Università degli studi di Pavia
Dipartimento di Matematica
Corso di Laurea in Matematica

UNIVERSITÀ
DI PAVIA

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

Tesi di Laurea in Matematica

Relatore:
Prof. Emanuele Dolera

Tesi di Laurea di:
Gianluca Covini
Matricola 481021

Anno Accademico 2021-2022

Abstract

Uno dei problemi più rilevanti nella nostra epoca di cambiamenti climatici è lo studio della biodiversità delle popolazioni animali, cioè l'analisi della varietà delle specie all'interno di un ambiente. Un contributo alla trattazione di questo problema può venire dalla statistica bayesiana: essa, infatti, fornisce degli strumenti per stimare la distribuzione delle specie in una popolazione a partire da un campione osservato; uno di questi è la formula di campionamento di Ewens. Nel corso di questa tesi ci occupiamo della deduzione, delle proprietà e delle applicazioni della formula di Ewens: essa consiste in una formula per le probabilità delle partizioni di un insieme $\{1, \dots, n\}$ e si deduce in maniera astratta ma trova, poi, numerose applicazioni, la più nota delle quali allo studio della dinamica delle popolazioni. Apriremo la nostra trattazione dando delle nozioni introduttive di statistica bayesiana non parametrica con l'obiettivo di definire il processo di Dirichlet a partire dalle leggi finito-dimensionali; dedurremo poi, a partire da quest'ultimo e dal processo del ristorante cinese, la formula di Ewens. Presenteremo anche un metodo Monte Carlo per la costruzione sperimentale della formula tramite simulazioni in MATLAB. Infine, descriveremo il modello di Wright-Fisher, un modello di dinamica delle popolazioni in cui il risultato di Ewens trova applicazioni.

Indice

Introduzione

"La formula di campionamento di Ewens esemplifica l'armonia della teoria matematica, dell'applicazione statistica e della scoperta scientifica" [5], queste sono le parole che usa Harry Crane per riassumere la potenza della formula di campionamento di Ewens, un risultato fondamentale della statistica bayesiana di cui quest'anno celebriamo i 50 anni dalla sua derivazione originaria [7]. Questa formula, infatti, è in grado di convogliare al suo interno uno degli aspetti più entusiasmanti della matematica cioè l'armonica convivenza di astrazione con risvolti applicativi a un ampio spettro di problemi, alcuni più concreti, come lo studio della genetica delle popolazioni, motivazione originale della nascita di questa formula, altri più simili a divertissement matematici: per esempio, si può utilizzare la formula di Ewens per contare il numero di cerchi ottenuti legando casualmente tra di loro gli estremi di n spaghetti cotti [17]. Come le teorie matematiche più interessanti, la formula di Ewens rivela il legame profondo che unisce sottotraccia problemi provenienti dagli ambiti più disparati: mette in luce, cioè, un pattern che la matematica permette di vedere astraendo dalla realtà delle cose. In questo caso, il fil rouge che unisce tutte le applicazioni di questa formula è un problema statistico. Lo illustriamo, però, a partire da un caso concreto.

Nel 1943 Fisher, Corbet e Williams decidono di studiare la distribuzione delle specie di farfalle in Malesia [9]. I tre studiosi raccolgono un campione di misurazioni delle specie di farfalle osservate: un obiettivo è quello di dedurre tramite inferenza, a partire da n osservazioni $x_1, \dots, x_n$, la distribuzione delle specie nell'intera popolazione studiata. Il problema che sorgeva in questi studi consisteva nell'impossibilità di compiere inferenza con tecniche tradizionali: a differenza dalla teoria standard, infatti, in questo caso lo spazio campionario $\mathbb{X}_1$ in cui le osservazioni assumono valori non è ben definito. $\mathbb{X}_1$, infatti, avrebbe dovuto contenere tutte le possibili specie di farfalle, tuttavia durante gli studi venivano osservati individui di tipi non ancora scoperti. Il problema statistico alla base, quindi, consiste nel voler compiere studi di frequenza su uno spazio a priori non noto. Il modo di agire che si segue in questi casi è quello di tradurre i dati in una struttura di partizione, cioè dalla n-upla di osservazioni $x_1, \dots, x_n$ si passa a una partizione dell'insieme $\{1, \dots, n\}$. Essa permette di classificare le osservazioni di uno stesso tipo assegnando gli indici corrispondenti alla stessa partizione. La formula di Ewens consente, poi, di definire un modello statistico sulle partizioni di $\{1, \dots, n\}$: essa, infatti, consiste in una formula che assegna un valore di probabilità a ciascuna partizione.

La versatilità della formula di Ewens riemerge nel fatto che la sua deduzione

può avvenire in vari modi: può avvenire per via diretta e astratta, attraverso alcune strutture del processo di Dirichlet, per via ricorsiva, tramite il processo del ristorante cinese, e per via sperimentale da studi sulla genetica delle popolazioni. La deduzione originale di Ewens è proprio quest'ultima: egli, infatti, ha introdotto la sua formula per misurare le probabilità di partizioni alleliche in una popolazione.

Nella tesi presentiamo diverse di queste costruzioni. Il capitolo iniziale si occupa di alcuni aspetti preliminari, in particolare ha come scopo la costruzione del processo di Dirichlet tramite la scritture di leggi finito-dimensionali: in questo capitolo tratteremo, in particolare, con variabili aleatorie scambiabili che modellizzano l'idea delle osservazioni sperimentali in condizioni analoghe. Nel capitolo successivo affronteremo il cuore della tesi, cioè la deduzione della formula di campionamento di Ewens a partire dal processo di Dirichlet e dal processo del ristorante cinese: dopo l'analisi dettagliata di alcuni casi elementari, procederemo a una costruzione generale della formula. Inoltre, introdurremo una costruzione "Monte Carlo" della formula tramite alcune simulazioni di partizioni casuali effettuate con l'utilizzo di MATLAB. Nel terzo capitolo, poi, proporremo un modello di dinamica delle popolazioni, il modello di Wright-Fisher, in cui la formula di Ewens trova applicazione per descrivere la distribuzione delle partizioni alleliche all'interno di una popolazione. Infine, nella sezione *Conclusioni e orizzonti* trarremo le conclusioni sull'elaborato e sul problema di apertura e presenteremo alcuni possibili orizzonti di ricerca sui temi trattati. Nelle tre appendici sono riportati alcuni risultati riguardanti i numeri di Stirling e i numeri di Bell e degli approfondimenti sugli aspetti implementativi della deduzione Monte Carlo della formula.

Capitolo 1

Processo di Dirichlet

1.1 Variabili aleatorie scambiabili

In questo capitolo preliminare introduciamo alcuni concetti di statistica bayesiana non parametrica [14]: in particolare, tratteremo con variabili aleatorie scambiabili, il cui interesse viene naturale dallo studio statistico che anima la ricerca presentata nell'introduzione. Come vedremo, infatti, il concetto di scambiabilità riflette l'indifferenza dell'ordine delle osservazioni effettuate e quindi ben si addice a descrivere osservazioni sperimentali in contesti analoghi. L'obiettivo del capitolo è, poi, la costruzione del processo di Dirichlet a partire dalle leggi finito-dimensionali.

Incominciamo definendo lo spazio in cui andremo a studiare i nostri risultati. Ci poniamo in uno spazio misurabile $(\mathbb{X}_1, \chi_1)$: nel corso della trattazione questo coinciderà, di fatto, con $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ o anche $([a,b] \subseteq \mathbb{R}, \mathcal{B}([a,b]))$. Chiameremo, poi, $(\mathbb{P}, \mathcal{L})$ lo spazio metrico delle misure di probabilità sullo spazio $(\mathbb{X}_1, \chi_1)$. Sarà $(\mathbb{P}, \mathcal{P})$ lo spazio misurabile associato dove $\mathcal{P}$ è il boreliano di $\mathbb{P}$ rispetto alla metrica $\mathcal{L}$. Sia, inoltre, $(\Omega, \mathbb{F}, P)$ uno spazio di probabilità nel senso di Kolmogorov.

Consideriamo, poi, $\{X_n\}_{n\geq 1}$ delle variabili aleatorie definite da $(\Omega, \mathbb{F})$ in $(\mathbb{X}_1, \chi_1)$.

Possiamo, quindi, dare la definizione di variabili aleatorie scambiabili:

Definizione 1.1.1 (scambiabilità)**.** *Data una successione di variabili aleatorie $\{X_n\}_{n\geq 1}$, diciamo che è formata da variabili aleatorie scambiabili se vale la seguente proprietà:*

$$P[X_1 = x_1, \ldots, X_n = x_n] = P[X_1 = x_{\sigma(1)}, \ldots, X_n = x_{\sigma(n)}]$$
$$\forall n \in \mathbb{N}, \quad \forall \sigma \in S^n, \quad \forall x_1, \ldots, x_n \in \{0,1\}$$

La definizione indica che per le variabili aleatorie $\{X_n\}_{n\geq 1}$ che andremo a considerare non è rilevante l'ordine con cui vengono raccolte le osservazioni ma solo il risultato delle osservazioni stesse; possiamo immaginare questo come il caso in cui le osservazioni vengono effettuate tutte in condizioni analoghe. Capiamo, quindi, il collegamento con il problema presentato nell'introduzione in cui le misure effettuate avvengono sempre nello stesso contesto.

Un caso particolarmente rilevante di variabili scambiabili è quello delle variabili indipendenti e identicamente distribuite. Tramite la scambiabilità si generalizza il concetto di variabili i.i.d. È possibile generalizzare anche uno dei risultati più importanti per variabili aleatorie i.i.d., cioè la legge dei grandi numeri, che assume la forma che segue.

Teorema 1.1.1 (legge forte dei grandi numeri)**.** *Siano* $\{X_n\}_{n\geq 1}$ *una successione di variabili aleatorie scambiabili definite da* Ω *in* $\mathbb{R}$*.*

Allora esiste una misura di probabilità

$$\mu : (\Omega, \mathbb{F}) \to (\mathbb{P}, \mathcal{L})$$

tale per cui per ogni funzione continua e limitata $g : \mathbb{R} \to \mathbb{R}$ *vale che*

$$\lim_{n\to+\infty} \frac{1}{n} \sum_{i=1}^{n} g(X_i) = \int_{\mathbb{R}} |g(t)|\, d\mu(t)$$

quasi certamente e in L^2*.*

Vale, inoltre, che la misura μ si può caratterizzare anche grazie al teorema di *portmanteau* come limite di convergenza debole di $\frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}$.

La legge dei grandi numeri porta alcune implicazioni immediatamente visibili: una caratterizzazione equivalente ma forse di più facile intuizione, per esempio, si ha riscrivendo l'integrale limite come valore atteso. Otteniamo quindi

$$\lim_{n\to+\infty} \frac{1}{n} \sum_{i=1}^{n} g(X_i) = \mathbb{E}[|g(X_1)|]$$

Si nota in questo modo anche la differenza rispetto al teorema analogo per variabili aleatorie i.i.d. che consiste nel fatto che il termine di destra è un termine aleatorio.

A partire dalla legge dei grandi numeri per variabili aleatorie scambiabili si ricava, poi, la legge per variabili aleatorie i.i.d. come caso particolare.

1.2 Costruzione di misure su $(\mathbb{P}, \mathcal{L})$

Una questione che vogliamo affrontare, a questo punto, è la definizione di una misura di probabilità q su $(\mathbb{P}, \mathcal{L})$, cioè, di fatto, la definizione di una misura di probabilità su uno spazio di misure di probabilità.

Possiamo caratterizzare la funzione q come legge di μ, la misura di probabilità da $(\Omega, \mathbb{F})$ a $(\mathbb{P}, \mathcal{L})$ della legge dei grandi numeri. Questo significa che, chiamata $\mathcal{P}$ l'algebra di Borel di $\mathbb{P}$, vale che:

$$\forall B \in \mathcal{P} \quad q(B) := P(\mu^{-1}(B)) = P(\mu \in B)$$

Dove ricordiamo che P è una misura di probabilità sullo spazio $(\Omega, \mathbb{F})$

Conseguenza di questa definizione è che q permette di caratterizzare lo spazio $(\mathbb{P}, \mathcal{P}, q)$ come spazio di probabilità nel senso di Kolmogorov. Fa uso di questa caratterizzazione il seguente risultato fondamentale:

Teorema 1.2.1 (di rappresentazione di De Finetti)**.** $\{X_n\}_{n\geq 1}$ *rappresenta una successione di variabili aleatorie scambiabili se e solo se*

$$P[X_1 \in A_1, \ldots, X_n \in A_n] = \int_{\mathbb{P}} \prod_{i=1}^{n} p(A_i)\, dq(p) \quad \forall A_1, \ldots, A_n \in \mathcal{B}(\mathbb{R})$$

Il teorema di rappresentazione è un risultato molto importante: esso, infatti, ci permette di caratterizzare la legge delle variabili $\{X_n\}_{n\geq 1}$ utilizzando solamente la misura q. Per rendere più intuitiva la caratterizzazione, la si può dare tramite la misura μ, sfruttando la definizione di q data precedentemente. Varrà, infatti, che, per ogni $h : \mathbb{P} \to \mathbb{R}$ continua e limitata, $\mathbb{E}[h(\mu)] = \int_{\mathbb{P}} h(p)\, dq(p)$ per definizione della speranza matematica.

Quindi, riscrivendo l'integrale del teorema, otteniamo che:

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

1.3 Esempi di misure su $(\mathbb{P}, \mathcal{L})$

La definizione di q data finora non ci aiuta nella costruzione concreta di misure di questo tipo dato che essa è stata caratterizzata tramite una misura μ di cui abbiamo riportato un risultato di esistenza non costruttivo.

Vogliamo, quindi, descrivere alcuni esempi di misure q. Per fare ciò usiamo l'approccio seguito da Ferguson in [8], cioè costruiamo delle leggi di probabilità finito-dimensionali e sfruttiamo, poi, un teorema astratto che garantisce esistenza e unicità di q associandola a una di queste leggi.

Per costruire le leggi finito-dimensionale procediamo nel seguente modo:

- Consideriamo una variabile $m \in \mathbb{N},\ m \geq 2$.
- Fissata m, consideriamo una partizione di $\mathbb{R}$ costituita da m sottoinsiemi a due a due disgiunti $C_1, \ldots, C_m$ con $C_i \in \mathcal{B}(\mathbb{R})$.
- Consideriamo il vettore aleatorio $(\mu(C_1), \ldots, \mu(C_m)) \in [0,1]^m$, con $\sum_{i=1}^{m} \mu(C_i) = 1$. Notiamo che $V_{C_1,\ldots,C_{m-1}} = (\mu(C_1), \ldots, \mu(C_{m-1}) \in \Delta_{m-1}$ e possiamo, quindi, considerare solamente il vettore aleatorio $V_{C_1,\ldots,C_m} : \Omega \to \Delta_{m-1}$.
- A questo punto, possiamo sfruttare il fatto che $\Delta_{m-1} \subseteq [0,1]^{m-1}$, su cui pongo la misura di Lebesgue $(m-1)$-dimensionale.
- Possiamo, quindi, infine, definire una densità $\phi_{C_1,\ldots,C_{m-1}}$ su Δ_{m-1} e costruire, in maniera analoga rispetto a quanto si faceva su $\mathbb{R}$, la legge del vettore aleatorio $V_{C_1,\ldots,C_{m-1}}$. Quindi, data P una probabilità su $(\Delta_{m-1}, \mathcal{B}(\Delta_{m-1}))$, avremo che

$$P[V_{C_1,\ldots,C_{m-1}} \in D] = \int_D \phi_{C_1,\ldots,C_{m-1}}(x)\, dx \quad \forall D \in \Delta_{m-1}$$

Il teorema di Ferguson permette, poi, di associare in maniera univoca una misura di probabilità q su $(\mathbb{P}, \mathcal{L})$ a una densità $\phi_{C_1,\dots,C_{m-1}}$ costruita come legge finito-dimensionale.

