**La distribution des noms de famille comme outil pour l'analyse des dynamiques migratoires**

Paolo Rossi – Dipartimento di Fisica « E.Fermi » - Università di Pisa
et I.N.F.N. – Sezione di Pisa – Italie

L'étude de l'origine et de la diffusion des noms de famille est une thématique fortement interdisciplinaire. Pas seulement les sciences historiques et démographiques peuvent y contribuer, mais aussi la génétique, qui a étudié pour longtemps la distribution des noms, et la physique statistique, qui s'est approchée à cette matière dans le dernier décennie.

L'objectif de mon exposé est la présentation des principaux résultats obtenus jusqu'ici dans les études de la génétique, mais aussi des thèmes et des techniques mises récemment en champ par la physique statistique, en indiquant les informations les plus importantes qui peuvent être obtenues grâce à ces méthodologies et en illustrant les résultats avec des exemples concrets.

L'emploi des noms de famille dans les études de biologie humaine a été proposé pour la première fois pendant la deuxième moitié du XIX siècle, mais l'étude systématique a été mis en train seulement dans les années Soixante du XX siècle et peut être reconduit à deux thèmes principaux, celui des études de consanguinité et celui de la distribution en fréquence des noms de famille et de sa possible origine.

L'idée d'étudier le degré de consanguinité au moyen des noms de famille remonte à George Darwin (fils de Charles) qu'en 1875 calcula (pour l'Angleterre) la probabilité que dans un mariage les conjoints pouvaient avoir fortuitement le même nom de famille et la confronta avec les données empiriques, en concluant que la consanguinité dû aux mariages parmi des cousins rejoignait le 4,5% dans l'aristocratie, avec des valeurs plus basses pour la bourgeoisie et la population urbaine.[1]

Ce genre d'étude fut repris seulement après 1965, lorsque Crow et Mange montrèrent que, pour la plupart des typologies de mariages entre parents, la probabilité que des individus apparentés aient le même nom est proportionnelle au degré de consanguinité. Par conséquent le pourcentage des mariages isonymiques constitue une mesure assez directe du niveau de consanguinité de la population.[2]

Plus généralement on établit le degré de parenté entre deux populations différentes à partir dès pourcentages de personnes ayant le même nom de famille.[3] Il faut pourtant corriger les résultats pour tenir compte des deux faits très importants:

a) le même nom peut avoir des origines différentes et n'être pas dû à un ancêtre commun
b) les phénomènes migratoires peuvent influencer les distributions des noms

Il vaut la peine de signaler que certains études empiriques, parmi les plus importants dans ce contexte, ont été réalisés sur des populations ibériques et ibéro-américaines, grâce au fait que le système espagnol des noms de famille fournit un immédiat et facile contrôle de l'isonymie des ancêtres.[4]

La seconde ligne d'enquête fait son début avec le travail de Galton et Watson (1874) sur la probabilité d'extinction des noms de famille.[5] Ce travail fut motivé en ce temps-là par l'observation soucieuse que nombreuses familles de l'aristocratie britannique étaient éteintes ou risquaient de s'éteindre. Cet étude entama une entière ligne de recherches mathématiques sur les procès de ramification (*branching processes*). Le premier résultat prévoyait erronément que chaque nom avait probabilité 1 de s'éteindre après un temps suffisamment long; en suite on éclaircit que telle probabilité n'est pas 1 dans une population croissante, cependant elle a une valeur pas négligeable et prochaine à 0.5.

La génétique ne s'intéressa pas trop aux noms de famille jusqu'aux premières années Soixante-dix, lorsque Yasuda, Cavalli-Sforza et leurs collaborateurs reconnurent que la transmission des noms obéit aux mêmes lois biologiques et statistiques de la transmission des allèles neutres (la large partie du patrimoine génétique qui ne conditionne pas le phénotype, et donc n'est pas assujettie à la pression évolutive).[6] En particulier les noms de famille européens se transmettent exactement comme le chromosome Y, en passant de père en fils dans la ligne masculine, et, comme pour les gènes, leur distribution se développe dans le temps pour effet des deux phénomènes concomitants: la mutation et la migration.

