International Statistical Review (1989), **57**, 2, pp. 161–167. Printed in Great Britain

On the Evolution of Surnames

John Panaretos

School of Engineering, Division of Applied Mathematics, University of Patras, P.O. Box 1325, Patras, Greece

1 Introduction

In a study of the distribution of surnames in the areas of Reading and Workingham, England, Fox & Lasker (1983) considered using the discrete Pareto distribution known also as the zeta distribution for describing the mechanism causing the occurrence of surnames. Their justification for the use of this distribution was empirical and amounted to the following fact. In all of the available sets of data the log of the proportion $p(x)$ of names occurring x times demonstrated a linear relationship to the $\log x$ of slope $-(c+1)$, so that $p(x)$ was reasonably assumed to be of the form

$$p(x) = x^{-(c+1)} \Big/ \sum_{x=1}^{\infty} x^{-(c+1)}. \tag{1.1}$$

The fit of the data by (1.1) was satisfactory in all the cases: in no case was the value of the chi-square goodness of fit statistic significant at the 5% level.

However, as the authors pointed out, a theoretical justification for their model would be desirable. In a forthcoming paper (Panaretos, 1989) a probability model is proposed leading to a Yule distribution, which, suitably truncated, gives a satisfactory fit to the data of Fox & Lasker (1983).

The purpose of this paper is to provide a theoretical justification for the observed results of Fox & Lasker, i.e. to develop probability models of the surname generation mechanism which will form the theoretical basis for the interpretation of the observed satisfactory agreement between empirical and discrete Pareto frequencies.

So §§ 2 and 3 consider alternative stochastic derivations of the theoretical surname distribution. It turns out that in both cases this distribution is the Yule distribution with parameter c and probability function (p.f.)

$$P(X=x) = c(x-1)!/(c+1)_{(x)} \quad (x=1, 2, \dots), \tag{1.2}$$

where the symbol $\alpha_{(\beta)}$ denotes the ratio $\Gamma(\alpha+\beta)/\Gamma(\alpha)$, $\alpha>0$, $\beta \in \mathbb{R}$.

This distribution obtained by Yule (1924) in the context of a biological problem is always J-shaped and long-tailed. Moreover, (1.2) approximates (1.1) in the tail since as $x \to +\infty$ the right-hand side of (1.2) is proportional to $x^{-(c+1)}$.

As demonstrated in § 4 the Yule distribution provides a fit to the data which is as good as that provided by (1.1). Finally, in § 5, some conclusions and remarks are made.

2 A Stochastic Derivation of the Surname Distribution—A Contagion Model

Suppose that at time $t = 0$ there exist k different surnames in a population and let X denote the number of occurrences of a given surname. Let $p_\lambda(x, t)$ denote the probability that a surname has had x occurrences by time t given λ, where λ is a parameter reflecting differences among geographic areas as far as the commonality of surnames is concerned. Suppose that during the time period from t to $t + dt$ a name having had x occurrences by time t can have, for given λ:

$$\begin{aligned} &0 \text{ occurrences with probability } 1 - f_\lambda(x, t)\, dt, \\ &1 \text{ occurrence with probability } f_\lambda(x, t)\, dt, \\ &>1 \text{ occurrences with probability } 0. \end{aligned}$$

Therefore, since $p_\lambda(1, 0) = 1$ (i.e. since we start off with k different surnames),

$$\begin{aligned} p_\lambda(1, t + dt) &= p_\lambda(1, t)(1 - f_\lambda(1, t)\, dt) \\ p_\lambda(2, t + dt) &= p_\lambda(2, t)(1 - f_\lambda(2, t)\, dt) + p_\lambda(1, t) f_\lambda(1, t)\, dt \\ &\vdots \qquad\qquad\qquad\qquad \vdots \\ p_\lambda(x, t + dt) &= p_\lambda(x, t)(1 - f_\lambda(x, t)\, dt) + p_\lambda(x - 1, t) f_\lambda(x - 1, t)\, dt \\ &\vdots \qquad\qquad\qquad\qquad \vdots \end{aligned}$$

