Formula di campionamento di Ewens e applicazioni allo studio della biodiversità delle popolazioni

Gianluca Covini

Relatore:
Prof. Emanuele Dolera

UNIVERSITÀ
DI PAVIA

26 Settembre 2022

Indice

Introduzione

”La formula di campionamento di Ewens esemplifica l’armonia della teoria matematica, dell’applicazione statistica e della scoperta scientifica”, H. Crane

The Sampling Theory of Selectively Neutral Alleles*

W. J. Ewens†

Department of Zoology, University of Texas at Austin, Austin, Texas, 78712

Problema di Fisher-Corbet-Williams

[42]

THE RELATION BETWEEN THE NUMBER OF SPECIES AND THE NUMBER OF INDIVIDUALS IN A RANDOM SAMPLE OF AN ANIMAL POPULATION

BY R. A. FISHER (*Galton Laboratory*), A. STEVEN CORBET (*British Museum, Natural History*) AND C. B. WILLIAMS (*Rothamsted Experimental Station*)

Problema statistico: lo spazio campionario $\mathbb{X}_1$ a priori non è noto.

Partizioni casuali

Idea: traduco i dati in struttura di partizione

Partizioni di $\{a, b, c\}$:

$$\{a, b, c\};\ \{a, b\}, \{c\};\ \{a, c\}, \{b\};\ \{b, c\}, \{a\};\ \{a\}, \{b\}, \{c\}$$

$$X_1, \ldots, X_n \in \mathbb{X}_1 \longrightarrow \{1, \ldots, n\}$$

Strumento: Formula di campionamento di Ewens, una distribuzione di probabilità sulla classe delle partizioni

Variabili aleatorie scambiabili

Scambiabilità

Le v.a. $X_i : \Omega \to [a, b]$, $i = 1, 2, \ldots$ sono scambiabili se

$$P[X_1 \in A_1, \ldots, X_n \in A_n] = P[X_1 \in A_{\sigma(1)}, \ldots, X_n \in A_{\sigma(n)}]$$

$$\forall n \in \mathbb{N}, \quad \forall \sigma \in S^n, \quad \forall A_1, \ldots, A_n \in \mathcal{B}([a, b])$$

Teorema di De Finetti

Sia $\mathbb{P}$ lo spazio delle misure di probabilità su $([a,b],\mathcal{B}([a,b]))$.

Teorema (di rappresentazione di De Finetti)

$\{X_i\}_{i\geq 1}$ rappresenta una successione di variabili aleatorie scambiabili se e solo se esiste un'unica m.d.p. $\mu : \Omega \to \mathbb{P}$ t.c.

$$P[X_1 \in A_1, \ldots, X_n \in A_n] = \mathbb{E}[\mu(A_1)\ldots\mu(A_n)]$$

$$\forall A_1, \ldots, A_n \in \mathcal{B}(\mathbb{R})$$

Processo di Dirichlet

Processo di Dirichlet

Sia α una misura finita su $[a, b]$ e $\theta = \alpha([a, b])$. Il processo di Dirichlet μ si definisce tramite la densità di Dirichlet, ovvero dati $C_1, \dots, C_n$ partizione di $[a, b]$

$$q(\{(\mu(C_1), \dots, \mu(C_{n-1})) \in R\})$$

$$= \int_R \frac{\Gamma(\theta)}{\prod_{i=1}^n \Gamma(\alpha_i)} z_1^{\alpha_1 - 1} z_2^{\alpha_2 - 1} \dots z_{n-1}^{\alpha_{n-1}-1} (1 - \sum_{i=1}^{n-1} z_i)^{\alpha_n - 1} dz$$

con $R \in \mathcal{B}(\Delta_{n-1}), \alpha_i = \alpha(C_i)$.

Caso $n = 2$

Obiettivo: formula per la probabilità delle partizioni di $\{1, 2\}$ corrispondenti alle partizioni di variabili aleatorie $\{X_1 = X_2\}$ e $\{X_1 \neq X_2\}$.

- Calcoliamo $P[X_1 \in A_1, X_2 \in A_2]$ per $A_1, A_2 \in \mathcal{B}([a, b])$
- Deduciamo da $P[X_1 \in A_1, X_2 \in A_2]$ il valore di $P[X_1 = X_2]$

Legge di probabilità

Legge di probabilità

$$\begin{aligned} P[X_1 \in A_1, X_2 \in A_2] &= \mathbb{E}[\mu(A_1)\mu(A_2)] \\ &= \gamma\bar{\alpha}(A_1)\bar{\alpha}(A_2) + (1-\gamma)\bar{\alpha}(A_1 \cap A_2) \end{aligned}$$

Dove $\bar{\alpha}$ è la misura α normalizzata e $\gamma = \theta/(\theta+1)$.

