OPEN ACCESS Freely available online

PLoS one

Evolution of Surname Distribution under Gender-Equality Measures

Luis F. Lafuerza, Raul Toral*

IFISC (Instituto de Fsica Interdisciplinar y Sistemas Complejos), CSIC-UIB, Campus UIB, Palma de Mallorca, Spain

Abstract

We consider a model for the evolution of surname distribution under a gender-equality measure currently being discussed by the Spanish Parliament (whereby children would adopt their mother's and father's surnames in alphabetical order). We quantify how this would bias the alphabetical distribution of surnames, and analyze its effect on the present distribution of surnames in Spain.

Citation: Lafuerza LF, Toral R (2011) Evolution of Surname Distribution under Gender-Equality Measures. PLoS ONE 6(4): e18105. doi:10.1371/journal.pone.0018105

Editor: Yamir Moreno, University of Zaragoza, Spain

Received December 22, 2010; **Accepted** February 21, 2011; **Published** April 14, 2011

Copyright: © 2011 Lafuerza, Toral. This is an open-access article distributed under the terms of the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original author and source are credited.

Funding: The authors acknowledge financial support by the MICINN (Spain) and FEDER (EU) through project FIS2007-60327. L.F.L. is supported by the JAEPredoc program of CSIC. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: raul@ifisc.uib-csic.es

Introduction

In Spain, as in many other countries, children usually inherit their father's surname. As a consequence, the mother's surname is lost in the child's generation (in Spain, however, the mother's surname is kept as a second surname, it is consequently totally lost in the grand-children's generation). Nowadays, in Spain, parents can agree upon whether it is the mother's or father's surname that is given to their children, but if parents do not reach an agreement, it will automatically be the father's surname that is inherited by the children. Due to gender-equality issues, a new law is under review whereby, if either, parents do not reach an agreement, or if no wish is expressed, the surname inherited by the children will be selected according to the alphabetical order of the parents' two surnames.

People have immediately realized that this implies a bias on the surnames favoring those beginning with the first letters in the alphabet (A,B,…) and could mean that surnames beginning with the last letters (…,Y,Z) disappear completely. In this short paper, we quantify the effect of this bias on the present distribution of surnames in Spain.

Evolution of surnames distribution (in the absence of any alphabetical preference) has been studied previously. The pioneering work of Galton and Watson used a branching process to study the probability of extinction of surnames in English aristocracy[1]. This problem turns out to be mathematically equivalent to that of the evolution of non-recombining neutral alleles [2] and several authors[3–5] have used similar ideas in the context of biological evolution. Later developments by mathematicians and physicists centered on the distribution of family sizes [6–9]. The novelty of our work is that we analyze how the distribution of surnames evolves when a preference depending on their alphabetical position is present. We derive an equation for the evolution of surnames distribution on these premises. The solution of the equations allows us to determine the timescale at which the surnames disappear.

Analysis

As a first order model aimed at capturing the essence of the process of surname inheritance we propose the following:

(*i*) Initially, a population of N individuals ($N/2$ male and $N/2$ female) is considered. Each individual has a surname chosen according to some prescribed distribution.

(*ii*) Males and females reproduce in random pairs in such a way that, on average, the total population remains constant.

(*iii*) With probability a it is assumed that parents reach an agreement, so that the surnames of their children are chosen at random between those of the parents (the proportion of whether the father's surname or the mother's is preferred is irrelevant on the results). With probability $1-a$, parents do not reach or do not express an agreement, and children adopt their surname by the alphabetical order rule.

We measure time t in average reproductions per person, or generations. In a generation, parents are replaced by their children in the population.

The population evolves according to a bisexual Galton and Watson branching process[10]. The statistics of the number of people as a function of time [11] and the distribution of the frequency of surnames in a model similar to this one when the surnames are chosen at random have been studied previously[8].