which imply that

$$\begin{aligned} \frac{\partial}{\partial t} p_\lambda(1, t) &= -f_\lambda(1, t) p_\lambda(1, t) \\ \frac{\partial}{\partial t} p_\lambda(2, t) &= -f_\lambda(2, t) p_\lambda(2, t) + f_\lambda(1, t) p_\lambda(1, t) \\ &\vdots \qquad\qquad\qquad\qquad \vdots \\ \frac{\partial}{\partial t} p_\lambda(x, t) &= -f_\lambda(x, t) p_\lambda(x, t) + f_\lambda(x - 1, t) p_\lambda(x - 1, t) \\ &\vdots \qquad\qquad\qquad\qquad \vdots \end{aligned}$$

Multiplying the ith equation by s^i, $i = 1, 2, \ldots$ and summing over i we obtain

$$\frac{\partial}{\partial t} G_\lambda(s; t) = (s - 1) \sum_{x=1}^{\infty} s^x f_\lambda(x, t) p_\lambda(x, t), \tag{2.1}$$

where

$$G_\lambda(s; t) = \sum_{x=1}^{\infty} p_\lambda(x, t) s^x. \tag{2.2}$$

Assume now that

$$f_\lambda(x, t) = \lambda m x \quad (x = 1, 2, \ldots; m > 0), \tag{2.3}$$

i.e. that the more occurrences a surname has had the more likely it is to have a further

occurrence. Then using (2.3), equation (2.1) reduces to

$$\frac{\partial}{\partial t} G_\lambda(s;t) = \lambda m s(s-1)\frac{\partial}{\partial s} G_\lambda(s;t).$$

A solution to this equation is

$$G_\lambda(s;t) = s/[e^{\lambda m t} - s(e^{\lambda m t} - 1)], \tag{2.4}$$

which under the initial conditions $G_\lambda(1;t) = s^{-1}G_\lambda(s;0) = 1$ is unique.

Thus, for given λ (i.e. when all the surnames are common to the same degree), the surname distribution is the geometric with parameter $e^{-\lambda m t}$. Assuming now that differences in the commonality of surnames reflected by the fluctuations of λ are effected through an exponential distribution with parameter α the probability generating function of the surname distribution is given by

$$\begin{aligned} G_X(s;t) &= \alpha s \int_0^{+\infty} e^{-\alpha\lambda}[e^{\lambda m t} - s(e^{\lambda m t} - 1)]^{-1}\, d\lambda \\ &= \frac{\alpha s}{mt(1-s)} \int_0^{+\infty} e^{-(1+\alpha/(mt))\lambda}\left(1 - \frac{s}{s-1}e^{-\lambda}\right)^{-1} d\lambda \\ &= \frac{\alpha s}{mt(1-s)} \frac{\Gamma(1+\alpha/(mt))}{\Gamma(2+\alpha/(mt))} {}_2F_1(1,\, 1+\alpha/(mt);\, 2+\alpha/(mt);\, s/(s-1)) \\ &= \frac{\alpha s}{\alpha + mt} {}_2F_1(1,\, 1;\, 2+\alpha/(mt);\, s), \end{aligned}$$

where

$${}_2F_1(\alpha,\, \beta;\, \gamma;\, z) = \sum_{r=0}^{\infty} \frac{\alpha_{(r)}\beta_{(r)}}{\gamma_{(r)}} \frac{z^r}{r!}$$

is the Gauss hypergeometric function whose radius of convergence is $[-1, 1]$, provided that $\gamma - \alpha - \beta > 0$.

Therefore

$$G_X(s;t) = \frac{\alpha s}{\alpha + mt} \sum_{r=0}^{\infty} \frac{r!\, s^r}{(2+\alpha/(mt))_{(r)}} = \frac{\alpha}{\alpha + mt} \sum_{r=1}^{\infty} \frac{(r-1)!\, s^r}{(2+\alpha/(mt))_{(r-1)}}$$

which implies that

$$P(X = x) = \frac{\alpha}{mt} \frac{(x-1)!}{(1+\alpha/(mt))_{(x)}} \quad (x = 1, 2, \ldots). \tag{2.5}$$

But this is the p.f. of the Yule distribution with parameter $c = \alpha/(mt)$. It is evident that in a unit time period the (p.f.) in (2.5) becomes proportional to $x^{-(1+\alpha/m)}$ as x increases, so that (2.5) can be thought of as a reasonable approximation to (1.1) in the tail.

