1^{'} \wedge E_2^{'} \wedge ..... \wedge E_n^{'}$, ove h eventi sono affermati ed n-h sono negati, in un ordinamento qualsiasi delle affermazioni e negazioni, e valutando la probabilità P($E_1^{'} \wedge E_2^{'} \wedge ..... \wedge E_n^{'}$) secondo la procedura sequenziale

P($E_1^{'} \wedge E_2^{'} \wedge ..... \wedge E_n^{'}$) = P($E_1^{'}$).P($E_2^{'} / E_1^{'}$)……P($E_n^{'} / E_1^{'} \wedge ..... \wedge E_{n-1}^{'}$)

ci si rende conto facilmente che dei prodotti logici subordinanti conta, per la valutazione, soltanto il numero degli eventi coinvolti e quello degli eventi affermati, o negati; l'ordine delle affermazioni e negazioni è irrilevante. Ciò comporta, come si può dimostrare, che ogni prodotto logico di n eventi distinti (non necessariamente i primi n) implicante l'evento { $S_n = h$} ha la medesima probabilità che dipende solo dagli interi n ed h:

$$P(E_1^{'} \wedge E_2^{'} \wedge ..... \wedge E_n^{'}) = \frac{(b+h-1)_h.(r+n-h-1)_{n-h}}{(b+r+n-1)_n} \ ,$$

ove il simbolo $(n)_h$ sta per $(n)_h = n.(n-1).(n-2)……..(n-h+1)$. Evidentemente, la probabilità dell'evento {$S_n = h$} è data dalla

$$P(S_n = h) = \binom{n}{h} . P(E_1^{'} \wedge E_2^{'} \wedge ..... \wedge E_n^{'}) = \binom{n}{h} . \frac{(b+h-1)_h.(r+n-h-1)_{n-h}}{(b+r+n-1)_n}$$

che viene denominata "distribuzione di Pòlya". I valori della speranza matematica e della varianza sono dati dalle

$n.b/(b+r)$ e, rispettivamente, $n.b.r.(b+r+n) / (b+r)^2.(b+r+1)$.

Considerando ora i processi {X_n ; $n \geq 1$} e {Z_n ; $n \geq 1$}, avendo posto $X_0 = b$ e $Z_0 = b/(b+r)$, si ha che il n.a. X_n può assumere tutti i valori interi compresi tra b e b+n con probabilità che, come facilmente si verifica, dipendono soltanto dal valore di S_n ; precisamente, l'evento {$X_n = b+h$} si verifica quando e solo quando $S_n = h$ e ciò accade con la probabilità espressa sopra. In corrispondenza è vero l'evento $Z_n = (b+h) / (b+r+n)$. Per quanto concerne le probabilità subordinate $P(X_{n+1} = j / X_n = b+h)$, queste sono diverse da zero solo se $j = b+h$ e $j = b+h+1$ e risulta:

$$P(X_{n+1} = j / X_n = b+h) = \frac{(r+n+h)/(b+r+n)}{(b+h)/(b+r+n)}$$

a seconda che $j = b+h$ oppure $j = b+h+1$.

Proveremo ora alcune proposizioni riguardanti il carattere dei tre processi indicati in precedenza:

Proposizione 1: il processo $\{Y_n ; n \geq 1\}$ è stazionario scambiabile.

Dobbiamo provare che le distribuzioni congiunte finite-dimensionali dei n.a. sono invarianti per traslazione rigida nei valori del parametro operativo (condizione di stazionarietà) e addirittura che ogni n-pla di variabili distinte ha la stessa distribuzione congiunta (condizione di scambiabilità). Poiché i n.a. Y_n sono indicatori di eventi, la distribuzione congiunta della sequenza $(Y_1,\ldots\ldots, Y_n)$ è costituita dalle probabilità dei prodotti logici $E_1^{'} \wedge E_2^{'} \wedge \ldots\ldots \wedge E_n^{'}$ che dipendono, come si è già stabilito, dalla coppia (n, h), oppure dalla coppia equivalente (h, n-h), e non da quali eventi sono stati considerati. Pertanto non solo è $P(E_1^{'} \wedge E_2^{'} \wedge \ldots\ldots \wedge E_n^{'}) = P(E_{i_1}^{'} \wedge E_{i_2}^{'} \wedge \ldots\ldots \wedge E_{i_n}^{'})$ ove $(i_1,\ldots\ldots,i_n)$ è un'arbitraria permutazione degli indici $(1,\ldots\ldots,n)$ ma anche $P(E_1^{'} \wedge E_2^{'} \wedge \ldots\ldots \wedge E_n^{'}) = P(E_{j_1}^{'} \wedge E_{j_2}^{'} \wedge \ldots\ldots \wedge E_{j_n}^{'})$ ove $(j_1,\ldots\ldots,j_n)$ è un'arbitraria n-pla di interi distinti, essendo comunque vero l'evento $\{\sum_{k=1}^{n} \left| E_{j_k} \right| = h\}$.

L'invarianza delle valutazioni di probabilità rispetto alla scelta dei numeri aleatori del processo rivela la scambiabilità dello stesso e, a maggior ragione, la sua stazionarietà.

Proposizione 2: il processo $\{X_n ; n \geq 1\}$ è markoviano.

Ricordiamo che la markovianità del processo sussiste se, per ogni intero n e ogni scelta dei valori dei n.a., le valutazioni di probabilità subordinate soddisfano le condizioni $P\left[X_{n+1} = x_{n+1} / \bigcap_{j=1}^{n} (X_j = x_j) \right] =$

$= P[X_{n+1} = x_{n+1} / X_n = x_n]$. Nel nostro caso porremo $x_j = b + h_j$ essendo $h_1 \leq h_2 \leq \ldots\ldots\ldots \leq h_{n+1}$. Nell'ipotesi che, dopo le prime n estrazioni, si sia arrivati ad avere nell'urna $b + h_n$ palline bianche, qualunque sia stata la sequenza $(h_1,\ldots\ldots\ldots,h_n)$, avremo

$$P\left[X_{n+1} = b + h_{n+1} / \bigcap_{j=1}^{n} (X_j = b + h_j) \right] = \frac{r + n - h_n}{b + r + n} \cdot \frac{b + h_n}{b + r + n},$$

INTRODUZIONE ALLA STATISTICA BAYESIANA (prof. A. Wedlin)

1) Generalità

Questi appunti raccolgono alcuni argomenti trattati nei miei corsi di Statistica bayesiana per studenti di Scienze statistiche che hanno già frequentato in precedenza altri corsi istituzionali di Statistica e di Calcolo delle probabilità. Pertanto le nozioni elementari di carattere probabilistico e statistico saranno considerate già acquisite; in alcune appendici saranno richiamate solo alcune di quelle nozioni a causa della loro importanza per lo svolgimento del tema principale.

Inferenza statistica

Poiché l'oggetto di questi appunti è l'inferenza statistica, seppure secondo il punto di vista bayesiano, è forse opportuno precisare fin d'ora come essa viene intesa. Tra le diverse possibili accezioni io preferisco quella secondo cui **l'inferenza statistica è un caso particolare di ragionamento induttivo ove la particolarità risiede nel fatto che il dato dell'esperienza è solitamente costituito dall'osservazione di eventi più o meno analoghi** (B. de Finetti). Poiché il ragionamento induttivo riguarda la trasformazione delle opinioni individuali dovuta ad un incremento di informazione io considero lo statistico più un "analizzatore di opinioni espresse probabilisticamente" che un mero "analista di dati osservati".

Cenni storici sull'approccio bayesiano all'inferenza statistica

Venendo ora all'approccio bayesiano all'inferenza statistica, esso risale al lontano XVIII Secolo e si assume che il suo atto di nascita sia la nota "An Essay Towards Solving a Problem in the Doctrine of Chances" del filosofo britannico Thomas Bayes (1702 – 1761) pubblicata nel 1764 a cura di R. Price. La stessa impostazione venne ripresa poco dopo, nel 1774, dal celebre Laplace, ma alcune sue discutibili applicazioni del metodo contribuirono a provocare l'insorgere di un forte atteggiamento critico verso l'approccio da parte degli altri studiosi. Soltanto dopo un secolo e mezzo da quell'epoca, e cioè negli anni 30 del XX Secolo, si verificherà una ripresa di interesse per tale impostazione soprattutto ad opera di F.P.Ramsey, B. de Finetti, e H. Jeffreys. L'effettiva affermazione della Statistica bayesiana si e' avuta pero' soltanto nell'ultimo trentennio del XX secolo.

Per spiegare queste difficoltà nell'accettazione dell'approccio bayesiano alla Statistica occorre sottolineare **la stretta interconnessione esistente tra l'induzione statistica e la nozione di probabilità soggettiva**. Nella nota di T. Bayes veniva preso in considerazione il seguente problema, che esporremo usando una terminologia moderna: noto che su p + q eventi tra loro omogenei se ne sono verificati p, cosa si puo' affermare sulla probabilita' di un altro evento dello stesso tipo? La soluzione di T. Bayes, sempre formulata in termini attuali, era che quest'ultima probabilita' andrebbe valutata al livello (p+1)/(p+q+2).

Tale espressione e' anche nota come "regola di successione di Laplace". Come si vedrà meglio nel seguito, uno degli assunti di T. Bayes per la detta soluzione era che **quando non si ha nessuna informazione sulla probabilità di un evento è ragionevole ritenere che il suo valore sia un qualunque numero reale dell'intervallo [0 , 1], essendo tutti questi infiniti numeri considerati ugualmente attendibili**. E' proprio su questo assunto che si concentrerà la principale discussione in merito alla validità dell'approccio bayesiano.

Infatti, molti studiosi si chiesero: **che tipo di probabilità può essere attribuita ad un evento quando l'informazione disponibile su di esso è nulla**? Bisogna ricordare che a quell'epoca ancora non si era raggiunto un effettivo accordo su una qualche definizione di probabilità: veniva sostanzialmente accettato che nelle situazioni caratterizzate da una ragionevole simmetria (uguale verosimiglianza dei casi possibili), come nel lancio di un dado o di una moneta dall'aspetto regolare o nei giochi di carte, la probabilità di un evento dovesse essere valutata come **rapporto tra il numero dei casi favorevoli all'evento e quello di tutti i casi possibili**; era ugualmente accettato che quando si avevano numerose informazioni su casi analoghi, come per esempio sulla mortalità entro l'anno di individui aventi un'età fissata, si dovesse valutare la probabilità in termini della **frequenza relativa di "successo"**. Ora nessuno dei due ricordati criteri di valutazione poteva venire applicato nell'assunto di T. Bayes! Pertanto doveva trattarsi di un problema mal posto.

Soprattutto in Inghilterra, dove per una possibile definizione di probabilità ci si orientò sempre più sul criterio di valutazione basato sulla frequenza relativa di successo su un grande numero di eventi osservati, l'impostazione di Bayes venne accantonata e prevalse al suo posto un approccio più tardi denominato "principio di induzione": nell'ipotesi di conoscere la frequenza di successo su un numero non trascurabile di casi ritenuti approssimativamente analoghi tra loro, la probabilità di un evento dello stesso tipo non dovrebbe scostarsi sostanzialmente dalla frequenza relativa osservata.

Alcuni autori (per esempio G. Castelnuovo nel suo trattato "Calcolo delle probabilita'" del 1918) ritennero che tra probabilita' e frequenza osservata sussistesse una relazione in qualche senso analoga a quella tra la misura effettiva di una grandezza fisica e le sue valutazioni approssimate effettuate con uno strumento di misura che introduce in esse almeno errori di misura accidentali; quegli autori formularono il "postulato empirico del caso": se un evento ha una probabilita' costante p in ogni prova e se esso si verifica m volte in n prove, il rapporto m/n da' un valore approssimato della probabilita' p e l'approssimazione e' ordinariamente tanto migliore quanto maggiore e' il numero n delle prove.

Oggi possiamo affermare che l'approccio bayesiano alla Statistica era decisamente in anticipo sui tempi perche' richiedeva una nozione di probabilita', quella soggettiva, che doveva comparire appena nel XX Secolo, ad opera di F.P. Ramsey (1926) e B. de Finetti (1928) , che lavorarono indipendentemente l'uno dall'altro. Fu de Finetti a dar vita ad una teoria organica della probabilità soggettiva, soprattutto con la monografia del 1930 dal titolo "Probabilismo. Saggio critico sulla teoria delle probabilita' e sul valore della scienza" , con la monografia "La prevision: ses lois logiques, ses sources subjectives" del 1937 e con il trattato "Teoria delle probabilità" del 1970.

Finalmente, accettando l'interpretazione della probabilità di un evento come **grado di fiducia di una persona nel verificarsi dell'evento sulla base di un determinato stato di conoscenza**, era possibile attribuire un significato all'assunzione di T. Bayes e interpretare il suo risultato (p+1)/(p+q+2) come una valutazione coerente conseguente ad un incremento di informazione che implichi p successi su p + q eventi osservabili. Nell'Appendice n. 1, anticipando considerazioni che faremo in seguito ed interpretando i p + q eventi omogenei del problema affrontato da Bayes come risultati di successive estrazioni da un'urna

contenente palline bianche e nere in numero e proporzione non noti, è data una giustificazione formale di quel risultato.

Concludiamo questo rapido richiamo ai primi passi della Statistica bayesiana ricordando, per sommi capi, alcuni contributi di P.S. Laplace (-): in un suo articolo del 1774 egli si occupò del problema della "probabilità delle cause", più tardi noto come "probabilità delle ipotesi". Si assuma di conoscere le probabilità condizionate $P(E/H_j)$ di un evento E rispetto ad ogni costituente (ipotesi o causa) H_j , j = 1,2…….,n , di una partizione finita dell'evento certo (una ed una sola delle H_j è vera, ma non è noto quale); noto che l'evento E si è verificato, qual'è la probabilità che a determinarlo sia intervenuta la causa H_i ?

La risposta di Laplace è: $P(H_i/E) = P(E/H_i) / \sum_{j=1}^{n} P(E/H_j)$, che si può giustificare mediante la applicazione del teorema di Bayes

$$P(H_i/E) = \frac{P(H_i).P(E/H_i)}{\sum_{j=1}^{n} P(H_j).P(E/H_j)} ,$$

assumendo che ad ogni causa o ipotesi sia stata attribuita una probabilità iniziale $P(H_j)$ pari a 1/n . Il Laplace non fece alcun riferimento esplicito a T. Bayes, ma in una occasione successiva osservò che se le cause non sono equiprobabili allora si deve avere $P(H_i/E) = \frac{\omega_i.P(E/H_i)}{\sum_{j=1}^{n} \omega_j.P(E/H_j)}$, ove le ω_j sono le probabilità iniziali delle cause H_j .

La piena giustificazione dell'approccio statistico bayesiano arrivò appena un secolo e mezzo dopo ed è dovuta a B. de Finetti che al Congresso Internazionale di Matematica tenutosi a Bologna nel 1928 presentò la comunicazione dal titolo "Funzione caratteristica di un fenomeno aleatorio" con la quale:

- introdusse la nozione di processo stocastico scambiabile,

- dimostrò il corrispondente teorema di rappresentazione e

- risolse il problema dell'inferenza statistica in ipotesi di scambiabilità delle variabili osservabili.

Tali nozioni e risultati saranno presentati nel secondo e terzo Capitolo.

Mentre la modesta frase iniziale della suddetta comunicazione – "Scopo di questa comunicazione è di mostrare come il metodo della funzione caratteristica, già così vantaggiosamente introdotto nella teoria delle variabili casuali, si presti pure assai utilmente allo studio dei fenomeni aleatori" – non lascia trasparire i notevoli contributi innovativi in essa contenuti, la frase finale - "Queste conclusioni e questi esempi possono chiarire l'influenza che sulla valutazione di una probabilità esercitano i dati dell'esperienza." – fa intravedere che la ragione principale della introduzione della nozione di scambiabilità era stata l'intenzione di chiarire le condizioni in cui l'osservazione di una frequenza su una sequenza di prove fornisce coerentemente la base per una valutazione di probabilità o per una previsione della frequenza su una sequenza di prove ancora da eseguire.

In altri termini, il lavoro appare motivato soprattutto dall'esigenza di carattere filosofico di giustificare il ragionamento per induzione, almeno nel caso in cui il dato dell'esperienza consiste in un certo numero di osservazioni di fatti più o meno analoghi (induzione statistica). Più precisamente de Finetti si riprometteva di giustificare il "principio di induzione" $P(E_{n+1} / \sum_{i=1}^{n} |E_i| = m) \cong \frac{m}{n}$ che, come si è già detto, suggeriva di valutare la probabilità di E_{n+1} in modo approssimativamente uguale alla frequenza relativa osservata su n eventi analoghi ad E_{n+1}.

Primo confronto dell'impostazione classica e bayesiana dell'inferenza statistica.

A questo punto, dopo aver sommariamente introdotto l'approccio bayesiano alla Statistica, può essere utile confrontarlo con quello non-bayesiano, o classico, almeno con riferimento ad uno dei più semplici problemi inferenziali: la stima (puntuale) di un parametro non noto θ presente nella comune funzione di ripartizione (f.d.r.) $F(x;\theta)$ delle variabili $X_j (j \geq 1)$ costituenti il processo stocastico osservabile (data generating process).

