Un risultato di approssimazione per spazi di probabilità finiti non uniformi

Marco Bramanti
Dipartimento di Matematica, Politecnico di Milano

22 marzo 2010

In questa nota si dimostra un risultato di approssimazione numerica in un problema di probabilità elementare nel finito che coinvolge numeri elevati. Questo risultato è stato utilizzato in [1], a cui si rimanda per la motivazione e l'illustrazione del contesto.

Problema. *Siano $X_1, X_2, ..., X_m$ variabili aleatorie indipendenti e identicamente distribuite (i.i.d.), la cui legge è la seguente: ogni X_i può assumere un numero finito N di valori ($N > m$), con probabilità $q_1, q_2, ..., q_N$. Qual è la probabilità π che almeno due delle X_i assumano lo stesso valore?*

Nel caso uniforme il problema si risolve con la combinatoria:

$$\pi = 1 - \frac{N\,(N-1)\,(N-2)\cdot ...\,(N-m+1)}{N^m}. \tag{1}$$

Nel caso non uniforme, applicazioni ripetute del teorema delle probabilità totali portano a scrivere:

$$\pi = 1 - \sum_{i_1=1}^{N} q_{i_1} \sum_{\substack{i_2=1,\\ i_2\neq i_1}}^{N} q_{i_2} \sum_{\substack{i_3=1,\\ i_3\neq i_1, i_3\neq i_2}}^{N} q_{i_3} ... \sum_{\substack{i_m=1,\\ i_m\neq i_1,...,i_m\neq i_{m-1}}}^{N} q_{i_m}. \tag{2}$$

Siamo interessati al calcolo di π in situazioni di probabilità non uniforme in cui N, m sono grandi e $m \ll N$ (nell'applicazione che faremo, $N = 10^{15}; m = 65000$). In questi casi la formula (2), per quanto esatta, è praticamente inservibile, per il numero molto elevato di addendi che richiederebbe di sommare. Vogliamo allora determinare una formula di calcolo approssimato per π che sia effettivamente utilizzabile e permetta di avere una stima a priori dell'errore commesso.

L'idea è approssimare il risultato nel caso non uniforme con quello del caso uniforme, ottenuto però per un diverso numero N di valori possibili. Precisamente, sia

$$\frac{1}{M} \equiv p \equiv \sum_{i=1}^{N} q_i^2, \tag{3}$$

interpretabile come il valore atteso di una variabile X che assume il valore q_i con probabilità q_i. Il ricorso a questa variabile X, che è *quantitativa*, ed è definita in termini della densità discreta delle variabili X_i di partenza, che sono qualitative, sarà utile in molti calcoli, nel seguito.

Il numero p è una sorta di probabilità media dei valori assunti dalle variabili X_i; se M fosse un intero e $Y_1, Y_2, ..., Y_m$ fossero m v.a. i.i.d. ognuna di legge uniforme con M valori possibili equiprobabili, la probabilità che almeno due delle Y_i assumano lo stesso valore sarebbe data da:

$$\pi' = 1 - \frac{M\,(M-1)\,(M-2)\cdot ...\,(M-m+1)}{M^m}. \tag{4}$$

Ovviamente la quantità (4) è ben definita anche se M non è un intero, ma è legato alla legge delle X_i dalla (3). Vogliamo usare π' come approssimazione di π.

Il problema è dimostrare una stima a priori per l'errore di approssimazione. I parametri che possiamo usare nella stima dell'errore sono anzitutto m, p e $q_{\max}$ (il massimo dei valori q_i); inoltre, per rendere la stima utile nel caso numerico che ci interessa, sarà opportuno stimare l'errore anche in termini dei momenti successivi della v.a. X che vale q_i con probabilità q_i, ossia in termini delle quantità:

$$E\left(X^k\right) = \sum_{j=1}^{N} \left(q_j\right)^{k+1}.$$

Dimostreremo prima (sez. 1) un risultato teorico generale, poi (sez. 2) mostreremo come applicarlo quantitativamente al problema numerico che motiva questa nota.

Notiamo anche che, come discusso in [1, sez. 3.4.1], la quantità π' in (4) può essere a sua volta approssimata al modo seguente

$$\pi' \simeq 1 - \exp\left(-\frac{1}{2}pm^2\right)$$

con p come in (3).

1 Un risultato teorico di approssimazione

Il risultato a cui arriveremo è il seguente:

Teorema 1 *Sia π come in (2) e π' come in (4). Allora:*

a. vale la seguente stima dell'errore:

$$|\pi - \pi'| \leq \frac{p}{q_{\max}} \left\{ \exp\left(\frac{m\,(m-1)}{2} q_{\max}\right) - 1 - \frac{m\,(m-1)}{2} q_{\max} \right\};$$

b. inoltre, per ogni intero n compreso tra 2 e $m-1$ vale la stima:

$$\begin{aligned}|\pi - \pi'| \leq & \sum_{k=2}^{n} \frac{1}{k!} \left[\frac{m\,(m-1)}{2}\right]^k E\left(X^k\right) + \\ & + \frac{E\left(X^n\right)}{\left(q_{\max}\right)^n} \left\{ \exp\left[\frac{m\,(m-1)}{2} q_{\max}\right] - \sum_{k=0}^{n} \frac{1}{k!} \left[\frac{m\,(m-1)}{2} q_{\max}\right]^k \right\} \\ & \equiv E_1^n + E_2^n;\end{aligned}$$

c. infine,

$$|\pi - \pi'| \leq E\left(\exp\left(\frac{m\,(m-1)}{2} X\right) - 1 - \frac{m\,(m-1)}{2} X\right).$$

Osservazione 2 *Affinché la stima del punto (a) risulti informativa, è necessario che sia $\left(\frac{m(m-1)}{2} q_{\max}\right) < 1$, ossia che m non sia troppo grande. Purtroppo, nel caso numerico a cui siamo interessati, questa condizione non è verificata, per cui si rende necessario raffinare la stima e utilizzare quella del punto (b). L'enunciato (a), più trasparente, costituisce una motivazione del processo attraverso cui si arriva a (b).*

Il punto (b) sfrutta il fatto che, anche se $\left(\frac{m(m-1)}{2} q_{\max}\right) > 1$, il resto della serie esponenziale è comunque piccolo; inoltre, ci si può aspettare che risulti $E\left(X^n\right) \ll \left(q_{\max}\right)^n$ (così che nella stima su E_2^n ciò che guadagniamo con il resto della serie esponenziale non vada perso a causa del fattore $\frac{E(X^n)}{(q_{\max})^n}$). Come si capisce meglio dal punto (c) (che è il caso limite di (b) per $n = m-1$), il termine E_1^n può essere piccolo se mediamente è $\frac{m(m-1)}{2} X \ll 1$, il che dà un vincolo su m oltre che sulla distribuzione: dev'essere almeno $m^2 \ll N$, come si capisce considerando il caso uniforme, in cui $E\left(X^k\right) = 1/N^{k-1}$.

Il contesto generale in cui immaginiamo di muoverci è quello di una distribuzione discreta in cui le probabilità q_i spaziano su molti ordini di grandezza diversi. In particolare, $p \ll q_{\max} \ll 1$ e i momenti successivi sono ciascuno molto più piccolo del precedente.