The above model assumes that the probability of a name to occur is affected not only by the number of its previous occurrences but also by a factor reflecting the commonality of the particular name in the area of study. It is essentially a hypothesis of contagious transmission of a surname within an area of a given degree of commonality. Such an assumption can be supported by the fact that, in the cultures of today, surnames of one locus are transmitted through the male or, occasionally, through the female line implying that the higher the number of previous occurrences of a name the higher the probability of a further occurrence.

3 An Alternative Probability Model

The model that was considered in § 2 was based on a 'contagion' hypothesis, namely that the probability of a given surname to occur next is proportional to the number of its previous occurrences.

In this section we will consider weakening this assumption by considering a stochastic model adopted by Simon (1955) in the context of a problem in linguistics.

Let $f(x, k)$ denote the number of different surnames that have occurred x times among k surnames that exist in a population at time t.

Assume that the probability $p(x)$ that the $(k+1)$st surname is a surname that has already occurred x times is proportional to $xf(x, k)$, that is to the total number of occurrences of all the surnames that have appeared exactly x times. Assume further that the probability that the $(k+1)$st surname is a new surname (that has not occurred previously) is equal to p, $0<p<1$. Assume also that the frequencies increase proportionally to k, that is

$$f(x, k+1)/f(x, k) = (k+1)/k, \tag{3.1}$$

and let E_r denote the event {the $(k+1)$st surname has already occurred r times among the k surnames}, $r = 1, 2, \ldots$.

Obviously

$$\{f(x, k+1) - f(x, k) = 1\} = E_{x-1}, \quad \{f(x, k+1) - f(x, k) = -1\} = E_x.$$

Therefore

$$\mathcal{E}\{f(x, k+1) - f(x, k) \mid f(x, k)\} = P(E_{x-1}) - P(E_x).$$

But

$$P(E_r) = c_k r f(r, k)$$

for some constant c_k dependent on k. Therefore

$$E\{f(x, k+1) - f(x, k) \mid f(x, k)\} = c_k\{(x-1)f(x-1, k) - xf(x, k)\} \quad (x = 2, 3, \ldots). \tag{3.2}$$

Similarly

$$E\{f(1, k+1) - f(1, k) \mid f(1, k)\} = p - c_k f(1, k). \tag{3.3}$$

Obviously,

$$c_k \sum_{i=1}^{k} i f(i, k) = 1 - p.$$

But,

$$\sum_{i=1}^{k} i f(i, k) = k,$$

i.e.

$$c_k = (1-p)/k. \tag{3.4}$$

From (3.1) it follows that the relative frequency of names with x occurrences in a total of k names is independent of k. Using this fact and combining (3.1) and (3.4), equation (3.2) can be written in terms of probabilities as

$$\frac{p(x)}{p(x-1)} = \frac{(1-p)(x-1)}{1+(1-p)x} \quad (x = 1, 2, \ldots).$$

Solving for $p(x)$ we obtain

$$p(x) = \frac{1}{1-p}\,\frac{(x-1)!}{(1/(1-p)+1)_{(x)}} \quad (x = 1, 2, \ldots).$$

That is the probabilities with which a given name will have $1, 2, \ldots$ occurrences form the Yule distribution with parameter $c = 1/(1-p)$.

The model of this section assumes that the probability with which an occurring name has already occurred i times is proportional to its previous occurrences and constant if the name occurs for the first time. This is a weaker hypothesis as compared to the 'contagion hypothesis' and is quite plausible in the context of the present analysis if one takes into account the fact that no 'new' surnames can be introduced through the male or female line in the given locus; a new surname can only occur through 'immigration' of new persons in the study area from a different geographical area. As a result, the probability of a new name can be thought of as being constant (possibly proportional to the immigration rate), while that of a name observed i times can be thought of as being proportional to the total number of names with i occurrences.