Prima di enunciare il teorema, però, è necessario dare la seguente definizione:

Definizione 1.3.1 (condizioni di compatibilità)**.** *Sia* $\Phi = \{\phi_{C_1,\dots,C_{m-1}},\ m \in \mathbb{N};\ m \geq 2;\ C_1, \dots, C_{m-1}, C_m$ *partizione di* $\mathbb{R}\}$ *un sistema di densità. Siano* $\mathcal{Q} = \{q_{C_1,\dots,C_{m-1}}\}$ *le misure di probabilità associate. Il sistema di densità si dice compatibile se valgono le seguenti condizioni:*

(1) $\forall A \in \mathcal{B}(\Delta_{m-1}) \quad q_{C_1,\dots,C_{m-1}}(A) = q_{C_{\sigma(1)},\dots,C_{\sigma(m-1)}}(\sigma(A))$
con $\sigma \in S^{m-1}$ *e* $\sigma(A) = \{(x_{\sigma(1)}, \dots, x_{\sigma(m-1)}) : (x_1, \dots, x_{m-1}) \in A\}$*;*

(2) $q_{\mathbb{R}} = \delta_1$*;*

(3) se si considera $B_1, \dots, B_n$ *una partizione di* $\mathbb{R}$ *più fine di* $C_1, \dots, C_m$ *è possibile passare da* $q_{C_1,\dots,C_{m-1}}$ *a* $q_{B_1,\dots,B_{n-1}}$ *tramite una trasformazione delle variabili aleatorie associate;*

(4) se $C_n \in \mathcal{B}(\mathbb{R})$ *per ogni* n *e se* $\{C_n\}_{n\geq 1}$ *è una successione decrescente di insiemi con* $\bigcap_{n=1}^{\infty} C_n = \emptyset$*, allora* q_{A_n} *converge debolmente a* δ_0*.*

Possiamo quindi enunciare il teorema di Ferguson:

Teorema 1.3.1 (di Ferguson)**.** *Se la famiglia di densità* $\Phi = \{\phi_{C_1,\dots,C_{m-1}},\ m \in \mathbb{N};\ m \geq 2;\ C_1, \dots, C_{m-1}, C_m$ *partizione di* $\mathbb{R}\}$ *è compatibile, allora per ogni densità* $\phi_{C_1,\dots,C_{m-1}}$ *esiste un'unica misura di probabilità* $\mu : (\mathbb{R}, \mathcal{B}(\mathbb{R})) \to (\mathbb{P}, \mathcal{P})$ *avente legge* $\phi_{C_1,\dots,C_{m-1}}$*.*

La probabilità q su $(\mathbb{P}, \mathcal{P})$ che cercavamo si definisce, poi, come legge di μ, in maniera analoga a quanto fatto precedentemente.

1.4 Misura di Dirichlet

A questo punto possiamo usare le leggi di dimensione finita per costruire una misura esplicita q su $(\mathbb{P}, \mathcal{L})$.

La strategia più comune per definire la misura q, infatti, è quella di dare la famiglia di densità compatibili $\Phi = \{\phi_{C_1,\dots,C_{m-1}},\ m \in \mathbb{N};\ m \geq 2;\ C_1, \dots, C_{m-1}, C_m\}$.

Costruiamo, quindi, queste leggi secondo la procedura illustrata precedentemente, andando, in particolare, a definire esplicitamente la distribuzione della probabilità P.

Ripercorrendo i passi precedenti definiamo esplicitamente la legge della probabilità P del vettore aleatorio $V_{C_1,\dots,C_{n-1}} = (\mu(C_1), \dots, \mu(C_{n-1}))$:

$$\forall n \in \mathbb{N},\ n \geq 2,\ \forall C_1, \dots, C_{n-1} \text{ partizione di } \mathbb{R} \text{ e } \forall D \in \mathcal{B}(\Delta_{n-1})$$

$$P[V_{C_1,\dots,C_{n-1}} \in D] = \int_D \phi_{C_1,\dots,C_{n-1}}(x)\, dx$$

Si tratta, quindi, di dare una definizione esplicita delle leggi $\phi_{C_1,\dots,C_{n-1}}$, richiedendo su di esse solamente le condizioni di compatibilità.

Una costruzione notevole che sfrutta questo metodo è quella della misura di Dirichlet. Essa parte assegnando una misura finita α su $\mathbb{R}$. Per comodità di notazione, introduciamo le seguenti quantità:

- $\theta := \alpha(\mathbb{R})$
- $\bar{\alpha}(C) := \alpha(C)/\theta \quad \forall C \in \mathcal{B}(\mathbb{R})$

Chiaramente varrà che $0 < \theta < +\infty$. Notiamo anche che $\bar{\alpha}$ non è altro che la misura normalizzata di α e sarà, quindi, una misura di probabilità. In particolare, varrà che $\alpha(C) = \theta\bar{\alpha}(C) \quad \forall C \in \mathcal{B}(\mathbb{R})$.

A questo punto, definiamo le seguenti leggi di probabilità:

Definizione 1.4.1 (leggi di Dirichlet)**.** *$\forall n \in \mathbb{N},\ n \geq 2,\ \forall C_1, \dots, C_n$ partizione di $\mathbb{R}$ tale per cui $\alpha(C_i) > 0, \quad \forall i = 1, \dots, n$ chiamiamo leggi di probabilità di Dirichlet le leggi di probabilità definite nel seguente modo:*

$$\phi_{C_1,\dots,C_{n-1}}(z_1, \dots, z_{n-1}) = \frac{\Gamma(\theta)}{\prod_{i=1}^{n}\Gamma(\alpha_i)} z_1^{\alpha_1-1} z_2^{\alpha_2-1} \dots z_{n-1}^{\alpha_{n-1}-1} (1 - \sum_{i=1}^{n-1} z_i)^{\alpha_n - 1}$$

con $(z_1, \dots, z_{n-1}) \in \Delta_{n-1}$ e $\alpha_i := \alpha(C_i)$

Ponendo $n = 2$ si ritrova un caso notevole: la legge di Dirichlet, infatti, assume la forma

$$\phi_{C_1}(z_1) = \frac{\Gamma(\theta)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} z_1^{\alpha_1-1}(1 - z_1)^{\alpha_2-1}, \quad z_1 \in \Delta_1 = [0, 1]$$

che coincide con la densità della distribuzione beta.

Possiamo, ora, studiare la compatibilità della famiglia delle leggi di Dirichlet appena definita.

La condizione (1) è ovvia: permutando contemporaneamente gli z_i e gli insiemi C_i, il valore della densità, infatti, rimane lo stesso. La condizione (2) chiediamo che sia soddisfatta per definizione: infatti abbiamo definito le leggi di Dirichlet solamente per $n \geq 2$; poniamo, quindi, per $n = 1$

$$q_{\mathbb{R}} = \delta_1$$

Verifichiamo, poi, la condizione (3) in un esempio.

Partiamo dalla partizione C_1, C_2, C_3 e costruiamo una nuova partizione

$$C_1' = C_1 \cup C_2, \quad C_2' = C_3$$

È semplice notare che il vettore aleatorio $(\mu(C_1), \mu(C_2))$ e $\mu(C_1')$ sono legati dalla seguente relazione

$$\mu(C_1') = \mu(C_1 \cup C_2) = \mu(C_1) + \mu(C_2)$$

Possiamo, quindi, verificare che le densità relative sono legate da una somma di variabili aleatorie.

In particolare le densità saranno rispettivamente:

$$\phi_{C_1,C_2}(z_1,z_2)=\frac{\Gamma(\theta)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)}z_1^{\alpha_1-1}z_2^{\alpha_2-1}(1-z_1-z_2)^{\alpha_3-1}$$

e

$$\phi_{C_1'}(z_1)=\frac{\Gamma(\theta)}{\Gamma(\alpha_1')\Gamma(\alpha_2')}z_1^{\alpha_1'-1}(1-z_1)^{\alpha_2'-1}\quad con\quad \alpha_1'=\alpha(C_1'),\alpha_2'=\theta-\alpha_1'$$

A questo punto, se prendiamo (X,Y) vettore aleatorio con densità

$$f(x,y)=\frac{\Gamma(\theta)}{\Gamma(\alpha_1)\Gamma(\alpha_2)\Gamma(\alpha_3)}x^{\alpha_1-1}y^{\alpha_2-1}(1-x-y)^{\alpha_3-1},\quad (x,y)\in\Delta_2$$

La legge della variabile aleatoria $X+Y$ si ricava essere

$$f(z)=\frac{\Gamma(\theta)}{\Gamma(\alpha_1+\alpha_2)\Gamma(\alpha_3)}z^{\alpha_1+\alpha_2-1}(1-z)^{\alpha_3-1},\quad z\in[0,1]$$

E riotteniamo il risultato voluto ricordando che $\alpha_1'=\alpha_1+\alpha_2$ e $\alpha_3'=\alpha_3$.

Quindi possiamo passare da una densità all'altra tramite la trasformazione di variabili aleatorie $(X,Y)\to X+Y$, il che verifica la condizione di compatibilità (3) in quest'esempio.

La verifica a mano funziona analogamente per casi simili ma più complessi. Ad esempio nel caso in cui invece che una somma abbiamo un sistema lineare $Y=\boldsymbol{M}X$ di variabili aleatorie. Anche in quel caso si calcola la legge di Y direttamente e si verifica che è uguale alla legge della partizione corrispondente.

Capitolo 2

Formula di campionamento di Ewens

Nel presente capitolo ci dedichiamo alla deduzione della formula di campionamento di Ewens. Essa fornisce una probabilità sullo spazio delle partizioni di $\{1, \ldots, n\}$; partiremo da casi semplici, con $n = 2$ e $n = 3$, e arriveremo poi a scrivere la costruire la formula nel caso generale e a mostrare il suo legame con la legge di probabilità delle variabili aleatorie scambiabili $X_1, \ldots, X_n$. Partiamo enunciando la formula cui vogliamo giungere:

Teorema 2.0.1 (formula di campionamento di Ewens)**.** *Siano $\theta > 0$ e $n \in \mathbb{N}$. Sia, inoltre, $B_1, \ldots, B_k$ una partizione dell'insieme $\{1, \ldots, n\}$ e n_j la cardinalità dell'insieme B_j per $j = 1, \ldots, k$, allora la probabilità assegnata alla partizione $B_1, \ldots, B_k$ è la seguente:*

$$Ewens^{(n)}(n_1, \ldots, n_k; \theta) = P[\Pi_n = B_1, \ldots, B_k] = \frac{\theta^k}{(\theta)_{n\uparrow}} \prod_{j=1}^{k} (n_j - 1)! \qquad (2.1)$$

Dove con Π_n indichiamo una variabile aleatoria che assume valori nell'insieme delle partizioni di $\{1, \ldots, n\}$.

La formula 2.1 prende il nome di formula di campionamento di Ewens.

L'interpretazione immediata è la seguente: sia $X_1, \ldots, X_n$ una successione di variabili aleatorie scambiabili sullo spazio $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ e sia $B_1, \ldots, B_k$ una partizione di $\{1, \ldots, n\}$. Se vogliamo classificare i valori assunti dalle variabili aleatorie, potremo dire che $X_i = X_j$ se i e j appartengono allo stesso insieme della partizione. La formula di Ewens ci fornisce, quindi, la probabilità che due o più variabili aleatorie siano uguali o distinte. Dunque, ad esempio, nel problema che ha animato questa tesi, le variabili aleatorie relative alle osservazioni di due individui distinti coincidono se questi appartengono alla stessa specie.

Esiste anche un altro risultato strettamente legato alla formula 2.1 che prende sempre il nome di formula di campionamento di Ewens ma che noi chiameremo, per evitare confusioni, *formula di campionamento di Ewens parziale*. La scelta di questo nome è dovuta al fatto che, nonostante contenga la stessa informazione della formula di Ewens, la sua deduzione si ferma a uno step precedente

rispetto ad essa. Essa, infatti, non assegna la probabilità alla singola partizione ma la assegna ai valori $n_1, \dots, n_k$ che rappresentano la cardinalità degli insiemi della partizione. Possiamo dire, però, che le due formule contengono la stessa informazione perché, come ovvio dalla definizione data, la formula 2.1 attribuisce la stessa probabilità a tutte le partizioni i cui insiemi hanno le stesse cardinalità $n_1, \dots, n_k$.

Per introdurre la formula di Ewens parziale, definiamo dei valori m_j. Chiamiamo $m_j,\ j = 1, \dots, n$ il numero di $n_i\ i = 1, \dots, k$ uguali a j, cioè il numero di insiemi nella partizione con esattamente j elementi. In maniera più formale possiamo definirli come segue:

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

Notiamo, anche, che per gli m_j vale la relazione

$$\sum_{j=1}^{n} j m_j = n$$

Essi, inoltre, forniscono tutte le informazioni che ci interessano sulla partizione di $\{1, \dots, n\}$: infatti, fornire gli $m_1, \dots, m_n$ è equivalente a fornire gli $n_1, \dots, n_k$: infatti, dati gli m_j avremo che gli $n_1, \dots, n_k$ corrispondenti saranno $\underbrace{1, \dots, 1}_{m_1 \text{ volte}}, \underbrace{2, \dots, 2}_{m_2 \text{ volte}}, \dots, \underbrace{n, \dots, n}_{m_n \text{ volte}}$.

Possiamo, ora, definire la formula di Ewens parziale nel seguente teorema:

Teorema 2.0.2 (formula di Ewens parziale)**.** *Siano $\theta > 0$ e $n \in \mathbb{N}$. Siano, inoltre, $m_1, \dots, m_n \in \{1, \dots, n\}$ tali per cui $\sum_{j=1}^{n} j m_j = n$, allora la probabilità assegnata alla n-upla $m_1, \dots, m_n$ è la seguente:*

$$p(m_1, \dots, m_n; \theta) = \frac{n!}{(\theta)_{(n)\uparrow}} \prod_{j=1}^{n} \frac{\theta^{m_j}}{j^{m_j} m_j!} \tag{2.2}$$

La dimostrazione del legame tra le due formule è riportata più avanti nell'equazione 2.7.

Precisiamo che d'ora in avanti quando parleremo di formula di Ewens o formula di campionamento di Ewens ci riferiremo alla formula 2.1, mentre per riferirci alla formula 2.2 diremo sempre formula di campionamento di Ewens parziale o formula di Ewens parziale.

2.1 Caso base: $n = 2$

Partiamo nella nostra costruzione da un caso semplice, fissando $n = 2$, e mostriamo come ricavare le formula a partire dal processo di Dirichlet.

In questo caso abbiamo che le possibili partizioni sono due: $\{1\}, \{2\}$ e $\{1, 2\}$, che corrispondono a $\{X_1 = X_2\}$ e $\{X_1 \neq X_2\}$. Quindi, scegliendo $x \in \mathbb{R}$, che - ricordiamo - è lo spazio su cui sono definite le variabili aleatorie considerate,

finirò nel primo o nel secondo blocco secondo le probabilità date dalla formula di Ewens.

Per la deduzione della formula, seguiamo questi due passi:

(A) Calcoliamo esplicitamente $P[X_1 \in A_1, X_2 \in A_2]$ per $A_1, A_2 \in \mathcal{B}(\mathbb{R})$;

(B) deduciamo da $P[X_1 \in A_1, X_2 \in A_2]$ il valore di $P[X_1 = X_2]$.

Il valore delle probabilità $P[X_1 = X_2]$ e del complementare $P[X_1 \neq X_2]$ corrispondono ai due valori della formula di Ewens nel caso $n = 2$.

Dato che X_1 e X_2 sono variabili aleatorie scambiabili, possiamo sfruttare, per il passo (A) il teorema di rappresentazione di De Finetti, che ricordiamo ponendo $n = 2$:

$$P[X_1 \in A_1, X_2 \in A_2] = \int_{\mathbb{P}} p(A_1)p(A_2)dq(p) = \mathbb{E}[\mu(A_1)\mu(A_2)]$$

Per calcolare il valore della probabilità a sinistra, quindi, è sufficiente calcolare il valore atteso a destra. Per farlo, sfruttiamo la linearità del valore atteso, partizionando lo spazio campionario $\mathbb{R}$ nel seguente modo:

$$C_1 = A_1 \backslash A_2, \quad C_2 = A_1 \cap A_2, \quad C_3 = A_2 \backslash A_1, \quad C_4 = (A_1 \cup A_2)^c$$

Chiaramente C_1, C_2, C_3, C_4 è una partizione di $\mathbb{R}$.