La stima del teorema non è ottimale in quanto non restituisce un errore nullo nel caso in cui si sappia a priori che la distribuzione è uniforme. In questo caso il punto (c) del teorema darebbe la stima:

$$|\pi - \pi'| \leq \exp\left(\frac{m\,(m-1)}{2N}\right) - 1 - \frac{m\,(m-1)}{2N}.$$

Poiché sappiamo che il primo membro è zero, ciò significa che l'informazione contenuta nel teorema può essere utile solo se $m^2 \ll N$, come già anticipato. Più in generale, la piccolezza della quantità a secondo membro dell'ultima disuguaglianza si può prendere come un indice della bontà dell'approssimazione fornita dal teorema per i valori in esame di m, N.

L'argomentazione che porta a costruire la stima dell'errore e dimostrare il teorema precedente si svolge in vari passi.

Passo 1. Cominciamo a riscrivere la formula per π' come segue:

$$\pi' = 1 - \left(1 - \frac{1}{M}\right)\left(1 - \frac{2}{M}\right) \cdot \ldots \cdot \left(1 - \frac{(m-1)}{M}\right).$$

Osservando quest'identità, si vede che se la riscriviamo in forma di polinomio in $\frac{1}{M}$ otteniamo:

$$\pi' = \sum_{k=1}^{m-1} (-1)^{k+1} c_{m-1,k} \frac{1}{M^k} = \sum_{k=1}^{m-1} (-1)^{k+1} c_{m-1,k} p^k \tag{5}$$

con

$$c_{m-1,k} = \sum_{1 \le j_1 < j_2 < \ldots < j_k \le m-1} j_1 j_2 \cdot \ldots \cdot j_k. \tag{6}$$

Stabiliamo subito un semplice ma utile risultato riguardo a questi coefficienti:

Lemma 3 *Vale la seguente stima per i coefficienti* $c_{m-1,k}$*:*

$$c_{m-1,k} \le \frac{1}{k!}\left[\frac{m\,(m-1)}{2}\right]^k \quad \textit{per } 1 \le k \le m-1.$$

Inoltre

$$c_{m-1,1} = \frac{m\,(m-1)}{2};$$
$$c_{m-1,m-1} = (m-1)!$$

Dimostrazione. Si ha:

$$\begin{aligned} c_{m-1,k} &= \frac{1}{k!} \cdot \sum_{\substack{j_1,j_2,\ldots,j_k=1 \\ j_1,j_2,\ldots,j_k \text{ tutti diversi}}}^{m} j_1 j_2 \cdot \ldots \cdot j_k \le \\ &\le \frac{1}{k!} \cdot \sum_{\substack{j_1,j_2,\ldots,j_k=1 \\ j_1,j_2,\ldots,j_k \text{ qualsiasi}}}^{m} j_1 j_2 \cdot \ldots \cdot j_k = \\ &= \frac{1}{k!} \sum_{j_1=1}^{m-1} j_1 \cdot \sum_{j_2=1}^{m-1} j_2 \cdot \ldots \cdot \sum_{j_k=1}^{m-1} j_k = \\ &= \frac{1}{k!}\left[\frac{m\,(m-1)}{2}\right]^k. \end{aligned} \tag{7}$$

I valori di $c_{m-1,1}, c_{m-1,m-1}$ seguono immediatamente dalla definizione (6). ■

Passo 2. Ora riscriviamo la formula (2) per π in in forma di polinomio nelle $m-1$ variabili

$$\sum_{i=1}^{N} q_i^2, \sum_{i=1}^{N} q_i^3, ..., \sum_{i=1}^{N} q_i^m.$$

Il fatto che π abbia quest'espressione si può vedere iterativamente. Vediamolo ad esempio per $m = 3$:

$$\begin{aligned}
1-\pi &= \sum_{i_1=1}^{N} q_{i_1} \sum_{\substack{i_2=1,\\ i_2\neq i_1}}^{N} q_{i_2} \sum_{\substack{i_3=1,\\ i_3\neq i_1, i_3\neq i_2}}^{N} q_{i_3} = \sum_{i_1=1}^{N} q_{i_1} \sum_{\substack{i_2=1,\\ i_2\neq i_1}}^{N} q_{i_2} \left[1-(q_{i_1}+q_{i_2})\right] \\
&= \sum_{i_1=1}^{N} q_{i_1} \sum_{\substack{i_2=1,\\ i_2\neq i_1}}^{N} q_{i_2} - \sum_{i_1=1}^{N} q_{i_1}^2 \sum_{\substack{i_2=1,\\ i_2\neq i_1}}^{N} q_{i_2} - \sum_{i_1=1}^{N} q_{i_1} \sum_{\substack{i_2=1,\\ i_2\neq i_1}}^{N} q_{i_2}^2 = \\
&= \sum_{i_1=1}^{N} q_{i_1}(1-q_{i_1}) - \sum_{i_1=1}^{N} q_{i_1}^2 (1-q_{i_1}) - \sum_{i_1=1}^{N} q_{i_1} \left[\sum_{i=1}^{N} q_i^2 - q_{i_1}^2\right] = \\
&= 1 - \sum_{i=1}^{N} q_i^2 - \sum_{i=1}^{N} q_i^2 + \sum_{i=1}^{N} q_i^3 - \sum_{i=1}^{N} q_i^2 + \sum_{i=1}^{N} q_i^3 = \\
&= 1 - 3\sum_{i=1}^{N} q_i^2 + 2\sum_{i=1}^{N} q_i^3,
\end{aligned}$$

e

$$\pi = 3\sum_{i=1}^{N} q_i^2 - 2\sum_{i=1}^{N} q_i^3.$$

All'aumentare di m la struttura del polinomio si complica (senza mostrare una regola evidente di formazione), ma si trova sempre un polinomio in queste variabili. Ad esempio per $m = 4, m = 5$, calcoli laboriosi danno:

$$\begin{aligned}
\pi_4 &= 6\sum_{i=1}^{N} q_i^2 - \left[8\sum_{i=1}^{N} q_i^3 + 3\left(\sum_{i=1}^{N} q_i^2\right)^2\right] + 6\sum_{i=1}^{N} q_i^4 \\
\pi_5 &= 10\sum_{i=1}^{N} q_i^2 - \left[20\sum_{i=1}^{N} q_i^3 + 15\left(\sum_{i=1}^{N} q_i^2\right)^2\right] + \\
&\quad + \left[30\sum_{i=1}^{N} q_i^4 + 20\left(\sum_{i=1}^{N} q_i^2\right)\left(\sum_{i=1}^{N} q_i^3\right)\right] - 24\sum_{i=1}^{N} q_i^5.
\end{aligned}$$

Diamo ora una dimostrazione generale dell'affermazione appena fatta.