The model introduced above is a minimal model and does not consider some realistic issues such as geographical distribution of surnames, etc. but those are expected to be second order effects with little impact in the overall trend. The effect of immigration and population growth is analyzed later in the text.

Let us define $p(n,t)$ as the proportion of individuals (both males and females) with surname in the alphabetical position $n=1,\ldots,M$, being M the total number of surnames. It evolves according to:

$$\frac{\partial p(n,t)}{\partial t} = (1-a)p(n,t)\left[\sum_{k=n+1}^{M} p(k,t) - \sum_{k=1}^{n-1} p(k,t)\right] \qquad (1)$$

$$=(1-a)p(n,t)[1-P(n,t)-P(n-1,t)], \quad (2)$$

where $P(n,t)=\sum_{k=1}^{n} p(k,t)$ is the cumulative distribution. The first term in the square brackets of Eq.(1) represents the increase in probability of surname n due to the pairing with surnames $k\in[n+1,M]$ which are further forwards the end in the alphabetical order, while the second term represents the loss in probability due to pairings with surnames $k\in[1,n-1]$ earlier in the alphabetical order. It follows that:

$$\frac{\partial P(n,t)}{\partial t}=(1-a)P(n,t)[1-P(n,t)], \quad (3)$$

whose solution is:

$$P(n,t)=\frac{P(n,0)e^{(1-a)t}}{1+P(n,0)(e^{(1-a)t}-1)}. \quad (4)$$

The distribution of surnames at time t is then $p(n,t)=P(n,t)-P(n-1,t)$ for $n\geq 1$ with the convention $P(0,t)=0$. Approximating the difference by a derivative $p(n,t)\simeq\frac{\partial P(n,t)}{\partial n}$, we obtain:

$$p(n,t)=\frac{p(n,0)e^{(1-a)t}}{[1+P(n,0)(e^{(1-a)t}-1)]^2}. \quad (5)$$

Eq. (4) shows that the distribution of surnames approaches a Kronecker-delta at $n=1$ ($P(n,t)=1,\forall n$) exponentially quickly with a characteristic time $1/(1-a)$. Assuming, for instance, that couples reach, and express, an agreement in 50% of the cases ($a=1/2$), we find from Eq.(5) that the frequency of a surname towards the end of the alphabetical table would be decreased by a factor 10 in around 4.6 generations($\sim$115 years). If, on the other hand, couples do not reach an agreement in 5% of the cases ($a=0.95$), then the decrease by a factor 10 occurs in 46 generations.

Evolution of current distribution

We have applied the above results to the current distribution of Spanish surnames. Besides the analytical result of Eq.(5), we have performed a numerical simulation of the model by which $N=10^7$ couples have probabilities (0.05,0.2,0.5,0.2,0.05) of having (0,1,2,3,4) children (average value is 2). The probability of parents reaching an agreement is set at $a=0.5$. Whether an agreement has been reached or not, the rule applied to the first-born child is used for all subsequent children. We have used as the initial condition $p(n,0)$ the distribution of the $M=100$ most common surnames in Spain, after ordering them in alphabetical order. The data appear in the INE webpage www.ine.es (INE stands for "Instituto Nacional de Estadística"). Similar data is available for other countries. Our simulation results only consider those 100 surnames for which data are publicly available. In figure 1 we plot (symbols) the probability distribution $p(n,t)$ resulting from this numerical simulation after $n=4$ (top panel) and $n=10$ (bottom panel) generations. In the same figure we also plot the theoretical prediction, Eq. (5) using the same initial condition and for the same number of generations. As it can be seen in the figure (note the logarithmic scale for better viewing of data in the case $t=10$) the concurrence between the simulation and the analytical result is excellent. It can be noted that the relative importance of surnames moves towards the surnames which are earlier in the alphabet as time increases.