A questo punto, essendo μ una misura, varranno le seguenti uguaglianze:

$$\mu(A_1) = \mu(C_1 \cup C_2) = \mu(C_1) + \mu(C_2)$$

$$\mu(A_2) = \mu(C_2 \cup C_3) = \mu(C_2) + \mu(C_3)$$

Quindi possiamo riscrivere il valore atteso nel seguente modo:

$$\begin{aligned}\mathbb{E}[\mu(A_1)\mu(A_2)] &= \mathbb{E}[(\mu(C_1) + \mu(C_2))(\mu(C_2) + \mu(C_3))] \\ &= \mathbb{E}[\mu(C_1)\mu(C_2) + \mu(C_1)\mu(C_3) + (\mu(C_2))^2 + \mu(C_2)\mu(C_3)] \\ &= \mathbb{E}[\mu(C_1)\mu(C_2)] + \mathbb{E}[\mu(C_1)\mu(C_3)] + \mathbb{E}[(\mu(C_2))^2] + \mathbb{E}[\mu(C_2)\mu(C_3)]\end{aligned}$$

Si tratta, quindi, di studiare il valore atteso dei prodotti dei $\mu(C_i) \quad i = 1, 2$.

Calcoliamo esplicitamente, ad esempio, $\mathbb{E}[\mu(C_1)\mu(C_2)]$. Per farlo, vorremmo usare la formula per il prodotto di variabili aleatorie X e Y su Ω aventi densità congiunta $f(x, y)$:

$$\mathbb{E}[XY] = \int_{\Omega} xyf(x, y)dxdy$$

Abbiamo, quindi, bisogno di una funzione di densità f che dipenda solamente da due variabili. Tuttavia, nel nostro caso, la densità è la legge di Dirichlet per una partizione con 4 elementi che dipende, però, da tre variabili. L'idea per ricondurci a una funzione di sole due variabili è quella di definire una nuova partizione nel seguente modo:

$$C_1' = C_1, \quad C_2' = C_2, \quad C_3' = C_3 \cup C_4$$

Chiaramente $\mathbb{E}[\mu(C_1)\mu(C_2)] = \mathbb{E}[\mu(C_1')\mu(C_2')]$.

Inoltre la legge del vettore aleatorio $(\mu(C_1'), \mu(C_2'))$ è quella di Dirichlet per la partizione C_1', C_2', C_3', cioè

$$f(z_1, z_2) = \frac{\Gamma(\theta)}{\Gamma(\alpha_1')\Gamma(\alpha_2')\Gamma(\alpha_3')} z_1^{\alpha_1'-1} z_2^{\alpha_2'-1}(1 - z_1 - z_2)^{\alpha_3'-1}, \quad (z_1, z_2) \in \Delta_2$$

Quindi il calcolo del valore atteso di $(\mu(C_1), \mu(C_2))$ si riduce al seguente integrale:

$$\mathbb{E}[\mu(C_1), \mu(C_2)] = \int_{\Delta_2} z_1 z_2 \frac{\Gamma(\theta)}{\Gamma(\alpha_1')\Gamma(\alpha_2')\Gamma(\alpha_3')} z_1^{\alpha_1'-1} z_2^{\alpha_2'-1}(1 - z_1 - z_2)^{\alpha_3'-1} dz_1 dz_2$$

Sfruttiamo, ora, il seguente lemma:

Lemma 2.1.1 (formula di Dirichlet)**.** *Dati $n \in \mathbb{N}$ e $\beta_1, \ldots, \beta_n > 0$, vale che*

$$\int_{\Delta_{n-1}} z_1^{\beta_1-1} z_2^{\beta_2-1} \ldots z_{n-1}^{\beta_{n-1}-1}(1 - z_1 - \cdots - z_{n-1})^{\beta_n-1} dz_1 \ldots dz_{n-1} = \frac{\Gamma(\beta_1)\Gamma(\beta_2)\ldots\Gamma(\beta_n)}{\Gamma(\beta_1 + \cdots + \beta_n)}$$

Attraverso la formula di Dirichlet si giunge immediatamente alla conclusione che

$$\mathbb{E}[\mu(C_1), \mu(C_2)] = \frac{\alpha(C_1)\alpha(C_2)}{(\theta+1)\theta}$$

Analogamente si calcolano anche gli altri termini e si ottiene:

$$\mathbb{E}[\mu(C_1), \mu(C_3)] = \frac{\alpha(C_1)\alpha(C_3)}{(\theta+1)\theta}$$

$$\mathbb{E}[\mu(C_2), \mu(C_3)] = \frac{\alpha(C_2)\alpha(C_3)}{(\theta+1)\theta}$$

$$\mathbb{E}[(\mu(C_2))^2] = \frac{\alpha(C_2)(\alpha(C_2)+1)}{(\theta+1)\theta}$$

Ricostruendo la formula risulterà, quindi:

$$\begin{aligned}
&P[X_1 \in A_1, X_2 \in A_2] = \\
&= \frac{\alpha(C_1)\alpha(C_2)}{(\theta+1)\theta} + \frac{\alpha(C_1)\alpha(C_3)}{(\theta+1)\theta} + \frac{\alpha(C_2)\alpha(C_3)}{(\theta+1)\theta} + \frac{\alpha(C_2)(\alpha(C_2)+1)}{(\theta+1)\theta} = \\
&= \frac{\theta}{\theta+1}[\bar{\alpha}(C_1)\bar{\alpha}(C_2) + \bar{\alpha}(C_1)\bar{\alpha}(C_3) + \bar{\alpha}(C_2)\bar{\alpha}(C_3) + \bar{\alpha}(C_2)^2] + \frac{1}{\theta+1}\bar{\alpha}(C_2) = \\
&= \gamma\bar{\alpha}(A_1)\bar{\alpha}(A_2) + (1-\gamma)\bar{\alpha}(A_1 \cap A_2)
\end{aligned}$$

Che riassumiamo come segue:

$$P[X_1 \in A_1, X_2 \in A_2] = \gamma\bar{\alpha}(A_1)\bar{\alpha}(A_2) + (1-\gamma)\bar{\alpha}(A_1 \cap A_2) \tag{2.3}$$

dove abbiamo definito $\gamma := \theta/(\theta+1)$

Possiamo, ora, procedere al punto (B). Vogliamo, cioè, dimostrare che $P[X_1 = X_2] = \gamma$ e $P[X_1 \neq X_2] = 1-\gamma$, quindi che la formula di Ewens fornisce i pesi γ dell'equazione 2.3.

Enunciamo e dimostriamo la seguente proposizione:

Proposizione 2.1.1.

$$P[X_1 = X_2] = 1 - \gamma$$

Dimostrazione. Per procedere con la dimostrazione notiamo che, se il nostro spazio è $\mathbb{X}_1 = \mathbb{R}$, oppure $\mathbb{X}_1 = [a,b] \subseteq \mathbb{R}$, la probabilità dell'evento $\{X_1 = X_2\}$ è pari alla probabilità che il vettore aleatorio bidimensionale (X_1, X_2) appartenga alla diagonale dello spazio $\mathbb{X}_1 \times \mathbb{X}_1$. Formalizziamo la dimostrazione prendendo, per esempio, $\mathbb{X}_1 = [a,b]$. Definiamo, allora, una tassellazione attorno alla diagonale $D_n = \bigcup_{i=1}^{n}(E_i \times E_i)$, con $E_i = [a + b(i-1)/n, a + bi/n]$.

Chiaramente per n che tende all'infinito D_n si assottiglia attorno alla diagonale, quindi $P[X_1 = X_2] = P[(X_1, X_2) \in diag([a,b]\times[a,b])] = lim_{n\to\infty} P[(X_1, X_2) \in D_n]$.

Sfruttando la relazione 2.3, allora:

$$\begin{aligned}
P[(X_1, X_2) \in D_n] &= \sum_{i=1}^{n} P[(X_1, X_2) \in E_i \times E_i] \\
&= \sum_{i=1}^{n} P[X_1 \in E_i, X_2 \in E_i] \\
&= \sum_{i=1}^{n} [\gamma\bar{\alpha}(E_i)^2 + (1-\gamma)\bar{\alpha}(E_i)] \\
&= \gamma \sum_{i=1}^{n} \bar{\alpha}(E_i)^2 + (1-\gamma) \sum_{i=1}^{n} \bar{\alpha}(E_i)] \\
&\sim_{n\to+\infty} \gamma n \frac{1}{n^2} + (1-\gamma) n \frac{1}{n}
\end{aligned}$$

Risulterà, di conseguenza, che

$$P[X_1 = X_2] = 1 - \gamma$$

□

Abbiamo ricavato, quindi, la formula di Ewens nel caso $n = 2$, che possiamo riportare esplicitamente:

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

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

2.2 Caso base: $n = 3$

Studiamo, ora, anche il caso $n = 3$, per intuire il pattern che ci permetterà, poi, di generalizzare al caso n generico.

Nel caso $n = 3$ le partizioni possibili dell'insieme $\{1, 2, 3\}$ sono 5 e corrispondono, in termini delle variabili aleatorie, agli insiemi $\{X_1 = X_2 = X_3\}$, $\{X_1 = X_2 \neq X_3\}$, $\{X_2 = X_3 \neq X_1\}$, $\{X_1 = X_3 \neq X_2\}$,$\{X_1 \neq X_2 \neq X_3\}$.

L'obiettivo iniziale è anche in questo caso quello di ricavare una formula per la probabilità $P[X_1 \in A_1, X_2 \in A_2, X_3 \in A_3] = \mu(A_1 \times A_2 \times A_3)$, con A_1, A_2, A_3 arbitrari in $\mathcal{B}(\mathbb{R})$ e poi ricavare da essa la probabilità delle partizioni, che, come vedremo, corrispondono alla formula di Ewens.

Anche in questo caso stiamo trattando con variabili aleatorie scambiabili X_1, X_2, X_3 e quindi possiamo sfruttare nuovamente la formula di De Finetti:

$$P[X_1 \in A_1, X_2 \in A_2, X_3 \in A_3] = \int_{\mathbb{P}} p(A_1)p(A_2)p(A_3)dq(p) = \mathbb{E}[\mu(A_1)\mu(A_2)\mu(A_3)]$$

Introduciamo, ora, a partire da A_1, A_2, A_3 la seguente partizione di $\mathbb{R}$:

$$\begin{aligned}
C_1 &= A_1 \backslash (A_2 \cup A_3),\\
C_2 &= A_2 \backslash (A_1 \cup A_3),\\
C_3 &= A_3 \backslash (A_1 \cup A_2),\\
C_4 &= (A_1 \cap A_2) \backslash A_3),\\
C_5 &= (A_2 \cap A_3) \backslash A_1),\\
C_6 &= (A_1 \cap A_3) \backslash A_2),\\
C_7 &= A_1 \cap A_2 \cap A_3,\\
C_8 &= (A_1 \cup A_2 \cup A_3)^c.
\end{aligned}$$

Posto $\mu(C_i) = \mu_i$, valgono le seguenti formule che per i valori di $\mu(A_i)$.

$$\begin{aligned}
\mu(A_1) &= \mu_1 + \mu_4 + \mu_6 + \mu_7\\
\mu(A_2) &= \mu_2 + \mu_4 + \mu_5 + \mu_7\\
\mu(A_3) &= \mu_3 + \mu_5 + \mu_6 + \mu_7
\end{aligned}$$

Si tratta, a questo punto, di calcolare la seguente speranza matematica:

$$\mathbb{E}[(\mu_1 + \mu_4 + \mu_6 + \mu_7)(\mu_2 + \mu_4 + \mu_5 + \mu_7)(\mu_3 + \mu_5 + \mu_6 + \mu_7)]$$

Per farlo ricaviamo i valori di

$$\mathbb{E}[\mu_i\mu_j\mu_k],\ \mathbb{E}[\mu_i^2\mu_j],\ \mathbb{E}[\mu_i^3] \text{ con } i, j, k \text{ distinti.}$$

Per ottenere il primo consideriamo la partizione $C_i, C_j, C_k, (C_i, C_j, C_k)^c$ e rinominiamo i suoi insiemi come C'_1, C'_2, C'_3, C'_4.

La legge di p, conseguentemente, sarà la densità di Dirichlet con $n = 4$ e varranno, allora, le seguenti uguaglianze

$$\begin{aligned}\mathbb{E}[\mu_i\mu_j\mu_k] &= \int_{\Delta_3} z_1 z_2 z_3 \frac{\Gamma(\theta)}{\Gamma(\alpha'_1)\Gamma(\alpha'_2)\Gamma(\alpha'_3)\Gamma(\alpha'_4)} z_1^{\alpha'_1-1} z_2^{\alpha'_2-1} z_3^{\alpha'_3}(1-z_1-z_2-z_3)^{\alpha'_4-1} dz_1 dz_2 dz_3 \\ &= \frac{\alpha'_1\alpha'_2\alpha'_3}{(\theta)_{3\uparrow}} = \frac{\alpha_i\alpha_j\alpha_k}{(\theta)_{3\uparrow}}\end{aligned}$$

Dove abbiamo posto $\alpha'_i = \alpha(C'_i)$ e $\alpha_i = \alpha(C_i)$.
Analogamente per gli altri valori attesi:

$$\begin{aligned}\mathbb{E}[\mu_i^2\mu_j] &= \int_{\Delta_2} z_1^2 z_2 \frac{\Gamma(\theta)}{\Gamma(\alpha'_1)\Gamma(\alpha'_2)\Gamma(\alpha'_3)} z_1^{\alpha'_1-1} z_2^{\alpha'_2-1}(1-z_1-z_2)^{\alpha'_3} dz_1 dz_2 \\ &= \frac{(\alpha'_1)_{2\uparrow}\alpha'_2}{(\theta)_{3\uparrow}} = \frac{(\alpha_i)_{2\uparrow}\alpha_j}{(\theta)_{3\uparrow}}\end{aligned}$$

$$\mathbb{E}[\mu_i^m] = \int_0^1 z_1^m \frac{\Gamma(\theta)}{\Gamma(\alpha'_1)\Gamma(\alpha'_2)} z_1^{\alpha'_1-1}(1-z_1)^{\alpha'_2} dz_1 = \frac{(\alpha'_1)_{m\uparrow}}{(\theta)_{m\uparrow}} = \frac{(\alpha_i)_{m\uparrow}}{(\theta)_{m\uparrow}}$$

A questo punto, si tratta di sfruttare quanto appena calcolato per scrivere esplicitamente il valore della probabilità delle partizioni di tre elementi, in maniera analoga a quanto fatto nel caso $n = 2$.

Come prima cosa, riorganizziamo le terne di indici (i, j, k) che appaiono nella speranza matematica precedente: il primo indice corrisponde agli insiemi C_i contenuti in A_1, il secondo agli insiemi C_j in A_2 e il terzo ai C_k in A_3. Se vogliamo scriverlo esplicitamente, $i \in 1, 4, 6, 7$, $j \in 2, 4, 5, 7$ e $k \in 3, 5, 6, 7$.

Introduciamo anche tre classi per le terne di indici (i, j, k): la prima contenente le terne con tre indici distinti, la seconda con due indici uguali e il terzo distinto e la terza con tre indici uguali (che corrisponde solo al caso dell'intersezione dei tre insiemi).

La somma che risulterà sarà la seguente:

$$\begin{aligned}&P[X_1 \in A_1, X_2 \in A_2, X_3 \in A_3] = \\ &= \sum_{\substack{(i,j,k)\\ \text{prima classe}}} \frac{\alpha_i\alpha_j\alpha_k}{(\theta)_{3\uparrow}} + \sum_{\substack{(i,j,k)\\ \text{seconda classe}}} \frac{(\alpha_i)_{2\uparrow}\alpha_j}{(\theta)_{3\uparrow}} + \sum_{\substack{(i,j,k)\\ \text{terza classe}}} \frac{(\alpha_i)_{3\uparrow}}{(\theta)_{3\uparrow}} = \\ &= \frac{1}{(\theta)_{3\uparrow}}\Big[\sum_{\substack{(i,j,k)\\ \text{distinti}}} \alpha_i\alpha_j\alpha_k + \sum_{\substack{(i,j,k)\\ i=j\neq k}} (\alpha_i)^2\alpha_k + \sum_{\substack{(i,j,k)\\ i=j\neq k}} \alpha_i\alpha_k + \\ &\quad + \sum_{\substack{(i,j,k)\\ i\neq j=k}} (\alpha_j)^2\alpha_i + \sum_{\substack{(i,j,k)\\ i\neq j=k}} \alpha_j\alpha_i + \sum_{\substack{(i,j,k)\\ i=k\neq j}} (\alpha_i)^2\alpha_j + \sum_{\substack{(i,j,k)\\ i=k\neq j}} \alpha_i\alpha_j + \alpha_7^3 + 3\alpha_7^2 + 2\alpha_7\Big]\end{aligned}$$

Studiamo, ora, i singoli termini, raggruppandoli nei vari ordini. La somma

dei termini di ordine tre produce

$$\frac{\theta^3}{(\theta)_{3\uparrow}}\bar{\alpha}(A_1)\bar{\alpha}(A_2)\bar{\alpha}(A_3)$$

La somma dei termini di ordine due produce

$$\frac{\theta^2}{(\theta)_{3\uparrow}}[\bar{\alpha}(A_1)\bar{\alpha}(A_2 \cap A_3) + \bar{\alpha}(A_2)\bar{\alpha}(A_1 \cap A_3) + \bar{\alpha}(A_3)\bar{\alpha}(A_1 \cap A_2)]$$

Infine, abbiamo solo un termine di ordine 1 che è

$$\frac{2\theta}{(\theta)_{3\uparrow}}\bar{\alpha}(A_1 \cap A_2 \cap A_3)$$

Otteniamo, quindi, la formula:

$$\begin{aligned}
&P[X_1 \in A_1, X_2 \in A_2, X_3 \in A_3] = \\
&= \gamma_1\bar{\alpha}(A_1)\bar{\alpha}(A_2)\bar{\alpha}(A_3) + \\
&\quad + \gamma_2[\bar{\alpha}(A_1)\bar{\alpha}(A_2 \cap A_3) + \bar{\alpha}(A_2)\bar{\alpha}(A_1 \cap A_3) + \bar{\alpha}(A_3)\bar{\alpha}(A_1 \cap A_2)] + \\
&\quad + \gamma_3\bar{\alpha}(A_1 \cap A_2 \cap A_3)
\end{aligned}$$

dove $\gamma_1 = \theta^3/(\theta)_{3\uparrow}$, $\gamma_2 = \theta^2/(\theta)_{3\uparrow}$, $\gamma_3 = 2\theta/(\theta)_{3\uparrow}$.