Lemma 4 *La funzione π è un polinomio nelle $m-1$ variabili*

$$\sum_{i=1}^{N} q_i^2, \sum_{i=1}^{N} q_i^3, ..., \sum_{i=1}^{N} q_i^m.$$

Dimostrazione. Poniamo

$$u_\alpha = \sum_{i=1}^{N} q_i^\alpha \text{ per } \alpha = 1, 2, ..., m.$$

Per la dimostrazione è più comodo ignorare il fatto che $u_1 = 1$, e provare che π è un polinomio (omogeneo, di grado m) nelle m variabili u_α. Ponendo poi $u_1 = 1$ si avrà la tesi. Introduciamo il simbolo $Q(k, \alpha)$ per indicare una qualunque espressione del tipo

$$\sum_{i_1=1}^{N} (q_{i_1})^{\alpha_1} \sum_{\substack{i_2=1,\\ i_2 \neq i_1}}^{N} (q_{i_2})^{\alpha_2} ... \sum_{\substack{i_k=1,\\ i_k \neq i_1, ..., i_k \neq i_{k-1}}}^{N} (q_{i_k})^{\alpha_k}$$

per $\alpha_1 + \alpha_2 + ... + \alpha_k = \alpha$. L'indice k indica quindi il numero di sommatorie mentre α è il grado complessivo del polinomio nelle q_i. Se nella sommatoria più interna poniamo

$$\sum_{\substack{i_k=1,\\ i_k \neq i_1, ..., i_k \neq i_{k-1}}}^{N} (q_{i_k})^{\alpha_k} = u_{\alpha_k} - \left[(q_{i_1})^{\alpha_k} + (q_{i_2})^{\alpha_k} + ... + \left(q_{i_{k-1}}\right)^{\alpha_k} \right]$$

otteniamo

$$\begin{aligned}
&\sum_{i_1=1}^{N} (q_{i_1})^{\alpha_1} \sum_{\substack{i_2=1,\\ i_2 \neq i_1}}^{N} (q_{i_2})^{\alpha_2} ... \sum_{\substack{i_k=1,\\ i_k \neq i_1, ..., i_k \neq i_{k-1}}}^{N} (q_{i_k})^{\alpha_k} \\
&= u_{\alpha_k} \left[\sum_{i_1=1}^{N} (q_{i_1})^{\alpha_1} \sum_{\substack{i_2=1,\\ i_2 \neq i_1}}^{N} (q_{i_2})^{\alpha_2} ... \sum_{\substack{i_{k-1}=1,\\ i_{k-1} \neq i_1, ..., i_{k-1} \neq i_{k-2}}}^{N} \left(q_{i_{k-1}}\right)^{\alpha_{k-1}} \right] \\
&- \sum_{i_1=1}^{N} (q_{i_1})^{\alpha_1+\alpha_k} \sum_{\substack{i_2=1,\\ i_2 \neq i_1}}^{N} (q_{i_2})^{\alpha_2} ... \sum_{\substack{i_{k-1}=1,\\ i_{k-1} \neq i_1, ..., i_{k-1} \neq i_{k-2}}}^{N} \left(q_{i_{k-1}}\right)^{\alpha_{k-1}} \\
&- \sum_{i_1=1}^{N} (q_{i_1})^{\alpha_1} \sum_{\substack{i_2=1,\\ i_2 \neq i_1}}^{N} (q_{i_2})^{\alpha_2+\alpha_k} ... \sum_{\substack{i_{k-1}=1,\\ i_{k-1} \neq i_1, ..., i_{k-1} \neq i_{k-2}}}^{N} \left(q_{i_{k-1}}\right)^{\alpha_{k-1}} \\
&- ... - \sum_{i_1=1}^{N} (q_{i_1})^{\alpha_1} \sum_{\substack{i_2=1,\\ i_2 \neq i_1}}^{N} (q_{i_2})^{\alpha_2+\alpha_k} ... \sum_{\substack{i_{k-1}=1,\\ i_{k-1} \neq i_1, ..., i_{k-1} \neq i_{k-2}}}^{N} \left(q_{i_{k-1}}\right)^{\alpha_{k-1}+\alpha_k}
\end{aligned}$$

ossia troviamo la seguente identità iterativa:

$$Q(k, \alpha) = u_\beta Q(k-1, \alpha - \beta) - \sum Q_i(k-1, \alpha).$$

per qualche $\beta \in \{1, 2, ..., \alpha - 1\}$. Ad ogni passo iterativo la nuova espressione è somma di termini ognuno contenente $k-1$ sommatorie anziché k, mentre il grado complessivo nelle q_i rimane costante. Al passo iniziale dell'iterazione abbiamo

$$\pi = u_1^m - Q(m, m).$$

Iterativamente si ha:

$$\begin{aligned} u_1^m - \pi &= u_1 Q(m-1, m-1) - \sum Q_i(m-1, m) \\ &= u_1 \left[u_1 Q(m-2, m-2) - \sum Q_i(m-2, m-1) \right] \\ &- \sum_i \left[u_{\beta_i} Q_i(m-2, m-\beta_i) - \sum_j Q_{ij}(m-2, m) \right] \\ &= (...) \end{aligned}$$

Perciò dopo $m-1$ passi otteniamo un polinomio nelle u_α e nei termini del tipo $Q(1, \beta)$, cioè u_β. Il tutto è un polinomio nelle u_α. ■

Ora ragioniamo così. La formula trovata è vera qualunque siano i valori $q_1, q_2, ..., q_N$; se in particolare scegliamo i q_i tutti uguali a $1/N$ otteniamo

$$\begin{aligned} \sum_{i=1}^N q_i^2 &= N \cdot \frac{1}{N^2} = \frac{1}{N}; \\ \sum_{i=1}^N q_i^3 &= \frac{1}{N^2}; \\ &... \\ \sum_{i=1}^N q_i^m &= \frac{1}{N^{m-1}}, \end{aligned}$$

e il polinomio nelle variabili $\sum_{i=1}^N q_i^2, \sum_{i=1}^N q_i^3, ..., \sum_{i=1}^N q_i^m$ diventa un polinomio in $\frac{1}{N}$. Anzi, diventa lo stesso polinomio assegnato dalla formula (5) per $M = N$. In vista di questo, riscriviamo la formula per π raggruppando i vari addendi in polinomi omogenei rispetto alla seguente convenzione: la variabile $\sum_{i=1}^N q_i^k$ ha grado $k-1$. Troveremo che

$$\pi = \sum_{k=1}^{m-1} (-1)^{k+1} P_k \left(\sum_{i=1}^N q_i^2, \sum_{i=1}^N q_i^3, ..., \sum_{i=1}^N q_i^m \right) \tag{8}$$

dove P_k è un polinomio, nelle variabili $\sum_{i=1}^N q_i^2$, $\sum_{i=1}^N q_i^3$, ..., $\sum_{i=1}^N q_i^m$, di grado complessivo k (nel senso spiegato). In particolare, π non contiene il termine costante, perché non lo contiene il polinomio π'. Inoltre, per il principio di identità dei polinomi,

$$P_k \left(\frac{1}{N}, \frac{1}{N^2}, ..., \frac{1}{N^{m-1}} \right) = c_{m-1,k} \frac{1}{N^k}. \tag{9}$$

In particolare

$$P_1\left(\frac{1}{N},\frac{1}{N^2},...,\frac{1}{N^{m-1}}\right)=\frac{m\left(m-1\right)}{2}\frac{1}{N}$$

da cui

$$P_1\left(\sum_{i=1}^{N}q_i^2,\sum_{i=1}^{N}q_i^3,...,\sum_{i=1}^{N}q_i^m\right)=\frac{m\left(m-1\right)}{2}\sum_{i=1}^{N}q_i^2. \tag{10}$$