Parametri incogniti

Per quanto concerne il parametro incognito θ, nell'approccio classico esso viene considerato alla stregua di una costante (numero certo) non nota, mentre nell'approccio bayesiano esso viene considerato, in quanto non noto, un numero aleatorio Θ sul quale una persona può esprimere un'opinione nella forma di una valutazione probabilistica. Per esempio, egli potrebbe affermare che in base alle sue informazioni ed opinioni è $P(a \leq \Theta \leq b) = p$, e ovviamente $P[(\Theta < a) \bigcup (\Theta > b)] = 1 - p$.

Variabili osservabili o campionarie.

Con riferimento ai numeri aleatori (n.a.) osservabili $X_j (j \geq 1)$, nell'ipotesi che il procedimento di osservazione (o di misurazione) dei singoli n.a. X_j possa avvenire almeno approssimativamente "nelle stesse condizioni e in modo tale che il risultato di ogni osservazione o misurazione non sia influenzato da quelli precedenti", essi vengono spesso assunti essere indipendenti e ugualmente distribuiti (i.i.d.) e la sequenza delle osservazioni o misurazioni ($x_1, ..., x_m$) viene conseguentemente detta "campione casuale" in analogia con la sequenza dei risultati ottenibili per esempio da lanci successivi e "regolari" di una stessa moneta. Nell'approccio bayesiano, poiché la comune f.d.r. dei n.a. X_j non è nota, o completamente specificata, a causa della presenza del parametro incognito θ, l'ipotesi di indipendenza e

uguale distribuzione viene riformulata come "i.i.d. rispetto ad ogni possibile valore del parametro incognito" o "indipendenza condizionata e uguale f.d.r. condizionata $F(x/\theta)$". Non si tratta di una mera riformulazione verbale: **la condizione di indipendenza condizionata tra n.a. X_j equivale, in generale, ad una situazione di dipendenza stocastica tra essi** e tale situazione corrisponde, come vedremo, in forza del teorema di rappresentazione di B. de Finetti, a quella detta "scambiabilità" dei n.a. X_j.

Stima puntuale e stimatori

Nell'impostazione non bayesiana, o classica, dell'inferenza statistica sul parametro incognito θ la sequenza potenziale delle osservazioni ($x_1, ..., x_m$), o l'evento osservabile $E = \left\{ \bigcap_{i=1}^{m} (X_i = x_i) \right\}$, determina il valore assunto dallo stimatore S_m adottato e quindi la valutazione numerica, o stima, di θ. E' noto che lo stimatore S_m di θ viene individuato dal metodo di stima puntuale stabilito e che la scelta del metodo di stima viene effettuata in base alle proprietà statisticamente rilevanti che esso conferisce allo stimatore S_m. Alcune delle proprietà più frequentemente richieste per gli stimatori sono la "correttezza" (o non distorsione) che richiede la condizione $E(S_m) \equiv \theta$, la "consistenza" che richiede la convergenza in probabilità, al divergere di m, della sequenza degli stimatori S_m a θ, la "sufficienza" di S_m che sostanzialmente richiede l'utilizzazione di tutta l'informazione rilevante presente nella sequenza ($x_1, ..., x_m$), l'efficienza relativa che richiede che l'errore quadratico medio $E(S_m - \theta)^2$ sia minimo rispetto a quello di ogni altro stimatore $S_m^{'}$ per θ e così via. Per esempio, si dimostra che lo stimatore di "massima verosimiglianza" S_m^* possiede le proprietà di sufficienza e di efficienza asintotica; esso non è invece sempre corretto.

Teorema di Bayes.

Nell'impostazione bayesiana dell'inferenza statistica sul parametro incognito, e quindi aleatorio, Θ il risultato dell'evento osservabile $E = \left\{ \bigcap_{i=1}^{m} (X_i = x_i) \right\}$ è la trasformazione coerente dell'opinione su Θ espressa da una valutazione probabilistica: se $G(\theta)$ denota la f.d.r. per Θ che esprime l'opinione di una persona in uno stato di informazione precedente l'osservabilità di E e $G(\theta/E)$ quella esprimente l'opinione nello stato di informazione che comprende E allora la coerenza impone che sia

$$dG(\theta/E) = \frac{\ell(\theta/E)}{\int_R \ell(\theta/E) dG(\theta)} dG(\theta) \ ,$$

ove $\ell(\theta / E)$ indica la funzione di verosimiglianza di θ corrispondente all'osservabilità di E.

Funzione di verosimiglianza

A causa dell'importanza centrale della funzione di verosimiglianza anche nell'approccio non bayesiano alla Statistica riteniamo opportuno a questo punto soffermarci brevemente su tale nozione. In corrispondenza all'evento osservabile $E = \left\{ \bigcap_{i=1}^{m} (X_i = x_i) \right\}$ i simboli $\ell(\theta / E)$ ed $\ell(\theta / x_1, ..., x_n)$ sono equivalenti, ma quest'ultimo non dev'essere confuso con la distribuzione congiunta delle variabili osservabili $X_1, ..., X_n$ dipendente dal parametro non noto θ : se, in particolare, tale distribuzione congiunta è dotata di densità, allora $\ell(\theta / x_1, ..., x_n)$ ed $f(x_1, ..., x_n; \theta)$ non devono essere confuse in quanto mentre la prima ha come argomento θ e la sequenza $x_1, ..., x_n$ va considerata nota, la seconda ha come argomenti le determinazioni $x_1, ..., x_n$ e θ va considerato noto.

Come conseguenza di quanto affermato si ha che $f(x_1, ..., x_n; \theta)$ è una funzione di n variabili, non negativa ed è $\int_{R^n} f(x_1, ..., x_n; \theta) \; dx_1 .....dx_n = 1$ mentre, in generale, è

$$\int_R \ell(\theta / x_1, ..., x_n) \; d\theta \neq 1 \; .$$

La definizione di verosimiglianza del parametro incognito θ corrispondente all'osservazione $x_1, ..., x_n$ viene solitamente data dicendo che essa è una funzione di θ proporzionale alla densità congiunta della sequenza $x_1, ..., x_n$ e scrivendo $\ell(\theta / x_1, ..., x_n) \propto f(x_1, ..., x_n; \theta)$.

Assumendo che la sequenza $x_1, ..., x_n$ provenga da un campionamento casuale, cioè assumendo che i n.a. osservabili $X_1, ..., X_n$ siano i.i.d., $\ell(\theta / x_1, ..., x_n)$ viene costruita per fattorizzazione:

$\ell(\theta / x_1, ..., x_n) = \prod_{j=1}^{n} \ell(\theta / x_j)$. E' bene però fin d'ora tener presente che per sequenze osservabili $X_1, ..., X_n$ più generali, in cui cioè i n.a. X_j sono mutuamente dipendenti, la costruzione di $\ell(\theta / x_1, ..., x_n)$ non può avvenire allo stesso modo: ritorneremo nel seguito su tale questione.

APPENDICE n. 1 : Giustificazione del risultato di T. Bayes

Indicando con Θ la proporzione non nota, e perciò aleatoria, di palline bianche nell'urna si assuma che:

1) in corrispondenza ad ogni possibile evento $\{\Theta = \theta\}$ i risultati aleatori di p + q estrazioni dall'urna, ognuna seguita dal reimbussolamento della pallina estratta, siano indipendenti e ugualmente distribuiti: ciò implica che, indicata con A una sequenza osservata di risultati di p + q estrazioni nella quale compaiono p palline bianche e q palline nere in un ordine determinato, sia $P\{A / \Theta = \theta\} = \theta^p.(1-\theta)^q$;
2) la probabilità che la proporzione Θ (avente determinazioni razionali) appartenga ad un qualunque sub-intervallo [a , b] di [0 , 1] sia b – a : chiaramente tale ipotesi corrisponde all'assunto di T. Bayes circa l'iniziale e assoluta mancanza di informazione su Θ. Si ottiene, formalmente, questo stesso risultato pensando di aver espresso l'opinione suddetta attribuendo a Θ una densità di probabilità iniziale $g(\theta)$ uniforme su [0 , 1] .

Con tali ipotesi la probabilità di estrarre una pallina bianca al colpo p + q + 1, evento E, e' data, per il teorema delle probabilità composte, dalla:

$$P(E / A) = \frac{P(E \cap A)}{P(A)} = \frac{\int_0^1 P(E \cap A/\theta).g(\theta)d\theta}{\int_0^1 P(A/\theta).g(\theta)d\theta} = \frac{p+1}{p+q+2} \quad .$$

L'ultima uguaglianza si prova facilmente tenendo presente che sussiste la relazione $\int_0^1 x^{\alpha-1}.(1-x)^{\beta-1}dx = \frac{\Gamma(\alpha).\Gamma(\beta)}{\Gamma(\alpha+\beta)}$, ove $\Gamma(\alpha)$ e' la funzione "gamma" di Eulero; per n intero positivo si ha $\Gamma(n) = (n-1)!$ ed inoltre si prova che per α reale ed n intero positivo e' $\Gamma(\alpha+n) =$

$[\alpha.(\alpha+1).....(\alpha+n-1)].\ \Gamma(\alpha)$.

Il suddetto rapporto di integrali può essere riformulato utilmente nel modo seguente:

$$\frac{\int_0^1 P(E \cap A/\theta).g(\theta)d\theta}{\int_0^1 P(A/\theta).g(\theta)d\theta} = \frac{\int_0^1 \theta^{p+1}.(1-\theta)^q.g(\theta)d\theta}{\int_0^1 \theta^p.(1-\theta)^q g(\theta)d\theta} = \int_0^1 \theta.\left[\frac{\theta^p.(1-\theta)^q.g(\theta)}{\int_0^1 \theta^p.(1-\theta)^q.g(\theta)d\theta}\right]d\theta$$

ove l'ultima espressione può interpretarsi come valor medio di Θ relativamente ad una densità di probabilità data dal contenuto della parentesi quadrata; avendo indicato in precedenza come **densità iniziale** la g(θ), quella in parentesi quadrata potrebbe denominarsi **densità finale**, con riferimento alla informazione ottenuta mediante le p + q estrazioni dall'urna. Denotando quest'ultima con g(θ/A) possiamo scrivere

$$g(\theta/A) = \frac{g(\theta).[\theta^p.(1-\theta)^q]}{\int_0^1 \theta^p.(1-\theta)^q.g(\theta)d\theta}$$

e tale espressione è nota come **teorema di Bayes** (per densità di probabilità). Esso esprime una regola di coerenza tra le due densità g(θ) e g(θ/A) ; il prodotto al numeratore in parentesi quadrata è detto **funzione di verosimiglianza** per Θ corrispondente all'incremento di informazione A.

Si osservi che mentre dal punto di vista che stiamo considerando, e che nel seguito chiameremo bayesiano, l'effetto dell'evento osservato A è quello di trasformare la valutazione iniziale dell'evento E

$$P(E) = \int_0^1 P(E/\Theta=\theta).g(\theta)d\theta = \int_0^1 \theta.g(\theta)d\theta = ½ ,$$

nella valutazione finale

$$P(E/A) = \int_0^1 P(E/\Theta=\theta).g(\theta/A)d\theta = \int_0^1 \theta.g(\theta/A)d\theta = \frac{p+1}{p+q+2} ,$$

il ben noto metodo di stima puntuale di "massima verosimiglianza", che consiste nella determinazione del valore di θ che massimizza $[\theta^p.(1-\theta)^q]$, porterebbe al valore $\hat{\theta} = \frac{p}{p+q}$, per cui si avrebbe P(E/A) ≈ $\hat{\theta} = \frac{p}{p+q}$.

APPENDICE n. 2 : Cenni sulla probabilità soggettiva.

Secondo B. de Finetti, la probabilità di un evento E è "l'espressione numerica del grado di fiducia che un individuo ripone nel verificarsi di E in base alle sue informazioni ed opinioni". Si tratta evidentemente di una nozione soggettiva e dinamica perché la probabilità P(E) dipende dall'individuo che la valuta e, per uno stesso individuo, dipende dal suo stato di conoscenza su E che generalmente varia con il tempo.

Tale definizione ricomprende, come casi particolari, le altre definizioni proposte storicamente quali, ad esempio, quella espressa dal "rapporto tra il numero dei casi favorevoli al verificarsi di E ed il numero di tutti i casi, giudicati ugualmente possibili" o quella espressa dalla "frequenza relativa di successo per E su un numero di prove giudicato sufficientemente elevato". Dal punto di vista soggettivo tali definizioni costituiscono semplicemente alcune possibili modalità di valutazione per P(E) in base al giudizio dell'individuo.

De Finetti ha proposto due criteri di valutazione operativi, tra loro equivalenti: il primo in termini di "quota di scommessa" ed il secondo in termini di "penalizzazione" ed entrambi consentono di evidenziare l'aspetto interpretativo della nozione di probabilità. In questa Appendice accenneremo soltanto al primo dei due.

Secondo il primo criterio la probabilità P(E) è interpretabile come la quota di scommessa che dà diritto all'importo aleatorio $|E|$, cioè all'importo unitario (un Euro) esigibile se l'evento E risulta vero; più in generale, se la posta della scommessa è l'importo S esigibile se E risulta vero, la corrispondente quota di scommessa è il prodotto S . P(E) . Pertanto la valutazione di P(E) da parte dell'individuo determina l'importo certo S . P(E) che egli giudica equivalente all'importo aleatorio che vale S se E risulta vero e 0 se E risulta falso. Ovviamente, l'equità della sua valutazione è garantita soltanto se l'individuo ignora all'atto della valutazione di P(E) quale parte gli spetterà nella scommessa: scommettitore o banco.

Nel caso di più valutazioni $P(E_1), P(E_2), P(E_3)$,……. da parte di uno stesso individuo e nello stesso stato di conoscenza de Finetti afferma che esse sono del tutto libere a condizione che risultino "coerenti" , o non contraddittorie; la condizione di coerenza è la seguente: se $S_1, S_2, S_3, \ldots\ldots\ldots$ sono le poste delle scommesse sugli eventi $E_1, E_2, E_3, \ldots\ldots\ldots$, le valutazioni $P(E_1), P(E_2), P(E_3)$,…… sono coerenti se, considerato un arbitrario sottoinsieme finito di n eventi $E_{j_1}, \ldots\ldots\ldots E_{j_n}$ e denotato con G = $\sum_{i=1}^{n} S_i.[|E_{j_i}| - P(E_{j_i})]$ il guadagno sulle n scommesse risulta min G $\leq 0 \leq$ max G per ogni intero n > 0 ed ogni sequenza di interi positivi $j_1, \ldots\ldots\ldots, j_n$ distinti tra loro. Al contrario, una valutazione $\{ P(E_j ; j \in J) \}$ è incoerente, o intrinsecamente contraddittoria, se risulta possibile individuare una combinazione di scommesse, cioè una sequenza di importi $\{ S_j ; j \in J \}$, che determini, con la fissata valutazione probabilistica, valori di G tutti dello stesso segno.

Da quanto detto si ricava che una valutazione probabilistica è possibile, in linea di principio, per qualsiasi evento e in qualunque stato di conoscenza ci si trovi con riferimento all'evento considerato. Ciò non dovrebbe però essere interpretato nel senso che, indipendentemente da quanto si conosce sull'evento E, sia sempre possibile specificare la valutazione precisa P(E) con assoluta convinzione. Una interpretazione di questo tipo è nota in letteratura come "dogma bayesiano di precisione" ed è stata oggetto di ampie discussioni ed obiezioni.

Da molto tempo e da parte di molti studiosi si ritiene che ad una valutazione probabilistica P(E) dovrebbe essere associato un indice della quantità di informazione che l'ha prodotta allo scopo di evidenziare il grado di convincimento del valutatore. Naturalmente non è questa la sede per approfondire tali argomenti: rinviamo il lettore interessato, per esempio, al testo di P. Walley dal titolo "Statistical Reasoning with Imprecise Probabilities" (Chapman and Hall, 1991).

Per un approfondimento dell'impostazione soggettiva della Teoria della probabilità suggeriamo al lettore interessato i seguenti riferimenti:

B. de Finetti: Teoria delle probabilità (2 vol.) – Giulio Einaudi editore, Torino 1970.

L. Crisma: Lezioni di Calcolo delle probabilità – Libreria Goliardica Editrice, Trieste 1997.

APPENDICE n. 3: Processi scambiabili e parzialmente scambiabili

Dal punto di vista formale, i processi scambiabili sono caratterizzati dalla condizione di invarianza delle distribuzioni congiunte rispetto a permutazioni arbitrarie degli indici dei n.a., cioè dall'invarianza delle distribuzioni rispetto all'ordine dei n.a. . Si richiede cioè che sia

$$F_{1,\ldots\ldots\ldots,n}(x_1,\ldots\ldots\ldots,x_n)=F_{j_1,\ldots\ldots\ldots,j_n}(x_1,\ldots\ldots\ldots,x_n)$$

per ogni intero positivo n $\geq$ 1, per ogni permutazione ($j_1,\ldots\ldots\ldots\ldots,j_n$) della sequenza (1,……,n) e per ogni sequenza di argomenti ($x_1,\ldots\ldots\ldots,x_n$). Per quanto concerne i momenti fino al secondo ordine si ha evidentemente $\varphi_X(t)=E(X_t)\equiv E(X_1)$ e $\psi_X(s,t)$ = $Cov(X_s,X_t)=\psi_X(s-t)=\sigma^2$ o $\psi_X(s-t)=\gamma>0$ a seconda che sia s – t = 0 o, rispettivamente, s – t $\neq$ 0 . La suddetta definizione implica che ogni insieme finito di n n.a. distinti X_t abbia la stessa distribuzione di probabilità congiunta, cioè

$$F_{t_1,\ldots\ldots\ldots,t_n}(x_1,\ldots\ldots\ldots,x_n)=F_n(x_1,\ldots\ldots\ldots,x_n),$$

qualunque sia l'intero positivo n, la sequenza di interi positivi distinti ($t_1,\ldots\ldots\ldots\ldots,t_n$) e la sequenza di argomenti ($x_1,\ldots\ldots\ldots\ldots,x_n$) e ove F_n indica la funzione di ripartizione congiunta per un qualunque insieme di n n.a. distinti.

