Statistica Applicata Vol. 8, n. 1, 1996

NUMEROSITÀ DELLE SPECIE E DISTRIBUZIONE DEI COGNOMI: DAL MODELLO STATISTICO ALL'INFORMAZIONE GENETICA

Gianna Zei, Antonella Lisa, Onella Fiorani, Carmela R. Guglielmino

IGBE–CNR, Pavia e DGM, Università di Pavia.

RIASSUNTO

Le distribuzioni dei cognomi, utilizzate per studiare la struttura genetica di popolazioni umane, si adattano molto bene a due distribuzioni teoriche, una di significato genetico, la distribuzione di Karlin e McGregor per geni selettivamente neutrali e una di significato entomologico, la distribuzione logaritmica di Fisher che rappresenta la variazione nell'abbondanza di specie in campioni di popolazioni animali. Il parametro α, stimato dalla distribuzione di Fisher, che misura la "diversità dei cognomi" (e quindi la diversità dei geni) in differenti sottopopolazioni, è indipendente dalla dimensione del campione ma è influenzato dal modo in cui le sottopopolazioni sono campionate. La stima dell'effetto del numero di suddivisioni di un'area sul valore di α misura la deviazione di una popolazione della panmixia e mette in luce i processi migratori del passato.

Parole chiave: cognomi, geni naturali, distribuzioni teoriche, struttura di popolazioni.

1. INTRODUZIONE

La regola, comune alla maggior parte dei paesi, secondo la quale i cognomi vengono trasmessi di padre in figlio e perciò seguono e "marcano" gli spostamenti e i mescolamenti degli individui (così come avviene per un gene del cromosoma Y con molti alleli), fa dei cognomi uno strumento che può essere ricco di informazioni su specifici aspetti della struttura genetica di popolazioni umane.

Uno di questi aspetti è l'evoluzione neutrale – intesa come evoluzione delle frequenze geniche in condizioni di equilibrio tra deriva genetica, mutazione e migrazione, in assenza di selezione – che i cognomi possono simulare per la loro natura di caratteri "neutri", natura che può essere loro ragionevolmente riconosciuta dal momento che non dovrebbe esserci riproduttività o sopravvivenza differenziale per persone che portano differenti cognomi, né matrimonio preferenziale per specifici cognomi.

Ma è solo dall'applicazione di tests statistici che la neutralità evolutiva dei cognomi può essere verificata, cosicché, stabilite le condizioni di applicabilità e apportate le eventuali correzioni, le stime ottenute dai cognomi possano sostituire quelle che si otterrebbero dai geni.

Due distribuzioni teoriche formulate per risolvere problemi posti in campi differenti – genetica teorica e entomologia – sono state confrontate e ricondotte a una stessa forma che permette di calcolare la distribuzione attesa dei cognomi.

2. LA DISTRIBUZIONE DI KARLIN E McGREGOR

Questa distribuzione descritta in un articolo dal titolo: "The number of mutant forms maintained in a population", apparso nel 1967, costituiva il termine di confronto teorico ideale per un test di neutralità dei cognomi, essendo derivata da un modello di equilibrio tra deriva genetica e mutazione per alleli selettivamente neutrali.

Il modello prevede una popolazione di dimensione costante di N geni (o dimensione aploide della popolazione), dove:

- il numero r di alleli è distribuito casualmente tra gli individui, ciascuno dei quali ha la stessa probabilità di riproduzione e di mutazione;
- il numero medio di figli per individuo è distribuito secondo Poisson con media uguale a 1;
- ogni individuo che muore è sostituito da un nuovo individuo che ha la probabilità $(1 - v)$ di portare lo stesso allele, o la probabilità v di portare uno degli altri r possibili alleli.