4 Fitting the Model to Actual Data

In this section the Yule distribution as defined by (1.2) has been fitted to the frequency distribution of surnames observed by Lasker et al. (1979). The same division of the study area in eight non-overlapping districts has been considered for comparison purposes to Fox & Lasker's (1983) results. (The 'ninth district' of Lasker et al. (1979) consisting of persons resident outside the study area has not been included in the present analysis since the assumptions of the two stochastic models concerning the transmission of surnames would not be relevant.) In a given district, x denotes the number of people in the district with a given surname and $f(x)$ the number of surnames occurring x times. The Yule distribution has been fitted by the method of moments which yielded estimates for the parameter c given by the statistic $\hat{c} = \bar{X}/(\bar{X}-1)$, where $\bar{X}$ denotes the mean of the given sample. The results are summarized in Table 1. Upper entries of the table denote observed frequencies while lower entries denote expected frequencies by the Yule distribution. The last four rows of the table provide the value of the parameter c, the value of the chi-square goodness of fit test statistic, the number of degrees of freedom on which it was calculated and the p-value.

An inspection of Table 1 shows that the fit of the observed distribution by the expected Yule frequencies is quite reasonable. With the exception of district 6 the p-value nearly equals or is above 0·25. For district 6 the value of p equals approximately 0·10.

Comparing the fit of the observed surname distribution by the Yule distribution to that by the discrete Pareto distribution given by Fox & Laker (1983) shows that the results are in close agreement as reflected by Table 2. The overall fit of the data by the two distributions shows no appreciable difference as far as the degree of the goodness of fit is concerned. Apart from one or two cases where one may say that the fit is noticeably better on one of the two models, e.g. district 8, the degree of agreement between observed and expected frequencies as judged by the p value is similar in both cases (greater than 0·10 and in some cases even greater than 0·25).

The results do not come as a surprise. They were expected and the reason for this obviously lies in the fact that the discrete Pareto distribution is a limiting form of the Yule distribution.

Table 1

Observed, upper entries, and expected, lower entries, frequencies $f(x)$, of the occurrences of surnames in the study area of Reading using the Yule distribution with parameter c.

Occurrence	District							
x	5	4	6	3	8	7	2	1
1	234	243	281	292	282	349	329	832
	232·702	240·86	276·475	289·490	279·391	345·893	328·945	819·478
2	19	17	23	28	34	30	43	151
	21·450	20·798	30·129	31·038	37·15	35·144	42·36	155·081
3	5	4	9	6	11	7	11	39
	3·621	3.306	5·921	6·011	8·72	6·483	9·665	49·356
4	0	2	1	2	2	3	1	20
	0·845	0·730	1·589	1·592	2·748	1·642	2·969	20·327
5	1	0	0	0	0	1	0	11
	0·244	0·200	0·522	0·517	1·045	0·512	1·103	9·815
6	0	0	0	0	0	0	1	2
	0·082	0·064	0·198	0·194	0·453	0·185	0·469	5·286
7	0	0	0	1	0	0	0	4
	0·031	0·023	0·084	0·081	0·217	0·075	0·220	3·084
8	0	0	1	0	0	0	0	5
	0·013	0·009	0·039	0·037	0·112	0·033	0·112	1·913
9	0	0	0	0	0	0	1	0
	0·006	0·004	0·019	0·018	0·062	0·016	0·061	1·246
10	0	0	0	0	0	0	0	1
	0·003	0·002	0·010	0·009	0·036	0·008	0·035	0·844
11	0	0	0	0	0	0	0	0
	0·001	0·001	0·006	0·005	0·022	0·004	0·021	0·591
12	0	0	0	0	1	0	0	2
	0·0009	0·0005	0·003	0·003	0·014	0·002	0·013	0·425
⩾13	0	0	0	0	0	0	0	2
	0·001	0·0006	0·005	0·005	0·030	0·004	0·027	1·554
c	3·284	9·581	7·176	7·327	5·521	7·842	5·765	3·284
x^2	1·35	1·35	3·45	0·35	1·51	1·25	1·00	6·85
d.f.	1	1	2	1	2	1	2	5
p	>0·25	0·25	0·10	>0·25	>0·25	>0·25	>0·25	0·25

Table 2

Values of the x^2 square goodness of fit test statistic for the discrete Pareto and the Yule models.