Dal punto di vista interpretativo la condizione di scambiabilita' traduce l'idea di **analogia o equivalenza** tra i n.a. osservabili X_t in modo ben piu' efficace della condizione di indipendenza stocastica e uguale distribuzione degli stessi: infatti, se gli X_t sono assunti mutuamente indipendenti l'apprendimento dall'esperienza non puo' avvenire. Per il caso X_t = $\left|E_t\right|$ l'assunzione di indipendenza tra gli eventi equivale ad assumere che sia $P\left(E_{n+1}/K\right)\equiv P\left(E_{n+1}\right)$ qualunque sia l'evento osservabile K riguardante gli eventi $E_1,\ldots\ldots\ldots\ldots,E_n$. Per contro, l'assunzione di scambiabilita' tra gli eventi introduce tipicamente una dipendenza stocastica tra essi e questa implica la $P\left(E_{n+1}/K\right)\cong\frac{m}{n}$, come facilmente si puo' provare.

Sussiste un importante teorema di rappresentazione per i processi scambiabili illimitati, cioè costituiti da una infinità numerabile di n.a. X_t .

Teorema di B. de Finetti: tutti e soli i processi scambiabili illimitati $\{X_t; t \geq 1\}$ sono combinazioni lineari (o misture) di processi stocastici $\{X_t^{(\omega)}; t \geq 1\}$, $\omega \in \Omega$, i cui n.a. sono indipendenti e dotati di una comune funzione di ripartizione $F^{(\omega)}(x)$ nel senso che esiste una funzione di ripartizione G(ω) tale che per ogni intero positivo n ed ogni sequenza finita di indici ($t_1, ..., t_n$) sussiste la

$$F_{t_1,.....,t_n}(x_1,....., x_n) = \int_{\Omega} F^{(\omega)}(x_1,....., x_n)\ dG(\omega) = \int_{\Omega}\left[\prod_{j=1}^{n} F^{(\omega)}(x_j)\right]\ dG(\omega).$$

In questa sede ci limitiamo a fornire due semplici esemplificazioni del suddetto teorema di rappresentazione: nel primo esempio le funzioni di ripartizione univariate $F^{(\omega)}$ siano tutte di tipo normale o Gaussiano con valor medio ω e varianza unitaria; supponiamo ancora che G(ω) sia la funzione di ripartizione di una distribuzione normale con parametri $\mu \quad e \quad \sigma^2$. Si ha allora:

$$F_{t_1,.....,t_n}(x_1,....., x_n) = \int_{R}\left[\prod_{j=1}^{n}(2\pi)^{-1/2}.\exp\{-(x_j-\omega)^2/2\}\right]\ dG(\omega) =$$

$$= (2\pi)^{-n/2}\int_{R}\left[\exp\{-\frac{1}{2}\sum_{j=1}^{n}(x_j-\omega)^2\}\right].\left[(2\pi\sigma^2)^{-1/2}.\exp\{-\frac{\sigma^2}{2}(\omega-\mu)^2\}\right]d\omega$$

dalla quale si ricava, con facili calcoli, che $F_{t_1,.....,t_n}(x_1,....., x_n)$ corrisponde ad una distribuzione normale n – dimensionale caratterizzata da valori medi tutti uguali a μ, varianze tutte uguali a $1+\sigma^2$ e covarianze tutte uguali a σ^2.

Nel secondo esempio i n.a. X_t siano indicatori di eventi $|E_t|$ scambiabili per i quali si considera il prodotto logico $A = \{E_1' \wedge E_2' \wedge ..... \wedge E_n'\}$ implicante l'evento $\{\sum_{j=1}^{n}|E_j| = m\}$; supponiamo anche che sia noto il valore logico (vero o falso) di ogni evento del prodotto logico A. Le funzioni di ripartizione $F^{(\omega)}(x)$ sono identicamente nulle per x < 0, uguali a ω per $0 \leq x < 1$ e identicamente uguali a 1 per x > 1.

Se assumiamo che $G(\omega)$ sia la funzione di ripartizione di una distribuzione Beta con parametri numerici $\alpha \ \ e \ \ \beta$ l'applicazione del teorema di rappresentazione fornisce la

$$P(A) = \int_0^1 \left[\omega^m .(1-\omega)^{n-m} \right] . \left[\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha).\Gamma(\beta)} \omega^{\alpha-1} .(1-\omega)^{\beta-1} \right] d\omega$$

dalla quale si ricava, con facili calcoli, che è $P(A) = \frac{(\alpha)_m .(\beta)_{n-m}}{(\alpha+\beta)_n}$, ove $(\alpha)_m = \prod_{j=0}^{m-1} (\alpha + j)$.

La condizione di **scambiabilità parziale** è stata introdotta sempre da B. de Finetti nel 1937 ed è più flessibile rispetto a quella di scambiabilità e quindi più adatta a modellizzare schemi nei quali l'analogia tra le unità campionarie non è considerata perfetta. Si pensi a misurazioni ripetute di una stessa grandezza effettuate con strumenti di misura aventi precisioni diverse, a rischi assicurativi (per esempio contratti R.C.A.) giudicati non omogenei (per esempio per la diversa cilindrata degli autoveicoli), ai risultati aleatori di lanci ripetuti di monete diverse e così via.

Il modello più semplice di processo parzialmente scambiabile consiste di un **insieme finito** (per esempio una coppia, o una terna,.....) **di processi stocastici scambiabili, tra loro correlati**. Si considerino, per semplicità, due processi stocastici distinti $\{X_t ; t \geq 1\}$ e $\{Y_t ; t \geq 1\}$ ciascuno dei quali è supposto essere scambiabile. Circa i legami di dipendenza stocastica tra i due processi si possono ipotizzare tante situazioni diverse: i casi limite sono quello di indipendenza tra essi e, all'opposto, quello di scambiabilità semplice generalizzata (in cui non c'è bisogno di considerare distinti i n.a. del primo da quelli del secondo processo). Tra queste situazioni limite c'è tutta la gamma di situazioni di scambiabilità parziale non banale: ciò che le accomuna è la condizione di uguale covarianza tra ogni n.a. X_t del primo processo e tutti i n.a. Y_s del secondo, cioè $\mathrm{Cov}(X_t , Y_S) = \gamma$ per ogni coppia di valori (t,s) degli indici. Ovviamente dovrà essere $|\gamma| \leq \sqrt{\sigma_X^2 .\sigma_Y^2}$.

Dal punto di vista formale, i processi parzialmente scambiabili (costituiti da due processi scambiabili) sono caratterizzati dalla seguente proprieta' di invarianza delle distribuzioni congiunte:

$$F\left(x_{t_1}, ..., x_{t_n} ; y_{s_1}, ..., y_{s_m}\right) = F\left(x_1, ..., x_n ; y_1, ..., y_m\right)$$

per ogni coppia di interi non negativi n, m e ogni coppia di sequenze di interi positivi distinti $t_1, ..., t_n \ \ e \ \ s_1 ....., s_m$.

Sussiste il seguente risultato dovuto a B. De Finetti:

Teorema di rappresentazione: tutti e soli i processi parzialmente scambiabili illimitati (costituiti da due processi scambiabili componenti) sono combinazioni lineari (o misture) di processi stocastici i cui n.a. sono indipendenti, con una funzione di ripartizione $F_1^{(\omega)}(x)$ comune a tutti i n.a. X_t e una funzione di ripartizione $F_2^{(\eta)}(y)$ comune a tutti i n.a. Y_s nel senso che esiste una distribuzione congiunta $G(\omega, \eta)$ tale che sia

$$F\left(x_{t_1},.....,x_{t_n};y_{s_1},.....,y_{s_m}\right)=\iint_{\{\omega,\eta\}}\left[\prod_{i=1}^{n}F_1^{(\omega)}(x_{t_i})\right].\left[\prod_{j=1}^{m}F_2^{(\eta)}(y_{s_j})\right]dG(\omega,\eta)$$

per ogni coppia di interi non negativi (n,m) e ogni coppia di sequenze di interi positivi $t_1,.....,$ $t_n \;\; e \;\; s_1.....,s_m$.

In un primo esempio di applicazione del teorema precedente in cui $X_t \;\; e \;\; Y_s$ sono indicatori di eventi sia $g(\omega,\eta)$, la densità congiunta corrispondente a $G(\omega,\eta)$, una densita' di probabilita' di tipo Beta bivariata con parametri reali positivi $\nu_1,\nu_2 \;\; e \;\; \nu_3$, cioè

$$g(\theta_1,\theta_2)=\frac{\Gamma(\nu_1+\nu_2+\nu_3)}{\Gamma(\nu_1).\Gamma(\nu_2).\Gamma(\nu_3)}\theta_1^{\nu_1-1}.\theta_2^{\nu_2-1}.(1-\nu_1-\nu_2)^{\nu_3-1}.$$

Si prova allora che i processi $\{X_t\} \;\; e \;\; \{Y_s\}$ riescono entrambi scambiabili con $P\left\{\sum_{t=1}^{n}X_t=n\right\}=\frac{(\nu_1)_n}{(\nu_1+\nu_2+\nu_3)_n}$ e, rispettivamente, $P\left\{\sum_{t=1}^{n}Y_t=n\right\}=\frac{(\nu_2)_n}{(\nu_1+\nu_2+\nu_3)_n}$; si ha infine

$$P\left\{[\sum_{t=1}^{n}X_t=n]\wedge[\sum_{s=1}^{m}Y_s=m]\right\}=\frac{(\nu_1)_n.(\nu_2)_m}{(\nu_1+\nu_2+\nu_3)_{n+m}}.$$

Quest'ultima probabilità è quella che n eventi del primo tipo (con indicatori X_t) ed m eventi del secondo tipo (con indicatori Y_s) siano tutti veri, cioè che la frequenza di successo su n eventi del primo tipo ed m eventi del secondo tipo sia pari a n + m.

In un secondo esempio di applicazione le funzioni di ripartizione univariate $F_1^{(\omega)}(x)$, comune a tutti i n.a. X_t, ed $F_2^{(\eta)}(y)$, comune a tutti i n.a. Y_s, siano entrambe di tipo Gaussiano con valori medi ω ed η e varianze unitarie; supponiamo ancora che $G(\omega,\eta)$ sia la funzione di ripartizione di una distribuzione Gaussiana bivariata con vettore medio $\mu=(\mu_1,\mu_2)^T$ e matrice di varianze e covarianze $\Gamma=\begin{bmatrix}\gamma_1 & \gamma\\ \gamma & \gamma_2\end{bmatrix}$. Si verifica allora facilmente che $F\left(x_{t_1},.....,x_{t_n};y_{s_1},.....,y_{s_m}\right)$ e' la frunzione di ripartizione di una distribuzione Gaussiana (n+m) – dimensionale caratterizzata dai momenti $E(X_t)\equiv\mu_1,\; E(Y_s)\equiv\mu_2,\; Var(X_t)\equiv 1+\gamma_1,\; Var(Y_s)\equiv 1+\gamma_2,\; Cov(X_t,Y_s)\equiv\gamma$.

Per ulteriori approfondimenti sui processi scambiabili e parzialmente scambiabili si possono consultare, per esempio:

1) B. de Finetti – "Teoria delle probabilita'" (G. Einaudi, 1970 oppure, la corrispondente edizione in lingua inglese edita da J. Wiley, 1974).

2) L. Daboni e A. Wedlin – "Statistica: un'introduzione all'impostazione neo-bayesiana" (UTET, 1982).

3)J.M. Bernardo and A.F.M. Smith – “Bayesian Theory” (J. Wiley, 1994).

2) Introduzione all’inferenza bayesiana

Partiamo dalla nozione di **modello statistico parametrico**: esso consiste di un “processo stocastico di osservazione” $\{ X_t ; t \geq 1 \}$ per il quale è specificata una descrizione probabilistica, completa o parziale, nella quale compaiono uno o più parametri incogniti che indicheremo con θ, grandezza scalare o, rispettivamente, vettoriale. I tipi di specificazione completa più frequenti utilizzano:

a) una famiglia di distribuzioni condizionate tra loro coerenti, dette anche distribuzioni campionarie, $\{ F_{1,\ldots\ldots,n} (\mathbf{x} / \theta) ; n \geq 1 \}$, ove θ indica l’insieme dei parametri incogniti;
b) un’equazione stocastica alle differenze finite, o un sistema di equazioni stocastiche alle differenze finite, la cui soluzione determina la suddetta famiglia di distribuzioni condizionate $\{ F_{1,\ldots\ldots,n} (\mathbf{x} / \theta) ; n \geq 1 \}$.

Il processo di osservazione $\{ X_t \}$ potrebbe anche essere a parametro continuo, cioè t potrebbe assumere valori in un insieme T avente la potenza del continuo, come nel caso dell’intervallo [a, b] oppure $[0, +\infty]$ o dell’intero insieme dei numeri reali R. In tal caso la specificazione di tipo b) richiede equazioni differenziali stocastiche o sistemi di equazioni differenziali stocastiche anziché equazioni alle differenze finite. Nel seguito tratteremo quasi soltanto di processi generatori di dati, o di osservazione, a parametro discreto.

Presentiamo ora alcuni esempi dei due tipi di specificazione chiarendo, in qualche caso, anche la loro eventuale relazione. Per quanto riguarda il primo tipo di specificazione, cominciamo con il seguente

Esempio n. 1: processo di osservazione stazionario e normale.

Si assuma che ogni distribuzione congiunta $F_{1,\ldots\ldots,n} (\mathbf{x} / \theta)$ sia di tipo normale o gaussiano con vettore medio $E(\mathbf{X}/\theta) = \theta_1 . \mathbf{1}_n$ e funzione di covarianza $Cov (X_s , X_t / \theta) = \theta_2$ se $s = t$ e $Cov (X_s , X_t / \theta) = \theta_3$ se $s \neq t$. Il vettore dei parametri incogniti è dunque, in questo caso, $\theta = (\theta_1, \theta_2, \theta_3)^T$.

Esempio n. 2: processo di osservazione scambiabile e normale.

Si assuma che le variabili osservabili X_t abbiano, come nel primo esempio, distribuzioni congiunte di tipo normale e che siano subordinatamente indipendenti rispetto ai primi due parametri θ_1 e θ_2, cioè in corrispondenza ad ogni possibile coppia di valori per i parametri incogniti θ_1 e θ_2 esse sono assunte indipendenti.. Siamo ancora in presenza di un processo stazionario, ma ora è Cov (X_s, X_t / $\boldsymbol{\theta}$) = 0 quando $s \neq t$; vedremo in seguito che la covarianza non subordinata

Cov (X_s, X_t) è diversa da zero e costante, cosicché le variabili osservabili risultano essere mutuamente dipendenti. Ovviamente qui è $\boldsymbol{\theta} = (\theta_1, \theta_2)^T$.

Esempio n. 3: processo di osservazione markoviano.

Assumiamo che le variabili osservabili siano indicatori di eventi, di modo che $X_t = |E_t|$ assume il valore 1 se l'evento E_t è vero, il valore 0 se E_t è falso. Assumiamo anche che esse costituiscano una catena di Markov del primo ordine caratterizzata dalle probabilità subordinate

$$P(X_t = 1/ X_{t-1} = 1) = \theta_1, \quad P(X_t = 0/ X_{t-1} = 1) = 1 - \theta_1,$$

$$P(X_t = 0/ X_{t-1} = 0) = \theta_2, \quad P(X_t = 1/ X_{t-1} = 0) = 1 - \theta_2,$$

e dalla distribuzione non condizionata di X_1 espressa dalle $P(X_1 = 1) = \theta_3$ e $P(X_1 = 0) = 1 - \theta_3$. In questo caso il vettore parametrico è $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)^T$.

Per quanto concerne il secondo tipo di specificazione (sub b) alcuni semplici esempi sono i seguenti:

Esempio n. 4: processo di osservazione costituito da misurazioni di una grandezza incognita θ effettuabili iterativamente con uno strumento di misura, avente precisione nota π, che introduce soltanto errori accidentali di misura.

Assumiamo che sia $X_t = \theta + e_t$, ove e_t indica l'errore accidentale aleatorio, non osservabile, introdotto nella t-esima misurazione; assumiamo ancora che tali errori siano indipendenti da θ e che essi costituiscano un processo stocastico normale caratterizzato da valori medi tutti nulli, da varianze tutte uguali a $1/\pi$ e da covarianze tutte nulle. Vedremo in seguito che il processo$\{X_t ; t \geq 1\}$ risulta essere scambiabile e normale.

Esempio n. 5: processo di osservazione generato da un modello autoregressivo del primo ordine.