Con queste assunzioni e per un alto numero di alleli ($r \to \infty$) il numero atteso di alleli portato da k individui, $N(k)$, è dato da:

$$E[N(k)] = \frac{1}{k} \frac{Nv}{1-v} \frac{\binom{\frac{N}{1-v} - (k+1)}{N-k}}{\binom{\frac{N}{1-v} - 1}{N}} \qquad (1)$$

La formula (1) può essere così riscritta:

$$\frac{1}{k}\frac{N\nu}{1-\nu}\frac{N(N-1)(N-2)\ldots(N-k+1)}{\left(\frac{N}{1-\nu}-1\right)\left(\frac{N}{1-\nu}-2\right)\ldots\left(\frac{N}{1-\nu}-k\right)}\approx$$

$$\frac{1}{k}\frac{N\nu}{1-\nu}\frac{N^k}{\left(\frac{N}{1-\nu}-1\right)^k}=\frac{1}{k}\frac{N\nu}{1-\nu}(1-\nu)^k\left[\frac{N}{N-1+\nu}\right]^k,$$

e l' approssimazione è valida se $N>>k$. In questo caso, quando n è piccolo, la quantità $\left[\frac{N}{N-1+\nu}\right]^k \cong 1$ e la formula (1) viene così semplificata:

$$E\left[N(k)\right]=\frac{N\nu}{1-\nu}\frac{(1-\nu)^k}{k} \qquad (2)$$

Il calcolo dei termini della distribuzione di Karlin e McGregor nella sua forma originale (1) è molto laborioso, ma la sua forma semplificata (2) viene bene approssimata da una distribuzione formulata per rappresentare la variazione nell'abbondanza delle specie in campioni di popolazioni animali.

3. LA DISTRIBUZIONE LOGARITMICA DI FISHER

In un articolo apparso nel 1943 dal titolo "The relation between the number of species and the number of individuals in a random sample of an animal population" R.A. Fisher, uno dei padri della biometria, formulava la teoria relativa all'abbondanza di differenti specie, teoria che i suoi collaboratori Corbet e Williams applicavano, con successo, a distribuzioni osservate di Lepidotteri.

Nel modello di Fisher vengono descritte le relazioni che legano tre diverse distribuzioni teoriche.

Si prevede, infatti, che in un habitat occupato da un numero S di specie animali la cui numerosità, all'interno di un campione di dimensione N, è distribuita come un χ^2:

- la probabilità di trovare k individui di una data specie sia data da una *distribuzione di Poisson;*
- se il campione è eterogeneo, cioè se l'abbondanza delle diverse specie è molto variabile, la miscela di distribuzioni di Poisson con differenti medie sia descritta dalla *distribuzione binomiale negativa;* e che,

– la situazione, molto comune, nella quale un grande numero di specie sono così rare che la loro probabilità di inclusione è piccola, sia ben rappresentata da una forma limite della distribuzione binomiale negativa che corrisponde alla *distribuzione logaritmica.*

Sotto queste condizioni il numero atteso di specie, ciascuna rappresentata da k individui (dove k non può essere = 0), è dato da:

$$E[N(k)] = \alpha x^k / k, \tag{3}$$

dove α rappresenta l'abbondanza delle specie ed è indipendente dalla dimensione del campione, N, e x è una costante con valore < 1, dipendente da N.

Dalla espansione della distribuzione logaritmica si può ricavare il numero totale atteso di specie, S:

$$S = \sum_k \frac{\alpha}{k} x^k = -\alpha \ln(1-x) \tag{4}$$

e il numero totale atteso di individui, N:

$$N = \sum_k \alpha x^k = \frac{\alpha x}{1-x} \tag{5}$$

cosicché, per ogni serie di dati osservati, noti S e N, si possono calcolare i valori di α e di x.