Notiamo, peraltro, che $\gamma_1 + 3\gamma_2 + \gamma_3 = 1$, come vogliamo.

Dalla formula precedente si ricava, poi, la formula delle probabilità delle partizioni:

$$\begin{gathered}
P[X_1 = X_2 = X_3] = \gamma_1 \\
P[X_i \neq X_j = X_k] = \gamma_2, \quad i, j, k \text{ indici distinti in } \{1, 2, 3\} \\
P[X_1 = X_2 = X_3] = \gamma_3
\end{gathered}$$

La dimostrazione di queste formule avviene per via geometrica, in maniera analoga al caso $n = 2$. Dimostriamo, ad esempio, che $P[X_1 = X_2 = X_3] = \gamma_3$. Consideriamo come spazio campionario $\mathbb{X}_1 = [a, b] \subseteq \mathbb{R}$. In questo caso il vettore aleatorio (X_1, X_2, X_3) assumerà valori su un cubo di lato $[a, b]$.

Notiamo immediatamente la seguente uguaglianza:

$$P[X_1 = X_2 = X_3] = P[(X_1, X_2, X_3) \in diag([a, b]^3)]$$

Procediamo, quindi, costruendo attorno alla diagonale una tassellazione $D_n = \bigcup_{i=1}^{n}(E_i \times E_i \times E_i)$, con $E_i = [a + b(i-1)/n, a + bi/n]$.

Varrà, dunque, che

$$
\begin{aligned}
P[(X_1, X_2, X_3) \in D_n] &= \sum_{i=1}^{n} P[(X_1, X_2, X_3) \in E_i \times E_i \times E_i] \\
&= \sum_{i=1}^{n} P[X_1 \in E_i, X_2 \in E_i, X_3 \in E_i] \\
&= \sum_{i=1}^{n} [\gamma_1 \bar{\alpha}(E_i)^3 + 3\gamma_2 \bar{\alpha}(E_i)^2 + \gamma_3 \bar{\alpha}(E_i)] \\
&= \gamma_1 \sum_{i=1}^{n} \bar{\alpha}(E_i)^3 + \gamma_2 \sum_{i=1}^{n} \bar{\alpha}(E_i)^2 + \gamma_3 \sum_{i=1}^{n} \bar{\alpha}(E_i) \\
&\sim_{n \to +\infty} \gamma_1 n \frac{1}{n^3} + \gamma_2 n \frac{1}{n^2} + \gamma_3 n \frac{1}{n}
\end{aligned}
$$

Da cui ricaviamo immediatamente la formula voluta nel seguente modo:

$$
\begin{aligned}
&P[X_1 = X_2 = X_3] = P[(X_1, X_2, X_3) \in diag([a, b]^3)] = \\
&= lim_{n \to \infty} P[(X_1, X_2, X_3) \in D_n] = \gamma_3
\end{aligned}
$$

Le altre due formule si ricavano analogamente: per il caso $X_i \neq X_j = X_k$, ad esempio, l'unica differenza sarà che la tassellazione non avverrà più sulla diagonale del cubo, ma sulla diagonale di una faccia del cubo.

Osserviamo, infine, che le formule trovate per le probabilità delle partizioni corrispondono effettivamente alla formula di Ewens nel caso $n = 3$.

2.3 Caso generico: costruzione diretta

Potremmo ora procedere con altre deduzioni particolari, ad esempio ponendo $n = 4$, tuttavia la situazione diventa poco gestibile a mano, dal momento che il numero di partizioni di un insieme di n elementi cresce molto rapidamente al crescere di n. Ad esempio mentre era 5 per $n = 3$, è 15 per $n = 4$, 52 per $n = 5$, 203 per $n = 6$. Questi numeri prendono il nome di numeri di Bell e sono trattati con maggiori dettagli nell'appendice B.

Ci occupiamo, quindi, di dedurre la formula di Ewens nel caso n generico.

Il passo di partenza è sempre lo stesso, cioè il Teorema di Rappresentazione di De Finetti:

$$
\begin{aligned}
P[X_1 \in A_1, X_2 \in A_2, \ldots, X_n \in A_n] &= \int_{\mathbb{P}} p(A_1) p(A_2) \ldots p(A_n) dq(p) \\
&= \mathbb{E}[\mu(A_1)\mu(A_2)\ldots\mu(A_n)]
\end{aligned}
$$

Per calcolare esplicitamente la probabilità nel termine di sinistra possiamo riscrivere la speranza matematica del termine di destra.

Il ragionamento è lo stesso dei casi $n = 2$ e $n = 3$. Effettuiamo, quindi, una partizione dello spazio in cui si trovano gli insiemi A_i. Per farlo, introduciamo degli insiemi C_j, a due a due disgiunti, costruiti nel seguente modo:

$$A_i = \bigcup_{j \in I_i} C_j$$

Notiamo che, per ottenere una partizione, fissato n, avremo 2^n insiemi C_j: in particolare, C_{2^n} sarà $(\bigsqcup_{j=1}^{2^n-1} C_j)^c$.

Da questo seguirà che:

$$\begin{aligned} P[X_1 \in A_1, \dots, X_n \in A_n] &= \mathbb{E}[\mu(A_1) \dots \mu(A_n)] \\ &= \mathbb{E}[\mu\Big(\bigcup_{j_1 \in I_1} C_{j_1}\Big) \dots \mu\Big(\bigcup_{j_n \in I_n} C_{j_n}\Big)] \\ &= \mathbb{E}[\sum_{j_1 \in I_1} \mu(C_{j_1}) \cdots \sum_{j_n \in I_n} \mu(C_{j_n})] \\ &= \sum_{j_1 \in I_1, \dots j_n \in I_n} \mathbb{E}[\mu(C_{j_1}) \dots \mu(C_{j_n})] \end{aligned}$$

A questo punto fissiamo il vettore di indici $(j_1, \dots, j_n) \in \{1, \dots, 2^n\}$ e introduciamo il valore k che indica il numero di indici distinti. Quindi gli indici $(j_1, \dots, j_n)$ saranno n_1 uguali a un certo valore $r_1 \in \{1, \dots, 2^n\}$, n_2 uguali a un certo valore $r_2 \in \{1, \dots, 2^n\}$, $\dots$, n_k uguali a un certo valore r_k. Possiamo, allora, riscrivere nel seguente modo il nostro set di indici:

$$(j_1, \dots, j_n) = (\underbrace{r_1, \dots, r_1}_{n_1 \text{ volte}}, \underbrace{r_2, \dots, r_2}_{n_2 \text{ volte}}, \dots, \underbrace{r_k, \dots, r_k}_{n_k \text{ volte}})$$

La speranza matematica $\mathbb{E}[\mu(C_{j_1}) \dots \mu(C_{j_n})]$ ora diventa:

$$\mathbb{E}[\mu(C_{j_1}) \dots \mu(C_{j_n})] = \mathbb{E}[\mu(C_{r_1})^{n_1} \dots \mu(C_{r_k})^{n_k}]$$

Per semplicità rinominiamo gli indici $C_{r_i} = C'_i$ e definiamo $C'_{k+1} = (\bigcup_{j=1}^{k} C_j)^c$. Usiamo anche la notazione $\mu(C_{r_i}) = \mu_i$ e $\alpha'_i = \alpha(C'_i)$. Ci occupiamo, ora, di calcolare la speranza matematica come integrale della densità di Dirichlet.

$$\begin{aligned} &\mathbb{E}[\mu(C'_1)^{n_1} \dots \mu(C'_k)^{n_k}] = \\ &= \int_{\Delta_k} z_1^{n_1} \dots z_k^{n_k} \frac{\Gamma(\theta)}{\Gamma(\alpha'_1) \dots \Gamma(\alpha'_k)\Gamma(\alpha'_{k+1})} z_1^{\alpha'_1 - 1} \dots z_k^{\alpha'_k} (1 - z_1 - \cdots - z_k)^{\alpha'_{k+1} - 1} dz_1 \dots dz_k = \\ &= \frac{(\alpha'_1)_{n_1\uparrow} \dots (\alpha'_k)_{n_k\uparrow}}{(\theta)_{n\uparrow}} = \frac{\alpha(C_{r_1})_{n_1\uparrow} \dots \alpha(C_{r_k})_{n_k\uparrow}}{(\theta)_{n\uparrow}} \end{aligned}$$

Grazie alla formula delle probabilità totali è possibile, dopo aver opportunamente ricostruito gli insiemi A_j dagli insiemi C_i, di esprimere la legge della probabilità $P[X_1 \in A_1, \dots, X_n \in A_n]$ con una formula di questo tipo:

$$P[X_1 \in A_1, \ldots, X_n \in A_n] =$$
$$= \sum_{k=1}^{n} \gamma_k^{(n)} \sum_{\substack{(n_1,\ldots,n_k) \\ \in \mathcal{P}_{k,n}}} \Gamma_k^{(n)}(n_1,\ldots,n_k) \frac{1}{P_k^{(n)}(n_1,\ldots,n_k)} \sum_{\substack{I_1,\ldots,I_k \\ \in \Pi_k^{(n)}(n_1,\ldots,n_k)}} \prod_{j=1}^{k} \bar{\alpha}\Big(\bigcap_{i \in I_j} A_i\Big) \tag{2.4}$$

Dove k indica il numero di variabili aleatorie distinte, $\mathcal{P}_{k,n}$ è l'insieme delle k-uple di numeri in $\{1,\ldots,n\}$ che sommano n, n_i indica il numero di variabili uguali a X_i e $\Pi_k^{(n)}(n_1,\ldots,n_k)$ è l'insieme delle partizioni $I_1,\ldots,I_k$ dell'insieme $\{1,\ldots,n\}$ in cui I_j ha n_j elementi. $\gamma_k^{(n)}$, $\Gamma_k^{(n)}(n_1,\ldots,n_k)$, e $P_k^{(n)}(n_1,\ldots,n_k)$ sono dei pesi che vogliamo ora studiare. Essi forniscono la probabilità con cui viene scelta la partizione $I_1,\ldots,I_k \in \Pi_k^{(n)}(n_1,\ldots,n_k)$. k, infatti, è il numero di insiemi della partizione ed è scelto con probabilità $\gamma_k^{(n)}$, mentre $n_1,\ldots,n_k$ sono le cardinalità degli insiemi della partizione e sono scelte con probabilità $\Gamma_k^{(n)}(n_1,\ldots,n_k)$. Fissati, poi, $n_1,\ldots,n_k$ la partizione $I_1,\ldots,I_k$ è scelta con probabilità uniforme $1/P_k^{(n)}(n_1,\ldots,n_k)$.

Dato che la formula di Ewens fornisce proprio la probabilità delle partizioni, e anche in analogia a quanto visto nei casi $n = 2, 3$, i coefficienti della legge di probabilità corrispondono alla formula di Ewens. Vogliamo, quindi, vedere che:

$$Ewens^{(n)}(n_1,\ldots,n_k;\theta) = \sum_{k=1}^{n} \gamma_k^{(n)} \sum_{\substack{(n_1,\ldots,n_k) \\ \in \mathcal{P}_{k,n}}} \Gamma_k^{(n)}(n_1,\ldots,n_k)/P_k^{(n)}(n_1,\ldots,n_k)$$

Prima di farlo, però, è utile introdurre la struttura ricorsiva della formula di Ewens.

2.4 Costruzione ricorsiva

Come suggerito dal modo in cui abbiamo introdotto la formula di Ewens, a partire da casi elementari per poi generalizzare, la distribuzione di Ewens emerge da una struttura ricorsiva.

In particolare, possiamo incominciare dando una definizione ricorsiva della legge di probabilità (2.4). D'altra parte risulta molto più semplice studiare la probabilità predittiva

$$P[X_{n+1} \in A_{n+1} | X_1, \ldots, X_n]$$

che non la probabilità $P[X_1 \in A_1, \ldots, X_n \in A_n]$. Se considero, infatti, delle variabili aleatorie $X_1, \ldots, X_n$ come nella precedente sezione, quindi distribuite

con densità di Dirichlet n-dimensionale, vale la seguente formula [18]:

$$P[X_{n+1} \in B | X_1, \ldots, X_n] = \frac{1}{\theta + n}(\theta\bar{\alpha}(B) + \sum_{i=1}^{n} \delta_{X_i}(B))$$
$$= \frac{\theta}{\theta + n}\bar{\alpha}(B) + \frac{n}{\theta + n}(\frac{1}{n}\sum_{i=1}^{n} \delta_{X_i}(B)) \quad \forall B \in \mathcal{B}(\mathbb{R})$$

La formula (2.4) la si ricava poi per disintegrazione utilizzando l'uguaglianza:

$$P[X_1 \in A_1, \ldots, X_n \in A_n] =$$
$$= \int_{A_1} \cdots \int_{A_n} P[X_n \in A_n | X_1 = x_1, \ldots, X_{n-1} = x_{n-1}] d\bar{\alpha}(x_1, \ldots, x_n)$$

Il carattere ricorsivo delle formule di Ewens, d'altra parte, ha una motivazione più profonda legata al *processo del ristorante cinese* [1].

Dal punto di vista intuitivo, il processo del ristorante cinese funziona nel seguente modo: immaginiamo un ristorante cinese con infiniti tavoli, ciascuno con infiniti posti. Al tempo 1 arriva il primo cliente e si siede a uno dei tavoli, al tempo 2 arriva il secondo cliente e può decidere se sedersi nel tavolo già occupato, o in un tavolo libero, al tempo 3 il terzo cliente ha di fronte una scelta analoga, e così via per ogni cliente. Al tempo n, quindi, avremo che gli n clienti sono partizionati in $k \leq n$ tavoli. Notiamo subito l'analogia con il problema presentato nell'introduzione della tesi: basta pensare, infatti, i clienti come gli animali di cui sto studiando la specie e i tavoli come gli insiemi che raggruppano animali di una stessa specie.

Si osserva facilmente che il processo del ristorante cinese fornisce ricorsivamente delle partizioni dell'insieme $\{1, \ldots, n\}$, aggiungendo un elemento alla volta, cioè facendo crescere n di uno alla volta. Le probabilità di una partizione Π_n di $\{1, \ldots, n\}$ sono, quindi, definite come probabilità predittive, una volta assegnata la partizione Π_{n-1} di $n-1$ elementi. In particolare, dato θ un parametro positivo:

- $\theta/(\theta + n - 1)$ è la probabilità che l'elemento n-esimo costituisca un nuovo insieme nella partizione;
- $n_i/(\theta + n - 1)$ è la probabilità che l'elemento n-esimo finisca in un insieme della partizione che contiene già n_i elementi.

Queste probabilità si possono dimostrare per induzione.