Supponiamo che sia $X_t = \theta_1 . X_{t-1} + e_t$, ove il processo stocastico $\{ e_t\}$ è un processo normale con valori medi tutti nulli, varianze tutte uguali a θ_2 e covarianze tutte nulle; in letteratura un tale processo è denominato "normal white noise" e indicato sinteticamente con $\{ e_t\} \approx$ NWN $(0 ; \theta_2)$. Si assume anche che la variabile X_0, detta "condizione iniziale", sia stocasticamente indipendente da ogni e_t . Supponendo che sia $X_0 = k$ e risolvendo l'equazione stocastica alle differenze $X_t = \theta_1 . X_{t-1} + e_t$ (per esempio con un procedimento ricorsivo) si ottiene il processo stocastico soluzione

$$X_t = \theta_1^t .k + (\theta_1^{t-1} .e_1 + \theta_1^{t-2} .e_2 + ... + e_t)$$

che è ancora un processo subordinatamente normale rispetto ad ogni ipotesi su θ_1 e θ_2 .

Se si può ritenere che sia $|\theta_1| < 1$ allora $\{ X_t ; t \geq 1 \}$ risulta approssimativamente stazionario con funzione valor medio approssimativamente nulla e funzione di covarianza espressa approssimativamente dalla Cov $(X_{t+h} , X_t / \theta_1 , \theta_2) = \frac{\theta_1^h .\theta_2}{(1-\theta_1^2)}$.

Esempio n. 6: modello di regressione lineare semplice con parametri variabili nel tempo.

Esso è costituito dalle due seguenti equazioni lineari:

$$X_t = \theta_{0t} + \theta_{1t} . Y_t + u_t \quad ; \quad \boldsymbol{\theta}_t = A . \boldsymbol{\theta}_{t-1} + e_t$$

ove la prima rappresenta un modello di regressione lineare semplice e la seconda un modello autoregressivo vettoriale del primo ordine che descrive l'evoluzione del parametro vettoriale $\boldsymbol{\theta}_t = (\theta_{0t} , \theta_{1t})^T$. I due processi di errore $\{u_t\}$ e $\{e_t\}$, nel caso più semplice, sono assunti indipendenti tra loro e ciascuno di essi è supposto essere di tipo "normal white noise" con varianze note $\sigma_u^2 e \sigma_e^2$.

Tale modello è un caso particolare del "modello lineare dinamico", modello costituito da una "equazione di misurazione" che mette in relazione le variabili osservabili (nel nostro caso X_t e Y_t) con le variabili non osservabili, o variabili di stato, (nel nostro caso θ_{0t} e θ_{1t}) e da una "equazione di evoluzione" concernente le variabili di stato .

In una impostazione bayesiana dell'inferenza statistica, alle suddette specificazioni occorre aggiungere una valutazione dell'incertezza riguardante i parametri incogniti del modello statistico. Ciò si realizza,

nell'approccio bayesiano standard, attribuendo alla totalità dei parametri incogniti una distribuzione congiunta di probabilità basata sull'informazione disponibile. Riteniamo importante, a questo punto, fare un'osservazione. E' abbastanza diffusa l'opinione che mentre la specificazione del modello condizionato, nella forma a) o b), di cui si è parlato finora ha un carattere sostanzialmente oggettivo perché discende direttamente dal problema concreto che vogliamo affrontare, invece la specificazione probabilistica dell'incertezza sui parametri ha carattere essenzialmente soggettivo. E' invece nostra convinzione che entrambe le specificazioni abbiano un carattere soggettivo e che la maggiore facilità che talvolta si trova nello scegliere il modello condizionato dipenda dal fatto che esso concerne le variabili direttamente osservabili anziché dei parametri che non sono osservabili e che hanno soltanto una relazione indiretta con le variabili osservabili.

In qualche caso anche la scelta del modello condizionato può essere problematica. Per esempio, con riferimento all'Analisi delle serie temporali, si pensi al problema della specificazione di un modello entro la classe ARMA (p,q) per una fissata serie storica. Se la nostra incertezza riguardasse la scelta di un modello AR (2) piuttosto che un modello MA (1) potremmo venirne fuori abbastanza facilmente ricorrendo a ben noti strumenti statistici; quando invece si trattasse di dover scegliere tra un modello ARMA (2,1) o un modello ARMA (2,2) è ben difficile pensare di poter giungere ad una scelta convinta attraverso l'uso di quegli strumenti, in modo tale cioè che non sussista alcun dubbio sulla decisione presa.

Con la specificazione del modello condizionato $\{ F_{1,\ldots\ldots,n}(\mathbf{x} / \theta) ; n \geq 1 \}$ e della funzione di ripartizione della distribuzione di probabilità congiunta $G(\theta)$ per i parametri condizionanti si è specificata completamente la famiglia di distribuzioni congiunte non condizionate $\{ F_{1,\ldots\ldots,n}(\mathbf{x}) ; n \geq 1 \}$ del processo di osservazione, famiglia che denomineremo nel seguito "legge temporale iniziale" del processo $\{ X_t ; t \geq 1 \}$. Si osservi che questa famiglia di distribuzioni fornisce una conoscenza probabilistica completa del processo di osservazione.

A titolo di esempio determineremo la legge temporale iniziale per il modello dell'Esempio n. 4.

Supponiamo che in base alle informazioni disponibili venga attribuita al parametro incognito θ una distribuzione iniziale di tipo normale con parametri numerici μ e σ^2 fissati, cioè si assuma che sia $g(\theta) \sim N(\mu, \sigma^2)$. Subordinatamente ad ogni possibile valore di θ le variabili X_t hanno la stessa densità condizionata $f(x / \theta) \sim N(\theta, \pi^{-1})$ e sono condizionatamente indipendenti, pertanto considerata la sequenza finita $(X_1,\ldots\ldots,X_n)$ la sua densità congiunta è data dalla

$$f(\mathbf{x} / \theta) \sim N^{(n)}(\theta . \mathbf{1} ; \pi^{-1} . I_n) .$$

Ricorrendo a note proprietà delle distribuzioni congiunte normali si può provare che la densità non condizionata della sequenza $(X_1,\ldots\ldots,X_n)$ è ancora normale con parametri costituiti dal comune valor medio

E (X_t) = μ , dalla comune varianza Var (X_t) = $\sigma^2 + \pi^{-1}$ e dalla comune covarianza Cov (X_s , X_t) = σ^2 . Il processo generatore di dati { X_t ; $t \geq 1$ } è quindi un processo normale e scambiabile la cui legge temporale è costituita dalle suddette distribuzioni congiunte di tipo normale.

Non è sempre così semplice individuare la legge temporale del processo generatore di dati a partire dalla specificazione del modello condizionato e della distribuzione di probabilità dei parametri subordinanti. Ad esempio, con riferimento al modello dell'Esempio n. 1, già la specificazione della distribuzione congiunta del parametro subordinante $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)^T$ crea qualche problema in quanto il valore del parametro θ_3, il cui significato è quello di una covarianza, è necessariamente limitato ad assumere valori nell'intervallo $[-\theta_2 , \theta_2]$. Una volta che in qualche modo si sia scelta la distribuzione g ($\theta_1, \theta_2, \theta_3$) riesce comunque problematica la determinazione della corrispondente legge temporale per il processo { X_t ; $t \geq 1$ }, almeno se l'obiettivo è quello di far riferimento a famiglie di distribuzioni aventi forma funzionale abbastanza semplice.

Definiamo riassuntivamente a questo punto la **legge temporale iniziale** del processo di osservazione { X_t ; $t \geq 1$ } come la famiglia di distribuzioni congiunte, non condizionate, { $F_{1,.....,n}$ (**x**) ; $n \geq 1$ } che si ottiene a partire dalle distribuzioni campionarie { $F_{1,.....,n}$ (**x** / $\boldsymbol{\theta}$) ; $n \geq 1$ } e dalla funzione di ripartizione della distribuzione di probabilità congiunta G ($\boldsymbol{\theta}$) dell'insieme di parametri condizionanti $\boldsymbol{\theta}$. Ovviamente è, per ogni intero n :

$$F_{1,.....,n}(\mathbf{x}) = \int_{\{\theta\}} F_{1,.....,n}(x_1,....., x_n / \theta)\ dG(\boldsymbol{\theta}),$$

ove nell'espressione suddetta si è impiegato l'integrale di Stieltjes. Se G ($\boldsymbol{\theta}$) è dotata di densità,

g($\boldsymbol{\theta}$), allora si può scrivere l'espressione di $F_{1,.....,n}$ (**x**) in termini di un integrale di Cauchy-Rieman secondo la

$$F_{1,.....,n}(\mathbf{x}) = \int_{\{\theta\}} F_{1,.....,n}(x_1,....., x_n / \theta)\ g(\boldsymbol{\theta})\, d\boldsymbol{\theta} .$$

L'aggettivo "iniziale" sta semplicemente ad indicare "anteriore all'incremento di informazione di cui vogliamo trattare" nell'attuale procedimento di inferenza statistica e che concretamente riguarderà l'evento K = [$\bigcap_{t=1}^{n} (X_t = x_t)$] concernente i valori dei primi n numeri aleatori del processo di osservazione. Analogamente la distribuzione G ($\boldsymbol{\theta}$) sarà denominata "distribuzione iniziale di $\boldsymbol{\theta}$ " in quanto basata su informazioni e opinioni su $\boldsymbol{\theta}$ disponibili anteriormente alla conoscenza di eventi del tipo K; per maggiore chiarezza essa sarà indicata con G_0($\boldsymbol{\theta}$).

Fin qui ci siamo limitati a presentare le nozioni di modello statistico parametrico e di legge temporale (iniziale) del processo generatore di dati $\{ X_t ; t \geq 1 \}$. Abbiamo anche accennato ad un eventuale incremento di informazione K che rivestirà un ruolo centrale nell'inferenza statistica. Avvertiamo subito il lettore che esso va interpretato come **un evento possibile** e non come **un evento effettivamente osservato**, rinviando però tutte le considerazioni in proposito ad un discorso successivo. La possibilità che il nostro stato di informazione iniziale venga ampliato dalla conoscenza dell'evento K ci costringe, per coerenza, a chiederci quale sarebbe la legge temporale corrispondente al nuovo stato di informazione complessivo: una breve riflessione sulle prime nozioni di calcolo delle probabilità ci conduce alla constatazione che è necessario trasformare la legge temporale iniziale in una nuova legge temporale, che chiameremo "finale" e che dev'essere interpretata come **un insieme di distribuzioni condizionate all'evento K**.

Chiaramente, se l'evento K riguarda le prime n variabili osservabili X_t , o più esattamente i loro possibili valori, la nuova legge temporale (quella finale) riguarderà le variabili osservabili da X_{n+1} in poi, cioè il processo di osservazione "residuo" $\{ X_t ; t \geq n+1 \}$. La trasformazione che opera sulla legge iniziale determinando quella finale discende dalla norma di coerenza che collega tra loro le distribuzioni congiunte a quelle marginali e a quelle condizionate del processo $\{ X_t ; t \geq 1 \}$ e cioè il teorema delle probabilità composte. Supponendo che le distribuzioni $F_{1,\ldots\ldots,n}(\mathbf{x})$ siano dotate di densità, che indicheremo con $f_{1,\ldots\ldots,n}(\mathbf{x})$, si ha dunque:

$$f_{n+1,\ldots\ldots\ldots,n+m}(x_{n+1},\ldots\ldots,x_{n+m}/K)=\frac{f(x_1,\ldots\ldots,x_n,x_{n+1},\ldots\ldots\ldots,x_{n+m})}{f(x_1,\ldots\ldots,x_n)}\ ,\quad m\geq 1\ ,$$

ove le densità presenti nel rapporto a secondo membro appartengono entrambe alla legge iniziale; si osservi anche che mentre la sequenza $x_{n+1},\ldots\ldots,x_{n+m}$ è generica, quella $x_1,\ldots\ldots,x_n$ coincide con la sequenza dei valori che caratterizza l'evento K.

E' forse il caso di soffermarsi brevemente a considerare la differenza di significato tra la densità congiunta non condizionata

$f_{n+1,\ldots\ldots\ldots,n+m}(x_{n+1},\ldots\ldots,x_{n+m}) = \int\ldots\int f_{1,\ldots\ldots\ldots,n+m}(x_1,\ldots\ldots,x_{n+m})dx_1\ldots\ldots dx_n$ appartenente alla legge iniziale e l'analoga densità condizionata $f_{n+1,\ldots\ldots\ldots,n+m}(x_{n+1},\ldots\ldots,x_{n+m}/K)$, fornita dal precedente rapporto e appartenente alla legge temporale finale.

Se $f_{n+1,\ldots\ldots\ldots,n+m}(x_{n+1},\ldots\ldots,x_{n+m})$ esprime la nostra valutazione probabilistica sull'evento K* = $[\bigcap_{t=n+1}^{n+m}(X_t = x_t)]$ nello stato di informazione iniziale allora la densità condizionata $f_{n+1,\ldots\ldots\ldots,n+m}(x_{n+1},\ldots\ldots,x_{n+m}/K)$ rappresenta la valutazione che necessariamente, per coerenza, esprime la nostra opinione nello stato di informazione ampliato dall'ipotesi ulteriore che l'evento K sia vero. Se ritenessimo che quest'ultima densità non corrisponde alla nostra effettiva opinione allora dovremmo rivedere la legge temporale iniziale correggendola per qualche aspetto. Anche su questo punto dovremo soffermarci ulteriormente nel seguito.

Riassumendo, abbiamo definito **l'induzione statistica** in termini di una trasformazione che opera sulla legge temporale iniziale, che nel seguito indicheremo con $\Lambda(H)$ ove H rappresenta lo stato di informazione iniziale, determinando la legge temporale finale $\Lambda(H \bigcap K)$ ove $H \bigcap K$ indica lo stato di informazione ampliato dall'ipotesi che K sia vero. La trasformazione è determinata dalla norma di coerenza rappresentata dal teorema delle probabilità composte. Denomineremo la trasformazione $\Lambda(H) \Rightarrow \Lambda(H \bigcap K)$ sopra descritta "procedimento inferenziale diretto"; in letteratura le distribuzioni congiunte di $\Lambda(H \bigcap K)$ vengono spesso denominate "distribuzioni previsionali" e in effetti se nello stato di informazione $H \bigcap K$ volessimo effettuare una previsione su X_{n+1} e su X_{n+2} dovremmo utilizzare a tale scopo la distribuzione congiunta $f_{n+1,n+2}(x_{n+1}, x_{n+2} / K) \in \Lambda(H \bigcap K)$.

A questo punto occorre aggiungere che in taluni casi, e precisamente quando il processo di osservazione è supposto scambiabile o parzialmente scambiabile, è praticabile un diverso approccio all'inferenza statistica, che denomineremo "indiretto" e che porta alle stesse conclusioni dell'approccio diretto già presentato. Si tratta quindi di una diversa forma di inferenza agente sul parametro θ che riesce equivalente all'approccio diretto in quanto conduce alla stessa legge finale $\Lambda(H \bigcap K)$. Essa si concretizza nella trasformazione della distribuzione iniziale $G_0(\theta)$ in una nuova distribuzione, che chiameremo "finale", $G_n(\theta)$ mediante l'applicazione della stessa norma di coerenza nella forma del "teorema di Bayes". Nel simbolo $G_n(\theta)$ il deponente n rappresenta il numero di variabili osservabili che interviene nell'incremento di informazione K, ma allo scopo di evidenziare meglio il ruolo di K in questa trasformazione si potrebbe raffigurare il passaggio da $G_0(\theta)$ a $G_n(\theta)$ nella forma $G_0(\theta) = G(\theta) \Rightarrow G(\theta / K)$.

Quest'ultima trasformazione può essere sinteticamente espressa dalla

$$d\,G(\theta / K) \propto l(\theta / K)\, d\,G(\theta)$$

oppure, se esiste la densità corrispondente alla $G(\theta)$, dalla

$$g(\theta / K) \propto l(\theta / K) \,.\, g(\theta),$$

ove $l(\theta / K)$ in entrambe le relazioni rappresenta la funzione di verosimiglianza per θ corrispondente all'incremento di informazione K. Il segno di proporzionalità $\propto$ indica la mancanza di una costante di normalizzazione pari al reciproco di $\int_{\{\theta\}} l(\theta / K).g(\theta)d\theta$ nel senso che è, con riferimento alla seconda espressione :

$$g(\theta / K) = l(\theta / K) \cdot g(\theta) / \int_{\{\theta\}} l(\theta / K).g(\theta)d\theta \propto l(\theta / K) \cdot g(\theta).$$

Nei prossimi esempi, tutti relativi a processi di osservazione scambiabili, proveremo l'equivalenza dei due approcci inferenziali.

3) Analisi statistica bayesiana su proporzioni

Prendiamo in considerazione il problema dell'inferenza statistica sulla proporzione incognita Θ degli "individui" di una collettività aventi una fissata caratteristica di interesse: può trattarsi di lampadine elettriche prodotte in un dato stabilimento alcune delle quali presentano un certo difetto di fabbricazione o di esseri umani ricoverati in una data struttura ospedaliera alcuni dei quali hanno il gruppo sanguigno di tipo RH negativo o altro.

Allo scopo di costruire uno schema generale gli "individui" verranno assimilati a palline contenute in un'urna, alcune di colore bianco e altre di colore nero, e si supporrà di indagare sulla proporzione di bianche procedendo ad esperimenti consistenti in estrazioni successive dall'urna di una pallina alla volta, con reimbussolamento della pallina estratta. In questo modo la composizione dell'urna (cioè la percentuale di palline bianche) rimane la stessa dopo ogni estrazione. Indicheremo con $\{X_t ; t \geq 1\}$ il processo di osservazione, ove X_t rappresenta l'indicatore dell'evento E_t = "il risultato della t-esima estrazione è una pallina bianca". Supporremo, in generale, di non conoscere il numero totale delle palline nell'urna e di poter effettuare un fissato numero m di estrazioni.