Par conséquent on appliqua aux noms de famille les mêmes conclusions et les analyses déjà effectués pour les allèles neutres, et en particulier la théorie mathématique formulée en 1967 par Karlin et McGregor, qui tenait explicitement compte de la possibilité que, avec le temps, le patrimoine génétique était sujet à des mutations.[7] La version simplifiée de la théorie, applicable dans le cas de populations très nombreuses (distribution de Fisher), devint rapidement le moyen ordinaire de paramétrisation des distributions de noms.[8]

Un grand intérêt fut aussi soulevé par l'observation de Fox et Lasker (1983), qui relevèrent sur une base purement empirique que la fréquence des noms était décrite avec remarquable précision par la courbe de Pareto, originairement introduite pour décrire la distribution des revenus, mais plus tard réutilisée dans l'analyse quantitative de nombreux différents phénomènes sociaux.[9]

Un des aspects plus intéressants des analyses sur la distribution des noms de famille est la possibilité de calculer, sur la base des données empiriques relatifs à la distribution actuelle, une estime statistique pour la fraction de noms qui se sont

modifiés ou ajoutés par rapport à une hypothétique population initiale. Ainsi on peut estimer l'importance des phénomènes migratoires si le nombre des mutations peut être considère relativement négligeable.

L'étude des phénomènes migratoires a attrait beaucoup d'attention par les génétistes, qui ont étudié les dynamiques migratoires en Sardaigne (1984)[10], ensuite en Italie (1987)[11] et en France (1992)[12], pendant que le groupe dirigé par J. Barrai, en exploitant, à partir de 1995, les possibilités offertes par les versions digitalisées des listes téléphoniques, a étendu l'analyse à tous les principaux Pays européens, jusqu'à couvrir moyennement le 9% de l'entière population, et a fourni des estimes intéressantes sur l'entité des migrations à niveau continentale et sur les effets de séparation et d'isolation dus à la distance spatiale et aux barrières géographiques et linguistiques.[13]

Une autre ligne de recherche a été engendrée par l'observation de l'analogie entre le nom de famille et le chromosome Y, et concerne le problème de l'origine (monophylétique ou polyphylétique) des noms de famille. L'étude que Sykes et Irven (2000) ont conduit sur le DNA d'un groupe de porteurs mâles du nom Sykes, pour lequel les sources documentaires fouillées par les historiens semblaient indiquer une origine multiple, a montré que les données génétiques sont par contre cohérentes avec l'hypothèse que, compte tenu des inévitables cas de non-paternité, tous les Sykes descendent d'un seul ancêtre mâle commun.[14]

Autres intéressantes études sont ceux relatives au nom hébreu Cohen, pour lequel l'hypothèse biblique d'une descendance patrilinéaire directe à partir d'Aaron frère de Moïse résulte génétiquement plausible, et ceux sur l'ascendance séparée des irlandais ayant un nom de famille d'origine gaélique et de ceux ayant un nom anglo-saxon, et tout récemment (2007) les recherches sur l'influence génétique des Vikings dans le nord-ouest de l'Angleterre.

Prenons enfin en considération les aspects qui ont attrait particulièrement l'attention des mathématiciens et des physiciens statistiques, qui sont beaucoup attirés par l'existence de structures universelles en phénomènes très différents et cherchent à expliquer leur origine dynamique par la construction de modèles mathématiques (dans ce cas modèles d'évolution des populations).