Abbiamo tracciato uno schema in figura 2.1 per chiarificare la ricorsione delle partizioni nel processo del ristorante cinese. A ogni livello corrisponde n crescente a partire da 1; in ogni blocco abbiamo una k-upla $n_1, \ldots, n_k$ che somma a n. n_j è il numero di persone nel j-esimo tavolo occupato o, equivalentemente, la cardinalità del j-esimo insieme della partizione. Abbiamo colorato le ultime due righe dello schema in modo che due k-uple che hanno lo stesso k e lo stesso n abbiano lo stesso colore; ad esempio, sono colorate allo stesso

modo $(2,2)$ e $(1,3)$ perché corrispondo a partizioni diverse di $\{1,\dots,4\}$ in due sottoinsiemi. Per le prime tre righe non abbiamo ritenuto necessario aggiungere i colori dal momento che per ogni k ed n fissati esiste sempre un unico modo per scegliere $n_1,\dots,n_k$.

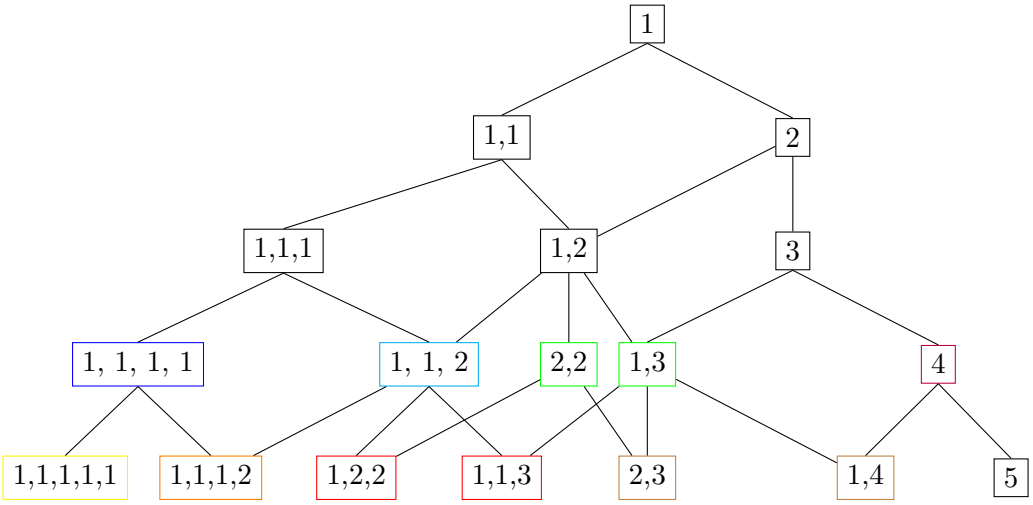

Figura 2.1: Albero del processo del ristorante cinese

Un processo analogo a quello del ristorante cinese è il processo dell'urna di Hoppe. Nell'urna di Hoppe poniamo inizialmente una palla nera di peso $\theta > 0$. A ogni passo $n = 1, 2, \dots$ estraiamo una palla dall'urna con probabilità proporzionale al suo peso. Se la palla estratta è quella nera, allora aggiungiamo una palla di peso 1 di un colore non presente nell'urna, se la palla estratta è di un altro colore, allora aggiungiamo una palla di peso 1 di quel colore. Dopo n passi, avrò un risultato corrispondente al processo del ristorante cinese con n clienti, con la sola differenza che le partizioni tra tavoli saranno partizioni tra colori.

A questo punto possiamo ricavare la formula di Ewens come dicevamo, cioè dai pesi della formula 2.4.

Studiamo, in primo luogo, il coefficiente $\gamma_k^{(n)}$: esso è la probabilità di avere k insiemi considerando una partizione di $\{1,\dots,n\}$. Detto in altri termini, sia K_n una variabile aleatoria che indica il numero di diversi insiemi di una partizione di $\{1,\dots,n\}$, allora vale che

$$\gamma_k^{(n)} = P[K_n = k]$$

Rende più intuitiva l'analisi immaginare K_n come il numero di tavoli occupati al tempo n (quando vi sono n clienti) nel processo del ristorante cinese. Notiamo che K_n è una catena markoviana perché possiede la proprietà dell'assenza di memoria, cioè la proprietà secondo cui

$$P[K_{n+1} = k | K_n = j_n, \dots, K_1 = j_1] = P[K_{n+1} = k | K_n = j_n]$$

Questo è evidente dalla natura del processo. Infatti, nel momento in cui arriva l'$(n+1)$-esimo commensale, la probabilità di assegnazione del posto dipende solo dal numero di tavoli occupati e da quanti clienti vi sono per tavolo. In particolare, la catena di Markov K_n è descritta da una legge ricorsiva contenuta nel seguente risultato:

Teorema 2.4.1. *$\{K_n\}_{n\geq 1}$ è una catena di Markov per cui vale:*

$$P[K_1 = 1] = 1, \quad P[K_{n+1} = j | K_n = k] = \begin{cases} \theta/(\theta+n) & \text{se } j = k+1 \\ n/(\theta+n) & \text{se } j = k \\ 0 & \text{se } j \neq k, k+1 \end{cases}$$

Dove $\theta > 0$ è il parametro del processo del ristorante cinese corrispondente.

Possiamo, per esempio, vedere quali sono i valori di $\gamma_k^{(n)}$ per $n = 2$:

$$\gamma_1^{(2)} = P[K_2 = 1] = P[K_2 = 1 | K_1 = 1] = \frac{1}{\theta+1}$$

$$\gamma_2^{(2)} = P[K_2 = 2] = P[K_2 = 2 | K_1 = 1] = \frac{\theta}{\theta+1}$$

Dimostriamo, ora, la scrittura generica di $\gamma_k^{(n)}$. Per farlo ci serviremo dei numeri di Stirling di prima specie $s(n,k)$ la cui teoria è approfondita nell'appendice A; in particolare useremo il teorema A.0.1.

Diamo, quindi, il seguente risultato:

Teorema 2.4.2. *Se $\{K_n\}_{n\geq 1}$ è una catena di Markov con probabilità assegnata*

$$P[K_1 = 1] \text{ e } P[K_{n+1} = j | K_n = k] = \begin{cases} \theta/(\theta+n) & \text{se } j = k+1 \\ n/(\theta+n) & \text{se } j = k \\ 0 & \text{se } j \neq k, k+1 \end{cases}$$

Allora vale che

$$P[K_n = k] = \frac{|s(n,k)|\theta^k}{(\theta)_{(n)\uparrow}}$$

Dimostrazione. La dimostrazione avviene per induzione su n:

- $n = 1$, ovvio.

- Sia $n > 1$ e valga l'ipotesi induttiva per n.

Sfruttiamo la formula delle probabilità totali:

$$\begin{aligned}
P[K_{n+1}=k] &= P[K_{n+1}=k|K_n=k]P[K_n=k] + P[K_{n+1}=k|K_n=k-1]P[K_n=k-1] \\
&= \frac{n}{\theta+n}\frac{|s(n,k)|\theta^k}{\theta_{(n)\uparrow}} + \frac{\theta}{\theta+n}\frac{|s(n,k)|\theta^{k-1}}{\theta_{(n)\uparrow}} \\
&= \frac{\theta^k[n|s(n,k)| + |s(n,k-1)|]}{\theta_{(n+1)\uparrow}} \\
&= \frac{\theta^k|s(n+1,k)|}{\theta_{(n+1)\uparrow}}
\end{aligned}$$

□

Abbiamo ricavato, quindi, che

$$\gamma_k^{(n)} = \frac{\theta^k|s(n,k)|}{\theta_{(n)\uparrow}} \tag{2.5}$$

Studiamo, ora, i coefficienti $\Gamma_k^{(n)}(n_1,\dots,n_k)$.

Dato $(\nu_1,\nu_2,\dots,\nu_k)$ il vettore aleatorio a valori in $\mathcal{P}_{n,k}$, $\Gamma_k^{(n)}(n_1,\dots,n_k)$ è definito nel seguente modo:

$$\Gamma_k^{(n)}(n_1,\dots,n_k) = P[\nu_1=n_1,\dots,\nu_k=n_k|K_n=k]$$

Ci interessa, quindi, calcolare la cardinalità dell'insieme $\mathcal{P}_{n,k}$, che contiene tutte le possibili k-uple di numeri in $\{1,\dots,n\}$ che sommano a n. Queste k-uple entrano in gioco nel teorema A.0.1 sui numeri di Stirling. Da esso segue che

$$\Gamma_k^{(n)}(n_1,\dots,n_k) \propto \frac{1}{n_1\cdot\dots\cdot n_k}$$

Tuttavia, dobbiamo osservare che, a differenza del teorema, noi stiamo considerando solamente le k-uple ordinate $(n_1,\dots,n_k)$, quindi per calcolare $\Gamma_k^{(n)}(n_1,\dots,n_k)$ dobbiamo dividere per il numero di permutazioni di ogni k-upla.

Notiamo che per fare ciò è sufficiente dividere per $\prod_{j=1}^n m_j!$, dove gli m_j, $j=1,\dots,n$ sono definiti come in 2.2. Quindi, sfruttando il teorema A.0.1 come dicevamo, varrà che

$$\Gamma_k^{(n)}(n_1,\dots,n_k) = \frac{n!}{|s(n,k)|}\frac{1}{\prod_{i=1}^k n_i \prod_{j=1}^n m_j!} \tag{2.6}$$

Si tratta ora, infine, di calcolare gli ultimi coefficienti $1/P_k^{(n)}(n_1,\dots,n_k)$ che corrispondono a una probabilità uniforme sugli insiemi $\Pi_k^{(n)}(n_1,\dots,n_k)$, quindi varrà semplicemente che

$$P_k^{(n)}(n_1,\dots,n_k) = card(\Pi_k^{(n)}(n_1,\dots,n_k))$$

Da un semplice conteggio segue che

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

Sostituiamo, quindi, tutti i coefficienti nella formula 2.4:

$$P[X_1 \in A_1,\dots,X_n \in A_n] =$$

$$= \sum_{k=1}^{n} \frac{|s(n,k)|\theta^k}{(\theta)n\uparrow} \sum_{\substack{(n_1,\dots,n_k)\\ \in \mathcal{P}_{n,k}}} \frac{n!}{|s(n,k)|} \frac{1}{\left(\prod_{i=1}^k n_i!\right)\left(\prod_{j=1}^n m_j!\right)} \frac{\left(\prod_{i=1}^k n_i\right)\left(\prod_{j=1}^n m_j!\right)}{n!} \cdot$$

$$\cdot \sum_{\substack{I_1,\dots,I_k\\ \in\Pi_k^{(n)}(n_1,\dots,n_k)}} \prod_{j=1}^{k} \bar{\alpha}(\bigcap_{i\in I_j} A_i)$$

$$= \sum_{k=1}^{n} \frac{\theta^k}{(\theta)n\uparrow} \sum_{\substack{(n_1,\dots,n_k)\\ \in \mathcal{P}_{n,k}}} \prod_{i=1}^{k}(n_i-1)! \sum_{\substack{I_1,\dots,I_k\\ \in\Pi_k^{(n)}(n_1,\dots,n_k)}} \prod_{j=1}^{k} \bar{\alpha}(\bigcap_{i\in I_j} A_i)$$

$$= \sum_{k=1}^{n} \sum_{\substack{(n_1,\dots,n_k)\\ \in\Pi_k^{(n)}}} \frac{\theta^k}{(\theta)n\uparrow} \prod_{i=1}^{k}(n_i-1)! \sum_{\substack{I_1,\dots,I_k\\ \in\Pi_k^{(n)}(n_1,\dots,n_k)}} \prod_{j=1}^{k} \bar{\alpha}(\bigcap_{i\in I_j} A_i)$$

$$= \sum_{k=1}^{n} \sum_{\substack{(n_1,\dots,n_k)\\ \in\Pi_k^{(n)}}} Ewens^{(n)}(n_1,\dots,n_k;\theta) \sum_{\substack{I_1,\dots,I_k\\ \in\Pi_k^{(n)}(n_1,\dots,n_k)}} \prod_{j=1}^{k} \bar{\alpha}(\bigcap_{i\in I_j} A_i)$$

Abbiamo, quindi, ricavato, come volevamo, la formula di Ewens come peso della legge di probabilità 2.4 delle $X_1,\dots,X_n$.

Notiamo che, senza effettuare le semplificazioni, abbiamo ottenuto una definizione "estesa" della formula di Ewens:

$$Ewens^{(n)}(n_1,\dots,n_k;\theta) = \sum_{k=1}^{n} \frac{|s(n,k)|\theta^k}{(\theta)n\uparrow} \cdot$$

$$\cdot \sum_{\substack{(n_1,\dots,n_k)\\ \in \mathcal{P}_{n,k}}} \frac{n!}{|s(n,k)|} \frac{1}{\left(\prod_{i=1}^k n_i!\right)\left(\prod_{j=1}^n m_j!\right)} \frac{\left(\prod_{i=1}^k n_i\right)\left(\prod_{j=1}^n m_j!\right)}{n!}$$

Essa, infatti, ci permette di dimostrare in maniera semplice il legame tra la formula di Ewens 2.1 e la formula di Ewens parziale 2.2. Per ottenere il 2.2 da 2.1 dobbiamo moltiplicare la formula di Ewens per la probabilità delle singole

partizioni $P_k^{(n)}(n_1, \ldots, n_k)$ che non è considerata in 2.2:

$$\frac{\theta^k}{\theta_{(n)\uparrow}} \prod_{i=1}^{k} (n_i - 1)! \binom{n}{n_1 \ldots n_k} \frac{1}{m_1! \ldots m_k!} = \frac{\prod_{j=1}^{k} \theta^{m_j}}{\theta_{(n)\uparrow}} \frac{n!}{\left(\prod_{j=1}^{n} m_j!\right) \prod_{j=1}^{n} j^{m_j}}$$
$$= \frac{n!}{\theta_{(n)\uparrow}} \prod_{j=1}^{n} \frac{\theta^{m_j}}{m_j! j^{m_j}} \tag{2.7}$$

Possiamo anche notare che la struttura ricorsiva trova riscontro nella seguente relazione fondamentale:

Proposizione 2.4.1. *Definita* $\mu_n(A_1 \times \cdots \times A_n) = P[X_1 \in A_1, \ldots, X_n \in A_n]$, *vale che*

$$\mu_{n+1}(A_1 \times \cdots \times A_n \times \mathbb{X}_1) = \mu_n(A_1 \times \cdots \times A_n)$$

La dimostrazione segue dalla linearità della misura μ rispetto alla somma.

Possiamo, infine, fare una semplice verifica della normalizzazione della formula di Ewens.

Per farlo usiamo il teorema (A.0.1).

$$\sum_{k=1}^{n} \sum_{\substack{\text{partizione} \\ \{B_1,\ldots,B_k\} \\ \text{di } \{1,\ldots,n\}}} P[\Pi_n = B_1, \ldots, B_k] = \sum_{k=1}^{n} \sum_{\substack{\text{partizione} \\ \{B_1,\ldots,B_k\} \\ \text{di}\{1,\ldots,n\}}} \frac{\theta^k}{(\theta)_{n\uparrow}} \prod_{j=1}^{k} (n_j - 1)!$$
$$= \sum_{k=1}^{n} \sum_{\substack{\text{partizione} \\ \{B_1,\ldots,B_k\} \\ \text{di}\{1,\ldots,n\}}} \frac{\theta^k}{(\theta)_{n\uparrow}} \prod_{j=1}^{k} (n_j - 1)! =$$
$$= \sum_{k=1}^{n} \frac{\theta^k}{(\theta)_{n\uparrow}} \sum_{\substack{(n_1,\ldots,n_k) \in \mathbf{N}^k: \\ n_1 + \cdots + n_k = n}} \frac{1}{k!} \binom{n}{n_1 \ldots n_k} \prod_{j=1}^{k} (n_j - 1)!$$
$$= \sum_{k=1}^{n} \frac{\theta^k}{(\theta)_{n\uparrow}} \sum_{(n_1,\ldots,n_k)} \frac{n!}{k!} \frac{1}{n_1 \ldots n_k}$$
$$= \sum_{k=1}^{n} \frac{\theta^k |s(n,k)|}{(\theta)_n \uparrow} = 1$$

2.5 Costruzione Monte Carlo

La costruzione che abbiamo dato precedentemente della formula di Ewens evidenzia aspetti computazionali interessanti che ci hanno suggerito l'inserimento di una deduzione "Monte Carlo" della formula. Nelle sezioni precedenti, infatti, abbiamo osservato che la formula di Ewens fornisce le probabilità per le partizioni dell'insieme $\{1, \ldots, n\}$. È possibile, quindi, sfruttare un metodo Monte Carlo simulando la generazione casuale delle partizioni e confrontando la

probabilità in senso frequentista ottenuta dagli esperimenti con le aspettative teoriche della formula di Ewens. La simulazione, peraltro, restituisce il senso sperimentale racchiuso nella formula di Ewens dal momento che ne ricaviamo i valori da dei dati di frequenza.