Evidentemente si è considerato il più semplice degli schemi di estrazione possibili e ciò allo scopo di concentrare tutta l'attenzione sull'essenza dell'impostazione bayesiana e cioè sulla trasformazione coerente dell'opinione sulla proporzione incognita Θ espressa mediante una valutazione probabilistica iniziale.

Sembra plausibile assumere che nelle suddette condizioni sperimentali tutte le sequenze di m estrazioni, $X_{t_1}, \ldots\ldots\ldots, X_{t_m}$, qualunque siano i valori degli indici purchè diversi, siano tra loro equivalenti dal punto di vista dell'informazione arrecabile sulla proporzione di palline bianche presenti nell'urna (e ciò per ogni valore di m); in tal caso, come si è già detto, riesce giustificabile assumere la condizione di scambiabilità per il processo $\{X_t ; t \geq 1\}$ poiché è ragionevole supporre che in corrispondenza ad ogni ipotesi sul valore θ di Θ le variabili X_t costituiscano un processo bernoulliano, cioè siano i.i.d. . Attribuiremo ad ogni X_t una distribuzione condizionata (o campionaria) del tipo $P(X_t = x / \Theta = \theta) = \theta^x.(1-\theta)^{1-x}$, con x = 0 oppure x = 1, mentre esprimeremo l'opinione iniziale sul valore di Θ mediante una distribuzione di tipo Beta avente una densità di probabilità del tipo $g_0(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha).\Gamma(\beta)} \cdot \theta^{\alpha-1}.(1-\theta)^{\beta-1}$, ove α e β sono parametri reali e positivi.

Il modello statistico accolto è quindi costituito dall'ipotesi di scambiabilità del processo di osservazione $\{X_t ; t \geq 1\}$, che consente di ritenere le X_t condizionatamente i.i.d., e dalle scelte della comune distribuzione condizionata $P(X_t = x / \Theta = \theta) = \theta^x.(1-\theta)^{1-x}$ (detta distribuzione di Bernoulli) e della distribuzione iniziale di tipo Beta con densità $g_0(\theta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha).\Gamma(\beta)} . \theta^{\alpha-1}.(1-\theta)^{\beta-1}$ per la proporzione incognita Θ . Tale modello è indicato in letteratura come "schema Bernoulli – Beta".

Ci sembra opportuna, a questo punto, una breve riflessione sul significato da attribuire alle scelte che hanno condotto al nostro modello. Si potrebbe pensare che la densità $g_0(\theta)$ debba esprimere con precisione l'opinione iniziale su Θ nel senso che, fissato un qualunque intervallo (a, b) contenuto in [0, 1], il valore dell'integrale $\int_a^b g_0(\theta)d\theta$ rappresenti effettivamente e con precisione la valutazione iniziale della $P\{a \leq \Theta \leq b\}$. Senza voler negare che in qualche caso, quando cioè le conoscenze disponibili su Θ sono basate su numerosissime esperienze precedenti, ciò possa corrispondere alla situazione concreta ci sembra che nella maggior parte dei problemi concreti la scelta di un modello quale quello descritto vada interpretata piuttosto come una plausibile ipotesi di lavoro ritenuta accettabile in relazione agli scopi che si vogliono conseguire.

Una considerazione analoga andrebbe fatta circa l'assunzione di scambiabilità dei n.a. osservabili nei casi concreti: essa può costituire uno schema probabilistico adeguato in certi casi, ma in altri potrebbe dar adito a sospetti di inadeguatezza. Si pensi, per esempio, al problema del gruppo sanguigno RH negativo: se fosse plausibile, dal punto di vista medico, pensare che il gruppo RH negativo è leggermente più frequente nelle donne che negli uomini potrebbe porsi l'alternativa di assumere l'ipotesi di scambiabilità o quella di scambiabilità parziale per il processo $\{X_t ; t \geq 1\}$.

In definitiva, dal punto di vista teorico, la scelta di un modello statistico, come anche ogni valutazione probabilistica, andrebbe motivata associando ad essa un indice della quantità di informazione su cui quella scelta si basa. Se ciò non è possibile è bene ritenere che il modello statistico adottato sia un'ipotesi di lavoro provvisoria più o meno adeguata a seconda dei casi e suscettibile di eventuali modifiche al sopraggiungere di ulteriori informazioni significative in proposito.

Immaginando di effettuare m estrazioni dall'urna e di ottenere s palline bianche ed m – s nere, disposte in un qualche ordine, la corrispondente verosimiglianza per la proporzione incognita Θ sarebbe data dalla $\ell(\theta / x_1, ..., x_m) = \prod_{j=1}^{m} \ell(\theta / x_j) = \theta^s.(1-\theta)^{m-s}$ in forza del modello accolto. Dal risultato si ricava che gli elementi di informazione presenti nella verosimiglianza corrispondono all'evento $K = \{ \sum_{t=1}^{m} X_t = s \}$ implicato dalla sequenza $x_1, ..., x_m$ dei risultati dell'osservazione.

Incominciando con l'approccio inferenziale indiretto, l'applicazione del teorema di Bayes alla densità iniziale è espressa dalla trasformazione:

$$g(\theta / K) \propto g_0(\theta).\left[\theta^s.(1-\theta)^{m-s}\right] \propto \frac{\Gamma(\alpha+\beta+m)}{\Gamma(\alpha+s).\Gamma(\beta+m-s)}.\theta^{\alpha+s-1}.(1-\theta)^{\beta+m-s-1} \ ,$$

che fornisce una distribuzione finale per Θ ancora di tipo Beta, con parametri $\alpha^* = \alpha + s$ e $\beta^* = \beta + m - s$. La forma funzionale della distribuzione di Θ è rimasta dello stesso tipo a causa dell'analogia formale dei nuclei della densità iniziale $\theta^{\alpha-1}.(1-\theta)^{\beta-1}$ e della funzione di verosimiglianza $\left[\theta^s.(1-\theta)^{m-s}\right]$. Diremo di più su questo punto nel seguito.

I nuovi parametri sono costituiti dalle somme dei parametri iniziali α e β con il numero osservato di palline bianche s e, rispettivamente, quello di palline nere m – s ; poiché questi due soli numeri compaiono nell'espressione della densità finale possiamo affermare che la coppia (s, m – s) o equivalentemente quella (s , m), costituisce un **riassunto esaustivo** (o statistica sufficiente) dell'osservazione statistica.

Si osservi che il nostro riassunto esaustivo è equivalente alla sequenza di m – s zeri ed s uni, (0,0,……,0, 1,1,……,1), nel senso che noto l'uno si ricava l'altra e viceversa, e che tale sequenza si ottiene mediante un'opportuna permutazione della sequenza osservata $x_1, ..., x_m$. Il lettore avrà certamente riconosciuto nella trasformazione $(x_1, ..., x_m) \rightarrow$ (0,0,……,0, 1,1,……,1) la **statistica d'ordine** (o statistica ordinata): affermiamo che quest'ultima costituisce il riassunto esaustivo tipico per qualunque processo stocastico di osservazione scambiabile, indipendentemente dalla forma funzionale delle distribuzioni di probabilità della legge temporale. La ragione di ciò risiede nella proprietà di invarianza, rispetto all'ordine degli argomenti, di tutte le distribuzioni di probabilità congiunte di un processo stocastico scambiabile.

In base alla densità finale g(θ / K) la valutazione della probabilità di ottenere una pallina bianca alla (m+1)-esima estrazione è determinabile secondo la

$$P(E_{m+1} / K) = E(\Theta / K) = \frac{\alpha^*}{(\alpha^*+\beta^*)} = \frac{\alpha+s}{\alpha+\beta+m} = \frac{(\alpha+\beta).\frac{\alpha}{\alpha+\beta}+m.\frac{s}{m}}{(\alpha+\beta)+m} \ .$$

L'ultima espressione assume la forma di una combinazione lineare convessa della probabilità

$P_0(E_{m+1}) = \frac{\alpha}{\alpha+\beta}$ e della frequenza relativa di successo osservata s/m , con pesi dati da (α + β) e, rispettivamente, la numerosità m delle estrazioni effettuate. Al crescere del numero m di osservazioni la valutazione finale $P(E_{m+1} / K)$ tende a coincidere con la frequenza relativa di successo osservata m/n che, ricordiamolo, è la stima di massima verosimiglianza della probabilità comune degli eventi osservabili.

E' bene sottolineare che questa sostanziale concordanza tra le conclusioni a cui conduce l'approccio neo-bayesiano dell'inferenza statistica e quelle degli altri approcci non bayesiani è un fatto generale in presenza di numerose osservazioni statistiche; quando cioè l'incremento di informazione rispetto allo stato di conoscenza iniziale è rilevante, quest'ultimo incide poco sulle valutazioni probabilistiche finali. Le differenze tra le conclusioni dei diversi approcci alla Statistica sono invece sensibili nelle situazioni in cui le osservazioni campionarie disponibili sono poche e scarsamente omogenee tra loro e ciò si verifica tipicamente in campi di studio quali la sperimentazione clinica medica, l'economia, la meteorologia, etc.

Aggiungiamo che, come si può verificare agevolmente in base al teorema di rappresentazione già utilizzato, le probabilità finali $P\{\sum_{t=m+1}^{m+n} X_t = n / K\}$ risultano determinate dalla relazione:

$$P\{\sum_{t=m+1}^{m+n} X_t = n / K\} = E(\Theta^n / K) = [(\alpha^*)_n / (\alpha^* + \beta^*)_n],$$

ove $(\alpha^*)_n = \alpha^*.(\alpha^*+1).(\alpha^*+2)\ldots\ldots(\alpha^*+n-1)$ è il simbolo di Pochhammer .

Poiché la legge temporale iniziale del processo scambiabile di indicatori di eventi $\{X_t ; t \geq 1\}$ è rappresentabile mediante la successione, al variare di n, delle probabilità $\{(\alpha)_n / (\alpha + \beta)_n\}$ e la legge finale è rappresentata dalla analoga successione $\{(\alpha^*)_n / (\alpha^* + \beta^*)_n\}$ viene spontaneo chiedersi quale trasformazione porta dalla prima alla seconda e tale questione costituisce l'essenza del procedimento bayesiano diretto. Come già detto, la suddetta trasformazione si realizza mediante

l'applicazione del teorema delle probabilità composte.

Indicando, per semplicità, le probabilità $P\{\sum_{t=1}^{n} X_t = n\}$ col simbolo $\omega(n,n)$ e le probabilità $P\{\sum_{t=1}^{n} X_t = r\}$, $r \leq n$, col simbolo $\omega(n,r)$, si può facilmente dimostrare che sussiste la seguente relazione tra esse:

$$\omega(n,r) = \binom{n}{r} . (-1)^{n-r} . \Delta^{n-r} \omega(r,r) \ ,$$

ove con $\Delta^{n-r}\omega(r,r)$ si è indicata la differenza di ordine n – r per le $\omega(r,r)$ avente l'espressione

generale

$$\Delta^{n-r}\omega(r,r) = \sum_{j=0}^{n-r} \binom{n-r}{j} . (-1)^{n-r-j} \omega(r+j, r+j) \ .$$

Si ottiene allora che in corrispondenza alle $\omega(r,r) = [(\alpha)_r / (\alpha + \beta)_r]$ sussistono le :

$$\omega(n,r) = \binom{n}{r} . \frac{(\alpha)_r . (\beta)_{n-r}}{(\alpha+\beta)_n} \ .$$

Considerato ora un generico evento $A = (S_k = k)$, riguardante k eventi distinti tra loro e dagli m eventi osservati, si ottiene:

$$P\{A / K\} = P(K \wedge A) / P(K) = \{[\omega(m+k, s+k) / \binom{m+k}{s+k}] / [\omega(m,s) / \binom{m}{s}]\} =$$

$$= \left[\frac{(\alpha)_{s+k} . (\beta)_{m-s}}{(\alpha+\beta)_{m+k}}\right] / \left[\frac{(\alpha)_s . (\beta)_{m-s}}{(\alpha+\beta)_m}\right] = \frac{(\alpha)_{s+k} . (\alpha+\beta)_m}{(\alpha)_s . (\alpha+\beta)_{m+k}} = \frac{(\alpha+s)_k}{(\alpha+\beta+m)_k}$$

e, chiaramente, l'ultimo rapporto coincide con il momento k-esimo di una densità di tipo Beta con parametri α+s e β+m-s .

Si è verificata così l'equivalenza del procedimento inferenziale diretto con quello indiretto in quanto quest'ultimo avrebbe fornito la suddetta probabilità P{A / K} mediante la

$$P\{A / K\} = \int_0^1 \theta^k . g(\theta / K) d\theta = \frac{(\alpha + s)_k}{(\alpha + \beta + m)_k} .$$

E' bene ripetere che entrambi i procedimenti inferenziali descrivono il passaggio da una legge temporale iniziale $\Lambda(H)$ per il processo di osservazione scambiabile $\{X_t ; t \geq 1\}$ ad una nuova legge temporale $\Lambda(H \wedge K)$, che denomineremo "finale", per il processo residuo $\{X_t ; t \geq m+1\}$ costituito dalle variabili osservabili non ancora osservate. In quanto precede si è indicato con H lo stato di informazione iniziale sulla base del quale è stata specificata la legge $\Lambda(H)$; si può pensare H come un prodotto logico di tanti eventi ciascuno dei quali corrisponde ad un'informazione elementare sul processo $\{X_t ; t \geq 1\}$. Poiché la trasformazione $\Lambda(H) \Rightarrow \Lambda(H \wedge K)$ è stata effettuata mediante l'applicazione della norma di coerenza costituita dal teorema delle probabilità composte (o dal teorema di Bayes) si può affermare che $\Lambda(H \wedge K)$ è l'unica legge temporale coerente con $\Lambda(H)$ e con l'evento osservabile K.

Nel problema inferenziale finora considerato, che potremmo denominare "apprendimento statistico su proporzioni", si è specificata una legge temporale iniziale $\Lambda(H)$ costituita dalla successione delle probabilità $\{\omega(n,n), n \geq 1\}$; la sua scelta deve fondarsi sullo stato di informazione iniziale H che può essere più o meno ampio. Nell'esempio da noi considerato si è scelta la successione $\omega(n,n) = (\alpha)_n / (\alpha + \beta)_n$ dei momenti di una densità Beta (α , β), con parametri fissati.

La particolare densità Beta con parametri $\alpha = \beta = 1$, uniforme sull'intervallo (0 , 1), è detta "non informativa" in quanto assegna ad ogni sub-intervallo di (0 , 1) avente lunghezza λ una stessa probabilità, appunto λ, indipendentemente dalla sua collocazione all'interno dell'intervallo (0 , 1). In tal modo, si afferma, non si esprimerebbe alcuna opinione sul valore del parametro incognito Θ o, equivalentemente, sulla probabilità subordinata $P\{X_t = 1/ \Theta = \theta\} = \theta$. Alla densità Beta uniforme corrisponde la successione dei momenti $\omega(n,n) = (n + 1)^{-1}$ ed è facile verificare che risulta anche $\omega(n,m) = (n + 1)^{-1}$, per ogni $m \leq n$. Per quanto si è già visto, in presenza dell'incremento di informazione K, l'applicazione del teorema di Bayes a tale densità porta ad una densità finale Beta (s+1 , m-s+1) il cui valor medio è

$$E(\Theta / K) = \frac{s+1}{m+2} \cong \frac{s}{m} ,$$

praticamente uguale alla stima di massima verosimiglianza, s/m , del parametro Θ. Tale risultato appare abbastanza ragionevole in quanto è logico aspettarsi che la combinazione dell'opinione iniziale con il risultato dell'osservazione statistica realizzata dalla norma di coerenza debba praticamente ridursi all'informazione campionaria quando l'informazione iniziale è inesistente.

Peraltro, la nozione di distribuzione iniziale "non informativa" è soggetta a varie critiche; la principale è che ogni valutazione probabilistica, e quindi anche la densità iniziale di Θ, può essere basata su quantità di informazione molto diverse tra loro. Per un esempio elementare si pensi che la valutazione della probabilità di testa in un lancio di una moneta pari a 0,5 può essere giustificata sia dalla sola vaga opinione che la moneta in questione, vista per la prima volta, è apparentemente "regolare" come anche dall'esperienza che su moltissimi lanci eseguiti in precedenza con quella moneta si sono ottenute frequenze di testa e croce approssimativamente uguali. Ritornando alla densità iniziale uniforme di Θ, essa può rappresentare

sia un'informazione praticamente nulla che anche, invece, un'informazione ben fondata su dati dell'esperienza.

Vogliamo soltanto accennare ad una diversa impostazione del problema dell'apprendimento bayesiano in condizioni di scarsa informazione iniziale: quella che si basa sulla "sensitivity analysis", o "analisi di robustezza". Partendo dalla constatazione che la specificazione di una distribuzione iniziale è, in generale, un'operazione molto complicata (si ricordi che essa equivale, per numeri aleatori con un insieme limitato di determinazioni numeriche, alla scelta coerente di infiniti momenti) e addirittura impossibile quando l'informazione iniziale disponibile è scarsa, si suggerisce di individuare, al posto di un'unica distribuzione iniziale, una famiglia di distribuzioni, per esempio la famiglia B(H) di densità Beta con parametri verificanti le condizioni $\alpha' \le \alpha \le \alpha''$ e $\beta' \le \beta \le \beta''$. Applicando idealmente il teorema di Bayes a ciascuna densità di B(H) si perviene ad una nuova famiglia B(H∧K) di densità Beta i cui parametri verificano le $\alpha' + s \le \alpha^* \le \alpha'' + s$ e $\beta' + m - s \le \beta^* \le \beta'' + m - s$. La valutazione della probabilità di successo in una nuova prova non potrà ovviamente essere univoca: si può solo affermare che essa è compresa tra l'estremo inferiore e superiore del rapporto $\alpha^* / (\alpha^* + \beta^*)$.