Ora ci occorre raffinare l'informazione su P_k nel senso seguente:

Lemma 5 *Per ogni $k=1,2,...,m-1$, il polinomio P_k ha coefficienti interi positivi, perciò è la somma di $c_{m-1,k}$ monomi monici nelle variabili*

$$\sum_{i=1}^{N}q_i^2,\sum_{i=1}^{N}q_i^3,...,\sum_{i=1}^{N}q_i^m.$$

Dimostrazione. Come nella dimostrazione del Lemma 4 introduciamo le quantità u_k e non supponiamo a priori che sia $u_1=1$. Facendo riferimento alla procedura con cui, per passaggi successivi, si arriva a riscrivere π in forma di polinomio nelle variabili $u_1,u_2,...,u_m$, assegniamo ad ogni espressione

$$u_{\beta_1}u_{\beta_2}...u_{\beta_h}\sum_{i_1=1}^{N}\left(q_{i_1}\right)^{\alpha_1}\sum_{\substack{i_2=1,\\ i_2\neq i_1}}^{N}\left(q_{i_2}\right)^{\alpha_2}...\sum_{\substack{i_k=1,\\ i_k\neq i_1,...,i_k\neq i_{k-1}}}^{N}\left(q_{i_k}\right)^{\alpha_k}$$

il grado:

$$\left(\beta_1-1\right)+\left(\beta_2-1\right)+...+\left(\beta_h-1\right)+\left(\alpha_1-1\right)+\left(\alpha_2-1\right)+...+\left(\alpha_k-1\right).$$

(Si noti che questa convenzione è coerente con quella con cui abbiamo assegnato il grado a P_k). Proviamo che π è un polinomio in $u_1,u_2,...,u_m$ in cui ogni monomio di grado pari (dispari) ha coefficiente negativo (positivo, rispettivamente). Al passo iniziale abbiamo:

$$\pi=1-\sum_{i_1=1}^{N}q_{i_1}\sum_{\substack{i_2=1,\\ i_2\neq i_1}}^{N}q_{i_2}...\sum_{\substack{i_k=1,\\ i_k\neq i_1,...,i_k\neq i_{k-1}}}^{N}q_{i_k}$$

dove la sommatoria in base alla nostra convenzione ha grado zero (pari) e difatti ha segno negativo. Mostriamo che ad ogni passo iterativo si creano nuovi monomi in modo che il segno cambia se e solo se cambia la parità del grado. Difatti nel procedimento iterativo, in un monomio

$$\sum_{i_1=1}^{N}\left(q_{i_1}\right)^{\alpha_1}\sum_{\substack{i_2=1,\\ i_2\neq i_1}}^{N}\left(q_{i_2}\right)^{\alpha_2}...\sum_{\substack{i_k=1,\\ i_k\neq i_1,...,i_k\neq i_{k-1}}}^{N}\left(q_{i_k}\right)^{\alpha_k}$$

di grado

$$\alpha_1 + \alpha_2 + ... + \alpha_k - k$$

si sostituisce

$$\sum_{\substack{i_k=1,\\ i_k\neq i_1,...,i_k\neq i_{k-1}}}^{N} (q_{i_k})^{\alpha_k} = u_{\alpha_k} - (q_{i_1})^{\alpha_1} - (q_{i_2})^{\alpha_2} ... - (q_{i_k})^{\alpha_k}$$

ottenendo un termine

$$u_{\alpha_k} \sum_{i_1=1}^{N} (q_{i_1})^{\alpha_1} \sum_{\substack{i_2=1,\\ i_2\neq i_1}}^{N} (q_{i_2})^{\alpha_2} ... \sum_{\substack{i_{k-1}=1,\\ i_k\neq i_1,...,i_{k-1}\neq i_{k-2}}}^{N} (q_{i_k})^{\alpha_{k-1}}$$

di grado

$$(\alpha_k - 1) + (\alpha_1 + \alpha_2 + ... + \alpha_{k-1} - (k-1)),$$

quindi con lo stesso grado del monomio di partenza e lo stesso segno, e k termini del tipo

$$-\sum_{i_1=1}^{N} (q_{i_1})^{\alpha_1} \sum_{\substack{i_2=1,\\ i_2\neq i_1}}^{N} (q_{i_2})^{\alpha_2} ... \sum_{\substack{i_j=1,\\ i_j\neq i_1,...,i_{j-1}}}^{N} \left(q_{i_j}\right)^{\alpha_j+\alpha_k} ... \sum_{\substack{i_{k-1}=1,\\ i_k\neq i_1,...,i_{k-1}\neq i_{k-2}}}^{N} (q_{i_k})^{\alpha_{k-1}}$$

che hanno grado

$$\alpha_1 + \alpha_2 + ... + \alpha_k - (k-1),$$

quindi un grado in più del monomio di partenza, e segno opposto.

Alla fine dell'iterazione poniamo $u_1 = 1$ e vediamo che π è somma di monomi di grado pari con segno negativo e grado dispari con segno positivo. Perciò nella scrittura (8) ogni P_k è somma di monomi a coefficienti interi positivi. ■

Passo 3. Con queste in formazioni su π e π', ora ne stimiamo la differenza. Ricordiamo che

$$\frac{1}{M} = p = \sum_{i=1}^{N} q_i^2$$

perciò

$$\begin{aligned}\pi - \pi' &= \sum_{k=1}^{m-1} (-1)^{k+1} \left[P_k \left(\sum_{i=1}^{N} q_i^2, \sum_{i=1}^{N} q_i^3, ..., \sum_{i=1}^{N} q_i^m \right) - c_{m-1,k} \frac{1}{M^k} \right] \\ &= \sum_{k=2}^{m-1} (-1)^{k+1} \left[P_k \left(\sum_{i=1}^{N} q_i^2, \sum_{i=1}^{N} q_i^3, ..., \sum_{i=1}^{N} q_i^m \right) - c_{m-1,k} \frac{1}{M^k} \right]\end{aligned}$$

cioè il primo termine si elide per la (10). Ora maggioriamo

$$|\pi - \pi'| \leq \sum_{k=2}^{m-1} \left| P_k \left(\sum_{i=1}^{N} q_i^2, \sum_{i=1}^{N} q_i^3, ..., \sum_{i=1}^{N} q_i^m \right) - c_{m-1,k} p^k \right|. \qquad (11)$$

In base al Lemma 5, il polinomio P_k è la somma di $c_{m-1,k}$ monomi a coefficiente 1 nelle variabili $\sum_{i=1}^N q_i^2$, $\sum_{i=1}^N q_i^3$,..., $\sum_{i=1}^N q_i^m$. Indicando con

$$M_k\left(\sum_{i=1}^N q_i^2, \sum_{i=1}^N q_i^3, ..., \sum_{i=1}^N q_i^m\right) \tag{12}$$

uno qualsiasi di questi monomi di grado k, vedremo ora come stimare $\left|M_k - p^k\right|$; basterà poi moltiplicare la quantità trovata per $c_{m-1,k}$ per ottenere una stima per ciascun addendo di (11). Cominciamo ad osservare che:

