Applicazione di modelli statistici alla descrizione (e alla comprensione) di processi biologici e sociali

Introduzione

- Le leggi di scala in fisica statistica compaiono in particolare nei fenomeni critici, e sono associate all'esistenza di punti fissi del gruppo di rinormalizzazione.
- Punto fisso repulsivo => Necessità di *fine tuning* dei parametri (temperatura, pressione) per raggiungere il punto critico.

- Esistono tuttavia leggi di scala associate a numerosi fenomeni geologici, biologici, ecologici, nella teoria delle reti, ma anche in contesti economici (Pareto, Stanley), sociali, e perfino artistici e linguistici (Zipf).
- In tutti questi casi il fenomeno si manifesta senza *fine tuning.*
- Una possibile spiegazione (non universalmente accettata) è la *criticità autoorganizzata*, in altri termini l'esistenza, in uno spazio di parametri di non facile individuazione, di un punto fisso *attrattivo.*

Processi evolutivi e leggi di scala

- Nei casi in cui il fenomeno studiato possa essere pensato come esito di un processo evolutivo (biologico, economico, sociale) una spiegazione di tipo fenomenologico, meno impegnativa in quanto basata solo su calcolo delle probabilità, fa emergere le leggi di scala, in un limite opportuno, dai *processi di Jule*.
- Un processo di Jule si ha quando in un sistema caratterizzato da tipi ("generi") e individui ("specie") possono saltuariamente comparire nuovi tipi, ma nell'intervallo tra la comparsa di due tipi il numero degli individui di ciascun tipo già presente cresce in proporzione al numero degli individui già presenti nel sistema.
- Con ovvi adattamenti semantici questa descrizione può coprire la crescita di un ecosistema o quella di una rete informatica, il popolamento delle città o la frequenza dei nomi di famiglia, etc…
- Da questi processi emergono naturalmente *distribuzioni binomiali negative.*

Il problema

- Data una popolazione finita di N individui appartenenti a S tipi diversi, la loro distribuzione non potrà soddisfare in modo esatto una legge di scala (se non nel limite $N \to \infty$)
- A maggior ragione un campionamento di $n << N$ individui di tale popolazione non solo non soddisferà una legge di scala, ma in generale non sarà distribuito come la popolazione originaria: basti pensare al caso $n<S$ in cui alcuni tipi saranno necessariamente assenti dal campione.

- PROBLEMA: Qual è la forma probabile della distribuzione del campione, essendo già nota la forma della distribuzione originaria? Viceversa, quali parametri della distribuzione originaria possiamo ricostruire a partire dalla distribuzione dei campioni, e come dobbiamo procedere a tal fine?

Risultati generali

- Usando l'indice α per identificare i differenti tipi presenti, la probabilità di estrarre uno specifico campione da una popolazione data è

$$P_{\{n_a\}} = \binom{N}{n}^{-1} \prod_{a=1}^{S} \binom{N_a}{n_a}.$$

- Noi siamo però interessati alle distribuzioni in frequenza N(k), dove N è il numero dei tipi caratterizzati dalla presenza di esattamente k individui nella popolazione, e alle corrispondenti distribuzioni n(l) risultanti dai campionamenti. Se definiamo N(k,l) il numero di tipi caratterizzati da k individui nella popolazione e da l individui nel campione, la probabilità di una specifica configurazione è

$$P_{\{N_{kl}\}} = \binom{N}{n}^{-1} \prod_{k=1}^{N} \left[N_k! \prod_{l=0}^{k} \frac{1}{N_{kl}!} \binom{k}{l}^{N_{kl}} \right],$$

Risultati generali (2)

- La probabilità di una distribuzione in frequenza n(l) nel campione si ottiene in linea di principio sommando su tutte le configurazioni che soddisfano la condizione che la somma degli N(k,l) sia n(l).
- Questa somma non si può calcolare in forma chiusa, ma in realtà noi siamo interessati soltanto al valore atteso <n(l)>, per il quale abbiamo ottenuto i risultati

$$< N_{kl} >= N_k \frac{\binom{k}{l}\binom{N-k}{n-l}}{\binom{N}{n}}, \qquad < n_l >= \frac{\sum_{k=1}^{N} N_k \binom{k}{l}\binom{N-k}{n-l}}{\binom{N}{n}}.$$

- Molto importante in pratica è il limite k,l << N,n, per cui vale

$$< n_l >= \sum_{k=1}^{N} N_k \binom{k}{l} (\frac{n}{N})^l (1-\frac{n}{N})^{k-l},$$

Risultati generali (3)

- Molto utile si rivelerà l'uso di funzioni generatrici;

$$F(t) \equiv \sum_{k=1}^{N} N_k t^k, \qquad f(t) \equiv \sum_{l=0}^{n} < n_l > t^l,$$

- Nel limite suddetto possiamo dimostrare che vale:

$$f(t) = \sum_{k=1}^{N} N_k (1 - \frac{n}{N} + \frac{n}{N} t)^k = F(1 - \frac{n}{N} + \frac{n}{N} t).$$

- Posto poi: $G(z) \equiv F(1 - \frac{z}{N}), \qquad g(z) \equiv f(1 - \frac{z}{n})$