L'analyse de la consanguinité a reçu une importante contribution de la publication du travail de Derrida, Manrubia et Zanette (1999), dans lequel on a comparé un modèle théorique avec les valeurs empiriques de la répétition des ancêtres dans l'arbre généalogique d'Edouard III, roi d'Angleterre.[15]

Un example très clair du phénomène de la répétition des ancêtres est donné par Don Carlos d'Habsburg, petit-fils de l'Empereur Charles V. À la quatrième generation

ses ancêtres étaient seulement quatre (en lieu d'huit), à la sixième generation ils étaient douze (en principe 32), et à la dixième ils étaient 83 (en principe 512)

Malheureusement, ne sont pas encore suivis autres études plus systématiques, mais aujourd'hui, avec l'énorme diffusion de matériels généalogiques dans Internet et avec la création de bases de données contenant des centaines de milliers de noms et leurs connexions familiales, les temps sont mûrs pour un étude beaucoup plus étendu, qui pourrait parvenir à une validation définitive du modèle.

Pour ce qui concerne la distribution des noms, le résultat le plus avancé est dû au groupe coréen de Baek, de Tiet et de Kim, qui ont formulé en 2007 un modèle très général de dynamique des populations, qui englobe les modèles antécédents, en introduisant les probabilités de naissance, de mort, de mutation et d'immigration et en tirant des lois statistiques pour la croissance du nombre des noms de famille par rapport à la croissance globale de la population, et pour la probabilité qui un nom paraisse dans la population en fonction du nombre des individus porteurs du nom.[16]

Dans tous les cas on a montré que la distribution tend vers une courbe de Pareto, mais le paramètre caractéristique de la courbe dépend dès modalités avec lesquelles les nouveaux noms sont ajoutés, et prend des valeurs significativement différents dans le cas d'absence de mutation et dans celui d'absence de migration, avec des valeurs intermédiaires dans les situations mixtes.
L'analyse des données empiriques, en particulier pour Pays, comme la Corée et la Chine, où les mutations des noms ont été toujours négligeables, montre un remarquable accord entre le modèle et la réalité historique.

Ces résultats ont été tout récemment reproduits dans la thèse d'A. De Luca, qui a utilisé le groupe de renormalisation, un outil tout propre de la physique théorique, ce qui a permis de reconduire la dynamique des noms de famille au contexte plus général des phénomènes de criticité auto-organisée.[17] Ces phénomènes semblent caractériser beaucoup de systèmes complexes, soit physiques que biologiques, économiques et sociaux, dans lesquels ne soit pas présente une échelle caractéristique et les conditions d'auto similarité soient vérifiées.

---

[1] G.H. Darwin, *Marriages between first cousins in England and their effects*, Journal of the Statistical Society 38 (1875) 153-184
[2] J.F. Crow and A.P. Mange, *Measurement of inbreeding from the frequency of marriages between persons of the same surname,* Eugenics Quarterly 12 (1965) 199-203
[3] G.W. Lasker, *A coefficient of relationship by isonymy: A Method for Estimating the genetic Relationship between Populations,* Human Biology 49 (1977) 489-493; G.W. Lasker, *Surnames and Genetic Structure,* Cambridge University Press 1985
[4] J. Pinto-Cisternas et al., *Estimation of Inbreeding from Isonymy in Iberoamerican Populations*, Am.J. Hum. Gen. 37 (1985) 373-385; G.W.Lasker, *Datos sobre los apellidos hispanoamericanos en los estudios de biologia humana*, Ann. Antrop. (Mex) 28 (1993) 107-128;

S.E. Colantonio, G.W. Lasker, B.A. Kaplan and V. Fuster, *Use of Surname Models in Human Population Biology: A Review of Recent Developments,* Human Biology 75 (2003) 785-807

[5] F. Galton and H.W. Watson, *On the Probability of the Extinction of Families,* Journal of the Anthropological Institute of Great Britain and Ireland 4 (1874) 138-144

[6] N. Yasuda, L.L. Cavalli-Sforza, M. Skolnick and A. Moroni, *The Evolution of Surnames: An Analysis of Their Distribution and Extinction*, Theor. Pop. Biol. 5 (1974) 123