Legge di probabilità

Legge di probabilità

$$P[X_1 \in A_1, X_2 \in A_2] = \mathbb{E}[\mu(A_1)\mu(A_2)]$$
$$= \gamma\bar{\alpha}(A_1)\bar{\alpha}(A_2) + (1-\gamma)\bar{\alpha}(A_1 \cap A_2)$$

Dove $\bar{\alpha}$ è la misura α normalizzata e $\gamma = \theta/(\theta+1)$.

Formula di Ewens con $n = 2$

$$P[\Pi_2 = \{1,2\}] = P[X_1 = X_2] = 1 - \gamma = \frac{1}{\theta+1}$$

$$P[\Pi_2 = \{1\},\{2\}] = P[X_1 \neq X_2] = \gamma = \frac{\theta}{\theta+1}$$

Caso generico

Deduciamo $P[X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n]$ tramite il processo di Dirichlet.

Legge di probabilità generica

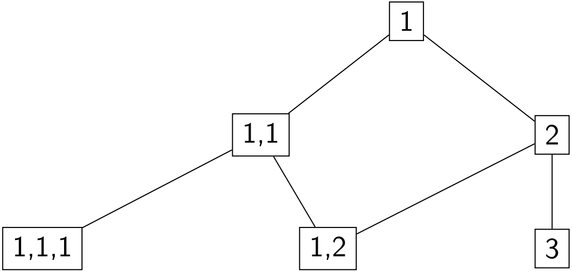
$$P[X_1 \in A_1, \ldots, X_n \in A_n] = \sum_{B \in \mathcal{P}_n} w_B \prod_{i=1}^{|B|} \bar{\alpha}(\bigcap_{j \in B_i} A_j)$$

Con $B = (B_1, \ldots, B_k) \in \mathcal{P}_n$, insieme delle partizioni di $\{1, \ldots, n\}$.

La formula di Ewens è la seguente:

$$P[\Pi_n = B] = w_B = \frac{\theta^{|B|}}{(\theta)_n \uparrow} \prod_{i=1}^{|B|} (\#B_j - 1)!$$

Processo del ristorante cinese



- $\theta/(\theta + n - 1)$: probabilità che l'n-esimo cliente occupi un tavolo vuoto;
- $n_i/(\theta + n - 1)$: probabilità che l'n-esimo cliente si sieda in un tavolo occupato già da n_i commensali.

Numero di insiemi della partizione

Probabilità di avere k insiemi

$\gamma_k^{(n)}$ è la probabilità di avere k insiemi in una partizione di $\{1, \ldots, n\}$.

$$\gamma_k^{(n)} = \frac{|s(n,k)|\theta^k}{(\theta)_{(n)\uparrow}}$$

Per i numeri di Stirling $|s(n,k)|$ vale la seguente formula:

$$(\theta)_{n\uparrow} = \theta(\theta+1)\ldots(\theta+n-1) = \sum_{k=1}^{n} |s(n,k)|\theta^k$$

Cardinalità degli insiemi della partizione

Siano $(n_1, \dots, n_k)$ tali che $n_1 \leq n_2 \leq \dots \leq n_k$ e $\sum_{j=1}^k n_j = n$.

Probabilità delle cardinalità degli insiemi

$\Gamma_k^{(n)}(n_1, \dots, n_k)$ è la probabilità che gli insiemi della partizione abbiano cardinalità $n_1, \dots, n_k$.

$$\Gamma_k^{(n)}(n_1, \dots, n_k) = \frac{n!}{|s(n,k)|} \frac{1}{\prod_{i=1}^k n_i \prod_{j=1}^n m_j!}$$

Con $m_j = \sum_{i=1}^k \mathbb{1}_{n_i = j}$

Probabilità di una data partizione

Probabilità della partizione

Fissati k e $(n_1, \ldots, n_k)$, la partizione è scelta con probabilità uniforme $1/P_k^{(n)}(n_1, \ldots, n_k)$.

$$P_k^{(n)}(n_1, \ldots, n_k) = \binom{n}{n_1 \ldots n_k} \frac{1}{\prod_{j=1}^{n} m_j!}$$

Formula di campionamento di Ewens

Teorema (formula di Ewens)

$$Ewens^{(n)}(n_1, \dots, n_k; \theta) = P[\Pi_n = B_1, \dots, B_k] = \frac{\theta^k}{(\theta)_{n\uparrow}} \prod_{j=1}^{k} (n_j - 1)!$$

Dove $n_j = \# B_j$.

Metodo Monte Carlo

Partizioni	Frequenza relativa	Valore atteso teorico
$\{1,3\},\{2\},\{4\},\{5\}$	0.0221	0.0222
$\{1,2\},\{3\},\{4\},\{5\}$	0.0215	0.0222
$\{1\},\{2,3,4,5\}$	0.0325	0.0333
$\{1\},\{2\},\{3\},\{4\},\{5\}$	0.0408	0.0444
$\{1,5\},\{2\},\{3\},\{4\}$	0.0231	0.0222
...	...	...