- Si dimostra la fondamentale relazione $g(z) = G(z)$
- Vedremo poi alcune importanti conseguenze

Momenti invarianti

- Notiamo che è possibile, grazie alle proprietà già individuate, definire combinazioni di valori attesi che non dipendono dalle dimensioni del campione, e quindi riflettono in modo diretto le proprietà della distribuzione in frequenza originaria. Si dimostra in generale che:

$$< m_p^{(n)} > \equiv < \frac{(n-p)!}{n!} \sum_{l=p}^{n} n_l \binom{l}{p} > = \frac{(N-p)!}{N!} \sum_{k=p}^{N} N_k \binom{k}{p} \equiv M_p$$

- Di conseguenza, purché p (minore di n), sia ancora un numero abbastanza grande, saremo in grado di ricostruire un grande numero di momenti della distribuzione originaria, e quindi con buona approssimazione la distribuzione stessa, a partire dal valore medio dei momenti dei suoi campioni, anche relativamente piccoli (ovvero per n<<N)

Momenti invarianti (2)

- Notiamo in particolare:

$$M_0 \equiv \sum_{k=1}^{N} N_k = \sum_{l=0}^{n} < n_l >= S,$$

$$M_1 \equiv \frac{1}{N} \sum_{k=1}^{N} k N_k = \frac{1}{n} \sum_{l=0}^{n} l < n_l >= 1,$$

$$2M_2 \to \frac{1}{N^2} \sum_{k=1}^{N} k(k-1) N_k = \frac{1}{n^2} \sum_{l=0}^{n} l(l-1) < n_l >,$$

$$2M_2 \equiv \frac{1}{\alpha} = \sum_{a=1}^{S} (\frac{N_a}{N})^2 - \frac{1}{N} = \sum_{a=1}^{S} < (\frac{n_a}{n})^2 > - \frac{1}{n},$$

Correlazione tra i campioni

- Spesso è importante poter verificare l'indipendenza statistica dei campioni. I risultati precedenti ci permettono di calcolare la correlazione attesa tra campioni casuali in termini dei parametri del sistema e del secondo momento invariante.
- Avendo definito la correlazione:

$$C = \frac{\sum_{a=1}^{S} n_a m_a}{\sqrt{\sum_{a=1}^{S} n_a^2}\sqrt{\sum_{a=1}^{S} m_a^2}}.$$

- Per due campioni casuali della stessa dimensione n vale:

$$< C >= \frac{n}{\alpha + n}\frac{\alpha + N}{N}.$$

Distribuzioni binomiali negative

- Nel caso delle distribuzioni binomiali negative vale per la funzione generatrice:

$$F_c(t) = \frac{N}{x}\frac{(1-x)^{1-c}}{c}\left[1-(1-xt)^c\right] = \sum_{k=1}^{\infty}\frac{N}{x}\frac{(1-x)^{1-c}}{\Gamma(1-c)}\frac{\Gamma(k-c)}{k!}x^k,$$

- Il comportamento asintotico della distribuzione in frequenza è:

$$N_k \to \frac{N}{x}\frac{(1-x)^{1-c}}{\Gamma(1-c)}\frac{x^k}{k^{1+c}}.$$

- Si riconosce che la distribuzione tende a una legge di scala nel limite per cui 1-x << 1, che è strettamente legato al limite di grande N
- Si dimostra in questi casi che la distribuzione dei campioni ha la stessa forma di quella originaria se si rimpiazza N con n e si sostituisce al posto di x la variabile y = n/N x /(1-x+ n/N x)

Distribuzioni binomiali negative (2)

- Introducendo il parametro $\beta = N(1-x)/x = n(1-y)/y$ è possibile esprimere in forma semplice i momenti invarianti e la loro forma asintotica (per grandi p):

$$M_p = \frac{\beta^{1-p}}{\Gamma(1-c)} \frac{\Gamma(p-c)}{p!} \rightarrow \frac{1}{\Gamma(1-c)} \frac{\beta^{1-p}}{p^{1+c}},$$

- Sotto opportune ipotesi si dimostra che un grande numero di distribuzioni che possono essere associate a leggi di scala ammette come forma asintotica dei propri momenti invarianti, per p abbastanza grande, espressioni analoghe, a meno di un fattore moltiplicativo, a quella trovata per le binomiali negative

Applicazioni

- Abbiamo applicato i risultati qui presentati allo studio della distribuzione in frequenza dei cognomi italiani, ottenendo un *fit* alla distribuzione empirica molto migliore di quello offerto da una legge di scala di tipo asintotico.
- Abbiamo considerato le “popolazioni” della docenza universitaria come campioni della popolazione italiana, verificando la coerenza dei momenti invarianti dei campioni con quelli dell’intera popolazione, e quindi la correttezza delle ipotesi
- Abbiamo utilizzato la formula della correlazione tra campioni per verificare (con piccole deviazioni) l’ipotesi dell’indipendenza statistica delle popolazioni di PO, PA e RU (contro l’ipotesi di nepotismo generalizzato)

Questa presentazione è coperta, per le sue parti originali, da licenza Creative Commons:

attribuzione, non commerciale, condividi allo stesso modo

http://creativecommons.org/licenses/by-nc-sa/2.5/it/legalcode

rossi@df.unipi.it