Lemma 6 *Per ogni $k \geq 2$ si ha*

$$\left|M_k - p^k\right| \leq p\left(q_{\max}\right)^{k-1}. \tag{13}$$

Dimostrazione. E' sufficiente maggiorare un fattore di M_k al modo seguente

$$\sum_{i=1}^N q_i^{2+h} \leq \left(q_{\max}\right)^h \sum_{i=1}^N q_i^2 = p\left(q_{\max}\right)^h,$$

e poi maggiorare tutti gli altri fattori q_i semplicemente con $q_{\max}$. Questo dimostra che

$$M_k \leq p\left(q_{\max}\right)^{k-1}.$$

D'altro canto è ovviamente $p^k \leq p\left(q_{\max}\right)^{k-1}$, e poiché M_k è positivo, segue la tesi. ■

La stima precedente, semplice anche se grossolana, permette di stabilire subito una stima dell'errore:

Dimostrazione del Teorema 1, punto a. Per il Lemma 6, il Lemma 3 e la (11) si ha:

$$\begin{aligned}
|\pi - \pi'| &\leq \sum_{k=2}^{m-1} \frac{1}{k!}\left(\frac{m\left(m-1\right)}{2}\right)^k p\left(q_{\max}\right)^{k-1} \\
&\leq \frac{p}{q_{\max}} \sum_{k=2}^{\infty} \frac{1}{k!}\left(\frac{m\left(m-1\right)}{2} q_{\max}\right)^k \\
&= \frac{p}{q_{\max}}\left\{\exp\left(\frac{m\left(m-1\right)}{2} q_{\max}\right) - 1 - \frac{m\left(m-1\right)}{2} q_{\max}\right\}.
\end{aligned}$$

■

Per dimostrare il punto (b) del Teorema 1, anziché il Lemma 6 ci occorrono stime di diversa raffinatezza per valori diversi di k:

Lemma 7 *Per ogni monomio monico di "grado" k nei momenti, vale la disuguaglianza:*

$$M_k\left(\sum_{i=1}^N q_i^2, \sum_{i=1}^N q_i^3, ..., \sum_{i=1}^N q_i^m\right) \leq \sum_{i=1}^N \left(q_i\right)^{n+1} \cdot \left(q_{\max}\right)^{k-n} \quad \textit{per ogni } k \geq n \geq 2. \tag{14}$$

Inoltre:

$$p^k \leq \sum_{i=1}^{N} (q_i)^{n+1} \cdot (q_{\max})^{k-n} \quad \textit{per ogni } k \geq n \geq 2 \tag{15}$$

e quindi

$$\left| M_k - p^k \right| \leq E\left(X^n\right) \cdot (q_{\max})^{k-n}$$

dove si è posto sinteticamente

$$E\left(X^n\right) = \sum_{i=1}^{N} (q_i)^{n+1}.$$

Dimostrazione. Proviamo la (14) per induzione su n. Per $n = 2$, sia M_k un monomio monico di grado 2. Può essere solo di due tipi:

$$\left(\sum_{i=1}^{N} q_i^2 \right)^2 \text{ oppure } \sum_{i=1}^{N} q_i^3.$$

e la tesi è mostrare che questo è $\leq \sum_{i=1}^{N} q_i^3$. Nel secondo caso la tesi è ovvia, nel primo la disuguaglianza

$$\left(\sum_{i=1}^{N} (q_i)^2 \right)^2 \leq \sum_{i=1}^{N} (q_i)^3$$

si può interpretare come disuguaglianza integrale nello spazio di probabilità:

$$\left(\int X dP \right)^2 \leq \int X^2 dP$$

e questa è la disuguglianza di Hölder. Se ora M_k è un monomio di grado maggiore di 2, contiene almeno un termine

$$\sum_{i=1}^{N} q_i^{3+h}$$

con $h > 0$, che si maggiora con

$$(q_{\max})^h \sum_{i=1}^{N} q_i^3,$$

e l'intero monomio si maggiora con

$$\sum_{i=1}^{N} (q_i)^3 \cdot (q_{\max})^{k-2}.$$

Supponiamo ora che la tesi sia vera per $n-1$ e dimostriamolo per n. Sia dunque M_k un monomio di grado $k \geq n \geq 3$. Supponiamo prima che il monomio sia esattamente di grado n. Primo caso: il monomio è

$$\sum_{i=1}^{N} q_i^{n+1}$$

allora la tesi è banale. Secondo caso: il monomio contiene soltanto potenze di termini del tipo

$$\sum_{i=1}^{N} (q_i)^k$$

con $k \leq n$. Allora il monomio è prodotto di due monomi di gradi n_1, n_2 con $2 \leq n_1 \leq n_2 \leq n-1$, a ciascuno dei quali possiamo applicare l'ipotesi induttiva, ottenendo

$$\begin{aligned} M_n &\leq \left[\sum_{i=1}^{N} (q_i)^{n_1+1}\right] \left[\sum_{i=1}^{N} (q_i)^{n_2+1}\right] \\ &= \left[\int X^{n_1} dP\right] \left[\int X^{n_2} dP\right] \\ &\leq \int X^{n_1+n_2} dP = \sum_{i=1}^{N} q_i^{n+1} \end{aligned}$$

dove l'ultima disuguaglianza si dimostra così:

$$\begin{aligned} \int X^{n_1} dP &\leq \left(\int X^n dP\right)^{\frac{n_1}{n}}; \\ \int X^{n_2} dP &\leq \left(\int X^n dP\right)^{\frac{n_2}{n}}, \end{aligned}$$

e moltiplicando termine a termine

$$\left[\int X^{n_1} dP\right] \left[\int X^{n_2} dP\right] \leq \int X^{n_1+n_2} dP.$$

La tesi quindi è provata se il monomio ha grado esattamente n. Se ha grado $k > n$, si maggiora con

$$(q_{\max})^{k-n} \cdot M_n,$$

dove

$$M_n \leq \sum_{i=1}^{N} (q_i)^{n+1}$$

per l'argomento precedente. Quindi la (14) è dimostrata. Quanto alla (15), si ha, per $k \geq n$:

$$p^k = E(X)^k = \left(\int X dP \right)^k \leq \int X^k dP \leq \int X^n \left(q_{\max} \right)^{k-n} dP$$
$$= \left(q_{\max} \right)^{k-n} E(X^n) = \sum_{i=1}^{N} \left(q_i \right)^{n+1} \cdot \left(q_{\max} \right)^{k-n}$$

ed il lemma è dimostrato. ■

Passo 4. Torniamo ora alla stima a priori dell'errore. Possiamo scrivere, per un intero $n \in [2, m-1]$ da scegliersi in seguito:

$$|\pi - \pi'| \leq \sum_{k=2}^{m-1} \left| P_k \left(\sum_{i=1}^{N} q_i^2, \sum_{i=1}^{N} q_i^3, ..., \sum_{i=1}^{N} q_i^m \right) - c_{m-1,k} p^k \right|$$
$$\leq \sum_{k=2}^{n} (...) + \sum_{k=n+1}^{m} (...) \equiv A_n + B_n.$$