In particolare, dal rapporto tra S e N:

$$S/N = -(1-x)\ln(1-x)/x \tag{6}$$

si ottiene, attraverso un' iterazione numerica, una stima di x e, quindi, dalla (5) una stima di α:

$$\alpha = N(1-x)/x$$

Questa quantità, indipendente dalla dimensione del campione, costituisce una misura della *ricchezza in specie* di un'area, che può essere messa a confronto con quella di altre aree, quando siano noti gli errori di campionamento:

$$V_{(\alpha)} = \frac{\alpha^3 \left\{ (N+\alpha)^2 \ln \frac{2N+\alpha}{N+\alpha} - \alpha N \right\}}{(SN + S\alpha - N\alpha)^2}$$

4. IL CONFRONTO TRA LE DUE DISTRIBUZIONI TEORICHE E L'ADATTAMENTO ALLE DISTRIBUZIONI DEI COGNOMI

Confrontando le due soluzioni (2) e (3) date da Karlin e McGregor e da Fisher per stimare $E[N(k)]$, dove $N(k)$ rappresenta rispettivamente il numero di alleli neutri o il numero di specie animali, si ricava:

$$x = 1 - v,$$
$$\alpha = Nv/(1-v),$$
$$v = \alpha/(N+\alpha)$$

e viene messa in luce la semplice relazione funzionale che lega x, v e α.

Collegando i due approcci si può sostituire al significato *ecologico* dei parametri della distribuzione logaritmica di Fisher il significato *genetico* proprio del modello di Karlin e McGregor che ben si adatta a una popolazione di N individui maschi i cui cognomi si comportano come alleli di un gene del cromosoma *Y*. Perciò:

- i termini della distribuzione del numero di specie diventano il numero atteso di cognomi portati da un numero k di individui;
- il rapporto tra il numero di cognomi, S, e il numero di individui, N, permette di risolvere l'equazione (6) in termini di v:

  $$S/N = v \ln v/(v-1), \qquad (7)$$

  e di stimare, così, il parametro fondamentale della distribuzione di Karlin e McGregor il cui significato di *mutazione,* che per i cognomi è trascurabile, può essere sostituito o completato dal significato di *immigrazione,* che per i cognomi è più rilevante;
- da v si può ottenere il valore di α che diventa una stima dell'*abbondanza* o *diversità* dei cognomi all' interno di un campione.

5. DALL'ABBONDANZA DELLE SPECIE ALL'INFORMAZIONE GENETICA

La stima di v dalla formula (7) rappresenta il tasso di mutazione (e immigrazione) in una popolazione di alleli neutri in *stato di equilibrio* tra mutazione e deriva genetica, così come descritto dal modello di Karlin e McGregor. Data e verificata sperimentalmente l'analogia geni–cognomi, la distribuzione attesa dei cognomi che si ricava da questa stima serve a controllare, in primo luogo, la situazione di equilibrio genetico della popolazione.

I dati della Sardegna, che sono stati il banco di prova dello studio dell'evoluzione neutrale per mezzo dei cognomi, hanno mostrato che, nelle aree in cui era

stata suddivisa la Sardegna, la distribuzione osservata dei cognomi non sempre si adattava alla distribuzione attesa in base alla teoria. Nella maggior parte dei casi, le classi che maggiormente contribuivano a rendere significativo il χ^2 erano le prime – quelle dei cognomi più rari – che presentavano una frequenza molto più grande dell'atteso.

Una spiegazione di questa discrepanza può essere fornita da notizie sullo sviluppo economico delle aree sotto studio. Una zona recentemente e fortemente industrializzata attira migrazione dall'esterno e questa improvvisa immissione di nuovi cognomi (e geni) può determinare una temporanea perturbazione dell'equilibrio. La conferma di questa ipotesi si ha, quando, adattando ai dati osservati una distribuzione troncata a livello della prima classe e, quindi, stimando un nuovo tasso di immigrazione che riflette la situazione (di equilibrio) precedente l'industrializzazione, si ottiene un buon adattamento alla distribuzione teorica (Zei *e coll.*, 1983; tabella I).

Tab. I: χ^2 per l'adattamento della distribuzione logaritmica di Fisher alle distribuzioni dei cognomi nelle 11 Diocesi della Sardegna (da Zei e coll., 1983).