L'implementazione, in linea concettuale, non è complessa: essa si basa sulla formula 2.4, cioè sulla seguente scrittura della formula di Ewens che si ricava dalla legge di probabilità:

$$Ewens^{(n)}(n_1,\dots,n_k;\theta) = \sum_{k=1}^{n} \gamma_k^{(n)} \sum_{\substack{(n_1,\dots,n_k)\\ \in \mathcal{P}_{k,n}}} \Gamma_k^{(n)}(n_1,\dots,n_k)/P_k^{(n)}(n_1,\dots,n_k)$$

Questa formula regola, infatti, la probabilità delle generazioni casuali delle partizioni di $\{1,\dots,n\}$ con cardinalità degli insiemi delle partizioni $n_1,\dots,n_k$ e parametro di probabilità θ.

Essa fornisce una procedura algoritmica in tre step per la costruzione di una partizione casuale con probabilità assegnata:

(A) Generazione di $k \in \{1,\dots,n\}$ numero di insiemi della partizione. Come abbiamo ricavato precedentemente, k viene generato con probabilità $\gamma_k^{(n)} = \theta^k |s(n,k)|/\theta_{(n)\uparrow}$.

(B) Fissato k, generazione di $n_1,\dots,n_k \in \{1,\dots,n\}$ con $n_1+\dots+n_k = n$, cardinalità degli insiemi della partizione. Questi numeri vengono generati con probabilità $\Gamma_k^{(n)}(n_1,\dots,n_k) = n!/(|s(n,k)| \prod_{i=1}^n n_i \prod_{j_1}^n m_j!)$.

(C) Fissati $n_1,\dots,n_k$, generazione della partizione $B_1,\dots,B_k$ di $\{1,\dots,n\}$ con insiemi di cardinalità $n_1,\dots,n_k$ con probabilità uniforme.

L'implementazione avviene, quindi, tramite la costruzione di tre procedure che corrispondono ai tre passi dell'algoritmo. L'esecuzione delle tre procedure in sequenza permette di generare una partizione casuale con la probabilità desiderata. L'implementazione del metodo Monte Carlo consiste, poi, nell'iterazione per un numero alto di volte di questa procedura e nella misurazione della frequenza di apparizione di ciascuna partizione. Normalizzando, poi, per il numero totale di iterazioni si ottiene la probabilità "frequentista" di ciascuna partizione. Per la legge dei grandi numeri, all'aumentare del numero di iterazioni la probabilità frequentista converge al valore teorico della speranza matematica che, in questo caso, corrisponde ai valori della formula di Ewens.

Riportiamo, ora, alcuni risultati che abbiamo ottenuto effettuando gli esperimenti.

Ciò che otteniamo tramite le nostre simulazioni sono tre vettori contenenti il primo la lista delle partizioni, il secondo i valori teorici attesi, ottenuti tramite la formula di Ewens, e il terzo le frequenze, normalizzate per il numero di esperimenti, di ciascuna partizione.

Ad esempio, effettuando 10000 esperimenti, fissati $n = 5$ e $\theta = 2$, otteniamo la seguente tabella delle partizioni (la tabella integrale comprenderebbe

52 entrate, noi riportiamo le prime 5 associando a ogni partizione la stima di frequenza e il valore atteso teorico):

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
...	...	...

Possiamo aspettarci, come conseguenza della legge dei grandi numeri, che all'aumentare del numero di esperimenti l'errore effettuando stimando le probabilità con il metodo Monte Carlo diminuisca. Ciò effettivamente avviene, come mostrato nella figura 2.2 dove abbiamo riportato l'errore in norma due in relazione al numero di iterazioni. Notiamo anche che la decrescita dell'errore non è monotona per via del fatto che i risultati di ogni esperimento sono aleatori.

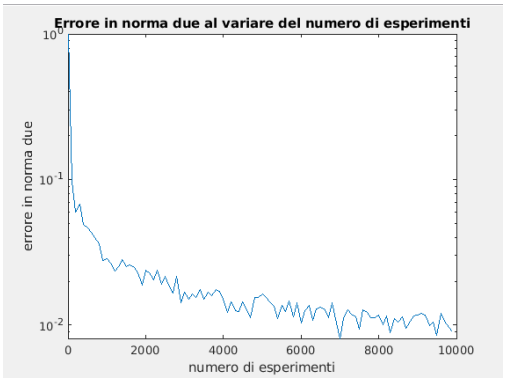

Figura 2.2: Errore metodo Monte Carlo

I dettagli sull'implementazione e il codice usato sono riportati nell'appendice C.

Le conseguenze computazionali della formula di Ewens, d'altra parte, sono molto rilevanti dal momento che lo studio delle partizioni di un insieme ha riscontri importanti in combinatoria e permette di costruire algoritmi implementativi per applicazioni matematiche [11]. Ad esempio, la formula di Ewens si può dedurre anche tramite la formula di *Faà di Bruno* che restituisce il valore della derivata k-esima mista $f(g(x))^{(k)}$. L'ordine delle derivate nello sviluppo della derivata mista è lo stesso degli $n_1,\dots,n_k$ che abbiamo considerato nella costruzione della formula di Ewens e questo suggerisce un'applicazione computazionale della formula.

Capitolo 3

Modello di Wright-Fisher

Una delle applicazioni più interessanti della formula di Ewens è allo studio della genetica delle popolazioni. La formula, infatti, nasce proprio in quest'ambito, dal momento che nella derivazione originale di Ewens essa è usata per analizzare il campionamento di diversi alleli di un singolo gene all'interno di una popolazione.

La formula nella sua interpretazione originaria emerge in un modello di dinamica delle popolazioni, il modello di Wright-Fisher [6], che è l'oggetto del presente capitolo.

3.1 Versione base del modello

Per descrivere il modello introduciamo le seguenti definizioni:

Definizione 3.1.1 (locus genico)**.** *Intendiamo per locus genico la posizione di un gene all'interno del genoma di un organismo, per esempio possiamo immaginarlo come la sequenza di nucleotidi che compongono il gene.*

Definizione 3.1.2 (allele)**.** *Intendiamo come allele una manifestazione di un gene, per esempio negli esperimenti di Mendel sui piselli due alleli di uno stesso gene sono l'essere ruvidi o l'essere lisci.*

Definizione 3.1.3 (fitness di un individuo)**.** *La fitness di un individuo è la sua abilità a sopravvivere e riprodursi; nel nostro caso parliamo di evoluzione neutra, cioè facciamo l'ipotesi che le mutazioni non abbiano influenza sulla fitness dell'individuo.*

Definizione 3.1.4 (individui diploidi)**.** *Sono diploidi tutti quegli individui che, come gli esseri umani, possiedono due copie del materiale genetico in ogni cellula: in una popolazione di N individui diploidi avremo quindi $2N$ copie del materiale genetico.*

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. Nel modello si suppone che questi abbiano tutti la stessa fitness. Inoltre, inizialmente faremo l'ipotesi che esistano due soli possibili alleli A e a e non possano avvenire mutazioni.

Dal punto di vista probabilistico possiamo rappresentare la popolazione alla n-esima generazione come un insieme di $2N$ palle in un'urna: i di queste saranno segnate come A e $2N-i$ come a, in base all'allele relativo ad ogni copia genica. A questo punto, per costruire la $(n+1)$-esima generazione estraiamo $2N$ palle dall'urna con reimmissione.

Definiamo, ora, X_n una variabile aleatoria che indica il numero di alleli A nella generazione n-esima: vogliamo studiare il comportamento di X_n. Notiamo, innanzitutto, che X_n è una catena di Markov, dato che vale la proprietà di assenza di memoria; inoltre X_n ha distribuzione binomiale con parametro $p_i = i/2N$, quindi varrà che:

$$P[X_{n+1} = j | X_n = i] = \binom{2N}{j} p_i^j (1-p_j)^{2N-j}$$

Una proprietà che può essere interessante studiare è l'eterozigosi, che definiamo come segue:

Definizione 3.1.5 (eterozigosi)**.** *Chiamiamo eterozigosi la probabilità che due copie distinte dello stesso locus genico corrispondano, al tempo n, a due alleli differenti. Il valore di eterozigosi è descritto da una variabile aleatoria che nel nostro modello è la seguente:*

$$H_n^0 = \frac{2X_n(2N - X_n)}{2N(2N-1)}$$

Il valore atteso dell'eterozigosi è descritto dal seguente teorema:

Teorema 3.1.1. *Sia $h(n) = \mathbb{E}(H_n^0)$ il valore atteso dell'eterozigosi al tempo n, nel modello di Wright-Fisher varrà che:*

$$h(n) = \left(1 - \frac{1}{2N}\right)^n \cdot h(0)$$

3.2 Coalescenza e genealogia

Vogliamo studiare, ora, la genealogia degli individui. Immaginiamo sempre ogni individuo come una palla segnata con A o a in base all'allele corrispondente. Ricordiamo che la generazione di un individuo avviene tramite l'estrazione di una palla all'interno dell'urna della generazione precedente: chiamiamo genitore di un individuo la palla della generazione precedente corrispondente. In termini genetici, indichiamo come genitore l'individuo da cui viene ereditato il carattere studiato, cioè nel nostro caso l'allele A o a. Dato che le estrazioni avvengono con reimmissione, è possibile che due individui abbiano lo stesso genitore, corrispondente all'estrazione della stessa palla. Indichiamo poi come lignaggio la sequenza degli "antenati" di un dato individuo; notiamo che due lignaggi possono fondersi quando due individui possiedono un antenato comune.

Innanzitutto possiamo osservare che se la popolazione è sufficientemente ampia, cioè N è sufficientemente grande, la probabilità che in un campione di k individui due abbiano lo stesso genitore è circa $k(k-1)/2 \cdot 2N$.

Inoltre vale il seguente teorema sul tempo per ottenere k lignaggi:

Teorema 3.2.1. *Il tempo t_k per ottenere k lignaggi è una variabile aleatoria con distribuzione esponenziale di media 2/k(k-1).*

Un concetto fondamentale nello studio della genealogia è quello di coalescenza: parliamo di coalescenza quando abbiamo la fusione di due lignaggi per via dell'individuazione di un antenato comune.

Possiamo indicare con T_j il primo momento in cui abbiamo j lignaggi. In tal caso T_1 indica il tempo corrispondente al più recente antenato comune di tutta la popolazione. Facciamo attenzione al fatto che stiamo misurando i tempi all'indietro, cioè fissiamo come tempo 0 il momento attuale e facciamo scorrere in avanti il tempo mentre ripercorriamo all'indietro le generazioni, quindi $T_i > T_j$ se $i < j$.

Possiamo fare un esempio calcolando il tempo medio trascorso con k lignaggi per $k = 2, 3, 4, 5$ sfruttando il teorema 3.2.1:

$$\mathbb{E}[t_2] = 1 \quad \mathbb{E}[t_3] = \frac{1}{3} \quad \mathbb{E}[t_4] = \frac{1}{6} \quad \mathbb{E}[t_5] = \frac{1}{10}$$

Per un campione di n individui varrà che $T_1 = t_2 + \cdots + t_n$, quindi il valore medio del tempo trascorso dal più recente antenato comune è

$$\mathbb{E}[T_1] = \sum_{k=2}^{n} \frac{2}{k(k-1)} = 2\sum_{k=2}^{n}\Big(\frac{1}{k-1} - \frac{1}{k}\Big) = 2\Big(1 - \frac{1}{n}\Big)$$

Questa quantità tende a 2 per $n \to \infty$: questo significa che il tempo impiegato per l'ultima coalescenza è almeno la metà di quello impiegato per la coalescenza dell'intera popolazione.

Le analogie con la formula di campionamento di Ewens iniziano ora dall'osservazione che lo stato della coalescenza a un tempo T_j di una popolazione di n individui può essere visto come una partizione dell'insieme $\{1, \ldots, n\}$. Al tempo 0 la partizione consisterà in n singoletti $\{1\}, \ldots, \{n\}$ e ogni coalescenza corrisponderà all'unione di due partizioni.

Possiamo fare già alcune considerazioni sul comportamento di queste partizioni: chiamiamo $\mathcal{E}_n$ l'insieme delle partizioni di $\{1, \ldots, n\}$ e, dato $\xi \in \mathcal{E}_n$, sia $|\xi|$ il numero di insiemi in ξ. Indichiamo, inoltre, con ξ_i^n la partizione di $\{1, \ldots, n\}$ al tempo T_i. Allora vale il seguente teorema:

Teorema 3.2.2. *Sia ξ una partizione di $\{1, \ldots, n\}$ con $|\xi| = i$, allora*

$$P[\xi_i^n = \xi] = c_{n,i} w(\xi)$$

Dove $w(\xi) = \lambda_1! \ldots \lambda_i!$ con $\lambda_i, \ldots, \lambda_i$ sono le cardinalità degli i insiemi della partizione e la costante $c_{n,i}$ è scelta in modo da far sì che la somma delle probabilità faccia 1.

Valgono anche i seguenti due risultati:

Teorema 3.2.3. *La probabilità che l'antenato comune più recente di un campione di n individui sia lo stesso dell'intera popolazione tende a $(n-1)(n+1)$ al tendere dell'ampiezza della popolazione a infinito.*

Teorema 3.2.4. *Per ricostruire la partizione* ξ_i^n *da* $\xi_{i-1}^n = \{A_1, \ldots, A_{i-1}\}$ *scegliamo casualmente un insieme selezionando* A_j *con probabilità* $(\lambda_j - 1)(n - i - 1)$ *dove* $\lambda_j = |A_j|$ *e lo dividiamo in due sottoinsiemi di dimensioni* k *e* $\lambda_j - k$ *con* k *scelto uniformemente in* $\{1, 2, \ldots, \lambda_j - 1\}$.

3.3 Modello a infiniti alleli

Per introdurre le mutazioni nel nostro modello facciamo l'ipotesi degli "infiniti alleli": secondo questa supposizione, infatti, esiste un numero di possibili alleli talmente elevato (idealmente infinito) che ci permette di supporre che ogni mutazione porti a un allele di un nuovo tipo non ancora registrato. Questa assunzione è motivata da un semplice calcolo probabilistico: immaginiamo che un gene consista in 500 nucleotidi, ciascuno dei quali può avere una base diversa. Il numero di possibili sequenze di DNA sarà

$$4^{500} = 10^{500\ log4/log10} = 10^{301}$$

Per ciascuna di queste sequenze, ve ne sono $3 \cdot 500 = 1500$ che possono essere ottenute da un singolo cambiamento di base, quindi la probabilità di tornare alla sequenza iniziale in 2 mutazioni è $1/1500$, da cui l'assunzione che il numero degli alleli possibili è essenzialmente infinito.

Il modello a infiniti alleli si sfrutta quando è necessario usare metodi indiretti per dedurre differenze tra gli individui. Per esempio, si sfrutta per degli studi sul numero di alleli differenti in un campione di popolazione studiato. Il punto di partenza è la cosiddetta partizione allelica, in cui si danno dei valori m_j pari al numero di alleli con j manifestazioni. Ad esempio, negli studi [4] e [16] si è osservata per la *Drosophila* la seguente partizione allelica:

$$m_1 = 10, \quad m_2 = 3, \quad m_4 = 1, \quad m_{32} = 1$$

Cioè sono stati trovati 23 diversi alleli per 60 individui: 10 alleli avevano una sola manifestazione, 3 alleli avevano 2 manifestazioni, un solo allele appariva 4 volte e un ultimo allele appariva 32 volte.

Notiamo che gli m_j corrispondono esattamente ai valori definiti in 2.2 e infatti ci condurranno alla stessa formula applicata al calcolo della probabilità di una certa partizione allelica.

Il modello a infiniti alleli si usa anche, ad esempio, per studiare le sequenze di DNA senza ricombinazione: in questo caso, non conteremo più il numero di alleli differenti ma il numero di aplotipi, cioè le possibili varianti delle sequenze nucleotidiche.

È nel modello a infiniti alleli che trova applicazione la formula di campionamento di Ewens che può essere usata per prevedere il comportamento della partizione allelica di una popolazione.