Metodo Monte Carlo

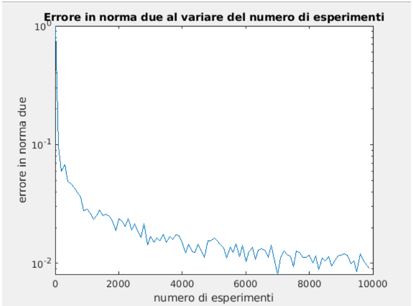
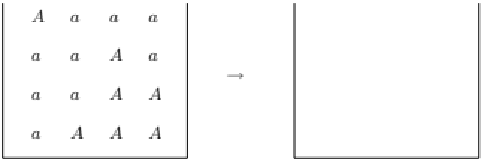


Figure: Errore metodo Monte Carlo

Modello di Wright-Fisher

Modello di Wright-Fisher

Dato un locus genico, il modello di Wright-Fisher studia la distribuzione degli alleli per il locus in una popolazione diploide di dimensione costante N con generazioni non sovrapponibili e accoppiamenti casuali.



generation n → generation $n+1$

Modello a infiniti alleli

Supponiamo che avvengano mutazioni con tasso μ e sia $\theta = 4N\mu$.

Ipotesi

Ogni mutazione genera un allele di un tipo non ancora registrato

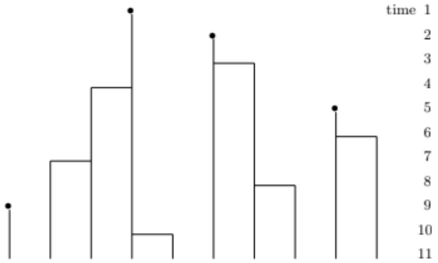
Partizione allelica

$(m_1, \dots, m_n)$ è una partizione allelica se m_j indica il numero di alleli che appaiono in j individui.

Modello a infiniti alleli

Teorema

La relazione genealogica tra k lignaggi nel modello a infiniti alleli può essere simulata attraverso un processo del ristorante cinese con k clienti.



Formula di campionamento di Ewens

Data $(m_1, \ldots, m_n)$ una partizione allelica e $\theta = 4N\mu$ tasso di mutazione scalato, allora la probabilità di ogni partizione allelica in un campione di n elementi è la seguente

Formula di campionamento di Ewens

$$p(m_1, \ldots, m_n; \theta) = \frac{n!}{(\theta)_{(n)\uparrow}} \prod_{j=1}^{n} \frac{\theta^{m_j}}{j^{m_j} m_j!}$$

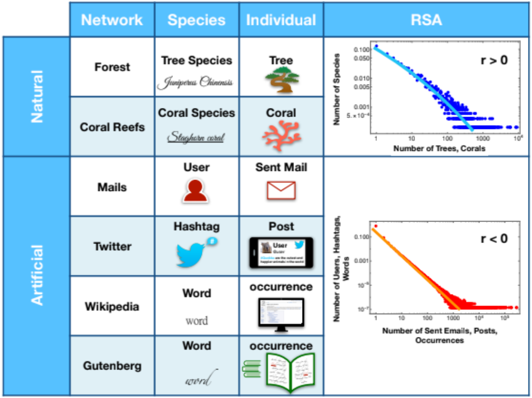
Conclusioni

- Problema di Fisher-Corbet-Williams
- Ridefinizione del problema in termini di partizioni
- Formula di campionamento di Ewens per la probabilità delle partizioni

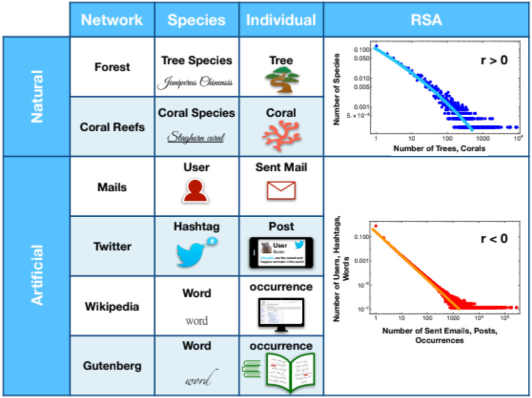
Linguistica e biodiversità

Upscaling human activity data: an ecological perspective

Anna Tovo [*a,b], Samuele Stivanello [†a], Amos Maritan[b], Samir Suweis[b,d], Stefano Favaro [‡c], and Marco Formentin [§a,d]



Linguistica e biodiversità



"Intraprendere azioni efficaci ed immediate per ridurre il degrado degli ambienti naturali, arrestare la distruzione della biodiversità", ONU, Agenda 2030