	Distribuzione completa			Distribuzione troncata	
Diocesi	χ^2	signif.	g.l.	χ^2	signif.
Ampurias	38.95	***	11	8.92	–
Sassari	39.36	***	13	16.68	–
Alghero	17.42	–	10	9.20	–
Ozieri	10.17	–	12	8.72	–
Bosa	6.91	–	7	6.27	–
Nuoro	35.05	**	17	26.63	*
Oristano	23.53	–	15	13.42	–
Ogliastra	11.46	–	14	11.45	–
Ales	12.07	–	6	5.98	–
Iglesias	6.74	–	5	5.42	–
Cagliari	22.48	*	11	7.78	–

Significatività:
* P < 5%; ** P< 1%; *** P< 0.1%.

Quando lo studio della struttura genetica è esteso a una popolazione suddivisa in gruppi o sottopopolazioni, come sono la maggior parte delle popolazioni umane, le stime di v tra gruppi non sono confrontabili, perché fortemente dipendenti dalla dimensione del campione, N. Al contrario, il parametro α di Fisher, indipendente da N, consente di mettere a confronto l'*abbondanza* di cognomi o geni tra le sottopopolazioni di una stessa area e, quindi, di studiare le cause della sua variabilità.

Tuttavia, il controllo preliminare della proprietà di invarianza di α rispetto ad N costituisce un'altra verifica dello stato di equilibrio delle sottopopolazioni, che è la condizione necessaria per un loro corretto confronto.

Un esempio viene da un riesame dei dati della Sardegna. I valori di α stimati dai cognomi degli sposi per ognuno dei comuni appartenenti a un campione di 115 unità sono correlati significativamente (P=0.002) con la dimensione N dei campioni. Ma se si escludono i grossi comuni di Cagliari e Sassari, la cui grande ricchezza in cognomi è giustificata dalla condizione di "città", α varia all'interno di un intervallo di valori che non risente della grandezza del comune (P=0.239).

E, per questi campioni che rappresentano sottopopolazioni in situazione di equilibrio, si può correttamente verificare che il grado di sviluppo economico è uno dei più importanti fattori che influenzano il valore di α (figura 1).

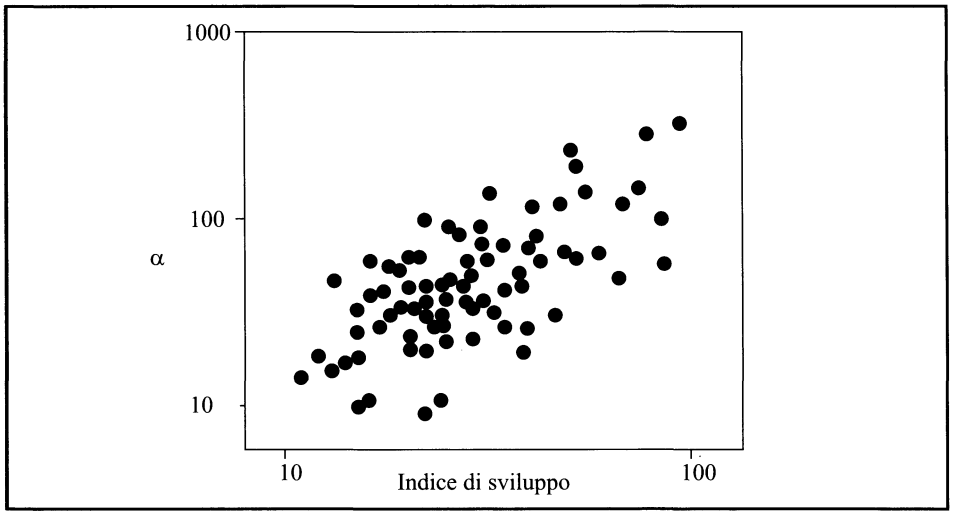

Fig. 1: Correlazione fra la diversità dei cognomi (α) e l'indice di sviluppo (100–% addetti all'agricoltura) in 113 comuni della Sardegna: r = 0.68, P < 0.001.