Applicando ora il Lemma 7 abbiamo:

$$A_n \leq \sum_{k=2}^{n} c_{m-1,k} E\left(X^k \right)$$
$$B_n \leq \sum_{k=n+1}^{m-1} c_{m-1,k} E\left(X^n \right) \left(q_{\max} \right)^{k-n}$$
$$|\pi - \pi'| \leq \sum_{k=2}^{n} c_{m-1,k} E\left(X^k \right) + \frac{E\left(X^n \right)}{\left(q_{\max} \right)^n} \sum_{k=n+1}^{m-1} c_{m-1,k} \cdot \left(q_{\max} \right)^k . \qquad (16)$$

L'utilità di questo spezzamento sta nel fatto seguente: la somma

$$\sum_{k=2}^{m-1} c_{m-1,k} \cdot \left(q_{\max} \right)^k$$

può avere i primi termini molto grandi; è bene quindi stimare almeno i primi termini di $|\pi - \pi'|$ in maniera più accurata piuttosto che maggiorando semplicemente ogni q_i con il massimo; per far questo, utilizziamo i momenti di ordine superiore (che dovrebbero essere molto piccoli perché la variabile X ha valori $\ll 1$). La quantità $\frac{E(X^n)}{(q_{\max})^n}$ che compare davanti al resto della somma può dare un ulteriore guadagno fintanto che $E(X^n) \ll \left(q_{\max} \right)^n$; d'altro canto non possiamo pensare realisticamente di calcolare il valore numerico di tutti i momenti di X fino a m. Questo il motivo per cui pensiamo di scegliere un $n > 2$ ma $\ll m$.

Per il Lemma 3 abbiamo allora:

$$
\begin{aligned}
|\pi - \pi'| &\leq \sum_{k=2}^{n} \frac{1}{k!} \left[\frac{m\,(m-1)}{2}\right]^k E\left(X^k\right) + \\
&+ \frac{E\left(X^n\right)}{(q_{\max})^n} \sum_{k=n+1}^{m-1} \frac{1}{k!} \left[\frac{m\,(m-1)}{2} q_{\max}\right]^k \\
&\leq \sum_{k=2}^{n} \frac{1}{k!} \left[\frac{m\,(m-1)}{2}\right]^k E\left(X^k\right) + \\
&+ \frac{E\left(X^n\right)}{(q_{\max})^n} \left\{ \exp\left[\frac{m\,(m-1)}{2} q_{\max}\right] - \sum_{k=0}^{n} \frac{1}{k!} \left[\frac{m\,(m-1)}{2} q_{\max}\right]^k \right\} \\
&\equiv E_1^n + E_2^n
\end{aligned}
$$

Con ciò il punto (b) del teorema è dimostrato. Si tratterà ora di valutare nel caso concreto in esame, in base al valore numerico di $p, q_{\max}, m$ e dei momenti $E\left(X^n\right)$, quale scelta di n rende l'errore accettabile.

Infine, se prima dell'ultima disuguaglianza poniamo $n = m - 1$ otteniamo

$$
\begin{aligned}
|\pi - \pi'| &\leq \sum_{k=2}^{m-1} \frac{1}{k!} \left[\frac{m\,(m-1)}{2}\right]^k E\left(X^k\right) = E\left(\sum_{k=2}^{m-1} \frac{1}{k!} \left[\frac{m\,(m-1)}{2}\right]^k X^k\right) \\
&\leq E\left(\exp\left(\frac{m\,(m-1)}{2} X\right) - 1 - \left(\frac{m\,(m-1)}{2} X\right)\right),
\end{aligned}
$$

che è il punto (c) del teorema. Con ciò il teorema è completamente dimostrato.

2 Applicazione al problema dei profili coincidenti in un database

Vediamo ora un'applicazione del risultato generale al seguente problema, considerato in [1]: calcolare la probabilità che, all'interno di un database dei profili DNA di 65000 individui ce ne siano almeno due uguali, sapendo che il database contiene profili su 9 loci, e la popolazione da cui il database è tratto segue la

distribuzione allelica riportata nella tabella seguente.

n° alleli	Locus 1	Locus 2	Locus 3	Locus 4	Locus 5	Locus 6	Locus 7	Locus 8	Locus 9
1	0,258	0,271	0,184	0,316	0,251	0,179	0,389	0,319	0,287
2	0,241	0,227	0,174	0,225	0,229	0,167	0,336	0,275	0,227
3	0,217	0,212	0,154	0,137	0,160	0,140	0,160	0,111	0,142
4	0,148	0,109	0,143	0,111	0,089	0,121	0,068	0,087	0,140
5	0,114	0,084	0,142	0,085	0,085	0,119	0,027	0,082	0,140
6	0,010	0,082	0,080	0,080	0,078	0,119	0,015	0,070	0,038
7	0,005	0,014	0,067	0,024	0,036	0,061	0,002	0,056	0,020
8	0,005	0,002	0,031	0,012	0,032	0,031	0,002		0,007
9	0,002		0,009	0,010	0,017	0,019			
10			0,007		0,012	0,012			
11			0,007		0,005	0,012			
12			0,002		0,002	0,007			
13			0,002		0,002	0,005			
14					0,002	0,005			
15						0,002			

Ricordiamo brevemente il contesto. Lo spazio di probabilità in cui ci muoviamo è quello dei possibili *profili* DNA degli individui di una certa popolazione. Un profilo è una 9-upla di *genotipi*, uno per ognuno dei 9 *loci*, e si suppone l'indipendenza dei genotipi che si presentano in loci diversi. A sua volta, un genotipo è una coppia non ordinata di *alleli*; in ogni locus c'è un certo numero di alleli, ognuno con una certa probabilità (v. tabella precedente). La probabilità di un genotipo costituito da due alleli uguali è il quadrato della probabilità dell'allele; la probabilità di un genotipo costituito da due alleli diversi è il doppio prodotto delle probabilità dei due alleli. A partire da queste premesse si capisce che:

è agevole calcolare la probabilità di ciascun genotipo in ciascun locus, e quindi la probabilità di un particolare profilo;

è agevole calcolare a partire dalla tabella precedente le quantità

$$\sum_{j=1}^{N} (q_j)^n$$

(almeno per i primi valori di n) perché, se pure N è molto grande, tale calcolo si può ridurre in realtà ad un calcolo sulle frequenze degli alleli (vedremo tra poco in dettaglio come si esegue questo calcolo);

invece, non è agevole eseguire calcoli che coinvolgano in modo essenziale la totalità dei profili, come è l'evento di cui ci stiamo occupando, ossia: scegliendo m profili tra gli N, almeno due di essi coincidono.

2.1 Calcolo dei momenti della variabile aleatoria X

Consideriamo lo spazio dei profili possibili, di probabilità q_i $(i = 1, 2, ..., N)$. Sia X la v.a. che vale q_i con probabilità q_i. Allora

$$E\left(X^{n-1}\right) = \sum_{j=1}^{N} (q_j)^n .$$

Vogliamo calcolare questi momenti in termini di probabilità degli alleli.

