

LA DISTRIBUZIONE DEL RANGO DI OSSERVAZIONI TRATTE DA UNA POPOLAZIONE CON FUNZIONE DI RIPARTIZIONE INCERTA E RETTA DA UN PROCESSO DI DIRICHLET

1. INTRODUZIONE E SOMMARIO

Per introdurre la problematica di cui intendiamo occuparci consideriamo il caso in cui, per motivi di salvaguardia, si sia interessati all'entità delle piene di un fiume che sappiamo averne avute già un certo numero b con portate $x_1, x_2, \dots, x_b$.

In situazioni di questo tipo può risultare di primaria importanza, la conoscenza:

a) della probabilità che tra le successive $(n - b)$ portate di piena ve ne siano K_0 inferiori, L_0 uguali ed M_0 maggiori di una qualunque tra le b .

b) della probabilità che le b osservazioni note occupino posti assegnati nella graduatoria delle n osservazioni considerate.

c) della probabilità che una delle piene successive abbia portata superiore alla massima o inferiore alla minima registrata in passato.

In un contesto bayesiano si può rispondere alle domande precedenti qualora si esplicitino i legami stocastici che sussistono tra i numeri aleatori $X_1, X_2, \dots, X_b, X_{b+1}, \dots, X_n$.

Se non conosciamo le leggi di distribuzione dei numeri aleatori X_i ($i = 1, 2, \dots, n$) ci troviamo di fronte ad un problema di natura tipicamente non parametrico. Ad un livello molto generale si può dire che un problema non parametrico, come è stato messo in evidenza da D.M. Cifarelli e E. Regazzini [1] sorge: "... tutte le volte in cui le distribuzioni degli enti osservabili non sono indicizzate da uno o più parametri, ma sono genericamente gli elementi di una classe di distribuzioni più o meno ampia".

L'applicazione del meccanismo bayesiano ai problemi non parametrici è stata piuttosto scarsa e la ragione deve ricercarsi nel fatto che in tali problemi è assai difficoltosa la scelta della distribuzione iniziale essenziale per affrontare in modo corretto il ragionamento induttivo. In tali problemi, infatti, lo spazio dei parametri è solitamente uno spazio funzionale cioè quello di tutte le funzioni di ripartizione.

Il problema dell'individuazione di una misura di probabilità iniziale è stato affrontato nel 1973 da T.S. Ferguson [2], il quale ha introdotto una particolare misura di probabilità aleatoria retta da un processo di Dirichlet che bene si adatta come distribuzione iniziale sullo spazio delle misure di probabilità. Data quindi la misura di probabilità iniziale diviene possibile analizzare problemi non parametrici da un punto di vista bayesiano, basando la risposta sulla distribuzione finale.

Nell'esempio introdotto possiamo considerare un campione di ampiezza n , $X_1, X_2, \dots, X_n$, estratto da una distribuzione F , in cui solamente b valori ($b < n$) sono noti.

Assumiamo ora che la funzione di ripartizione F non sia nota ma aleatoria e retta da un processo di Dirichlet. Le definizioni di base ed i risultati riguardanti il processo di Dirichlet che utilizzeremo in questo lavoro saranno presentate nel paragrafo 2.

La risposta ai punti a) e b) dell'esempio iniziale è sostanzialmente la stessa trattandosi di fornire la distribuzione del rango delle b osservazioni note sulla graduatoria delle n osservazioni considerate. Il risultato a cui perverremo nel paragrafo 3 ($b > 1$) è una estensione di quello fornito recentemente da G. Campbell e M. Hollander [3]. In [3] è possibile trovare altri problemi che possono essere affrontati con la stessa metodologia.

Nel paragrafo 4 vedremo come dare una risposta al quesito del punto c). Numerose applicazioni in cui è importante stabilire la probabilità che una futura osservazione superi il più alto livello registrato in passato si possono trovare nel libro di E.J. Gumbel [4] che contiene, altresì, una completa bibliografia sull'impostazione classica di tali problemi.

Nel paragrafo 5, infine, dedurremo, sulla base dei risultati dati in precedenza, alcune notevoli proprietà del campione estratto da un processo di Dirichlet. Tali proprietà sono state messe in evidenza, con un diverso approccio, da C. Antoniuk [5].

2. PRELIMINARI SUL PROCESSO DI DIRICHLET

In questo numero vogliamo presentare le definizioni ed i risultati di base riguardanti il processo di Dirichlet che poi utilizzeremo per trattare la problematica a cui si è alluso in precedenza. Nel dare tali nozioni ci avvarremo della simbologia introdotta da T.S. Ferguson nel già citato articolo del 1973.