	Discrete Pareto distribution		Yule distribution	
District	X^2	d.f.	X^2	d.f.
5	0·20	1	0·30	1
4	0·19	2	1·35	1
6	3·34	2	3·45	2
3	0·91	2	0·35	1
8	4·46	2	1·51	2
7	0·31	2	1·25	1
2	6·11	3	1·00	2
1	11·75	8	6·85	5

5 Concluding Remarks

In the preceding sections an attempt was made to provide a theoretical justification for the satisfactory description of surname frequency data by the discrete Pareto distribution as demonstrated by Fox & Lasker (1983). A reasonable explanation has been provided in terms of two probability models that lead to a distribution of a more general functional form than that of the discrete Pareto distribution known as the Yule distribution.

The first model is based on the assumption that the transmission of surnames is contagious and affected by the degree of commonality of the surnames in the various geographic areas. The second model is also based on a contagion hypothesis but weaker than that of the first model. At time t the probability of the occurrence of a surname is not affected by the number x of its previous occurrences, but by the total number of occurrences of all the surnames that have appeared x times up to time t. Moreover, new names are allowed to appear with a constant probability.

Both models lead to the same surname distribution: the Yule distribution. As already pointed out in § 1, this can be regarded as being of a more general functional form than the discrete Pareto which in effect is a limiting form of the former.

The fit of the Yule distribution to the Reading data was, as implied by Table 2, as good as that provided by the discrete Pareto. This is not surprising. It is the natural consequence of the fact that the discrete Pareto distribution can be regarded as a limiting solution of either of the proposed stochastic models.

The foregoing analysis therefore has shed some light onto the problem of determining what probability model(s) of surname generation would give rise to the discrete Pareto distribution. Two possible hypotheses were put forward and the good fit of the data by both the Yule and the discrete Pareto distribution constitutes supporting evidence of the validity of the two hypotheses. Of course the question of deciding which of these two stochastic assumptions is the underlying causing mechanism remains open.

References

Fox, W.R. & Lasker, G.W. (1983). The distribution of surname frequencies. *Int. Statist. Rev.* **51,** 81–87.

Lasker, G.W., Coleman, D.A., Aldridge, N. & Fox, W.R. (1979). Ancestral relationships within and between districts in the region of Reading, England as estimated by isonymy. *Human Biol.* **51,** 445–460.

Panaretos, J. (1989). A probability model involving the use of the zero-truncated Yule distribution of analyzing surname data. *I.M.A. J. Math. Applied in Med. and Biol.* To appear.

Simon, H.A. (1955). On a class of skew distribution functions. *Biometrika* **42,** 425–440.

Yule, G.U. (1924). A mathematical theory of evolution based on the conclusions of Dr. J.C. Willis, F.R.S. *Phil. Trans.* B **213,** 21–87.

Résumé

La détermination de l'évolution des noms est un problème qui a préoccupé la statistique ainsi que la biologie humaine. Plusieurs auteurs se sont penchés sur le problème de déterminer un modèle de probabilité convenable et bien justifié pour décrire la loi des noms. Dans cet article deux modèles stochastiques donnant lieu à la loi de Yule sont proposés pour décrire ainsi qu'adapter quelques répartitions de fréquences de noms. Le premier modèle est basé sur une hypothèse de contagion dans le sens que plus un nom apparaît dans le passé, plus il est probable d'apparaître dans le futur. Le second modèle est basé sur un ensemble de suppositions plus faibles qui permettent 'l'immigration' de noms. La loi qui émerge de ces modèles est ensuite adaptée à des données actuelles et le résultat est comparé à celui obtenu par l'adoption de la loi de Pareto discrète.

[*Received August* 1988, *accepted February* 1989]