La suddivisione della Sardegna in comuni rappresenta la situazione ottimale per lo studio della struttura genetica di questa regione. Infatti, i comuni, che rappresentano la più piccola unità amministrativa, corrispondono spesso (ad eccezione delle grandi città) a quello che i genetisti chiamano "effective population size", N_e, cioè l'area all'interno della quale viene scelto il coniuge. L'abbondanza di cognomi all'interno di ogni comune dovrebbe essere, perciò, una funzione inversa del grado di isolamento e di endogamia del passato.

Ma una popolazione può essere suddivisa secondo diversi criteri – amministrativi, geografici, linguistici, sociali, culturali – che definiscono aree i cui confini

in alcuni casi sono fittizi dal punto di vista genetico (come le province e le regioni che sono raggruppamenti di comuni, intesi come unità elementari), altre volte (come quelli linguistici) costituiscono barriere al "random mating". La varietà dei cognomi e dei geni risente del modo in cui la popolazione è stata suddivisa e della correlazione tra aree adiacenti.

La figura 2, ripresa da un lavoro di Zei e coll. (1983), mostra come i valori di α stimati dai cognomi della Sardegna suddivisa secondo diversi criteri (349 comuni, 44 raggruppamenti di comuni adiacenti, 18 zone storiche, 11 diocesi, 3 province), aumenta con il diminuire del numero di unità di ogni suddivisione e, quindi, con l'aumentare del numero di osservazioni per unità, N. La proprietà fondamentale dell'α di Fisher, quella di indipendenza da N, sembra, qui, non essere rispettata.

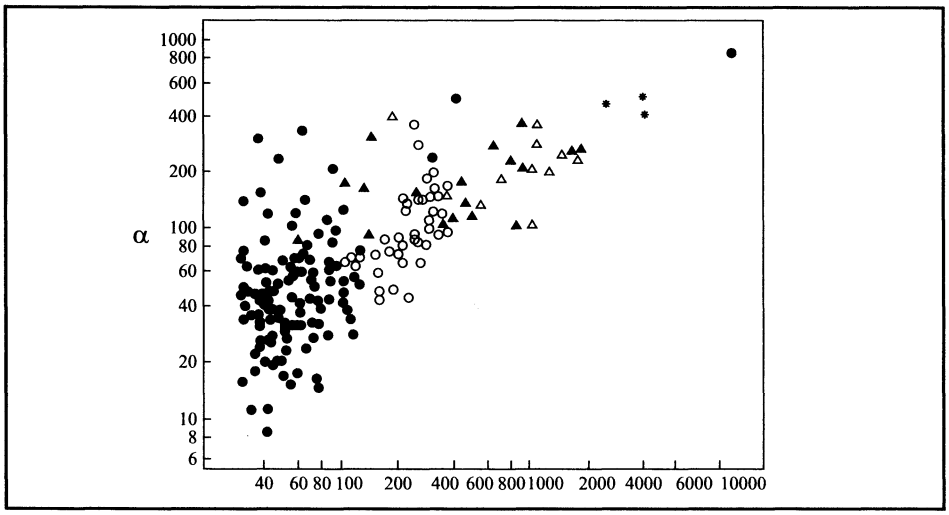

Fig. 2: Diversità dei cognomi (α) in differenti suddivisioni della Sardegna ● comuni, ○ zone geografiche, ▲ zone storiche, Δ diocesi, * province, ⊕ Sardegna (da Zei e coll. 1983).