3.4 Formula di Ewens nel modello di Wright-Fisher

Come visto precedentemente, anche nel modello a infiniti alleli se vi sono k lignaggi la coalescenza avviene con probabilità

$$\frac{k(k-1)}{2}\frac{1}{2N}$$

Tuttavia, ora, in seguito a una mutazione, un lignaggio può scomparire con probabilità $k\mu$, dove μ è il tasso di mutazione. Riscalando il tempo per $2N$, otteniamo che i tassi di coalescenza e mutazione sono rispettivamente $k(k-1)/2$ e $k\theta/2$ dove $\theta = 4N\mu$.

Questo modello si può rileggere come uno schema di urna di Hoppe o, equivalentemente, un processo del ristorante cinese. Infatti, è sufficiente immaginare le discendenze come la selezione da un'urna di Hoppe dove vi è una palla nera di massa θ, corrispondente al caso di una mutazione, e una palla colorata di massa 1 per ogni lignaggio. Le palle vengono scelte con reimmissione con probabilità proporzionale al loro peso: alla scelta di una palla colorata si associa la prosecuzione del lignaggio corrispondente, mentre alla palla nera una nuova a mutazione. Proseguendo a ritroso nell'urna di Hoppe incontriamo una mutazione con probabilità $\theta/(\theta + k)$ e una coalescenza con probabilità $k/(\theta + k)$ che corrispondono ai tassi precedenti. Infatti, posto che abbiamo $k + 1$ lignaggi nel momento in cui consideriamo la coalescenza, il rapporto tra il tasso di coalescenza e quello di mutazione è proprio pari alla probabilità di coalescenza.

Vale, quindi, il seguente risultato:

Teorema 3.4.1. *La relazione genealogica tra k lignaggi nel modello a infiniti alleli può essere simulata attraverso k ripetizioni dell'esperimento dell'urna di Hoppe.*

Per studiare la popolazione nel modello introduciamo la variabile aleatoria K_n che misura il numero di diversi alleli in un campione di dimensione n: essa corrisponde alla variabile che avevano introdotto per misurare il numero di partizioni quando abbiamo ricavato la formula di Ewens.

Asintoticamente il comportamento di K_n è descritto dal seguente teorema:

Teorema 3.4.2. *Fissato il valore di θ valgono le seguenti equivalenze asintotiche per $n \to \infty$*

$$\mathbb{E}[K_n] \sim \theta \log n \quad Var(K_n) \sim \theta \log n$$

Una conseguenza interessante del teorema è che $K_n/logn$ è uno stimatore asintoticamente normale del tasso di mutazione riscalato θ. Tuttavia, si dimostra che la deviazione standard dello stimatore è dell'ordine di $1/\sqrt{\log n}$: ciò significa che se, per esempio, il valore reale di θ è 1 e vogliamo stimarlo con un errore di $0,1$, allora dobbiamo avere un campione di dimensione e^{100}, portando quindi a tempi di calcolo molto grandi. Tuttavia non esiste un altro modo più veloce per stimare θ dai dati.

Mentre il teorema precedente descriveva il comportamento asintotico del numero di alleli, la distribuzione di essi ci è data proprio dalla formula di Ewens, che riassumiamo nel seguente teorema:

Teorema 3.4.3 (formula di campionamento di Ewens)**.** *Sia m_i il numero di alleli presenti i volte in un campione di dimensione n e quindi $(m_1, \dots, m_n)$ la partizione allelica della popolazione. Sia, inoltre, $\theta = 4N\mu$ il tasso di mutazione riscalato. Allora, la probabilità assegnata alla partizione allelica è data dalla seguente formula:*

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

La dimostrazione è ovvia per via delle analogie già mostrate con la formula di Ewens per le partizioni, in particolare abbiamo già rilevato che la costruzione della partizione allelica avviene tramite un processo di urna di Hoppe.

Possiamo fare un esempio di applicazione per un caso semplice, ponendo $n = 2$. Stiamo, quindi, studiando due individui: le possibili partizioni alleliche (m_1, m_2) sono quindi $(0, 1)$, che corrisponde a due alleli uguali, e $(2, 0)$, che corrisponde a due alleli differenti. Le probabilità calcolate tramite la formula di Ewens sono rispettivamente $1/(\theta + 1)$ e $\theta/(\theta + 1)$, quindi la probabilità di *omozigosi*, cioè due individui identici è di $1/(\theta + 1)$. Se assumiamo, quindi, che le mutazioni avvengano con probabilità 2μ, la probabilità di avere coalescenza prima di mutazione è effettivamente

$$\frac{\frac{1}{2N}}{2\mu + \frac{1}{2N}} = \frac{1}{1 + \theta}$$

come ci aspettavamo.

A questo punto, ricordiamo il seguente risultato, analogo a 2.5, il quale fornisce la distribuzione del numero di alleli distinti all'interno di un campione di n individui:

Teorema 3.4.4.

$$P_\theta[K_n = k] = \frac{\theta^k}{\theta_{(n)\uparrow}} |S(n, k)|$$

Combinando questo risultato con la formula di Ewens otteniamo il seguente teorema:

Teorema 3.4.5.

$$P_\theta[m_1, \dots, m_n | K_n = k] = \frac{n!}{|s(n, k)|} \prod_{j=1}^{n} \left(\frac{1}{j}\right)^{m_j} \frac{1}{m_j!}$$

La formula ci permette di notare che la distribuzione allelica $(m_1, \dots, m_n)$ non dipende dal parametro θ.

Infine, facciamo alcune osservazioni sulla stima di θ basata sul valore di K_n. Vale infatti il seguente risultato:

Teorema 3.4.6. *K_n è una statistica sufficiente per stimare θ.*

Per stimare θ a partire da K_n usiamo il seguente stimatore di massima verosimiglianza

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

Che è la likelihood di osservare $K_n = k$ quando il vero parametro è θ. Ricavando il valore massimo dello stimatore e calcolandone la derivata otteniamo che:

$$k = \frac{\theta}{\theta} + \frac{\theta}{\theta+1} + \cdots + \frac{\theta}{\theta+n-1} = \mathbb{E}[K_n]$$

Quindi lo stimatore di massima verosimiglianza $\hat{\theta}$ è il valore di θ che fa sì che il numero medio di alleli sia uguali al numero di alleli osservati.

Le proprietà degli stimatori di massima verosimiglianza ci dicono che asintoticamente vale che $\mathbb{E}[\hat{\theta}]$ e $Var(\hat{\theta}) = 1/I(\hat{\theta})$, dove $I(\theta)$ è l'informazione di Fisher. Si calcola facilmente che

$$I(\theta) = \frac{1}{\theta^2} Var(K_n)$$

Quindi osserviamo, infine, per il teorema 3.4.2, che se $n \to \infty$, $Var(\hat{\theta}) \to 0 \sim 1/\log n$, quindi aumentando il campione la varianza dello stimatore θ tende a 0 ma converge lentamente.

Conclusioni e orizzonti

In questa tesi abbiamo descritto le proprietà, la derivazione e le applicazioni della formula di campionamento di Ewens per lo studio della probabilità delle partizioni di un insieme discreto $\{1, \dots, n\}$. Nel primo capitolo introduttivo abbiamo presentato la costruzione del processo di Dirichlet a partire dalla struttura delle leggi finito-dimensionali, dando il concetto di variabili aleatorie scambiabili. Ci siamo poi dedicati, nel secondo capitolo, al cuore della tesi cioè la deduzione della formula di Ewens, che è avvenuta per via diretta tramite la struttura delle leggi finito-dimensionali del processo di Dirichlet e per via ricorsiva tramite il processo di ristorante cinese. Abbiamo, inoltre, inserito anche una deduzione Monte Carlo della formula a partire da alcuni esperimenti di generazione casuale delle partizioni di $\{1, \dots, n\}$. Nel terzo capitolo, infine, abbiamo presentato un'applicazione della formula di campionamento di Ewens a un modello di dinamica delle popolazioni, il modello di Wright-Fisher, dove la formula di Ewens fornisce la distribuzione di una data partizione allelica all'interno di una popolazione. L'applicazione agli studi genetici, oltre che fornire un utilizzo concreto della formula, permette anche di risolvere uno dei problemi di apertura, cioè lo studio della biodiversità di una popolazione a partire da un campione osservato: il parallelismo è semplice, invece che gli alleli è sufficiente considerare le specie degli individui e si riottiene che queste si distribuiscono secondo la formula di Ewens con un parametro θ che è legato al tasso di osservazione di nuove specie. In questo modo si può allora risolvere il problema degli studi statistici in una popolazione di cui non si conosce a priori lo spazio campionario: è sufficiente virare l'analisi dal campione in sé alle partizioni aleatorie dei suoi elementi; invece che associare, quindi, a ogni elemento osservato una manifestazione nello spazio campionario, si associano tra loro in una stessa partizione gli elementi osservati che condividono la stessa proprietà: lo spazio campionario studiato sarà quindi sempre noto perché sarà l'insieme delle partizioni di $\{1, \dots, n\}$, dove n è il numero di osservazioni.

La versatilità della formula di Ewens avrebbe permesso e meritato una trattazione ben più ampia di quella fatta in questo elaborato. Approfittiamo, però, di questa sezione per presentare alcune possibili orizzonti futuri, sia in ambito applicativo che di ricerca, a partire dal lavoro svolto in questa tesi. Una prospettiva di ricerca più astratta riguarda un argomento che abbiamo solamente sfiorato in questa tesi e che prende il nome di *calcolo umbrale* (in inglese *umbral calculus*). Si tratta di un tema di ricerca legato ai fattoriali ascendenti e, quindi, ai numeri di Stirling. Una domanda che ci si potrebbe porre a partire da quanto trattato, infatti, è se vi sono altri modi per scrivere la catena di Markov del

numero di insiemi in una partizione di n elementi K_n. In particolare avevamo ricavato che i suoi valori di probabilità sono i $gamma_k^{(n)}$ la cui espressione è contenuta in 2.5. Dall'osservazione della nostra deduzione si può, infatti, pensare che i fattoriali ascendenti rivestano un ruolo più centrale. Il calcolo umbrale si occupa di studiare il parallelismo tra le equazioni polinomiali e altre tecniche tra cui appunto il fattoriale ascendente. Per esso vale, ad esempio, la seguente relazione:

$$(x+y)_{n\uparrow} = \sum_{k=0}^{n} \binom{n}{k} (x)_{n\uparrow}(y)_{n\uparrow}$$

Da cui risulta immediatamente chiaro il parallelismo con i polinomi di potenze. Un'idea di ricerca, quindi, può essere tentare di sostituire i pesi della formula di Ewens $\gamma_k^{(n)}$ con un polinomio $p(\theta) = \sum_{k=1}^{n} \alpha_k^{(n)} \theta^k$, ottenendo qualcosa del tipo

$$\gamma_k^{(n)} = \frac{\alpha_k^{(n)} \theta^k}{p(\theta)}$$

La sfida è costruire $p(\theta)$ in modo da riottenere la ricorsività del teorema 2.4.1.

Un'applicazione recente e curiosa del processo di Dirichlet e della formula di Ewens è alla linguistica. La recente esplosione del numero di elaborati scritti prodotti tramite le nuove tecnologie (email, social network) ha fornito, infatti, numerosi dati da studiare [19]. In particolare risulta interessante analizzare la frequenza delle parole in un testo: il parallelismo con la formula di Ewens è immediato. Empiricamente si era notato che le parole si distribuivano in base alla loro frequenza lungo una curva $\sim 1/k^{\alpha}$. Tuttavia studi più recenti hanno mostrato che l'andamento segue un regime differente per le parole più frequenti che decade molto meno rapidamente e corrisponde alla distribuzione del processo di Dirichlet (come se il parametro α fosse infinto). L'obiettivo attuale di ricerca è quello di trovare un unico modello che descriva entrambi questi regimi e di costruirlo come la formula di Ewens. Un modello di questo tipo potrebbe avere applicazioni molto interessanti come lo studio della varietà di una lingua (intesa come il numero di parole) a partire dagli elaborati posseduti: ad esempio, si potrebbe applicare all'analisi delle lingue morte o delle lingue del passato di cui non conosciamo l'uso corrente ma di cui possediamo campioni di osservazioni dati dai testi giunti fino a noi.

Inoltre esiste un parallelismo molto attuale tra questo studio linguistico e uno studio ecologico della biodiversità delle popolazioni, tema che abbiamo già trattato nell'elaborato. Come mostrato in [19], infatti, la frequenza delle specie animali in un ecosistema segue la stessa distribuzione delle parole in un testo, quindi la formula di Ewens potrebbe intervenire nella definizione di indici di biodiversità. Il tema è particolarmente attuale dal momento che la conservazione della biodiversità è un argomento dell'agenda 2030 dell'ONU, un documento redatto nel 2015 per indicare i traguardi da perseguire per uno sviluppo sostenibile. Ad esempio, nell'obiettivo 15.5 si legge *"intraprendere azioni efficaci ed immediate per ridurre il degrado degli ambienti naturali, arrestare la distruzione della biodiversità"*[2] e la parola biodiversità appare ben 8 volte nell'agenda, a sottolineare che si tratta di un problema attuale e che sarà centrale ancora nei

prossimi anni. La definizione di strumenti statistici per la misurazione della varietà delle specie viventi, che permettano di identificare i trend e suggerire gli interventi possibili è, quindi, una questione che si continuerà ad affrontare anche nei prossimi anni e la formula di Ewens può giocare un ruolo fondamentale in questa sfida.

Appendice A

Numeri di Stirling

I numeri di Stirling di prima specie [3] sono i coefficienti dell'espansione dei fattoriali in potenze. Essi trovano varie applicazioni in probabilità e combinatoria: ad esempio, nella nostra trattazione abbiamo usato i numeri di Stirling di prima specie in merito al loro legame con i simboli di fattoriali crescenti e decrescenti $(x)_{n\uparrow}$ e $(x)_{n\downarrow}$.

Possiamo dare una definizione ricorsiva dei numeri di Stirling, come segue:

Definizione A.0.1 (Numeri di Stirling di prima specie senza segno)**.** *Dato $n \in \mathbb{N}$ e $k \in \mathbb{N}$, diamo la seguente definizione ricorsiva di $|s(n,k)|$:*

$$\begin{cases} |s(0,0)| = 1 \\ |s(n,0)| = 0, \quad n > 0 \\ |s(n,k)| = 0, \quad k > n \\ |s(n+1,k)| = |s(n,k-1)| + n|s(n,k)|, \quad n = 1,2,\ldots,n+1, \quad n = 0,1,\ldots \end{cases}$$

Definizione A.0.2 (Numeri di Stirling di prima specie)**.** *Dato $n \in \mathbb{N}$ e $k \in \mathbb{N}$, diamo la seguente definizione ricorsiva di $s(n,k)$:*

$$\begin{cases} s(0,0) = 1 \\ s(n,0) = 0, \quad n > 0 \\ s(n,k) = 0, \quad k > n \\ s(n+1,k) = s(n,k-1) - ns(n,k), \quad n = 1,2,\ldots,n+1, \quad n = 0,1,\ldots \end{cases}$$

Il legame fondamentale tra i fattoriali e i numeri di Stirling di prima specie sta nelle seguenti relazioni:

Proposizione A.0.1.

$$(x)_{n\downarrow} = \sum_{k=1}^{n} s(n,k)x^n$$

$$(x)_{n\uparrow} = \sum_{k=1}^{n} |s(n,k)|x^n$$

Dove $|s(n,k)|$ sono i numeri di Stirling senza segno.

Una proprietà fondamentale che abbiamo usato nella tesi è la seguente:

Teorema A.0.1. *Dato $n \in \mathbb{N}$, $k \in \mathbb{N}$ vale la seguente uguaglianza:*

$$|s(n,k)| = \frac{n!}{k!} \sum_{\substack{r_1,\dots,r_k \in \{1,\dots,n\}: \\ r_1+\cdots+r_k=n}} \frac{1}{r_1 \dots r_k}$$

Facciamo attenzione che nella somma le k-uple sono prese ordinate, quindi sommiamo più volte su ogni k-upla ottenuta permutando l'ordine degli addendi.