Riteniamo opportuno, a questo punto, porre in evidenza che tale processo dopo le prime applicazioni fatte dallo stesso Ferguson in problemi classici di inferenza statistica, è stato successivamente impiegato, per quanto ci risulta da V. Susarla e J. Van Ryzin [6], [7] nell'analisi di un problema predittivo con osservazioni censurate, da G. Campbell e M. Hollander nel lavoro già citato per la stima del rango di una osservazione in un gruppo di n future altre osservazioni, da D.M. Cifarelli e E. Regazzini [1] trattando l'impiego di una generica media associativa e da D.M. Cifarelli e A. Merlini [8] per la stima bayesiana della vita media residua.

DEFINIZIONE 2.1. (Ferguson [2])

Sia $\mathcal{X}$ un insieme (spazio campionario) ed $\mathcal{A}$ una σ -algebra di sottoinsiemi di $\mathcal{X}$. Sia α una misura non nulla, finita, non negativa e finitamente additiva su $(\mathcal{X}, \mathcal{A})$. Si dice che la probabilità aleatoria P su $(\mathcal{X}, \mathcal{A})$ è un processo di Dirichlet su $(\mathcal{X}, \mathcal{A})$ con parametro α , se, per ogni $k = 1, 2, \dots$ e per ogni partizione $\{B_1, B_2, \dots, B_k\}$ di $\mathcal{X}$, ($B_i \in \mathcal{A}$, $\forall i$) la distribuzione del vettore aleatorio $\{P(B_1), P(B_2), \dots, P(B_k)\}$ possiede densità di Dirichlet con parametri $(\alpha(B_1), \alpha(B_2), \dots, \alpha(B_k))$.

DEFINIZIONE 2.2. (Ferguson [2])

Sia P una misura di probabilità aleatoria su $(\mathcal{X}, \mathcal{A})$. Si dirà che il vettore aleatorio $(X_1, X_2, \dots, X_n)$ è un campione di ampiezza n estratto da P se per ogni $m = 1, 2, \dots$ e insiemi misurabili $A_1, A_2, \dots, A_m, C_1, C_2, \dots, C_n$,

$$\begin{aligned} & Pr \left\{ X_1 \in C_1, \dots, X_n \in C_n \mid P(A_1), \dots, P(A_m), P(C_1), \dots, P(C_n) \right\} = \\ &= \prod_{j=1}^n P(C_j) \quad \text{quasi certamente} \end{aligned}$$

PROPOSIZIONE 2.1. (Ferguson [2])

Se P è un processo di Dirichlet con parametro α e $(X_1, X_2, \dots, X_n)$ è un campione estratto da P , allora il processo $(P | X_1, \dots, X_n)$ è ancora di Dirichlet con parametro $\alpha + \sum_i^n \delta_{x_i}$, essendo δ_{x_i} la misura su $(\mathcal{X}, \mathcal{A})$ che assegna la massa unitaria al punto x_i , cioè

$$\delta_{x_i}(A) = \begin{cases} 1 & \text{se } x_i \in A \\ 0 & \text{se } x_i \notin A. \end{cases}$$

Nel nostro lavoro supporremo che $\mathcal{X} = \mathbf{R}$ e $\mathcal{A} = \mathcal{B}$ essendo $\mathcal{B}$ la σ -algebra dei boreliani di $\mathbf{R}$. Considereremo le misure α e P nella forma di funzioni di punto, vale a dire:

$$\alpha(x) = \alpha(-\infty, x], \quad F(x) = P(-\infty, x]$$

dove $F(x)$ è una funzione di ripartizione aleatoria.

Si può mostrare (cf. [1]) che se P ha distribuzione di Dirichlet allora il vettore aleatorio

$$\{F(x_1), F(x_2), \dots, F(x_k)\} \quad x_1 \leq x_2 \leq \dots \leq x_k$$

possiede come funzione di densità quella di Dirichlet ordinata (cf. S.S. Wilks [9] pag. 182), cioè