Se X_j la v.a. "probabilità del genotipo che si trova nel j-esimo locus", per l'indipendenza delle X_j è

$$E\left(X^{n-1}\right) = \prod_{j=1}^{9} E\left(X_j^{n-1}\right).$$

D'altro canto

$$E\left(X_j^{n-1}\right) = \sum_{i=1}^{k_j} \left(g_j^i\right)^n$$

dove g_j^i è la probabilità dell'i-esimo genotipo del j-esimo locus. Se ora a_j^i è la probabilità dell'i-esimo allele nel j-esimo locus, si vede facilmente, ricordando la diversa regola per il calcolo delle probabilità di genotipi omozigoti ed eterozigoti, che

$$\sum_{i=1}^{k_j} \left(g_j^i\right)^n = \sum_{i=1}^{n_j} \left(a_j^i\right)^{2n} + \sum_{h<k} \left(2a_j^h a_j^k\right)^n.$$

D'altro canto

$$\left[\sum_{i=1}^{n_j} \left(a_j^i\right)^n\right]^2 = \sum_{i=1}^{n_j} \left(a_j^i\right)^{2n} + 2\sum_{h<k} \left(a_j^h a_j^k\right)^n$$

da cui

$$\sum_{h<k} \left(2a_j^h a_j^k\right)^n = 2^{n-1}\left\{\left[\sum_{i=1}^{n_j} \left(a_j^i\right)^n\right]^2 - \sum_{i=1}^{n_j} \left(a_j^i\right)^{2n}\right\}$$

$$\sum_{i=1}^{k_j} \left(g_j^i\right)^n = 2^{n-1}\left[\sum_{i=1}^{n_j} \left(a_j^i\right)^n\right]^2 - \left(2^{n-1}-1\right)\sum_{i=1}^{n_j} \left(a_j^i\right)^{2n} \tag{17}$$

In definitiva, troviamo

$$E\left(X^{n-1}\right) = \sum_{j=1}^{N} \left(q_j\right)^n = \prod_{j=1}^{9}\left\{2^{n-1}\left[\sum_{i=1}^{n_j} \left(a_j^i\right)^n\right]^2 - \left(2^{n-1}-1\right)\sum_{i=1}^{n_j} \left(a_j^i\right)^{2n}\right\}. \tag{18}$$

La formula precedente, che generalizza ai momenti successivi la formula per il calcolo della RMP di una popolazione discussa in [1], può essere facilmente implementata in una tabella Excel per calcolare questi momenti per i primi valori di n, a partire dalla tabella delle frequenze alleliche. Riporteremo in seguito i valori numerici trovati.

2.2 Stima dell'errore

La formula (18) per $n = 2$ dà

$$EX \equiv p = 1.06 \cdot 10^{-11}.$$

Per questo valore di p e per

$$m = 65000$$

il valore approssimato della probabilità che ci interessa, in base alla (4) è

$$\pi' = 0.022143.$$

Applichiamo in questa situazione la stima dell'errore fornita dal Teorema 1, punto b.

Calcolo di E_1^n:

$$E_1^n = \sum_{k=2}^{n} \frac{1}{k!} \left[\frac{m\,(m-1)}{2}\right]^k E\left(X^k\right).$$

Sfruttando il valore dei momenti calcolati dalla tabella delle frequenze alleliche si può procedere al seguente calcolo:

n	$\sum_{j=1}^{N} (q_j)^{n+1} = E(X^n)$	$\frac{1}{n!}\left[\frac{m(m-1)}{2}\right]^n$	E_1^n
2	$1.64 \cdot 10^{-21}$	$2.23126 \cdot 10^{18}$	$3.6593 \cdot 10^{-3}$
3	$9.45 \cdot 10^{-31}$	$1.57115 \cdot 10^{27}$	$5.144 \cdot 10^{-3}$
4	$1.17 \cdot 10^{-39}$	$8.29753 \cdot 10^{35}$	$6.1148 \cdot 10^{-3}$
5	$2.32 \cdot 10^{-48}$	$3.50565 \cdot 10^{44}$	$6.9281 \cdot 10^{-3}$
6	$6.156 \cdot 10^{-57}$	$1.23426 \cdot 10^{53}$	$7.6879 \cdot 10^{-3}$
7	$1.976 \cdot 10^{-65}$	$3.72477 \cdot 10^{61}$	$8.4239 \cdot 10^{-3}$
8	$7.2256 \cdot 10^{-74}$	$9.83558 \cdot 10^{69}$	$9.1346 \cdot 10^{-3}$
9	$2.897 \cdot 10^{-82}$	$2.30859 \cdot 10^{78}$	$9.8034 \cdot 10^{-3}$
10	$1.24258 \cdot 10^{-90}$	$4.87683 \cdot 10^{86}$	$1.0409 \cdot 10^{-2}$

n	$\sum_{j=1}^{N} (q_j)^{n+1} = E(X^n)$	$\frac{1}{n!}\left[\frac{m(m-1)}{2}\right]^n$	E_1^n
11	$5.6076 \cdot 10^{-99}$	$9.36558 \cdot 10^{94}$	$1.0934 \cdot 10^{-2}$
12	$2.6322 \cdot 10^{-107}$	$1.64871 \cdot 10^{103}$	$1.1368 \cdot 10^{-2}$
13	$1.2745 \cdot 10^{-115}$	$2.67911 \cdot 10^{111}$	$1.1709 \cdot 10^{-2}$
14	$6.3271 \cdot 10^{-124}$	$4.04252 \cdot 10^{119}$	$1.1965 \cdot 10^{-2}$
15	$3.2055 \cdot 10^{-132}$	$5.69313 \cdot 10^{127}$	$1.2147 \cdot 10^{-2}$
16	$1.6515 \cdot 10^{-140}$	$7.51659 \cdot 10^{135}$	$1.2271 \cdot 10^{-2}$
17	$8.628 \cdot 10^{-149}$	$9.34033 \cdot 10^{143}$	$1.2352 \cdot 10^{-2}$
18	$4.5607 \cdot 10^{-157}$	$1.09617 \cdot 10^{152}$	$1.2402 \cdot 10^{-2}$
19	$2.4348 \cdot 10^{-165}$	$1.21875 \cdot 10^{160}$	$1.2432 \cdot 10^{-2}$
20	$1.3108 \cdot 10^{-173}$	$1.28729 \cdot 10^{168}$	$1.2449 \cdot 10^{-2}$

Calcolo di E_2^n. Ricordando che[1]

$$q_{\max} = 5.71 \cdot 10^{-9}$$

[1] Anche questo calcolo si può fare a partire dalla tabella delle frequenze alleliche, moltiplicando tra loro le massime probabilità dei genotipi sui 9 loci.