[7] S. Karlin and J. McGregor, *The number of mutant forms maintained in a population*, Proc. 5th Berkeley Symp. Math. Stat. Prob. 4 (1967) 415-438

[8] R.A. Fisher, A.S. Corbet and C.B. Williams, *The Relation Between the Number of Species and the Number of Individuals in a random sample of an Animal Population,* J. Animal Ecology **12** (1943) 42-58

[9] W.R. Fox and G.W. Lasker, *The Distribution of Surname Frequencies,* Int. Stat. Review 51 (1983) 81-87

[10] G. Zei, C.R. Guglielmino, E. Siri, A. Moroni and L.L. Cavalli-Sforza, *Surnames as Neutral Alleles: Observations in Sardinia,* Human Biology 55 (1983) 357-365; G. Zei, R. Guglielmino Matessi, E. Siri, A. Moroni and L. Cavalli-Sforza, *Surnames in Sardinia I. Fit of frequency distributions for neutral alleles and genetic population structure,* Ann. Hum. Genet. 47 (1983) 329-352; G. Zei, A. Piazza, A. Moroni and L.L. Cavalli-Sforza, *Surnames in Sardinia III. The spatial distribution of surnames for testing neutrality of genes,* Ann. Hum. Genet. 50 (1986) 169; E. Wijsman, G. Zei, A. Moroni and L.L. Cavalli-Sforza, *Surnames in Sardinia II. Computation of migration matrices from surname distributions in different periods,* Ann. Hum. Genet. 48 (1984) 65-78

[11] A. Piazza, S. Rendine, G. Zei, A. Moroni and L.L. Cavalli-Sforza, *Migration rates of human populations from surname distributions,* Nature 329 (1987) 714; A. Piazza, N. Cappiello, E. Olivetti and S. Rendine, *A genetic history of Italy*, Ann. Hum. Genet. 52 (1988) 203-213

[12] P. Darlu and J. Ruffié, *L'immigration dans les départements français étudiée par la méthode des patronymes,* Population 47 n.3 (1992) 719; P. Darlu and J. Ruffié, *Relationships between consanguinity and migration rate from surname distributions and isonymy in France*, Ann. Hum. Biol. 19 (1992) 133; P. Darlu, A. Degioanni, J. Ruffié, *Quelques statistiques sur la distribution des patronymes en France*, Population 52 n.3 (1997) 607; A. Degioanni, A. Lisa, G. Zei, P. Darlu, *Patronymes italiens et migration italienne en France*, Population 51 n.6 (1996) 1153

[13] C. Scapoli, A Rodriguez-Larralde, M. Beretta, C. Nesti, A. Lucchetti, I. Barrai, *Correlations between Isonymy Parameters*, Int. J. of Anthropology 12 (1997) 17; C. Scapoli, E. Mamolini, A. Carrieri, A. Rodriguez-Larralde, I. Barrai, *Surnames in Western Europe: A comparison of the subcontinental populations through isonymy,* Theor. Pop. Biol. 71 (2007) 37-48

[14] B. Sykes and C. Irven, *Surnames and the Y Chromosome,* Am. J. Hum. Genet. 66 (2000) 1417-1419; M. A. Jobling, *In the name of the father: surnames and genetics,* Trends in Genetics 17 (2001) 353-357

[15] B. Derrida, S.C. Manrubia and D.H. Zanette, *Statistical Properties of Genealogical Trees*, Phys. Rev. Lett. 82 (1999) 1987; B. Derrida, S.C. Manrubia and D.H. Zanette, *Distribution of repetitions of ancestors in genealogical trees,* Physica A 281 (2000) 1

[16] S.K. Baek, H.A.T. Kiet and B.J. Kim, *Family name distributions: Master equation approach,* Phys. Rev. E 76 (2007) 046113

[17] A. De Luca and P. Rossi, *Renormalization group evaluation of exponents in family-name distributions* (2008, à paraître)