International Statistical Review (1991), **59**, 3, pp. 271–278. Printed in Great Britain

Evolution of Surnames

P.C. Consul

Department of Mathematics, Statistics & Actuarial Science, University of Calgary, Calgary, Alberta, Canada

Summary

The determination of a suitable probability model to describe the distribution of surnames in various areas has been considered by many authors. In this paper a birth and death process model and a branching process model are proposed to explain the evolution of surnames. Both models give rise to the Geeta distribution which has been fitted to the actual data and the fit is compared to the fits with the discrete Pareto distribution and the Yule distribution. The three distributions have also been compared with regard to their domains of the mean and variance.

Key words: Birth and death process; Branching process; Discrete Pareto distribution; Geeta distribution; Surname distribution; Yule distribution.

1 Introduction

Fox & Lasker (1983) studied the distributions of surnames in nine areas of Reading and Wokingham, England and applied the discrete Pareto distribution (also called the zeta distribution)

$$P(X=x)=x^{-c-1}\Big/\sum_{x=1}^{\infty}x^{-c-1},\quad (x=1,2,3,\ldots), \tag{1.1}$$

to the actual data sets because the log of relative frequencies of surnames occurring x times had an approximate linear relationship to the $\log x$ of slope $-(c+1)$. Though the fit of the data by the discrete Pareto distribution was satisfactory in all nine cases, the authors felt that a theoretical justification of the model would be desirable.

To provide a theoretical justification for the observed results of Fox & Lasker (1983) based on the discrete Pareto distribution Panaretos (1989) has given two alternative derivations of Yule distribution with parameter c and probability function

$$P(X=x)=(x-1)!\,c/(c+1)_{(x)},\quad (x=1,2,3,\ldots) \tag{1.2}$$

where $c_{(x)}=\Gamma(c+x)/\Gamma(c)$, $c>0$. In both derivations he has tried to provide justification that the assumptions taken by him are appropriate to the surname distribution. He has demonstrated that the Yule distribution provides a good fit to eight sets of data considered by Fox & Lasker. In fact both, the Yule distribution and the discrete Pareto distribution, are always J-shaped and long tailed and approximate with each other in the tail as $x\to+\infty$. However, in the contagion model Panaretos (1989) assumes that there exist k different surnames in the area at time $t=0$ and that X, the number of occurrences of a given surname is 1 when $t=0$. Firstly, the value of k does not seem to change and play any part in the model and secondly the assumption of $X=1$ for each one of k surnames is rather artificial. Similarly, the assumption about $p(x)$ in the second model is somewhat questionable. Moreover, X seems to change from 1 to k which is the total

number of surnames in the population. Possibly, this has happened because Panaretos (1989) tried to restrict himself to a single parameter for the development of the model.

In this paper we consider a birth and death process model and a branching process model to develop the surnames distribution and show that both models yield the Geeta distribution, defined and studied by Consul (1990a,b), which is also J-shaped and is a long tailed distribution based on two parameters. The Geeta distribution has been compared with the discrete Pareto distribution and the Yule distribution in § 4 with regard to their mean and variance graphs. The fit of the Geeta distribution to the actual data sets of Fox & Lasker (1983) is given in § 5.

2 Surname Distribution by Birth and Death Process

Let $f(x)$ denote the number of different surnames each of which occurs x times so that x takes the values $1, 2, 3, \ldots, k$ in some area. Evidently, there are very few surnames which occur a large number of times. There will be a large number of surnames which will occur only once in any particular area (the largest number as evidenced by any telephone directory). Thus the distribution of surnames must be of J-shaped form.

Unless there is a mass scale movement of large families from one area to the other, the immigration of new persons into an area is more likely to increase the frequencies of surnames with smaller values of x. Also, when some people change their surname or stop writing their last name, their action increases the frequencies of surnames with smaller values of x. The number of surnames for large values of x increases when there is a high birth rate though the general trend is for smaller families. Accordingly, the number of new surnames will increase at a faster rate for smaller values of x. The addition of a surname can be considered as a birth and the deletion of a surname as a death. It may be noted that the birth rate of new surnames as well as the death rate of existing surnames will both decrease as x increases.

At time t ($t \geq 0$) let $N(t)$ denote the number of occurrences of given surnames in the system at time t. The birth and death process describes probabilistically how $N(t)$ changes as t increases. Though individual surnames are born and get eliminated randomly, their mean occurrence rates depend only on the current state of the system and on some parameters which vary from one area to another. The assumptions can be precisely stated as follows.