Il più delle volte l'incertezza nella specificazione iniziale del modello statistico concerne più la distribuzione dei parametri incogniti presenti nella distribuzione campionaria delle variabili osservabili che non quest'ultima; infatti la distribuzione campionaria è spesso suggerita, almeno per grandi linee, dal problema studiato. Come si è già detto, quando l'informazione iniziale è scarsa, alla specificazione di un'unica distribuzione iniziale per i parametri incogniti della distribuzione campionaria si sostituisce quella di una famiglia iniziale B(H) di distribuzioni.

Riprendendo lo schema precedente, anche l'individuazione degli intervalli $\alpha' \le \alpha \le \alpha''$ e $\beta' \le \beta \le \beta''$ può non essere agevole in quanto i parametri α e β non hanno una relazione immediata con le variabili osservabili per cui mancano di un'interpretazione concreta. Per superare, almeno in parte questa difficoltà, conviene ricorrere ad una riparametrizzazione della densità Beta secondo la formulazione:

$$g(\theta) = k.\theta^{st-1}.(1-\theta)^{s(1-t)-1} ,$$

ove i nuovi parametri s e t sono legati ai precedenti dalle relazioni $s=\alpha+\beta$ e $t=\alpha/(\alpha+\beta)$. Risulta evidente che il parametro t coincide con il valor medio della distribuzione ed anche, per quanto si è già visto, con la probabilità di successo in una prova; ciò facilita notevolmente la scelta di un intervallo di valori per esso. Il parametro s è detto "parametro di apprendimento" e il suo significato sarà chiarito in seguito.

Si supponga per momento di poter scegliere, sulla base dell'informazione iniziale, un valore preciso, diciamo s_0 , per s , mentre per t si scelga l'intervallo (t'_0 , t''_0): si ha quindi B(H) = {densità Beta (s,t) ; $s = s_0$, $t'_0 \le t \le t''_0$ }. In corrispondenza all'osservazione di m eventi dei quali r affermati in un'ordine noto (evento K), l'applicazione del teorema di Bayes determina la trasformazione di B(H) nella famiglia di densità finali B(H∧K) = {densità Beta (s,t) ; $s= s_0+m$, $t'_m \le t \le t''_m$ }, ove è

$$t'_m = (s_0 t'_0 + r)/(s_0 + m) \qquad \text{e} \qquad t''_m = (s_0 t''_0 + r)/(s_0 + m) .$$

Se definiamo la differenza $t_0'' - t_0'$ = P"(E_j /H) – P'(E_j /H) , $j \ge 1$, "grado iniziale di imprecisione" concernente il verificarsi di un successo in una prova non ancora eseguita, la corrispondente differenza $t_m'' - t_m'$ rappresenta il "grado finale di imprecisione" e facilmente si constata che è:

$t''_m - t'_m$ = P”(E $_{m+j}$ / H∧K) – P’(E $_{m+j}$ / H∧K) = $\frac{s_0.(t''_0 - t'_0)}{s_0 + m}$, j ≥ 1 .

Ponendo s_0 = m si trova $t''_m - t'_m = (t''_0 - t'_0)/2$ per cui il parametro s può essere interpretato come “numerosità delle osservazioni statistiche atta a dimezzare il grado iniziale di imprecisione”: da ciò il nome di “parametro di apprendimento”.

Quando l’informazione iniziale è molto scarsa s_0 andrebbe fissato ad un livello abbastanza elevato; viceversa quando quell’informazione è abbastanza consistente. Si noti che il grado di imprecisione iniziale $t''_0 - t'_0$ può variare da zero a uno, a seconda che l’informazione iniziale sia massimale o minimale: infatti se l’informazione disponibile è molto ampia si riesce solitamente a scegliere un valore preciso per il parametro t, mentre in caso contrario si ammettono come suoi valori possibili tutti i numeri dell’intervallo [0 , 1]. Almeno in linea di principio, secondo questa impostazione, è possibile associare ad una valutazione probabilistica sugli eventi che interessano un indice della quantità di informazione su cui quella valutazione si basa: il grado di imprecisione.

4) Analisi statistica bayesiana su processi di conteggio

Consideriamo ora processi di osservazione concernenti sequenze di eventi E_j che si verificano in epoche aleatorie: può trattarsi di ricoveri in una struttura ospedaliera, degli arrivi di chiamate su una linea telefonica, di utenti ad uno sportello di un ufficio informazioni, di particelle rivelate da un contatore Geiger, di denunce di sinistro presso una impresa di assicurazioni, di successivi arresti per guasto di una macchina utensile e così via.

L’analisi probabilistica della sequenza di eventi $\{E_j; j \geq 1\}$ può essere condotta da due distinti punti di vista tra loro complementari: si può considerare il processo stocastico di conteggio a parametro continuo $\{N(t); t \geq 0\}$, ove $N(t)$ è il numero aleatorio di eventi arrivati nell’intervallo di tempo (0, t], oppure il processo stocastico a parametro discreto dei tempi di attesa tra arrivi successivi $\{T_j; j \geq 1\}$, ove T_j misura l’ampiezza dell’intervallo aleatorio di tempo che intercorre tra l’arrivo di E_{j-1} e di E_j. La relazione

intercorrente tra i due processi è rivelata dall'uguaglianza $P\{N(t) < k\} = P\{\sum_{j=1}^{k} T_j > t\}$ dalla quale discende il seguente sviluppo:

$$P\{N(t) = k\} = P\{N(t) < k+1\} - P\{N(t) < k\} = P\{\sum_{j=1}^{k+1} T_j > t\} - P\{\sum_{j=1}^{k} T_j > t\} = F_k(t) - F_{k+1}(t) \ ,$$

ove $F_k(t)$ denota la funzione di ripartizione del n.a. $\sum_{j=1}^{k} T_j = S_k$ che misura il tempo di attesa per E_k.

Una notevole semplificazione della struttura stocastica di tali processi si ottiene se $\{T_j ; j \geq 1\}$ ha n.a. i.i.d.: in tal caso il processo $\{S_k ; k \geq 1\}$ è un "processo di rinnovamento". Il più noto dei processi di rinnovamento è quello in cui i n.a. T_j hanno una comune distribuzione esponenziale negativa, con densità $f(t) = \lambda.\exp(-\lambda.t)$. Il corrispondente processo di conteggio $\{N(t); t \geq 0\}$ è il ben noto processo di Poisson (si veda l'Esempio 2.7) come si può provare utilizzando la relazione $P\{N(t) = k\} = F_k(t) - F_{k+1}(t)$ e ricordando che se T_j ha densità $f(t) = \lambda.\exp(-\lambda.t)$ allora S_k ha densità $F_k^{'}(t) = \frac{\lambda^k}{\Gamma(k)} t^{k-1}.e^{-\lambda.t}$, di tipo Gamma (λ, k).

Nei successivi procedimenti inferenziali bayesiani assumeremo che il processo di osservazione sia costituito:

a) da un processo di conteggio di tipo poissoniano con intensità Θ non nota del quale sono osservabili gli incrementi N_h riguardanti intervalli di tempo unitari successivi;

b) da un processo dei tempi di attesa T_j tra arrivi successivi ove i n.a. sono assunti essere, per ogni possibile valore θ dell'intensità non nota Θ, i.i.d. con la comune densità di probabilità condizionata $f(t/\theta) = \theta.\exp(-\theta.t)$.

In entrambi i casi abbiamo dunque processi di osservazione a parametro discreto e scambiabili in quanto i n.a. osservabili N_h e, rispettivamente, T_j sono assunti essere i.i.d. subordinatamente ad ogni ipotesi {Θ = θ} concernente l'intensità aleatoria. I modelli statistici parametrici che discendono dalle precedenti ipotesi e che riguardano m' e, rispettivamente, m" osservazioni sono i seguenti:

a) per il processo di conteggio: $P\{ \bigcap_{h=1}^{m'} (N_h = n_h) / \theta \} = \frac{\theta^{\sum_h n_h}.e^{-m'.\theta}}{\prod_h (n_h!)}$,

b) per il processo dei tempi di attesa: $f(t_1, \ldots, t_{m"} / \theta) = \theta^{m"}.e^{-\theta.\sum_j t_j}$,

che si riferiscono all'evento $K_1 = \bigcap_{h=1}^{m'} (N_h = n_h)$ e, rispettivamente, all'evento $K_2 = \bigcap_{j=1}^{m"} (T_j = t_j)$.

E' facile ricavare le stime di massima verosimiglianza per θ corrispondenti agli eventi osservabili K_1 e K_2 ; si ottengono precisamente i valori $\hat{\theta} = \frac{\sum n_h}{m'}$ e, rispettivamente, $\hat{\theta} = \frac{m''}{\sum t_h}$.

Se, nell'approccio bayesiano dell'inferenza statistica su Θ per entrambi i modelli statistici assumiamo che la distribuzione iniziale $G_0(\theta)$ sia di tipo Gamma (α, β) con densità $g_0(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}.\theta^{\alpha-1}.e^{-\beta\theta}$, allora a tali modelli parametrici corrispondono le seguenti leggi temporali iniziali:

a) $P\{\bigcap_{h=1}^{m'}(N_h = n_h)\} = \frac{\Gamma(\alpha+\sum n_h)}{\Gamma(\alpha).[\prod(n_h!)]}.\frac{\beta^\alpha}{(\beta+m')^{\alpha+\sum n_h}}$, $m' \geq 1$;

b) $f(t_1, \ldots\ldots, t_{m''}) = \frac{\Gamma(\alpha+m'')}{\Gamma(\alpha)}.\frac{\beta^\alpha}{(\beta+\sum t_j)^{\alpha+m''}}$, $m'' \geq 1$.

Inferenza statistica sul processo { N_h ; $h \geq 1$}

Supporremo ora di voler migliorare la conoscenza iniziale del processo scambiabile degli arrivi mediante l'osservazione dei valori dei primi m' numeri aleatori osservabili N_h , cioè mediante l'osservazione dell'evento $K_1 = \bigcap_{h=1}^{m'}(N_h = n_h)$. Si è già visto che la funzione di verosimiglianza corrispondente è data dalla

$\ell(\theta / K_1) \propto P\{\bigcap_{h=1}^{m'}(N_h = n_h)/\theta\} = \frac{\theta^{\sum_h n_h}.e^{-m'.\theta}}{\prod_h(n_h!)}$ il cui nucleo, cioè la parte dipendente da θ, è data dal numeratore $\theta^{\sum n_h}.e^{-m'.\theta}$.

Secondo il procedimento inferenziale che abbiamo denominato "indiretto", l'incremento di informazione K_1 , attraverso la corrispondente funzione di verosimiglianza, viene utilizzato per adeguare la distribuzione iniziale $G_0(\theta)$ del parametro incognito. L'applicazione del teorema di Bayes fornisce la densità finale il cui nucleo è dato dalla:

$$g(\theta / K_1) \propto g_0(\theta) \,.\, \ell(\theta / K_1) \propto \theta^{(\alpha+\sum n_h)-1}.e^{-\theta(\beta+m')} \ .$$

La densità $g_0(\theta)$ è stata trasformata nella densità $g(\theta / K_1)$ che corrisponde ad una distribuzione Gamma (α^*, β^*) con parametri $\alpha^* = \alpha + \sum n_h$ e $\beta^* = \beta + m'$; si noti che il parametro iniziale α si combina additivamente col numero totale degli arrivi osservato, mentre β si combina con il numero m' dei periodi unitari di osservazione. Chiaramente, la coppia $(m', \sum_h n_h)$ costituisce la statistica sufficiente, o riassunto esaustivo, per la sequenza osservata $(n_1, \ldots\ldots, n_{m'})$.

A differenza dell'approccio non bayesiano che ricava dall'osservazione dell'evento K_1 la stima di massima verosimiglianza $\hat{\theta}_1 = \frac{\sum n_h}{m'}$, l'approccio bayesiano (indiretto) ricava la distribuzione finale dell'intensità incognita $g(\theta / K_1)$. E' detta "stima bayesiana per θ" un qualunque indice di posizione di questa distribuzione, anche se gli indici più usati sono il valor medio o il valore modale o la mediana. Scegliendo come stima bayesiana il valor medio della distribuzione finale e indicando per comodità $\sum_{h=1}^{m'} n_h = n$ si ha :

$$E(\Theta / K_1) = \frac{\alpha + n}{\beta + m'} = \frac{\beta.\frac{\alpha}{\beta} + m'.\frac{n}{m'}}{\beta + m'} ,$$

e dall'ultimo rapporto si osserva che $E(\Theta / K_1)$ è espresso da una combinazione lineare convessa del valor medio iniziale $E(\Theta) = \alpha / \beta$ e della stima di massima verosimiglianza n / m' . Al crescere del numero m' dei periodi di osservazione, cioè al crescere dell'informazione campionaria, la stima bayesiana tende a coincidere con quella di massima verosimiglianza.

Se fossimo interessati a fare una qualunque previsione sui valori delle variabili non ancora osservate N_h , $h > m'$, potremmo ricorrere al teorema di rappresentazione per processi scambiabili al modo seguente. Limitando la previsione alla sequenza $N_{m'+1}, \ldots\ldots\ldots, N_{m'+k}$ e indicando con $(\upsilon_1, \ldots\ldots, \upsilon_k)$ un insieme arbitrario dei loro possibili valori si ha:

$$P\{\bigcap_{i=1}^{k}(N_{m'+i} = \upsilon_i) / K_1\} = \int_{\{\theta\}}\left[\prod_{i=1}^{k} P[(N_{m'+i} = \upsilon_i)/\theta]\right].g(\theta / K_1)d\theta =$$

$$= \frac{\Gamma(\alpha + \sum n_h + \sum \upsilon_i)}{\Gamma(\alpha + \sum n_h).[\prod(\upsilon_i!)]}.\frac{(\beta + m')^{\alpha+\sum n_h}}{(\beta + m' + k)^{\alpha+\sum n_h+\sum \upsilon_i}} .$$

Il procedimento inferenziale diretto conduce a questa espressione della distribuzione previsionale per le variabili $N_{m'+1}, \ldots\ldots\ldots, N_{m'+k}$ mediante l'applicazione del teorema delle probabilità composte alle distribuzioni della legge temporale iniziale $\Lambda(H)$

$$P\{\bigcap_{h=1}^{m'}(N_h = n_h) / H\} = \frac{\Gamma(\alpha + \sum n_h)}{\Gamma(\alpha).[\prod(n_h!)]}.\frac{\beta^{\alpha}}{(\beta + m')^{\alpha+\sum n_h}} \quad , \quad m' \geq 1 .$$

Infatti si ha, come facilmente si verifica :

$$P\{\bigcap_{i=1}^{k}(N_{m'+i} = \upsilon_i) / H \cap K_1\} = \frac{P\left\{K_1 \cap [\bigcap_{i=1}^{k}(N_{m'+i} = \upsilon_i)] / H\right\}}{P\{K_1 / H\}} =$$

$$= \frac{\Gamma(\alpha + \sum n_h + \sum \upsilon_i)}{\Gamma(\alpha + \sum n_h).[\prod(\upsilon_i!)]}.\frac{(\beta + m')^{\alpha+\sum n_h}}{(\beta + m' + k)^{\alpha+\sum n_h+\sum \upsilon_i}} .$$

Inferenza statistica sul processo { T_j ; j ≥ 1}.

Ora l'incremento di informazione sul processo degli arrivi sarà costituito dall'evento $K_2 = \bigcap_{j=1}^{m''} (T_j = t_j)$ riguardante l'osservazione di m" tempi di attesa tra arrivi successivi T_j : in tal caso si osservano m" eventi per un tempo totale pari a $\sum_{j=1}^{m''} t_j$ = t . La corrispondente funzione di verosimiglianza è espressa dalla $\ell(\theta / K_2) \propto f(t_1, \ldots\ldots, t_{m''} / \theta) = \theta^{m''}.\exp\left(-\theta.\sum_{j=1}^{m''} t_j\right)$ cosicché la statistica sufficiente è costituita dalla coppia (m", t).

L'adeguamento della distribuzione iniziale G (θ) all'incremento di informazione K_2 si ottiene dalla:

$$g(\theta / K_2) \propto g_0(\theta) \,.\, \ell(\theta / K_2) \propto \theta^{\alpha+m''-1}.e^{-\theta(\beta+t)} \,,$$

in base alla quale si ha che la densità finale è di tipo Gamma (α**,β**) ove α** = α + m" e β** = β + t . Assumendo quale stima bayesiana per θ il valor medio della densità finale è :

$$E(\Theta / K_2) = \frac{\alpha + m''}{\beta + t} = \frac{\beta.\frac{\alpha}{\beta} + t.\frac{m''}{t}}{\beta + t} \,,$$

ove l'ultimo rapporto è ancora costituito da una combinazione lineare convessa del valor medio della distribuzione iniziale α / β e della stima di massima verosimiglianza m"/t. Al crescere del tempo totale di osservazione t = $\sum_{j=1}^{m''} t_j$ la stima bayesiana tende a coincidere con m" / t , in accordo con il prevalere dell'informazione campionaria sull'informazione iniziale.