$$\begin{aligned} d(y_1, \dots, y_k | \alpha(x_1), \alpha(x_2) - \alpha(x_1), \dots, \alpha(\mathbf{R}) - \alpha(x_k)) = \\ = \frac{\Gamma(\alpha(\mathbf{R}))}{\Gamma(\alpha(x_1)) \Gamma(\alpha(x_2) - \alpha(x_1)) \dots \Gamma(\alpha(\mathbf{R}) - \alpha(x_k))} y_1^{\alpha(x_1)-1} \cdot \\ \cdot (y_2 - y_1)^{\alpha(x_2) - \alpha(x_1) - 1} \dots (y_k - y_{k-1})^{\alpha(x_k) - \alpha(x_{k-1}) - 1} (1 - y_k)^{\alpha(\mathbf{R}) - \alpha(x_k) - 1} \cdot I_0 \end{aligned}$$

essendo I_0 l'indicatore dell'insieme

$$\{(y_1, \dots, y_k) : 0 \leq y_1 \leq y_2 \leq \dots \leq y_k \leq 1\}$$

La Proposizione 2.1 asserisce che il processo $(F | X_1, \dots, X_n)$ possiede densità k -dimensionale di Dirichlet ordinata con parametro $\alpha + \sum_i^n \delta_{x_i}$.

3. LA DISTRIBUZIONE DEL RANGO

Sia $X_1, X_2, \dots, X_n$ un campione di ampiezza n estratto da una distribuzione F e supponiamo di aver osservato solamente b delle n osservazioni $X_1 = x_1, X_2 = x_2, \dots, X_b = x_b$ ($b < n$).

Assumiamo che F sia selezionata in modo casuale da un processo iniziale di Dirichlet con parametro $\alpha(\cdot)$, una misura non nulla, finita, non negativa e finitamente additiva sull'asse reale con la σ -algebra di Borel.

Assumiamo, inoltre, la condizione di *scambiabilità* della successione $\{X_n\}$ ($n=1, 2, \dots$) dei numeri aleatori associati agli elementi del campione. Ciò equivale a ritenere tali numeri aleatori stocasticamente indipendenti ed identicamente distribuiti subordinatamente ad una assegnata funzione di ripartizione F . L'ipotesi di scambiabilità porta,

nella sostanza, a ritenere che le osservazioni agiscano simmetricamente sulle opinioni iniziali, implicando l'irrelevanza dell'ordine secondo cui si presentano i risultati dell'esperienza.

Nel procedere dobbiamo porre molta attenzione al fatto che, con probabilità uno, le misure di probabilità che il processo di Dirichlet seleziona sono discrete. Pertanto si possono ottenere, con probabilità positiva, due o più osservazioni coincidenti (si veda: Ferguson [2], Blackwell [10], Blackwell e Mac Queen [11]).

Siano K_0 , L_0 , M_0 il numero di osservazioni rispettivamente minori, uguali e maggiori di X_1 (una qualunque delle b osservazioni). Prima di iniziare lo studio della distribuzione di (K_0, L_0, M_0) diamo la seguente:

PROPOSIZIONE 3.1. (Campbell-Hollander [3])

Se $X_1, X_2, \dots, X_n$ è un campione di ampiezza n estratto da un processo di Dirichlet su $(\mathbf{R}, \mathcal{B})$ con parametro α e se $A \in \mathcal{B}^n$, allora

$$\begin{aligned} \Pr \{ (X_1, X_2, \dots, X_n) \in A \mid X_1 = x_1, X_2 = x_2, \dots, X_b = x_b \} = \\ = \int \Pr \{ (X_1, \dots, X_n) \in A \mid X_1, \dots, X_b, F \} dO_{\alpha'}(F) \end{aligned}$$

dove $O_{\alpha'}$ indica il processo di Dirichlet con parametro $\alpha' = \alpha + \sum_{i=1}^n \delta_{x_i}$

Definiamo, inizialmente, la distribuzione di (K_0, L_0, M_0) dati $X_1 = x_1$ ed F :

$$\begin{aligned} \Pr \{ (K_0, L_0, M_0) = (k_0, l_0, m_0) \mid X_1 = x_1, F \} = \\ = \frac{(n-1)!}{k_0! l_0! m_0!} [F(x_1^-)]^{k_0} [F(x_1) - F(x_1^-)]^{l_0} [1 - F(x_1)]^{m_0} \\ = \frac{(n-1)!}{k_0! l_0! m_0!} \sum_{j=0}^{l_0} \binom{l_0}{j} (-1)^{l_0-j} [F(x_1)]^j [F(x_1^-)]^{k_0+l_0-j} [1 - F(x_1)]^{m_0} \end{aligned} \quad (1)$$

La Proposizione 3.1 implica che per determinare la distribuzione di (K_0, L_0, M_0) dato $X_1 = x_1$ basterà limitarsi ad integrare la (1) rispetto alla distribuzione di F con parametro $\alpha(x_1) + \delta_{x_1}$.