The evolution in the frequency of a surname does not have to be monotonous, as it can first increase and then decrease in time. Let us take, for instance, the most common surname in Spain: "Garca". According to the INE data, there are 1,481,923 people bearing this surname and the cumulative distribution is $P(n,0)\approx 0.291$. Hence, if there is a 50% agreement, the frequency or this surname would first increase up to 1.8×10^6 in two generations, to then decrease to 1.1×10^5 in 10 generations. In the case of the surname "Toral", there are nowadays 3,190 people bearing this surname and the cumulative distribution is $P(n,0)\approx 0.97$. According to the previous analysis, and considering again 50% agreement, it would decrease to 1,980 in one generation and to 23 in 10 generations. The same study for "Lafuerza" (very rare surname, only 122 people bear it in Spain currently and the cumulative distribution is $P(n,0)\approx 0.465$), shows that it would stay practically constant in the first generation to then decrease to 4 (practical extinction) in 10 generations. Finally, if we take a surname high in the alphabetical order such as "Aguilar", of which there are 58,771 people at the moment, it would increase practically exponentially, as the cumulative distribution is very small and can be neglected in the denominator of Eq.(5). Of course, all these predictions are for the mean values, and significant statistical deviations could occur for low-frequency surnames.

Effect of immigration and population growth

In the previous analysis we assumed that the population remains constant (there is an average of two children per couple) and that the only changes in the distribution of surnames correspond to the application of the alphabetical order rule. We now consider the effect that, both, population growth and new surnames brought in by immigration have in the distribution of surnames. This implies modifying condition (*ii*) in the model, allowing the number of children per couple to have any average value r and setting immigration events with rate $\lambda_I N(t)$, proportional to the total population number. The alphabetical position of the surname of the immigrant is chosen at random according to some probability distribution $p_I(n)$. The total population increases exponentially as $N(t)=N(0)e^{(\lambda_I+\lambda_p)t}$ with $\lambda_p=(r-2)/2$.

Let $N(n,t)=p(n,t)N(t)$ be the number of people with surname in alphabetical position n. It evolves according to:

$$\begin{aligned} N(n,t+\Delta t)-N(n,t)=&\frac{1}{2}N(t)\Delta t\frac{N(n,t)}{N(t)}\frac{N(n,t)}{N(t)}\times(r-2)+\\ &\frac{1}{2}N(t)\Delta t\,2\frac{N(n,t)}{N(t)}\frac{N(t)-N(n,t)}{N(t)}\times\\ &\left[a\frac{r-2}{2}+(1-a)\left[\frac{N(t)(1-P(n,t))}{N(t)-N(n,t)}(r-1)+\frac{N(t)P(n-1,t)}{N(t)-N(n,t)}(-1)\right]\right]\\ &+\lambda_I N(t)p_I(n)\Delta t+O(\Delta t^2). \end{aligned} \quad (6)$$

The first term corresponds to mating of two people (one male, one female) both with surname in alphabetical position n, in this case the average increase in $N(n,t)$ equals the average increase due to the mating, which is $r-2$. The second term corresponds to the mating of a person with surname n with a person with a different surname: if, with probability a, they agree on the surname to be assigned to the children, the average increase in $N(n,t)$ equals $(r-2)/2$; otherwise, the average increase in $N(n,t)$ is either $r-1$ (if the other surname is later in the alphabet) or -1 (if it is earlier). This derivation neglects the possible fluctuations that can appear in the distributions of surnames in males