Per costruire una previsione sui successivi k tempi tra arrivi successivi $T_{m''+1}, \ldots\ldots\ldots, T_{m''+k}$ non ancora osservati, il procedimento è oramai noto: la densità previsionale congiunta per le suddette variabili si ottiene applicando il teorema di rappresentazione al modo seguente:

$$f(\tau_1, \ldots\ldots\ldots, \tau_k / K_2) = \int_{\{\theta\}} \left[\prod_{i=1}^{k} f(\tau_i / \theta)\right].g(\theta / K_2)d\theta = \frac{\Gamma(\alpha + m'' + k)}{\Gamma(\alpha + m'')}.\frac{(\beta + t)^{\alpha+m''}}{(\beta + t + \sum \tau_i)^{\alpha+m''+k}} \,.$$

Il procedimento inferenziale diretto conduce a questa espressione della distribuzione previsionale per le variabili $T_{m''+1}, \ldots\ldots\ldots, T_{m''+k}$ mediante l'applicazione del teorema delle probabilità composte alle distribuzioni della legge temporale iniziale Λ(H)

$$f(t_1, \ldots\ldots, t_{m''} / H) = \frac{\Gamma(\alpha + m'')}{\Gamma(\alpha)}.\frac{\beta^{\alpha}}{(\beta + \sum t_j)^{\alpha+m''}} \quad , \quad m'' \geq 1 \,.$$

Si ha dunque con facili calcoli:

$f(\tau_1,.....,\tau_k / H \bigcap K_2) = \frac{f(t_1,.....,t_{m''};\tau_1.....,\tau_k / H)}{f(t_1,.....,t_{m''} / H)} = \frac{\Gamma(\alpha+m''+k)}{\Gamma(\alpha+m'')} \cdot \frac{(\beta+t)^{\alpha+m''}}{(\beta+t+\sum\tau_i)^{\alpha+m''+k}}$.

Si osservi infine che, con riferimento al processo di osservazione residuo { N_h ; $h \geq m'+1$} e secondo l'impostazione classica, dopo l'utilizzazione dell'informazione statistica K_1 il processo è considerato essere un processo di Poisson con intensità $\hat{\theta} = \frac{\sum n_h}{m'}$. Nell'impostazione bayesiana, dopo l'utilizzazione di K_1, il processo { N_h ; $h \geq m'+1$} è invece una mistura di processi poissoniani caratterizzata dalla funzione peso g (θ / K_1). Ovviamente, la stessa considerazione riguarda anche il processo dei tempi di attesa residuo { T_j ; $j \geq m''+1$}, la stima di massima verosimiglianza $\hat{\theta} = \frac{m''}{t}$ e la distribuzione finale g (θ / K_2) .

IL PROCESSO DI POISSON COMPOSTO

Occupiamoci ancora di un processo di conteggio, quale quello appena discusso, con l'ulteriore complicazione costituita dal fatto che ad ogni evento E_j viene associato il valore Y_j di un "effetto aleatorio osservabile": si pensi per esempio di associare ad ogni comunicazione telefonica sulla stessa linea la sua durata o di associare ad ogni denuncia di sinistro presso un'impresa di assicurazione il valore economico del danno e così via. L'intendimento è quello di studiare l'andamento nel tempo dell'effetto complessivo: per le comunicazioni telefoniche si studia l'andamento della durata di occupazione complessiva della linea e per le denunce di sinistro si studia l'andamento del danno economico complessivo che l'impresa assicurativa deve rifondere.

Per questi problemi il modello stocastico più usato è quello che concerne **somme di un numero finito, ma non noto (e quindi aleatorio), di variabili aleatorie indipendenti e ugualmente distribuite.** Se indichiamo con N il numero aleatorio di eventi verificatisi in un intervallo di tempo unitario e con Y_j l'effetto aleatorio associato ad E_j, l'effetto complessivo è dato da $\sum_{j=0}^{N} Y_j$, ove nella somma si è inserito Y_0 = 0 per tener conto della possibilità che nell'intervallo temporale considerato non si verifichi alcun evento. Come già detto, gli effetti aleatori Y_j sono spesso assunti essere i.i.d.: più precisamente, assumeremo che condizionatamente ad ogni evento (N = n) i n.a. $Y_1,.....,Y_n$ siano mutuamente indipendenti e dotati di una comune distribuzione di probabilità non dipendente dall'intero n. Posto X = $\sum_{j=0}^{N} Y_j$, interessa conoscere la distribuzione dell'effetto complessivo X. Se in queste ipotesi indichiamo con {p_n ; $n \geq 0$} la distribuzione di

probabilità di N e con $F_Y(y)$ la funzione di ripartizione comune ai n.a. Y_j si ha che la distribuzione di X ha una funzione di ripartizione data dalla

$$\text{Prob}\{X \le x\} = F_X(x) = \sum_{n\ge 0} P(N=n).P(X \le x / N=n) = \sum_{n\ge 0} p_n.F_Y^{n*}(x) ,$$

ove il simbolo $F_Y^{n*}(x)$ indica la f.d.r. della distribuzione della somma $\sum_{j=0}^{n} Y_j$ di n addendi (poiché per ipotesi è $Y_0 = 0$) i.i.d con f.d.r. comune $F_Y(y)$. In termini matematici $F_Y^{n*}(x)$ è detta "convoluzione n-ma di $F_Y(y)$ " e, com'è noto, la sua funzione caratteristica è data dalla potenza n-ma della funzione caratteristica $\varphi_Y(\xi)$ di $F_Y(y)$.

Se invece di considerare un solo periodo di tempo unitario si considera una sequenza di periodi unitari ed N_t indica il numero aleatorio di eventi verificatisi nel periodo t-esimo, allora indicheremo con X_t l'effetto complessivo nel periodo t-esimo e sarà: $X_t = \sum_{j=0}^{N_t} Y_{tj}$, ove Y_{tj} indica l'effetto associato all'evento j-esimo nel periodo t-esimo. In questo schema intervengono due distinti processi stocastici: il processo di conteggio $\{N_t ; t \ge 1\}$ ed il processo degli effetti $\{Y_j ; j \ge 0\}$;

di norma, entrambi i processi sono supposti costituiti da numeri aleatori indipendenti e ugualmente distribuiti e si suppone anche che ogni elemento di uno dei due processi sia indipendente da tutti gli elementi dell'altro processo. Se si suppone che la comune distribuzione degli N_t, $\{p_n ; n \ge 0\}$, sia la distribuzione di Poisson, cioè se $p_n = \frac{\lambda^n.e^{-\lambda}}{n!}$, allora il processo $\{X_t ; t \ge 1\}$ è denominato "processo di Poisson composto".

Se inoltre si assume che la comune distribuzione degli Y_j sia di tipo Gamma (1, β), cioè una distribuzione di tipo "esponenziale negativo" con densità $f(y) = \beta.\exp(-\beta.y)$, in corrispondenza alla quale la convoluzione n-ma ha una densità di tipo Gamma (n , β), si ha:

$$F_{X_t}(x) = \sum_{n\ge 0} p_n.F_Y^{n*}(x) = \sum_{n\ge 0}\left(\frac{e^{-\lambda}.\lambda^n}{n!}\right).\int_0^x \frac{\beta^n}{\Gamma(n)} z^{n-1}.e^{-\beta.z}dz \text{ , ed anche } f_{X_t}(x) = \sum_{n\ge 0}\left(\frac{e^{-\lambda}.\lambda^n}{n!}\right).\left(\frac{\beta^n}{\Gamma(n)} x^{n-1}.e^{-\beta.x}\right).$$

Si ricavano facilmente, per i primi due momenti di X_t, i seguenti risultati :

$$E(X_t) = E(N_t).E(Y_j) = \lambda / \beta , \quad E(X_t^2) = E(N_t).\text{Var}(Y_j) + E(N_t^2).E^2(Y_j) = \lambda.(\lambda+2) / \beta^2 ,$$

dai quali si ottiene immediatamente Var (X_t) = $2\lambda / \beta^2$.

Supporremo ora di avere soltanto una conoscenza incerta dei valori di λ e β, che indicheremo con i simboli θ_1 e, rispettivamente, θ_2 ed esprimeremo le informazioni disponibili su essi mediante una distribuzione congiunta iniziale g(θ_1, θ_2 / H). Se **per ogni possibile ipotesi sui valori dei parametri aleatori Θ_1 e Θ_2** si assumono ancora le precedenti condizioni di indipendenza e uguale distribuzione per i numeri aleatori di ciascuno dei due processi {N_t} e {Y_j} e quella di mutua indipendenza tra di essi si ottiene un modello statistico per {X_t ; $t \geq 1$}, costituito dalla "composizione" dei due processi scambiabili {N_t} e {Y_j}, che rende formalmente semplice l'inferenza bayesiana su Θ_1 e Θ_2 .

Immaginiamo di poter osservare in m periodi di tempo consecutivi i numeri di arrivi n_i , $1 \leq i \leq m$, ed i valori y_j degli effetti associati a tutti gli arrivi osservati. Se, per esempio, nel periodo k-mo vengono osservati n_k arrivi ed i corrispondenti effetti ($y_1, \ldots\ldots, y_{n_k}$), la relativa verosimiglianza per Θ_1 e Θ_2 è data dalla:

$$\ell(\theta_1, \theta_2 / n_k ; y_1, \ldots\ldots, y_{n_k}) = \ell(\theta_1 / n_k).\ell(\theta_2 / y_1, \ldots\ldots, y_{n_k}) \propto (\theta_1.\theta_2)^{n_k}.\exp\{-\theta_1 - \theta_2.x_k\} ,$$

ove si è posto $x_k = \sum_{j=1}^{n_k} y_j$. La verosimiglianza corrispondente all'informazione K raccolta in tutti gli m periodi di osservazione è allora data dalla:

$$\ell(\theta_1,\theta_2 / K) = \prod_{k=1}^{m} \ell(\theta_1,\theta_2 / n_k ; y_1, \ldots\ldots, y_{n_k}) \propto (\theta_1.\theta_2)^{n}.\exp\{-m.\theta_1 - \theta_2.x\} ,$$

ove $x = \sum_{k=1}^{m} x_k$ ed $n = \sum_{k=1}^{m} n_k$, e le stime di massima verosimiglianza per i due parametri incogniti sono, come si verifica facilmente, $\hat{\theta}_1 = n/m$ e $\hat{\theta}_2 = n/x$.

Tenendo conto, oltre che della funzione di verosimiglianza, anche delle informazioni ed opinioni iniziali sui parametri espresse dalla densità iniziale g (θ_1,θ_2 / H) si può ottenere, mediante il teorema di Bayes, la corrispondente densità finale.

Assumeremo che sia g (θ_1,θ_2 / H) = $g_1(\theta_1$/ H). $g_2(\theta_2$/ H) $\propto \left(\theta_1^{\,a-1}.e^{-c.\theta_1}\right) . \left(\theta_2^{\,b-1}.e^{-d.\theta_2}\right)$ per cui la densità finale è :

$$g(\theta_1,\theta_2 / H \wedge K) \propto g(\theta_1,\theta_2 / H) . \ell(\theta_1,\theta_2 / K) \propto \left(\theta_1^{\,a+n-1}.e^{-\theta_1.(m+c)}\right) . \left(\theta_2^{\,b+n-1}.e^{-\theta_2(d+x)}\right)$$

e quindi ancora un prodotto di densità di tipo Gamma, conservandosi così l'indipendenza tra Θ_1 e Θ_2. Tale risultato implica che le "stime bayesiane" finali per i parametri incogniti siano

$E(\Theta_1 / H \wedge K) = (a+n)/(m+c)$ e $E(\Theta_2 / H \wedge K) = (b+n)/(d+x)$.

In molte situazioni concrete può essere plausibile assumere che Θ_1 e Θ_2 siano stocasticamente dipendenti, per cui dovrà essere $g(\theta_1,\theta_2 / H) \neq g_1(\theta_1 / H).\ g_2(\theta_2 / H)$. Onde garantire sviluppi formali non troppo complessi, scelte adeguate per $g(\theta_1,\theta_2 / H)$ potrebbero essere una distribuzione Gamma bivariata oppure una combinazione lineare convessa finita di prodotti di Gamma univariate del tipo $\sum_j u_j .Gamma(\theta_1 / a_j, c_j).Gamma(\theta_2 / b_j, d_j)$.

5) Analisi Bayesiana su processi Gaussiani di misurazione

Consideriamo il problema di stima di una grandezza fisica incognita Θ, per esempio una temperatura o una lunghezza o altro, che può venire misurata un numero indefinito di volte con uno strumento di misura che introduce soltanto errori accidentali. Ricordiamo che sono detti accidentali quegli errori di misura, non eliminabili, dovuti a moltissimi fattori indipendenti tra loro, ciascuno dei quali è ritenuto trascurabile, ma che nel loro insieme possono dar luogo a scostamenti sensibili del valore misurato dal valore effettivo della grandezza esaminata. In Teoria degli errori, applicando il teorema centrale limite, si stabilisce che gli errori accidentali di misura abbiano una distribuzione di probabilità di tipo gaussiano con valor medio nullo e varianza pari al reciproco della precisione dello strumento di misura.

Assumeremo che le ripetute misurazioni di Θ avvengano nelle stesse condizioni generali cosicché sia plausibile assumere scambiabile il processo di osservazione $\{X_t ; t \geq 1\}$ o, equivalentemente, che i risultati delle misurazioni X_t siano condizionatamente indipendenti e abbiano la stessa distribuzione condizionata rispetto ad ogni possibile valore θ di Θ. Per i motivi su esposti assumeremo anche che sia $F_{X_t}(x/\theta) \sim N(\theta, \sigma^2)$ ove σ^2 è il reciproco della precisione nota dello strumento di misura.

Il modello statistico parametrico che si è assunto risulta dunque costituito da una famiglia di densità campionarie $\{F_{1,\ldots\ldots,n}(\mathbf{x} / \theta) ; n \geq 1\}$ di tipo normale, ciascuna delle quali è caratterizzata dal vettore medio $\theta .\mathbf{1}$ e dalla matrice di dispersione $\sigma^2 . I_n$.

Ne discende che la funzione di verosimiglianza di Θ relativa ai risultati $\mathbf{x} = (x_1,\ldots\ldots,x_n)'$ di n misurazioni successive è espressa dalla:

$$\ell(\theta / x_1,\ldots\ldots,x_n) \propto f(x_1,\ldots\ldots,x_n / \theta) = \prod_{i=1}^{n} f(x_i / \theta) \propto \exp\left\{-n.(\theta - \bar{x})^2 / 2.\sigma^2\right\},$$

che fornisce la statistica sufficiente (n, $\bar{x}$) e la stima di massima verosimiglianza $\hat{\theta} = \frac{1}{n}.\sum_{i=1}^{n} x_i = \bar{x}$.

Da un punto di vista bayesiano occorre completare il modello statistico assunto esprimendo lo stato di informazione iniziale mediante una distribuzione probabilistica su Θ; supponiamo allora di scegliere una distribuzione normale $g_0(\theta)$ con parametri μ_0 e σ_0^2 fissati. E' facile verificare che la legge temporale iniziale $\Lambda(H)$ per il processo $\{ X_t ; t \geq 1\}$ è costituita da distribuzioni congiunte di tipo normale caratterizzate dal comune valor medio μ_0, dalla comune varianza $\sigma_0^2 + \sigma^2$ e dalla covarianza σ_0^2 comune a tutte le coppie di variabili osservabili tra loro diverse.

Se l'incremento di informazione è costituito dall'evento $K = \bigcap_{t=1}^{m} (X_t = x_t)$ l'inferenza bayesiana su Θ si realizza, nel procedimento indiretto, con l'adeguamento della distribuzione $g_0(\theta)$ mediante l'applicazione del teorema di Bayes:

$$g(\theta / K) \propto g_0(\theta) \,.\, \ell(\theta / x_1, ..., x_m) \propto \exp\{-(\theta - \mu_m)^2 / 2.\sigma_m^2\},$$

ove $$\mu_m = \frac{m.\sigma^{-2}.\bar{x} + \sigma_0^{-2}.\mu_0}{m.\sigma^{-2} + \sigma_0^{-2}} \quad ; \quad \sigma_m^2 = \frac{1}{m.\sigma^{-2} + \sigma_0^{-2}}.$$

La densità finale $g(\theta / K)$ è dunque Gaussiana con parametri μ_m e σ_m^2 e, come già osservato in precedenza, il valor medio finale μ_m di Θ è una combinazione lineare convessa del valor medio iniziale μ_0 e della stima di massima verosimiglianza $\bar{x}$; tale combinazione può essere interpretata come una stima bayesiana per Θ nello stato di informazione finale (rispetto a K) e tende a coincidere con la stima di massima verosimiglianza al crescere del numero n delle misurazioni eseguite. Si noti anche che la varianza della densità finale σ_m^2 decresce in modo monotono al crescere di m e non dipende dai risultati delle misurazioni: è bene sottolineare che quest'ultima caratteristica non ha carattere generale, ma dipende dal particolare modello statistico scelto e precisamente dalla normalità delle distribuzioni campionarie $f(x_1, ..., x_n / \theta)$ e della distribuzione iniziale $g_0(\theta)$.