Dopo alcuni calcoli, considerando $\alpha(x_1)$ misura non atomica e ricordando che

$$\sum_{j=0}^{l_0} (-1)^{l_0-j} \binom{l_0}{j} \frac{1}{\alpha(x_1) + m_0 + l_0 - j}$$

è una funzione Beta con parametri $(\alpha(x_1) + m_0)$ ed $(l_0 + 1)$ (si veda C. Jordan [12] pag. 50), otteniamo:

$$\begin{aligned} \Pr \{ (K_0, L_0, M_0) = (k_0, l_0, m_0) \mid X_1 = x_1 \} = \\ = \frac{(n-1)!}{k_0! (n-1-k_0-l_0)!} \frac{\Gamma(\alpha(\mathbf{R})+1)}{\Gamma(\alpha(\mathbf{R})+n)} \frac{\Gamma(\alpha(x_1)+k_0)}{\Gamma(\alpha(x_1))} \cdot \frac{\Gamma(\alpha(\mathbf{R})-\alpha(x_1)+n-1-k_0-l_0)}{\Gamma(\alpha(\mathbf{R})-\alpha(x_1))} \end{aligned} \quad (2)$$

Indicando, per ogni k intero positivo, con $x^{[k]}$ il fattoriale $x(x+1) \dots (x+k-1)$ e definendo $x^{[0]} \equiv 1$ la (2) risulta:

$$= \frac{(n-1)!}{(n-1-l_0)!} \binom{n-1-l_0}{k_0} \alpha(x_1)^{[k_0]} (\alpha(\mathbf{R}) - \alpha(x_1))^{[n-1-k_0-l_0]} / (\alpha(\mathbf{R}) + 1)^{[n-1]} \quad (3)$$

Siamo ora in possesso degli elementi per fornire la distribuzione di (K_0, L_0, M_0) .

Dalla (3), infatti, è immediato dedurre:

$$\begin{aligned} & Pr \{ (K_0, L_0, M_0) = (k_0, l_0, m_0) \} = \\ & = \frac{(n-1)!}{(n-1-l_0)!} \binom{n-1-l_0}{k_0} \frac{1}{(\alpha(R)+1)^{[n-1]}} \int_R \alpha(x_1)^{[k_0]} (\alpha(R)-\alpha(x_1))^{[n-1-k_0-l_0]} \frac{d\alpha(x_1)}{\alpha(R)} \end{aligned} \quad (4)$$

La (4) è il risultato che si deduce dalla formula (3.11) nel lavoro di G. Campbell e M. Hollander supponendo $\alpha(\cdot)$ misura non atomica.

Mostreremo, ora, come sia possibile esprimere la (4) utilizzando solamente le funzioni Beta.

Con la sostituzione $\frac{\alpha(x_1)}{\alpha(R)} = t$, si ha:

$$\begin{aligned} & Pr \{ (K_0, L_0, M_0) = (k_0, l_0, m_0) \} = \\ & = \frac{(n-1)!}{(n-1-l_0)!} \binom{n-1-l_0}{k_0} \frac{\alpha(R)^{[k_0]+[n-1-k_0-l_0]}}{(\alpha(R)+1)^{[n-1]}} \int_0^1 t^{[k_0]} (1-t)^{[n-1-k_0-l_0]} dt \end{aligned} \quad (5)$$

Ci sembra importante sottolineare il fatto che la (5) dipende in modo essenziale da $\alpha(R)$, cioè dal "peso" che si vuole assegnare all'opinione iniziale $\frac{\alpha(x)}{\alpha(R)}$ sul fenomeno statisticamente osservabile.

Definiamo una successione di polinomi $A_{k_0}(t)$ nel modo seguente:

$$A_1(t) = t$$

$$A_2(t) = (t+1)A_1(t) = t(t+1) = t^{[2]}$$

⋮

$$A_{k_0}(t) = (t+k_0-1)A_{k_0-1}(t) = t(t+1)\dots(t+k_0-1) = t^{[k_0]}$$

Quindi $A_{k_0}(t)$ è un polinomio di grado k_0 in t con coefficienti interi, vale a dire:

$$A_{k_0}(t) = k_0 a_1 t + k_0 a_2 t^2 + \dots + k_0 a_{k_0} t^{k_0}$$

I coefficienti $k_0 a_j$ del polinomio sono i valori assoluti dei numeri del primo tipo di Stirling (tabulati in Abramowitz e Stegun [13] pag. 833).