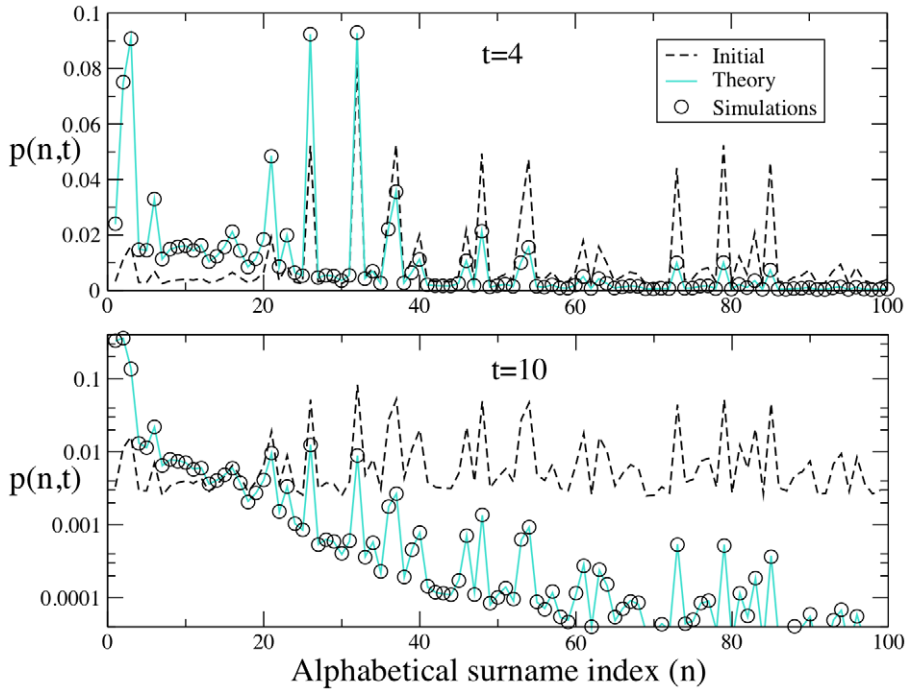

Figure 1. Evolution of the distribution of surnames after $n=4$ (top) and $n=10$ (bottom) generations, taking as initial condition $p(n,0)$ the actual distribution of the $M=100$ most common surnames in Spain. For $n=10$ we have used a logarithmic scale for a better viewing of the data. The dots are the result of the numerical simulation of the model described in the main text, and solid lines correspond to the analytical result (4).
doi:10.1371/journal.pone.0018105.g001

and females. The last term corresponds to the addition of new individuals brought in by immigration.

From this equation, and after some algebra, one can obtain the evolution of the number of people $N(s\leq n,t)=\sum_{s=1}^{n}N(s,t)$ with surname in alphabetical position smaller or equal to n:

$$N(s\leq n,t+\Delta t)-N(s\leq n,t)=\left[r\left(1-\frac{a}{2}\right)-1\right]\Delta t N(s\leq n,t)-\frac{r}{2}(1-a)\Delta t\frac{N(s\leq n,t)^2}{N(t)}+\lambda_I N(t)P_I(n)\Delta t+O(\Delta t^2), \quad (7)$$

being $P_I(n)$ the cumulative distribution of the immigrants surnames. Dividing by $N(t+\Delta t)=N(t)(1+(\lambda_p+\lambda_I)\Delta t)+O(\Delta t^2)$ and taking the limit $\Delta t\to 0$, we obtain:

$$\frac{\partial P(n,t)}{\partial t}=\lambda_I(P_I(n)-P(n,t))+(1-a)\frac{r}{2}P(n,t)[1-P(n,t)], \quad (8)$$

whose solution is:

$$P(n,t)=\frac{1}{2}-\frac{\lambda_I}{(1-a)r}+\frac{C(n)}{(1-a)r}\tanh\left[C(n)t+\operatorname{argtanh}\left(\frac{(1-a)r[P(n,0)-1/2]+\lambda_I}{C(n)}\right)\right], \quad (9)$$

with $C(n)\equiv\sqrt{2(1-a)r\lambda_I P_I(n)+[(1-a)r/2-\lambda_I]^2}$.