Poiché la densità finale di Θ differisce da quella iniziale soltanto per i valori dei due parametri, si intuisce, ma lo si verifica facilmente, che la legge temporale finale $\Lambda(H \wedge K)$ per il processo di misurazioni residuo $\{X_t ; t \geq m+1\}$ è costituita da distribuzioni congiunte di tipo normale caratterizzate dal comune valor medio μ_m, dalla comune varianza $\sigma_m^2 + \sigma^2$ e dalla comune covarianza σ_m^2 per coppie di variabili X_{m+t}, X_{m+s} con $t \neq s$.

Con la precedente osservazione abbiamo sostanzialmente riassunto il procedimento inferenziale diretto che ricava le distribuzioni di $\Lambda(H \wedge K)$ mediante l'applicazione del teorema delle probabilità composte alle densità della famiglia $\Lambda(H)$ in accordo con la:

$$f(\upsilon_{m+1}, ..., \upsilon_{m+k} / H \cap K) = \frac{f(x_1, ..., x_m, \upsilon_{m+1}, ..., \upsilon_{m+k} / H)}{f(x_1, ..., x_m / H)},$$

avendo indicato con $\upsilon_{m+1},.....,\upsilon_{m+k}$ gli argomenti della densità congiunta delle prime k variabili osservabili di $\{X_t ; t \geq m+1\}$ appartenente a $\Lambda(H \wedge K)$.

La trasformazione $\Lambda(H) \rightarrow \Lambda(H \wedge K)$, sia nell'approccio diretto che indiretto, può presentare notevoli complicazioni se uno degli obiettivi dell'analisi è la semplicità degli sviluppi formali. Finora, negli esempi di inferenza bayesiana presentati, la combinazione coerente dell'informazione iniziale e dell'informazione campionaria è stata effettuata senza alcuna complicazione grazie alla particolare scelta della forma funzionale delle distribuzioni condizionate $F(x_1,.....,x_m / \theta)$ e di quelle riguardanti i parametri incogniti $G_0(\theta)$: accadeva infatti, nei modelli statistici finora proposti (Bernoulli – Beta, Poisson – Gamma, Normale – Normale) che la distribuzione finale $G(\theta / H \bigcap K)$ avesse la stessa forma funzionale di $G_0(\theta)$ e ciò implicava il mantenimento della stessa forma funzionale per le distribuzioni congiunte di $\Lambda(H)$ e di $\Lambda(H \wedge K)$.

Se invece le forme funzionali di $F(x_1,.....,x_m / \theta)$ e di $G_0(\theta)$ fossero state scelte in modi diversi da quelli indicati le distribuzioni della legge temporale finale $\Lambda(H \wedge K)$ avrebbero dovuto essere determinate per via numerica. Si consideri, per esempio, l'ultimo problema proposto nel paragrafo precedente concernente densità campionarie $\{f_{1,.....,n}$ (**x** / θ); $n \geq 1\}$ di tipo Gaussiano, caratterizzate dal vettore medio incognito θ .**1** e dalla matrice di dispersione nota σ^2 .I_n ; se, nell'ipotesi che la grandezza incognita Θ fosse positiva come nel caso di una lunghezza, avessimo scelto di esprimere le informazioni iniziali su essa mediante una densità di tipo Gamma (α, β), $g_0(\theta) = \frac{\beta^\alpha}{\Gamma(\alpha)}.\theta^{\alpha-1}.e^{-\beta\theta}$, l'applicazione del teorema di Bayes

$$g(\theta / K) \propto g_0(\theta) \,.\, \ell(\theta / x_1,.....,x_m) \propto [\theta^{\alpha-1}.e^{-\beta\theta}] \,.\, \exp\left\{-n.(\theta - \bar{x})^2 / 2.\sigma^2\right\} =$$

$$= \theta^{\alpha-1}.\exp\{-[a.\theta^2 + b.\theta + c]\} ,$$

ove i coefficienti a, b, c si calcolano immediatamente, non condurrebbe ad alcuna distribuzione di tipo standard per cui l'unica via praticabile per la determinazione di $g(\theta / K)$ sarebbe l'integrazione numerica.

Questo tipo di difficoltà spiega l'attenzione dedicata in passato dagli statistici bayesiani all'individuazione di schemi o modelli statistici parametrici formalmente praticabili oltre che adeguati a rappresentare in modo sufficientemente plausibile concrete situazioni inferenziali. Più recentemente, questi schemi statistici standard hanno perduto importanza con l'avvento di potenti tecniche di integrazione numerica (metodi Monte Carlo, Gibbs sampling, algoritmo EM, etc.) facilmente implementabili su PC. Per tali motivi ci limiteremo a fornire sui modelli standard soltanto alcune nozioni di sicuro interesse concettuale senza però approfondirne la trattazione, poiché i riferimenti in letteratura sono molti e molto validi.

6) Alcune nozioni importanti

La famiglia esponenziale di distribuzioni

E' nota con questo nome una famiglia di distribuzioni di probabilità per le quali la densità (nel caso di n.a. con un insieme continuo di valori) o la probabilità (nel caso di n.a. con un insieme discreto di valori) sono date dall'espressione:

$$\left.\begin{matrix} f_X(x/\theta_1,.....,\theta_k) \\ P(X=x/\theta_1,.....,\theta_k) \end{matrix}\right\} = B(\theta_1,.....,\theta_k).h(x).\exp\left\{\sum_{j=1}^{k} Q_j(\theta_1,.....,\theta_k).U_j(x)\right\},$$

ove le funzioni B(.), h(.), $Q_j(.)$, $U_j(.)$ possono essere qualsiasi. Appartengono a tale famiglia le note distribuzioni binomiale, geometrica, poissoniana, normale, beta, gamma ed altre; non vi appartengono le distribuzioni uniforme sull'intervallo [0, θ], binomiale negativa, di Cauchy e altre.

In ipotesi di scambiabilità per il processo osservabile, la verosimiglianza dei parametri θ_j relativa ad m osservazioni ha l'espressione:

$$\ell(\theta_1,.....,\theta_k / x_1,.....,x_m) = \prod_{i=1}^{m} \ell(\theta_1,.....,\theta_k / x_i) =$$

$$= B^m(\theta_1,.....,\theta_k).\left[\prod_{i=1}^{m} h(x_i)\right].\exp\left\{\sum_{j=1}^{k} Q_j(\theta_1,.....,\theta_k).\sum_{i=1}^{m} U_j(x_i)\right\}$$

.

Il riassunto esaustivo

Si pensi di associare al processo di osservazione $\{X_j; j \geq 1\}$ un secondo processo stocastico $\{T_n; n \geq 1\}$ ove $T_n = t_n(X_1,.....,X_n)$ è una funzione scalare o vettoriale della sequenza di n.a. $X_1,.....,X_n$; l'ente aleatorio T_n è detto "riassunto campionario" se esso risulta "meno informativo" della sequenza $X_1,.....,X_n$ rispetto agli elementi incogniti del modello statistico. Presentiamo alcuni esempi:

a) $T_n = t_n(X_1,.....,X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$;

b) $T_n = t_n(X_1,.....,X_n) = \max_{1\leq i\leq n} X_i - \min_{1\leq i\leq n} X_i$;

c) $T_n = t_n(X_1,.....,X_n) = \left(n, \sum_{i=1}^{n} X_i, \sum_{i=1}^{n} X_i^2\right)$;

d) $T_n = t_n(X_1,.....,X_n) = \left(X_{(1)},.....,X_{(n)}\right)$, ove $X_{(j-1)} \leq X_{(j)}$ per ogni j = 1,……,n .

Evidentemente, l'uso di riassunti campionari anziché delle sequenze originarie $X_1,.....,X_n$ comporta un'utile semplificazione degli sviluppi formali purchè tale sostituzione non provochi una perdita di informazioni rilevanti per i nostri obiettivi: è questo il "problema dell'esaustività (o sufficienza) dei riassunti campionari".

Presentiamo ora la definizione di riassunto campionario esaustivo (o sufficiente) dal punto di vista dell'approccio bayesiano indiretto: il processo stocastico $\{T_n; n \geq 1\}$ è un riassunto esaustivo se per ogni $n \geq 1$ ed ogni sequenza osservabile $x_1,.....,x_n$ accade che sia verificata l'uguaglianza delle distribuzioni finali

$$G[\theta / H \bigcap (x_1,.....,x_n)] = G[\theta / H \bigcap t_n(x_1,.....,x_n)],$$

in corrispondenza ad ogni possibile specificazione della distribuzione iniziale $G(\theta / H)$.

Proveremo che in ipotesi di scambiabilità per il processo di osservazione $\{X_j; j \geq 1\}$ la definizione bayesiana di esaustività ora enunciata è equivalente alla classica definizione di R.A. Fischer secondo la quale $\{T_n; n \geq 1\}$ è esaustivo per il parametro incognito Θ della distribuzione campionaria $F(x_1,.....,x_n / \theta)$ se la distribuzione condizionata $F[x_1,.....,x_n / t_n(x_1,.....,x_n), \theta]$ non dipende da θ. Precisamente dimostreremo che la ben nota "condizione di fattorizzazione della verosimiglianza" di J. Neyman, che com'è noto è una condizione necessaria e sufficiente per l'esaustività di $\{T_n; n \geq 1\}$ secondo R.A. Fischer, è anche condizione necessaria e sufficiente per l'esaustività bayesiana.

Teorema: il riassunto campionario $\{T_n; n \geq 1\}$ è esaustivo per il parametro incognito Θ della distribuzione campionaria $F(x_1,.....,x_n / \theta)$ se e solo se esistono funzioni h(.) e k(.) tali che

$$\ell(\theta / x_1,....., x_n) = \prod_{i=1}^{n} \ell(\theta / x_i) = \; h\left[t_n(x_1,....., x_n), \theta\right].k(x_1,....., x_n),$$

ove k(.) non dipende dal parametro θ.

Dimostrazione.

Supponiamo che $G(\theta / H)$ sia dotata di densità e che $\{T_n; n \geq 1\}$ sia esaustivo, cioè che sia $g[\theta / H \bigcap (x_1,.....,x_n)] = g[\theta / H \bigcap t_n(x_1,.....,x_n)]$ per ogni possibile scelta di $g(\theta / H)$; allora, indicando $t_n(x_1,.....,x_n)$ con t, è :

$$\frac{g(\theta / H).\ell(\theta / x_1,....., x_n)}{\int\limits_{\{\theta\}} g(\theta / H).\ell(\theta / x_1,....., x_n)d\theta} = \frac{g(\theta / H).\ell(\theta / t)}{\int\limits_{\{\theta\}} g(\theta / H).\ell(\theta / t)d\theta}.$$

Semplificando l'espressione si ottiene $\ell(\theta / x_1,.....,x_n) = \ell(\theta / t).k(x_1,.....,x_n)$, ove $k(x_1,.....,x_n)$ è uguale al rapporto dei due integrali che dipende dalla sequenza $x_1,.....,x_n$ ma non da θ, mentre è $h[t_n(x_1,.....,x_n),\theta] = h(t,\theta) = \ell(\theta / t)$.

Supponiamo ora verificata la fattorizzazione della verosimiglianza e quindi l'esaustività del riassunto $\{T_n; n \geq 1\}$ secondo R.A. Fischer. Assumendo per semplicità espositiva che le distribuzioni campionarie dei n.a. osservabili X_j e dei riassunti T_n siano dotate di densità, poiché $T_n = t_n(X_1,.....,X_n)$ si ha $f(x_1,.....,x_n / \theta) = f(x_1,.....,x_n,t / \theta) = f(t / \theta).f(x_1,.....,x_n / t,\theta)$; allora, ricordando le uguaglianze formali $f(x_1,.....,x_n / \theta) = \ell(\theta / x_1,.....,x_n)$ e $f(t / \theta) = \ell(\theta / t)$ si ha:

$$g[\theta / H \cap (x_1,.....,x_n)] \propto g(\theta / H).\ f(x_1,.....,x_n / \theta) = g(\theta / H).\ f(t / \theta).f(x_1,.....,x_n / t,\theta).$$

Per ipotesi la densità $f(x_1,.....,x_n / t,\theta)$ non dipende da θ e poiché $g(\theta / H).\ f(t / \theta) \propto g(\theta / H \cap t)$ si ottiene la relazione $g[\theta / H \cap (x_1,.....,x_n)] \propto g(\theta / H \cap t)$, ma se due densità sono proporzionali esse sono anche uguali.

Ritornando ai quattro esempi di riassunti campionari a) – d), mentre i primi tre hanno una dimensione che rimane costante al variare della numerosità campionaria n (dimensione 1 per i primi due e 3 per il terzo) non è così per l'ultimo esempio, denominato in letteratura "statistica d'ordine" o "statistica ordinata": la sua dimensione coincide con n. E' facile rendersi conto che la variabilità della dimensione del riassunto determina notevoli appesantimenti per gli sviluppi formali: è quindi desiderabile ricorrere possibilmente a riassunti con dimensione costante. Sussiste in proposito un importante risultato riguardante una proprietà delle distribuzioni della famiglia esponenziale.

Teorema: in ipotesi di scambiabilità per il processo osservabile $\{X_t; t \geq 1\}$, se la comune distribuzione campionaria $F(x_t / \theta)$ appartiene alla famiglia esponenziale allora per l'inferenza sul parametro incognito Θ esiste un riassunto esaustivo $\{T_n; n \geq 1\}$ di dimensione costante rispetto ad n.

Dimostrazione: applicando il criterio di fattorizzazione di J. Neyman alla verosimiglianza

$$\ell(\theta_1,.....,\theta_k / x_1,.....,x_m) = B^m(\theta_1,.....,\theta_k).\left[\prod_{i=1}^{m} h(x_i)\right].\exp\left\{\sum_{j=1}^{k} Q_j(\theta_1,.....,\theta_k).\sum_{i=1}^{m} U_j(x_i)\right\}$$

si ha che $h[t_m(x_1,.....,x_m),\theta] = B^m(\theta_1,.....,\theta_k).\exp\left\{\sum_{j=1}^{k} Q_j(\theta_1,.....,\theta_k).\sum_{i=1}^{m} U_j(x_i)\right\}$ cosicché il riassunto esaustivo è $\boldsymbol{T_m} = \left(m, \sum_{i=1}^{m} U_1(x_i),.....,\sum_{i=1}^{m} U_k(x_i)\right)'$ di dimensione costante k + 1.

La famiglia coniugata di distribuzioni

E' nota con questo nome la famiglia di distribuzioni del parametro incognito Θ il cui nucleo ha la stessa forma funzionale della funzione di verosimiglianza: se si sceglie questo tipo di distribuzione allora

nell'applicazione del teorema di Bayes la distribuzione finale avrà la stessa forma funzionale di quella iniziale. La trasformazione coerente della distribuzione iniziale si riduce così ad una modificazione dei soli suoi parametri. E' quello che si è già visto negli schemi statistici usati in precedenza e cioè nei modelli statistici Bernoulli – Beta, Poisson – Gamma, Esponenziale – Gamma e Normale – Normale.

Il mantenimento della forma funzionale della $G(\theta / H)$ rende semplici gli sviluppi formali e le necessarie integrazioni; va però sottolineato che non c'è nessuna rilevanza teorica in questa scelta particolare di quella forma funzionale. Anzi bisogna dire che se l'opinione iniziale su Θ è meglio espressa mediante una distribuzione diversa da quella coniugata, il perseguimento della semplicità negli sviluppi matematici è sicuramente un falso obiettivo. L'uso di distribuzioni non coniugate con la verosimiglianza è del tutto fattibile grazie agli sviluppi del calcolo numerico e all'attuale potenza di calcolo dei personal computers che permette di eseguire facilmente integrazioni numeriche quando se ne presenti la necessità.

La stima parametrica bayesiana

Per stime puntuali bayesiane del parametro incognito Θ si intendono valori numerici estratti dalla distribuzione finale $G(\theta / H \wedge K)$: può trattarsi della speranza matematica, della mediana, del valore modale, se la distribuzione finale è unimodale, e così via. E' evidente che l'utilizzo della distribuzione finale per l'ottenimento di un solo valore numerico approssimato per Θ comporta una perdita di informazione enorme; si pensi, per esempio, che la speranza matematica di $G(\theta / H \wedge K)$ è solo uno degli infiniti momenti $\int_R \theta^n dG(\theta / H \wedge K), \quad n = 1, 2.....$ desumibili da $G(\theta / H \wedge K)$!

La scelta della particolare stima bayesiana (speranza matematica di Θ o mediana di $G(\theta / H \wedge K)$ o valore modale di $G(\theta / H \wedge K)$ o altro ancora) andrebbe fatta tenendo conto di una qualche *funzione di utilità o disutilità* (si parla in quest'ultimo caso anche di *funzione di danno o di perdita*) che renda possibile un'impostazione razionale del problema. Si tratta in fin dei conti di un problema di decisione ottimale in cui intervengono valutazioni probabilistiche ed anche valutazioni di merito circa l'importanza degli errori di stima. Ad esempio una funzione di perdita molto usata è la $\ell(\Theta, a) = (\Theta - a)^2$ che non distingue tra errori positivi e negativi e li pesa allo stesso modo; in questo caso la speranza matematica del quadrato dell'errore è $E\left[\ell(\Theta, a) / H \wedge K\right] = \int_R (\theta - a)^2 \, dG(\theta / H \wedge K)$ ed il valore di a che rende minima tale speranza matematica è $a = E(\Theta / H \wedge K)$. Per approfondimenti si veda J.O. Berger – Statistical Decision Theory and Bayesian Analysis (Springer – Verlag).