Possiamo quindi scrivere:

$$t^{[k_0]} = \sum_{j=1}^{k_0} k_0 a_j t^j$$

ed in modo analogo

$$(1-t)^{[n-1-k_0-l_0]} = \sum_{j=1}^{n-1-k_0-l_0} (n-1-k_0-l_0) b_j (1-t)^j$$

L'integrale che compare nella (5) può scriversi

$$\int_0^1 t^{[k_0]} (1-t)^{[n-1-k_0-l_0]} dt = \sum_{j=1}^{k_0} k_0 a_j \sum_{i=1}^{n-1-k_0-l_0} (n-1-k_0-l_0) b_i \int_0^1 t^j (1-t)^i dt =$$

$$= \sum_{j=1}^{k_0} \sum_{i=1}^{n-1-k_0-l_0} k_0 a_j (n-1-k_0-l_0) b_i B(j+1, i+1)$$

Possiamo quindi affermare che la probabilità che su n osservazioni ve ne siano rispettivamente k_0 minori, l_0 uguali ed m_0 maggiori del primo (o di un qualunque) elemento del campione è:

$$\begin{aligned} \Pr \{ (K_0, L_0, M_0) = (k_0, l_0, m_0) \} &= \\ &= \frac{(n-1)!}{(n-1-l_0)!} \binom{n-1-l_0}{k_0} \left(\frac{\alpha(R)^{[k_0]+[n-1-k_0-l_0]}}{(\alpha(R)+1)^{[n-1]}} \right) \sum_{j=1}^{k_0} \\ &\quad \sum_{i=1}^{n-1-k_0-l_0} k_0 a_j (n-1-k_0-l_0) b_i B(j+1, i+1) \end{aligned} \quad (6)$$

Il nostro scopo è ora quello di generalizzare quanto detto e di pervenire alla distribuzione del rango di b ($b > 1$) osservazioni sul totale delle n osservazioni considerate. Prima di dare la distribuzione dei ranghi delle b osservazioni $(x_1, x_2, \dots, x_b)$ riteniamo opportuno esporre la metodologia nel caso $b = 2$, $X_1 = x_1$, $X_2 = x_2$. Senza perdere in generalità supponiamo $x_1 < x_2$.

Siano K_0, L_0, K_1, L_1, M il numero di osservazioni del campione di ampiezza n rispettivamente minori di x_1 , uguali ad x_1 , maggiori di x_1 e minori di x_2 , uguali ad x_2 , maggiori di x_2 .

$$\begin{array}{c|c|c|c|c} k_0 & l_0 & k_1 & l_1 & m \\ \hline & X_1 = x_1 & & X_2 = x_2 & \end{array}$$

La distribuzione di (K_0, L_0, K_1, L_1, M) dati $X_1 = x_1, X_2 = x_2$, ed F è la seguente:

$$\begin{aligned} \Pr \{ (K_0, L_0, K_1, L_1, M) = (k_0, l_0, k_1, l_1, m) \mid X_1 = x_1, X_2 = x_2, F \} &= \\ &= \frac{(n-2)!}{k_0! k_1! l_0! l_1! m!} [F(x_1^-)]^{k_0} [F(x_1) - F(x_1^-)]^{l_0} [F(x_2^-) - F(x_1)]^{k_1} \cdot \\ &\quad \cdot [F(x_2) - F(x_2^-)]^{l_1} [1 - F(x_2)]^m \end{aligned} \quad (7)$$

La distribuzione di (K_0, L_0, K_1, L_1, M) dati $X_1 = x_1, X_2 = x_2$ si ottiene integrando la (7) rispetto alla distribuzione di F con parametro $\alpha + \sum_{i=1}^2 \delta_{x_i}$. Nel caso in cui si consideri $\alpha(\cdot)$ misura non atomica, otteniamo:

$$\begin{aligned} \Pr \{ (K_0, L_0, K_1, L_1, M) = (k_0, l_0, k_1, l_1, m) \mid X_1 = x_1, X_2 = x_2 \} &= \\ &= \frac{(n-2)!}{k_0! k_1! (n-2-k_0-k_1-l_0-l_1)!} \\ &\quad \frac{\Gamma(\alpha(R)+2)}{\Gamma(\alpha(R)+n)} \alpha(x_1)^{[k_0]} (\alpha(R) - \alpha(x_2))^{[n-2-k_0-k_1-l_0-l_1]} \\ &\quad \frac{\Gamma(\alpha(x_2) + k_0 + l_0 + k_1 + 1)}{\Gamma(\alpha(x_1) + k_0 + l_0 + 1) \Gamma(\alpha(x_2) - \alpha(x_1))} \end{aligned} \quad (8)$$

Il risultato a cui siamo pervenuti può essere facilmente generalizzato; non vi sono infatti difficoltà concettuali ma solamente alcune complicazioni formali.