We see from Eq.(8) that the intrinsic growth, r, of the population only changes the timescale of the dynamics of the surname distribution. Immigration, however, might have a greater impact. Let us focus on the asymptotic, $t\to\infty$ distribution, which has the form:

$$P(n)=\frac{1}{2}+\frac{C(n)-\lambda_I}{(1-a)r}. \quad (10)$$

An analysis of this expression, shows that a critical value $\lambda_c\equiv(1-a)r/2$ exists, such that the delta-like singularity that appeared in the non-immigration case disappears for $\lambda_I>\lambda_c$. The situation then, is that for $\lambda_I<\lambda_c$ there is an accumulation at surnames close to the lower limit $n=1$, such that a fraction $1-\frac{2\lambda_I}{(1-a)r}$ of people bear the surname first in the alphabetical order. So, the low immigration rate produces results which are the same, qualitatively, as in the case without immigration. For $\lambda_I>\lambda_c$, however, the fresh distribution of surnames brought in by immigration is enough to overcome the accumulation at $n=1$ towards which the probability distribution would tend in the asymptotic limit due to the alphabetical order rule. In both cases, the tail of the stationary probability distribution, behaves as $p(n)\sim p_I(n)/[P_I(n)]^{-1/2}$. If, for instance, the distribution of new surnames were uniform in the alphabet, $P_I(n)\sim n$, then the tail of the stationary distribution would behave as a power-law of slope $-1/2$.

Discussion

We have developed a mathematical model for the evolution of the surnames distribution when an alphabetical-order rule on the progenitor's surnames is applied. The premises of the model lead to a differential equation governing the evolution of the probability distribution. As initial condition we have considered the data for the present distribution of the 100 most common surnames in

Spain, obtained from the National Statistics Institute of Spain (INE), that is publicly available at its webpage www.ine.es. Similar data is available for other countries. We have also performed numerical simulations of an agent-based model, which agree with the analytical result.

In our minimal model for surname transmission, we prove that the adoption of the alphabetical rule leads to an exponential decrease in the surnames that begin with letters that are towards the end of the alphabet, with a characteristic decay time of $1/(1-a)$ generations, being a the fraction of parents that reach an agreement. This quantifies the decrease in the frequency of those surnames. We have also considered the effect of surnames brought in by immigration and found that, below a critical value of the immigration rate, the results are the same, qualitatively, as in the case without immigration. For large immigration rates, the delta-like singularity that appeared at names earlier in the alphabet, disappears.

We believe that this study offers an example in which statistical methods and mathematical modeling can be used to quantitatively calculate the consequences of a political measure and, consequently, it can serve as a guide to institutions and policy makers.

Acknowledgments

We thank Susan Meal for useful suggestions related to presentation.

Author Contributions

Analyzed the data: LFL RT. Contributed reagents/materials/analysis tools: LFL RT. Wrote the paper: LFL RT.

References

1. Galton F, Watson HW (1874) On the probability of the extinction of families. J Roy Anthropol Inst 4: 138–144.
2. Sykes B, Irven C (2000) Surnames and the y chromosome. Am J Hum Genet 66: 1417–1419.
3. Fisher R (1922) On the dominance ratio. Proc Roy Soc Edin 42: 321–341.
4. Haldane JBS (1927) A mathematical theory of natural and artificial selection, part v: selection and mutation. Proc Camb Philos Soc 26: 838–844.
5. Moran PAP (1962) The statistical processes of evolutionary theory. Oxford: Clarendon Press.
6. Panaretos J (1989) On the evolution of surnames. Int Stat Rev 57: 161–167.
7. Consul PC (1991) Evolution of surnames. Int Stat Rev 59: 271–278.
8. Zanette DH, Manrubia SC (2001) Vertical transmission of culture and the distribution of family names. Physica A 295: 1–8.
9. Manrubia SC, Zanette DH (2002) At the boundary between biological and cultural evolution: the origin of surname distributions. J Theor Biol 216: 461–477.
10. Harris TE (1989) The Theory of Branching Processes. Dover Pub., 2nd edition.
11. Gonzalez M, Molina M (1997) Some theoretical results on the progeny of a bisexual Galton-Watson branching process. J Serdica Math 23: 15–24.