e sfruttando anche il precedente calcolo dei momenti si ha:

n	$\sum_{j=1}^{N}(q_j)^{n+1} = E(X^n)$	$(q_{\max})^n$	$\frac{E(X^n)}{(q_{\max})^n}$
1	$p = 1.06 \cdot 10^{-11}$	$5.71 \cdot 10^{-9}$	$1.8564 \cdot 10^{-3}$
2	$1.64 \cdot 10^{-21}$	$3.26 \cdot 10^{-17}$	$5.0307 \cdot 10^{-5}$
3	$9.45 \cdot 10^{-31}$	$1.86 \cdot 10^{-25}$	$5.0806 \cdot 10^{-6}$
4	$1.17 \cdot 10^{-39}$	$1.06 \cdot 10^{-33}$	$1.1038 \cdot 10^{-6}$
5	$2.32 \cdot 10^{-48}$	$6.07 \cdot 10^{-42}$	$3.8221 \cdot 10^{-7}$
6	$6.156 \cdot 10^{-57}$	$3.466 \cdot 10^{-50}$	$1.7761 \cdot 10^{-7}$
7	$1.976 \cdot 10^{-65}$	$1.079 \cdot 10^{-58}$	$1.8313 \cdot 10^{-7}$
8	$7.2256 \cdot 10^{-74}$	$1.13 \cdot 10^{-66}$	$6.3943 \cdot 10^{-8}$
9	$2.897 \cdot 10^{-82}$	$6.45 \cdot 10^{-75}$	$4.4915 \cdot 10^{-8}$
10	$1.24258 \cdot 10^{-90}$	$3.68435 \cdot 10^{-83}$	$3.3726 \cdot 10^{-8}$

n	$\sum_{j=1}^{N}(q_j)^{n+1} = E(X^n)$	$(q_{\max})^n$	$\frac{E(X^n)}{(q_{\max})^n}$
11	$5.6076 \cdot 10^{-99}$	$2.10377 \cdot 10^{-91}$	$2.6655 \cdot 10^{-8}$
12	$2.6322 \cdot 10^{-107}$	$1.2012 \cdot 10^{-99}$	2.1913×10^{-8}
13	$1.2745 \cdot 10^{-115}$	$6.8591 \cdot 10^{-108}$	$1.8581 \cdot 10^{-8}$
14	$6.3271 \cdot 10^{-124}$	$3.9166 \cdot 10^{-116}$	$1.6155 \cdot 10^{-8}$
15	$3.2055 \cdot 10^{-132}$	$2.2364 \cdot 10^{-124}$	$1.4333 \cdot 10^{-8}$
16	$1.6515 \cdot 10^{-140}$	$1.277 \cdot 10^{-132}$	$1.2933 \cdot 10^{-8}$
17	$8.628 \cdot 10^{-149}$	$7.2914 \cdot 10^{-141}$	$1.1833 \cdot 10^{-8}$
18	$4.5607 \cdot 10^{-157}$	$4.5607 \cdot 10^{-149}$	$1.0 \cdot 10^{-8}$
19	$2.4348 \cdot 10^{-165}$	$2.3773 \cdot 10^{-157}$	$1.0242 \cdot 10^{-8}$
20	$1.3108 \cdot 10^{-173}$	$1.3574 \cdot 10^{-165}$	$9.6567 \cdot 10^{-9}$

Quindi, posto

$$Z(n) = \left\{ \exp\left[\frac{m(m-1)}{2} q_{\max}\right] - \sum_{k=0}^{n} \frac{1}{k!} \left[\frac{m(m-1)}{2} q_{\max}\right]^k \right\},$$

calcoliamo:

n	$Z(n)$	$\frac{E(X^n)}{(q_{\max})^n}$	E_2^n
4	171937	$1.1038 \cdot 10^{-6}$	0.18978
5	169810	$3.8221 \cdot 10^{-7}$	0.064903
6	165532	$1.7761 \cdot 10^{-7}$	0.0294
7	158160	$1.8313 \cdot 10^{-7}$	0.028964
8	147046	6.3943×10^{-8}	0.0094026
9	132150	4.4915×10^{-8}	0.0059355
10	114182	3.3726×10^{-8}	0.0038509
11	94478.8	$2.6655 \cdot 10^{-8}$	0.0025183

n	$Z(n)$	$\frac{E(X^n)}{(q_{\max})^n}$	E_2^n
11	94478.8	$2.6655 \cdot 10^{-8}$	0.0025183
12	74673.7	2.1913×10^{-8}	0.0016363
13	56297.3	$1.8581 \cdot 10^{-8}$	0.0010461
14	40464.5	$1.6155 \cdot 10^{-8}$	0.0006537
15	27732.6	$1.4333 \cdot 10^{-8}$	0.00039749
16	18134.2	$1.2933 \cdot 10^{-8}$	0.00023453
17	11323.8	$1.1833 \cdot 10^{-8}$	0.00013399
18	6759.95	$1.0 \cdot 10^{-8}$	0.0000676
19	3862.59	$1.0242 \cdot 10^{-8}$	0.000039561
20	2115.17	$9.6567 \cdot 10^{-9}$	0.000020426

In definitiva, otteniamo la seguente stima dell'errore:

n	E_1^n	E_2^n	$E_1^n+E_2^n$
4	$6.1148 \cdot 10^{-3}$	0.18978	0.19589
5	$6.9281 \cdot 10^{-3}$	0.064903	0.071831
6	$7.6879 \cdot 10^{-3}$	0.0294	0.037088
7	$8.4239 \cdot 10^{-3}$	0.028964	0.037388
8	$9.1346 \cdot 10^{-3}$	0.0094026	0.018537
9	$9.8034 \cdot 10^{-3}$	0.0059355	0.015739
10	$1.0409 \cdot 10^{-2}$	0.0038509	0.014260

n	E_1^n	E_2^n	$E_1^n+E_2^n$
11	$1.0934 \cdot 10^{-2}$	0.0025183	0.013452
12	$1.1368 \cdot 10^{-2}$	0.0016363	0.013004
13	$1.1709 \cdot 10^{-2}$	0.0010461	0.012755
14	$1.1965 \cdot 10^{-2}$	0.0006537	0.012619
15	$1.2147 \cdot 10^{-2}$	0.00039749	0.012544
16	$1.2271 \cdot 10^{-2}$	0.00023453	0.012506
17	$1.2352 \cdot 10^{-2}$	0.00013399	0.012486
18	$1.2402 \cdot 10^{-2}$	0.0000676	0.012470
19	$1.2432 \cdot 10^{-2}$	0.000039561	0.012472
20	$1.2449 \cdot 10^{-2}$	0.000020426	0.012469

$$|\pi - \pi'| \leq 0.012469$$

ossia

$$\pi = 0.022143 \pm 0.012469,$$

cioè

$$0.009674 < \pi < 0.034612.$$

Nell'ultima tabella si osserva che E_2^n è decrescente, mentre E_n^1 è ovviamente crescente, e la somma è (almeno fino a $n = 20$) decrescente. Apparentemente quindi aumentando n la precisione della stima può aumentare ancora (sia pure molto lentamente).

Per testare in questo caso particolare la bontà della stima, osserviamo che se, per questi valori di m e N, la distribuzione fosse uniforme, applicando il punto (c) si troverebbe:

$$0 = |\pi - \pi'| \leq \exp\left(\frac{m\,(m-1)}{2N}\right) - 1 - \frac{m\,(m-1)}{2N} = 2.23133 \cdot 10^{-12}.$$

Riferimenti bibliografici

[1] M. Bramanti: Valutazioni probabilistiche sui riscontri del DNA a scopo di identificazione criminale. La Matematica nella Società e nella Cultura - Rivista dell'Unione Matematica Italiana, Serie I, Vol.II, n. 3, Dicembre 2009, pp.447-493.