Supponiamo $x_1 < x_2 < x_3 < \dots < x_b$. La distribuzione dei ranghi delle b osservazioni note sul campione $X_1, X_2, \dots, X_n$ è data da:

$$\begin{aligned} & Pr\{k_0 < x_1, l_0 = x_1, x_1 < k_1 < x_2, \dots, x_{b-1} < k_{b-1} < x_b, l_{b-1} = x_b, \\ & \quad m > x_b \mid X_1 = x_1 \dots X_b = x_b\} = \\ & = \frac{(n-b)!}{\prod_{i=1}^{b-1} k_i! (n-b - \sum_{j=0}^{b-1} (k_j + l_j))!} \frac{\Gamma(\alpha(R) + b)}{\Gamma(\alpha(R) + n)} \alpha(x_1)^{[k_0]} \cdot \\ & \quad (\alpha(R) - \alpha(x_b))^{[n-b - \sum_{j=0}^{b-1} (k_j + l_j)]} \cdot \\ & \quad \cdot \prod_{i=2}^b \frac{\Gamma(\alpha(x_i) + \sum_{j=0}^{i-1} k_j + \sum_{j=0}^{i-2} l_j + (i-1))}{\Gamma(\alpha(x_{i-1}) + \sum_{j=0}^{i-2} (k_j + l_j) + (i-1)) \Gamma(\alpha(x_i) - \alpha(x_{i-1}))} \end{aligned} \quad (9)$$

In conclusione i ranghi delle b osservazioni, convenendo di assegnare come rango delle osservazioni coincidenti la media dei ranghi che avrebbero avuto se fossero state distinte, saranno:

$$r(X_1) = k_0 + 1/2 (l_0 + 1)$$

$$r(X_2) = k_0 + k_1 + 1/2 (l_0 + 1) + 1/2 (l_1 + 1)$$

⋮

$$r(X_b) = \sum_{j=0}^{b-1} k_j + 1/2 \left(\sum_{j=0}^{b-1} (l_j + 1) \right)$$

La loro distribuzione può, in linea di principio, essere dedotta dalla (9).

4. VALUTAZIONE DI $Pr(X_{b+1} > \max(X_1, X_2, \dots, X_b))$.

Come è noto, data una successione $X_1, X_2, \dots, X_b \dots$ di variabili aleatorie indipendenti con funzione di ripartizione $F(x)$, le funzioni di ripartizione, per ogni intero $b \geq 1$, di $Z_b = \max(X_1, X_2, \dots, X_b)$ e di $Z_1 = \min(X_1, X_2, \dots, X_b)$ sono:

$$a) \quad Pr(Z_b \leq x) = Pr(X_1 \leq x, X_2 \leq x, \dots, X_b \leq x) = F^b(x)$$

$$b) \quad Pr(Z_1 \leq x) = 1 - Pr(X_1 > x, X_2 > x, \dots, X_b > x) = 1 - (1 - F(x))^b$$

Sia, ora, $\{X_b, b \geq 1\}$ una successione "scambiabile" e $F_m(x_1, x_2, \dots, x_m)$ rappresenti, per ogni m , la distribuzione congiunta di m variabili aleatorie.

Dal teorema di rappresentazione di leggi di probabilità scambiabili, dovuto a B. de Finetti, segue che $F_m(x_1, x_2, \dots, x_m)$ può essere rappresentata come una misura

delle distribuzioni congiunte di m variabili aleatorie indipendenti ed identicamente distribuite.

In particolare, possiamo scrivere (si veda M. Loève [14], pag. 365)

$$F_m(x_1, x_2, \dots, x_m) = \int_{\Omega} F_{\omega}(x_1), \dots, F_{\omega}(x_m) dP(\omega) \quad (10)$$

dove, per ogni x fissato, $F_{\omega}(x)$ è una variabile aleatoria, e per ogni ω $F_{\omega}(x)$ è una funzione di ripartizione in x .