Ma è proprio lo scostamento dall'attesa invarianza rispetto al tipo di suddivisione che può dare una misura del grado di "strutturazione genetica" della popolazione. Se la popolazione in esame fosse completamente panmittica, cioè se i matrimoni fossero completamente casuali, senza ostacoli posti dai confini di qualunque tipo di suddivisione, la variazione dei valori di α rispetto ad N dovrebbe corrispondere a quella rappresentata nella parte alta della figura 3, ricavata da una distribuzione logaritmica teorica, per N uguale all'intera area campionata. I dati reali, sintetizzati dai valori modali di α e rappresentati nella parte inferiore della figura, permettono di misurare questa discrepanza stimando il valore di $\beta = 0.57$ dalla relazione osservata:

$$\alpha = kN^{\beta}$$

Il problema dell'effetto della dimensione dell'area sulla stima di parametri che descrivono la struttura genetica di popolazioni locali era già stato posto a proposito della *varianza di Wahlund* o F_{ST}, una misura della divergenza nelle frequenze geniche tra sottopolazioni.

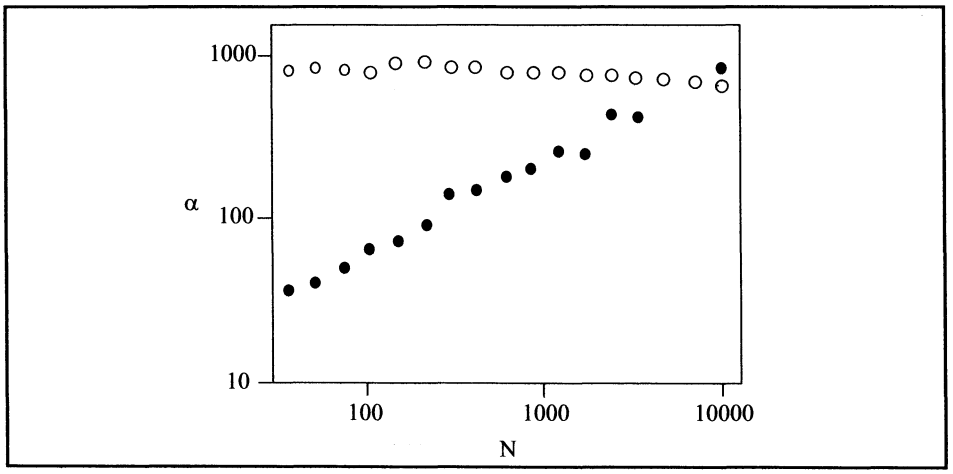

Fig. 3: Effetto del tipo di suddivisione dell'area sul valore di α: ○ valori di α ottenuti da campioni casuali estratti da una distribuzione logaritmica teorica con N uguale a quello della intera Sardegna, ● valori mediani di α dalla popolazione reale per classi di N (da Zei e coll. 1983).

Passando in rassegna i lavori in cui venivano confrontati i valori di F_{ST} in popolazioni diverse, Jorde (1980) consigliava cautela nell'interpretazione delle differenze proprio a causa della sensibilità di questa misura alla dimensione dell'area scelta. Più tardi Cavalli–Sforza e Feldman (1990) derivavano le formule adatte a predire il comportamento della varianza delle frequenze geniche in funzione del tipo di suddivisione geografica di una popolazione.

Dal punto di vista sperimentale, i valori di F_{ST} calcolati dalla stessa popolazione di cognomi della Sardegna, utilizzati come modelli per la stima dell'eterogeneità delle frequenze geniche della regione suddivisa secondo i criteri già descritti (Zei e coll. 1986), avevano mostrato una relazione inversa con il grado di suddivisione (figura 4), quantificata da un valore di $\beta = -0.67$, stimato dalla relazione osservata:

$$F_{ST} = kN^{-\beta}$$

È interessante notare come da due diversi approcci siano state ottenute stime di β molto simili, che misurano il grado di strutturazione della popolazione sarda, in un intervallo di variazione compreso tra $\beta = 0$ che indica completa panmixia e

$\beta = 1$ che indica una suddivisione totale della popolazione in isolati chiusi. Le due stime hanno segno opposto così come il significato di F_{ST} che rappresenta l'*isolamento* tra le sottopopolazioni è opposto a quello di α che rappresenta la *capacità ricettiva* delle singole sottopopolazioni.