Appendice B

Numeri di Bell

I numeri di Bell [10] contano le possibili partizioni di un insieme finito. In particolare li possiamo definire nel seguente modo:

Definizione B.0.1 (Numeri di Bell)**.** *Fissato $n \in \mathbb{N}$, chiamiamo n-esimo numero di Bell B_n il numero di partizioni di un insieme di n elementi.*

I numeri di Bell si definiscono ricorsivamente nel seguente modo:

$$\begin{cases} B_0 = 1 \\ B_{n+1} = \sum_{k=0}^{n} \binom{n}{k} B_k \end{cases}$$

Notiamo subito che i numeri di Bell crescono molto rapidamente dal momento che per ricavare il numero n-esimo si sommano tutti i numeri precedenti opportunamente pesati. Infatti, ad esempio, i primi numeri di Bell sono:

$$B_1 = 1,\ B_2 = 2,\ B_3 = 5,\ B_4 = 15,\ B_5 = 52,\ B_6 = 203,\ B_7 = 877, \ldots$$

I numeri di Bell trovano, poi, diverse applicazioni: contano, ad esempio, il numero di diverse fattorizzazioni di un numero. Una delle applicazioni più interessanti consiste nel fatto che il numero di Bell n-esimo B_n conta i possibili schemi rimici di un componimento di n righe.

Appendice C

Implementazione del metodo Monte Carlo

Nella presente appendice presentiamo alcuni approfondimenti sugli aspetti implementativi della deduzione Monte Carlo della formula di Ewens. Abbiamo trattato gli aspetti teorici nella sezione 2.5. In particolare, nella seguente appendice riportiamo i codici utilizzati e spieghiamo brevemente alcune scelte implementative.

Le implementazioni sono avvenute tutte in MATLAB senza l'utilizzo di toolbox aggiuntivi oltre a quelli già built-in nella sua versione base. Nella scrittura del codice non ci siamo occupati degli aspetti computazionali e numerici di efficienza di esecuzione.

Come prima cosa, è stato necessario definire una funzione `Stirling` che calcolasse i numeri di Stirling di prima specie senza segno, e abbiamo sfruttato la loro definizione ricorsiva, come segue:

```
function s=Stirling(n,k)
% La funzione prende in inputi i valori di n e k e restituisce il
    numero di Stirling di prima specie senza segno corrispondente
    |s(n,k)|
    if n==0 && k==0
        s = 1;
    elseif k==0
        s = 0;
    elseif k>n
        s = 0;
    else
        s = Stirling(n-1,k-1)+(n-1)*Stirling(n-1,k);
    end
end
```

Listing C.1: Funzione Stirling

Abbiamo anche definito una funzione `risingFactorial` che calcolasse ricorsivamente il fattoriale ascendente di $x_{(n)\uparrow}$.

```
function x_rising_n = risingFactorial(x, n)
% La funzione prende in input un intero x e un numero naturale n e
     restituisce il fattoriale ascendente di n di x
    if n == 1
```

```
        x_rising_n = x;
    else
        for i=1:n
            x_rising_n = risingFactorial(x,n-1)*(x+n-1);
        end
    end
end
```

Listing C.2: Funzione risingFactorial

Un'altra procedura che abbiamo definito è la funzione `m from n` che prende i valori delle cardinalità degli insiemi della partizione $n_1, \dots, n_k \in \{1, \dots, n\}$ con $n_1 + \cdots + n_k = n$ e restituisce i valori $m_1, \dots, m_n$ in cui m_j indica il numero di insiemi della partizione con j elementi.

```
function mVector = m_from_n(nVector)
% La funzione prende in input i valori n_1,...,n_k in un vettore
    nVector e resituisce i corrispondenti valori m_1,...,m_n in un
     vettore mVector
    n = sum(nVector);
    mVector = zeros(1,n);
    for v=nVector
        for j=1:n
            if v == j
                mVector(j) = mVector(j)+1;
            end
        end
    end
end
```

Listing C.3: Funzione m from n

A questo punto, abbiamo definito le tre procedure fondamentali che corrispondono ai tre step dell'algoritmo del metodo Monte Carlo.

La prima, chiamata `randomNumberOfPartitions`, prende in input n e θ e genera casualmente un numero $k \in \{1, \dots, n\}$ con probabilità data dai valori 2.5. Per la generazione casuale con probabilità assegnata abbiamo usato solamente la funzione di MATLAB `rand` che genera un numero casuale con probabilità uniforme. Per ottenere le probabilità volute abbiamo ragionato nel seguente modo: abbiamo diviso il segmento $[0, 1]$ in base alle probabilità $\gamma_k^{(n)}$ e abbiamo generato con `rand` un numero casuale x nell'intervallo $[0, 1]$. La funzione restituisce, poi, il valore k corrispondente al più piccolo tra i valori di $\sum_{i=1}^{k} \gamma_i^{(n)}$ più grandi di x.

```
function k = randomNumberOfPartitions(n,theta)
% La funzione prende in input n e theta e restituisce k il numero
    di insiemi della partizione con la probabilita' voluta
    dipendente dal parametro theta
    x = rand;
    gammaVector = zeros(1,n);
    for i=1:n
        gammaVector(i) = theta^i*Stirling(n,i)/risingFactorial(
    theta,n);
    end

    bestGamma = 1;
```

```
    k = n;
    for j=1:n
        if x <= sum(gammaVector(1:j))
            if sum(gammaVector(1:j)) <= bestGamma
                bestGamma = sum(gammaVector(1:j));
                k = j;
            end
        end
    end
end
```

Listing C.4: Funzione randomNumberOfPartition

Lo step successivo consiste nella generazione casuale, fissato k, delle cardinalità $n_1, \ldots, n_k$ degli insiemi della partizione. La generazione degli $n_1, \ldots, n_k$ avviene secondo le probabilità assegnate in 2.6. Abbiamo, come prima cosa, definito una funzione `createM` che genera una matrice M di k colonne, contenente nelle righe tutte le possibili k-uple di $n_1, \ldots, n_k$ che sommano a n. Successivamente la funzione `randomPartitionNumbers` assegna la probabilità a ogni k-upla, assegnandola all'indice di riga corrispondente in M e seleziona un indice casualmente con la probabilità voluta. Restituisce, quindi, la k-upla corrispondente. L'uso degli indici nella scelta randomica ha permesso, quindi, di evitare la complicazione che si poteva avere dal fatto che la scelta casuale avvenisse in $\mathbb{R}^k$.

```
function M=createM(n, k)
% La funzione prende in input n e k e restituisce una matrice le
    cui righe sono tutte le possibili k-uple in {1,...,n} che
    sommano a n
    B = [];
    u = ones(1,k);
    M = [];
    if k==n
        for j=1:k
            M = u;
        end
    else
        index = k;
        saved_u = zeros(1,k);
        while not(all(saved_u==u))
            saved_u = u;
            u = create_v(saved_u,n,index);
            B = [B; u];
        end

        [r, ~] = size(B);
        for i=1:r
            if sum(B(i,:))==n
                M = [M; B(i,:)];
            end
        end
    end
end
```

Listing C.5: Funzione createM

```
function nVector = randomPartitionNumbers(n,k)
% La funzione prende in input n e k naturali con k<=n e
    restituisce un vettore nVector con una k-upla di numeri in
    {1,...,n} che sommano a n con probabilita' voluta
    M = createM(n,k);
    [g, ~] = size(M);
    gammaVector = zeros(1, g);
    for i=1:g
        prod_n = prod(M(i,:));
        mVector = m_from_n(M(i,:));
        prod_m = prod(factorial(mVector));
        gammaVector(i) = factorial(n)/(Stirling(n,k)*prod_n*prod_m
    );
    end

    x = rand;

    index = k;
    [~, h] = size(gammaVector);
    bestGamma = Inf;
    for j=1:h
        if x <= sum(gammaVector(1:j))
            if sum(gammaVector(1:j)) <= bestGamma
                bestGamma = sum(gammaVector(1:j));
                index = j;
            end
        end
    end

    nVector = M(index,:);
end
```

Listing C.6: Funzione randomPartitionNumbers

Infine, abbiamo definito la procedura `randomPart` per il terzo step dell'algoritmo che, dati $n_1, \ldots, n_k$, genera con probabilità uniforme una partizione i cui insiemi hanno cardinalità $n_1, \ldots, n_k$. Per farlo abbiamo sfruttato la funzione `partitions` definita in [12] che, dati, n e k genera tutte le possibili partizioni di $\{1, \ldots, n\}$ aventi k elementi. La struttura dati usata per contenere le partizioni è quella delle celle. Ogni partizione corrisponde a una cella $1 \times k$ i cui elementi sono i vettori corrispondenti agli insiemi di ogni partizione. La funzione `randomPart` prende la cella con tutte le possibili partizioni di $\{1, \ldots, n\}$ aventi k elementi, elimina tutte le partizioni i cui insiemi non hanno cardinalità assegnata $n_1, \ldots, n_k$ e restituisce una partizione di quelle restanti con probabilità uniforme tramite la funzione `randi`.

```
function partition=randomPart(nVector)
% La funzione prende in input un vettore nVector contenente una k-
    upla n_1,...,n_k e restituisce una partizione di {1,...,n}  i
    cui elementi hanno cardinalita' n_1,...,n_k scelta con
    probabilita' uniforme
    n = sum(nVector);
    [~, k] = size(nVector);
    totalPartitions = partitions(n,k);
    [numPart, ~] = size(totalPartitions);
    correctPartition = [];
```

```
    for i=1:numPart
        nVector_handling = nVector;
        c = 0;
        singlePartition = totalPartitions{i};
        [~, dimSinglePart] = size(singlePartition);
        for j=1:dimSinglePart
            set = singlePartition{j};
            [~, dimSet] = size(set);
            d = 0;
            [~, dimNVector_handling] = size(nVector_handling);
            for t=1:dimNVector_handling
                if nVector_handling(t) == dimSet
                    index=t;
                    d = d+1;
                end
            end
            if d~=0
                c = c+1;
                nVector_handling = nVector_handling([1:index-1,
    index+1:end]);
            end
        end
            if c==dimSinglePart
                correctPartition{end+1} = singlePartition;
            end
    end
    [~, correctDim] = size(correctPartition);
    x = randi([1, correctDim]);
    partition = correctPartition(1, x);
end
```

Listing C.7: Funzione randomPart

Per ricavare la stima teorica abbiamo, poi, definito una funzione `Ewens` che prende in input θ e gli $n_1, \ldots, n_k$ e implementa la formula di Ewens.

```
function p=Ewens(nVector, theta)
% La funzione prende in input un vettore nVector con valori n_1
    ,...,n_k e un parametro theta e restituisce il valore  della
    formula di Ewens per n_1,...,n_k con parametro theta
    [~,k]=size(nVector);
    n = sum(nVector);
    p = theta^k/risingFactorial(theta,n) * prod(factorial(nVector
    -1));
end
```

Listing C.8: Funzione Ewens

Infine, abbiamo definito una funzione `Esperimenti` che, presi n, il parametro di probabilità θ e il numero di esperimenti it restituisce una cella `partitionCollection` con tutte le partizioni ottenute, un vettore `partitionCount` con le frequenze relative di apparizione corrispondenti e un altro vettore `EwensResults` che restituisce il valore teorico di probabilità per ogni partizione.

```
function [partitionCollection, partitionCount, EwensResults]=
    Esperimenti(n, theta, it)
% La funzione prende in input n, theta e it ed effettua it
    simulazioni di generazione di partizioni con parametri n e
```

```
    theta. Restituisce una cella 1xit con le partizioni ottenute,
    un vettore con le frequenze relative corrispondente, e un
    altro vettore con le stime teoriche corrispondenti
    partitionCollection = [];
    partitionCount = [];
    EwensResults = [];
    for i=1:it
        k = randomNumberOfPartitions(n,theta);
        nVector = randomPartitionNumbers(n,k);
        partition = randomPart(nVector);
        partition = partition{1};

        count = 0;
        [~, dimPartColl] = size(partitionCollection);
        for j=1:dimPartColl
            if isequal(partitionCollection{j},partition)
                count = count+1;
                partitionCount(j) = partitionCount(j)+1;
            end
        end
        if count==0
            partitionCollection{end+1} = partition;
            partitionCount(end+1) = 1;
            EwensResults(end+1) = Ewens(nVector, theta);
        end
    end

    partitionCount = partitionCount./it;
end
```

Listing C.9: Funzione Esperimenti

Lo script che chiama e plotta l'errore della funzione `Esperimenti` è, infine, il seguente:

```
clear
close all
clc

it = 10000;
theta = 2;
n = 5;
err_2 = [];
[partitionCollection, partitionCount, EwensResults]
=Esperimenti(n, theta, it);
for c=1:100:10000
    [~,expVec,teoVec]=Esperimenti(n,theta,c);
    err_2 = [err_2, norm(expVec-teoVec)];
end
```

Listing C.10: Script che chiama e plotta Esperimenti

Bibliografia

[1] D. J. Aldous. «Exchangeability and related topics». In: (1985).

[2] UN General Assembly. *Transforming our world: the 2030 Agenda for Sustainable Development*. Agenda. United Nations, ott. 2015.

[3] C. A. Charalambides. *Enumerative Combinatorics*. Chapman e Hall/-CRC., 2002.

[4] J.A. Coyne. «Lack of genic similarity between two sibling species of Drosophila as revealed by varied techniques». In: *Genetics* 84 (1976), pp. 593–607.

[5] Harry Crane. «The Ubiquitous Ewens Sampling Formula». In: *Statistical Science* 31.1 (2016), pp. 1–19. ISSN: 08834237, 21688745.

[6] Rick Durrett. *Probability Models for DNA Sequence Evolution*. Springer New York, NY, 2008.

[7] W.J. Ewens. «The sampling theory of selectively neutral alleles». In: *Theoretical Population Biology* 3.1 (1972), pp. 87–112. ISSN: 0040-5809.

[8] Thomas S. Ferguson. «A Bayesian Analysis of Some Nonparametric Problems». In: *The Annals of Statistics* 1.2 (1973), pp. 209–230. ISSN: 00905364.

[9] R. A. Fisher, A. Steven Corbet e C. B. Williams. «The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population». In: *Journal of Animal Ecology* 12.1 (1943), pp. 42–58. ISSN: 00218790, 13652656.

[10] Martin Gardner. «The Bells: versatile numbers that can count partitions of a set, primes and even rhymes». In: 238 (1978), pp. 24–30.

[11] Donald E. Knuth. «Two Thousand Years of Combinatorics». In: *Combinatorics: Ancient and Modern*. Oxford University Press, giu. 2013. ISBN: 9780199656592. DOI: 10.1093/acprof:oso/9780199656592.003.0001.

[12] Bruno Luong. *Set partition*. 2022. URL: https://www.mathworks.com/matlabcentral/fileexchange/24133-set-partition.

[13] MATLAB. *version 9.9.0 (R2020b)*. Natick, Massachusetts: The MathWorks Inc., 2020.

[14] Eugenio Regazzini. *Impostazione non parametrica di problemi di inferenza statistica bayesiana*. Consiglio nazionale delle ricerche, 1996.

[15] Laura M. Sangalli. «Some Developments of the Normalized Random Measures with Independent Increments». In: *Sankhyā: The Indian Journal of Statistics (2003-2007)* 68.3 (2006), pp. 461–487. ISSN: 09727671.

[16] R.S. Singh et al. «Genetic heterogeneity within electrophoretic "alleles" of xanthine dehydrogenase in Drosophila pseu- doobscura». In: *Genetics* 84 (1976), pp. 609–629.

[17] Simon Tavaré. «The magical Ewens sampling formula». In: *Bulletin of the London Mathematical Society* 53.6 (2021), pp. 1563–1582.

[18] Yee Teh. «Dirichlet Process». In: (gen. 2010). DOI: 10.1007/978-0-387-30164-8_219.

[19] Anna Tovo et al. «Upscaling human activity data: an ecological perspective». In: (dic. 2019).

Ringraziamenti

Volevo ringraziare, in quest'ultima pagina, tutte le persone che ho incontrato in questo percorso perché anche le influenze più impercettibili mi hanno reso la persona che sono: un grazie particolare va a chi, una volta entrato nella mia vita, ha deciso di continuare a camminare con me, sperando che io riesca a essere per loro un compagno di viaggio altrettanto buono.

Quindi ringrazio la mia famiglia, che mi accompagna nel mio viaggio da sempre, ringrazio Maria Gloria, perché la strada che stiamo percorrendo assieme è la più bella, ringrazio il mio relatore e i miei professori, perché il percorso che ha oggi il suo traguardo ha avuto loro come guide, ringrazio i miei amici, chi conosco da sempre e chi è riuscito a far sembrare pochi mesi come anni, perché ognuno dei loro cammini coincide, almeno in piccola parte, con il mio.

Quella di oggi è solo una tappa di un viaggio che ha ancora tanti capitoli da scrivere e spero di avere sempre al mio fianco una compagnia altrettanto bella.