Dalla rappresentazione (10) ne deriva che la funzione di ripartizione del massimo di b variabili è data da:

$$Pr(Z_b \leq x) = F_b(x, x, \dots, x) = \int_{\Omega} F_{\omega}^b(x) dP(\omega) = E[F_{\omega}^b(x)]$$

Se supponiamo che la distribuzione F sia selezionata casualmente da un processo di Dirichlet, otteniamo

$$Pr(Z_b \leq x) = E[F^b(x)] = \frac{\alpha(x)^{[b]}}{\alpha(R)^{[b]}} = \left(\frac{\alpha(x)}{\alpha(R)}\right)^{[b]}$$

In modo analogo si ha:

$$Pr(Z_1 \leq x) = 1 - E[1 - F(x)]^b = 1 - \frac{(\alpha(R) - \alpha(x))^{[b]}}{\alpha(R)^{[b]}}$$

Applicando la linea di dimostrazione del paragrafo precedente siamo in grado di determinare facilmente la probabilità che tra le future $(n-b)$ osservazioni ve ne siano K minori, L uguali ed M maggiori del $\max(X_1, X_2, \dots, X_b)$.

La distribuzione di (K, L, M) condizionata al $\max(X_1, X_2, \dots, X_b)$ è:

$$\begin{aligned} Pr\{(K, L, M) = (k, l, m) \mid \max(X_1, X_2, \dots, X_b) = x\} = \\ = \frac{(n-1)!}{k! m!} \frac{\alpha(x)^{[k]} (\alpha(R) - \alpha(x))^{[m]}}{(\alpha(R) + 1)^{[n-1]}} \end{aligned}$$

La distribuzione marginale di (K, L, M) si otterrà integrando su R rispetto alla distribuzione di $Z_b = \max(X_1, X_2, \dots, X_b)$.

$$\begin{aligned} Pr\{(K, L, M) = (k, l, m)\} = \\ = \frac{(n-1)!}{k! m!} \frac{1}{(\alpha(R) + 1)^{[n-1]}} \alpha(R)^{[k]+[m]} \int_0^1 t^{[k]} (1-t)^{[m]} dt^{[b]} \end{aligned} \quad (11)$$

Applicando, con leggere varianti, il procedimento adottato per la soluzione dell'integrale della (5), possiamo esprimere la (11) utilizzando solamente le funzioni Beta.

$$\begin{aligned} Pr\{(K, L, M) = (k, l, m)\} = \\ = \frac{(n-1)!}{k! m!} \frac{\alpha(R)^{[k]+[m]}}{(\alpha(R) + 1)^{[n-1]}} \sum_{j=1}^k \sum_{i=1}^m \sum_{s=1}^b k a_j m b_i b c_s B(j+s, i+1). \end{aligned} \quad (12)$$

Dalla (12) è immediato dedurre che:

$$Pr(X_{b+1} > \max(X_1, X_2, \dots, X_b)) =$$

$$= \frac{\alpha(R)}{\alpha(R) + 1} \sum_{s=1}^b b^s c_s B(s, 2)$$

Per la valutazione della $Pr(X_{b+1} < \min(X_1, X_2, \dots, X_b))$ la metodologia è analoga.

A conclusione di questo numero ci sembra utile ricordare che S. Berman [15] ha studiato la distribuzione limite del $\max(X_1, X_2, \dots, X_n)$ nell'ipotesi di variabili aleatorie soggette a diversi tipi di dipendenza ed in particolare nel caso di una successione scambiabile in generale.

5. PROPRIETÀ DEL CAMPIONE ESTRATTO DA UN PROCESSO DI DIRICHLET

I risultati dei numeri precedenti, grazie alla struttura del processo considerato, ci permettono di ricavare, in maniera assai semplice, alcune proprietà del campione estratto da un processo di Dirichlet.

Nel nostro modello è possibile pensare che il campione $X_1, X_2, \dots, X_n$ sia estratto in maniera sequenziale da un processo di Dirichlet P . Sia X_1 un campione di ampiezza 1 estratto da P ; dopo aver ottenuto X_1 , sia X_2 un campione di ampiezza 1 ottenuto dalla distribuzione condizionata P dato X_1 ; e così di seguito fino ad ottenere $X_1, X_2, \dots, X_n$.

Appare in questo modo evidente l'importanza che assume la relazione (6) nel determinare alcune proprietà del campione.