La misura dell'effetto del tipo di suddivisione della popolazione sulla variazione di α può, quindi, offrire un'ulteriore informazione utile per la comprensione dei processi migratori che hanno determinato l'eterogeneità genetica di una regione.

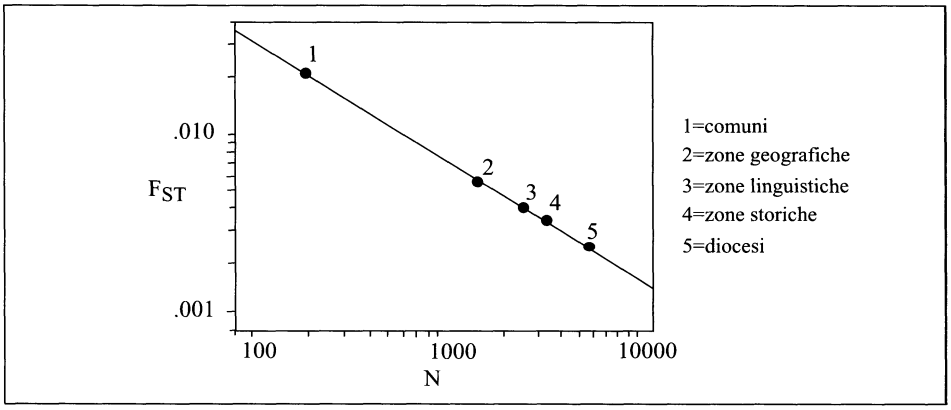

Fig. 4: Effetto del tipo di suddivisione dell'area sul valore della varianza di Wahlund o F_{ST} (da Zei e coll., 1986).

RIFERIMENTI BIBLIOGRAFICI

CAVALLI-SFORZA, L. L. and FELDMAN, M. W. (1990), Spatial subdivision of populations and estimates of genetic variation. *Theor. Pop. Biol.*, **37**: 3–25.

FISHER, R. A. (1943), The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Anim. Ecol.*, **12**:42–58.

JORDE, L. B. (1980), The genetic structure of subdived human populations. In *Current Developments in Anthropology* (ed. Mielke and Crawford), **1**: 135–208.

KARLIN, S. and McGREGOR, J. (1967), The number of mutant forms maintained in a population. *Proc. Fifth Berkeley Symp. Math. Stat. Prob.*, **4**: 415–438.

ZEI, G., GUGLIELMINO, C. R., SIRI, E., MORONI, A. and CAVALLI-SFORZA, L. L. (1983), Surnames in Sardinia. I. Fit of frequency distributions for neutral alleles and genetic population structure. *Ann. Hum. Genet.* **47**: 329–352.

ZEI, G., PIAZZA, A., MORONI, A. and CAVALLI-SFORZA, L. L. (1986), Surnames in Sardinia. III. The spatial distribution of surnames for testing neutrality of genes. *Ann. Hum. Genet.* **50**: 169–180.

ABUNDANCE OF DIFFERENT SPECIES AND SURNAME DISTRIBUTIONS: FROM STATISTICAL MODEL TO GENETIC INFORMATION

SUMMARY

Surname distributions, used for the studies on genetic structure of populations, fit very well Fisher's logarithmic distribution, by which expected number of species and individuals in samples of animal populations is estimated. The parameter α, measuring "surnames diversity" (gene diversity) in different subpopulations, is independent of the sample size but is influenced by the way in which each subpopulation area is sampled. The estimation of the effect of the number of subdvisions on α measures how a population deviates from panmixia and enlightens on the past migratory processes.

Keywords: surnames, neutral genes, theoretical distributions, population structure.