Nel caso di un campione di ampiezza $n = 2$, X_1, X_2 , se ne deduce immediatamente che:

$$\begin{aligned} Pr(X_2 = X_1) &= \frac{1}{\alpha(R) + 1} \\ Pr(X_2 > X_1) &= \frac{1}{2} \frac{\alpha(R)}{\alpha(R) + 1} \\ Pr(X_2 < X_1) &= \frac{1}{2} \frac{\alpha(R)}{\alpha(R) + 1} \\ Pr(X_2 \neq X_1) &= \frac{\alpha(R)}{\alpha(R) + 1} \end{aligned} \quad (13)$$

Dalla (6) è altresì possibile dedurre la seguente relazione generale:

$$Pr(X_n = X_{n-1} = \dots = X_1) = \frac{1^{[n-1]}}{(\alpha(R) + 1)^{[n-1]}}$$

Ci attenderemmo, invece, poiché abbiamo considerato $\alpha(\cdot)$ misura non atomica di ottenere $Pr(X_2 = X_1) = 0$. Il risultato ottenuto è giustificato dal fatto che sebbene $\alpha(\cdot)$ sia stata considerata non atomica la distribuzione condizionata di P dato X_1 è ancora un processo di Dirichlet con parametro $\alpha + \delta_x$ che è una misura atomica con atomo di misura 1 in x_1 . Appare evidente comunque che il valore prescelto per $\alpha(R)$ può attenuare questo inconveniente poiché se $\alpha(R)$ è molto grande, la probabilità della coincidenza sarà molto piccola.

A questo punto giova spendere una parola su di un possibile significato da attribuire ad $\alpha(R)$ in base alle relazioni (13). Poiché l'evento $\{X_1 = X_2\}$ è un evento

osservabile, allora una attenta analisi condurrebbe ad una risposta intorno alla sua probabilità e quindi sul valore da attribuire ad $\alpha(R)$. Nel caso in cui si ritenessero ugualmente probabili gli eventi $\{X_1 = X_2\}$ e $\{X_1 \neq X_2\}$, $\alpha(R)$ dovrebbe essere posto uguale ad uno. Nel caso in cui si ritenesse valida la disuguaglianza $Pr(X_1 = X_2) < Pr(X_1 \neq X_2)$ diviene naturale la scelta di $\alpha(R)$ maggiore di uno.

Istituto di Scienze Statistiche e Demografiche
Università degli Studi di Pavia

PIETRO MULIERE

BIBLIOGRAFIA

- [1] D.M. CIFARELLI e E. REGAZZINI, *Considerazioni generali sull'impostazione bayesiana di problemi non parametrici. Le medie associative nel contesto del processo aleatorio di Dirichlet*. "Rivista di Matematica per le Scienze Economiche e Sociali" 2, n. 1, pp. 39-52, 1979.
- [2] T.S. FERGUSON, *A Bayesian analysis of some nonparametric problems*, Ann. of Stat., 1, 209-230, 1973.
- [3] G. CAMPBELL, M. HOLLANDER, *Rank order Estimation with the Dirichlet prior*, Ann. of Stat., 6, 1, 142-153, 1978.
- [4] E.J. GUMBEL, *Statistics of Extremes*, Columbia Un. Press, 1960.
- [5] C.E. ANTONIAK, *Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems*, Ann. of Stat. 2, 6, 1152-1174, 1974.
- [6] V. SUSARLA, J. VAN RYZIN, *Nonparametric bayesian estimation of survival curves from incomplete observations*, J. Am. Stat. Assoc., 71, 897-902, 1976.
- [7] V. SUSARLA, J. VAN RYZIN, *Large sample theory of bayesian nonparametric survival curve estimator based on censored samples*, Ann. of Stat., 4, 775-768, 1978.
- [8] D.M. CIFARELLI, A. MERLINI, *La stima non parametrica bayesiana della vita media residua*, Metron, in corso di stampa.
- [9] S.S. WILKS, *Mathematical Statistics*, Wiley, New York, 1962.
- [10] D. BLACKWELL, *Discreteness of Ferguson selections*, Ann. of Stat. 1, 356-358, 1973.
- [11] D. BLACKWELL, J.B. MAC QUEEN, *Ferguson distributions via Pólya urn schemes*, Ann. of Stat., 1, 353-355, 1973.
- [12] C. JORDAN, *Calculus of finite differences*, Chelsea Publ. Co., New York, 2 ed. 1960.
- [13] M. ABRAMOWITZ, I.A. STEGUN, *Handbook of Mathematical functions*, Dover, N.Y. 1965.
- [14] M. LOËVE, *Probability Theory*, Van Nostrand, Princeton, 3 ed. 1963.
- [15] S.M. BERMAN, *Limiting distribution of the Maximum term in sequences of dependent random variables*, Ann. of Math. Stat., 894-908, 1962.