Assumption 1. Given $N(t) = n$, the probability distribution of the remaining time until the next birth of a surname is exponential with parameter λ_n ($n = 1, 2, 3, \ldots$) given by

$$\lambda_n = \frac{\theta}{(n+1)(n\beta - n)_{(n-1)}}, \quad 0 < \theta < 1, \quad \beta > 1, \tag{2.1}$$

which decreases with the increase in the value of n and where $a_{(n)} = \Gamma(a+n)\Gamma/(a)$. The state $n = 0$ is unobservable.

Assumption 2. Given $N(t) = n$, the probability distribution of the remaining time until the next death of a surname is exponential with parameter μ_n ($n = 2, 3, 4, \ldots$) given by

$$\mu_n = \frac{(1-\theta)^{1-\beta}}{(n\beta - n)_{(n-1)}}. \tag{2.2}$$

Assumption 3. Only one birth or death or none at all can occur during any small period of time.

On account of the relationship of the exponential density function with the Poisson distribution the quantities λ_n and μ_n become the mean rates of Poisson distributions and the rate diagram of surnames can be shown by Fig. 1.

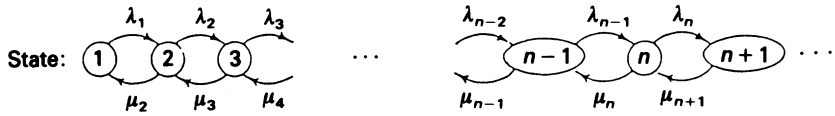

Figure 1. *Rate diagram for different states of the system.*

We assume that the process of the birth and the death of surnames has reached a steady state. Let P_n denote the probability that the system is in state n. In the steady state of the system for any state n ($n = 1, 2, 3, \ldots$) the mean rate (expected number of occurrences per unit time) at which the entering incidents occur must equal the mean rate at which the leaving incidents occur. This provides the principle of 'Rate in = Rate out' for each state.

Applying this principle for each state yields the balance equations for the birth and death process of surnames as shown in Table 1. The solutions of the balance equations are

$$P_2 = (\lambda_1/\mu_2)P_1, \quad P_3 = (\lambda_2/\mu_3)P_2 = (\lambda_1\lambda_2/\mu_2\mu_3)P_1, \ldots, \tag{2.3}$$

$$P_n = \frac{\lambda_1\lambda_2 \ldots \lambda_{n-1}}{\mu_2\mu_3 \ldots \mu_n} P_1. \tag{2.4}$$

Substitution of the values of λ's and μ's from (2.1) and (2.2) in (2.4) yields

$$P_n = \frac{\theta^{n-1}(1-\theta)^{n\beta-\beta-n+1}}{n!}(n\beta - n)_{(n-1)}P_1, \quad n = 2, 3, 4, \ldots. \tag{2.5}$$

Since the sum of all probabilities $P_1, P_2, \ldots, P_n, \ldots$ must be unity, the Lagrange expansion formula (Consul, 1990a),

$$1 = \sum_{n=1}^{\infty} \frac{\theta^{n-1}(1-\theta)^{n\beta-n}}{n!}(n\beta - n)_{(n-1)}, \tag{2.6}$$

gives $P_1 = (1-\theta)^{\beta-1}$.

Thus the probability distribution of the surnames becomes

$$P_n = \frac{\theta^{n-1}(1-\theta)^{n\beta-n}}{n!}(n\beta - n)_{(n-1)}, \quad n = 2, 3, \ldots, \tag{2.7}$$

which is the Geeta distribution (Consul, 1990a). The mean μ of the Geeta distribution is $\mu = (1-\theta)(1-\beta\theta)^{-1}$, $\beta\theta < 1$. Using this relation one can express the Geeta distribution in the form

$$P(X = x) = (x\beta - x)_{(x-1)}\left(\frac{\mu - 1}{\beta\mu - 1}\right)^{x-1}\left(\frac{\beta\mu - 1}{\beta\mu - 1}\right)^{\beta x - x}/x!, \quad (x = 1, 2, 3, \ldots), \tag{2.8}$$

where $\beta > 1$ and μ is the mean.

Table 1

Balance equations for each state

State	Balance equation
1	$\lambda_1 P_1 - \mu_2 P_2 = 0$
2	$(\lambda_1 P_1 - \mu_2 P_2) - \lambda_2 P_2 + \mu_3 P_3 = 0$
3	$(\lambda_2 P_2 - \mu_3 P_3) - \lambda_3 P_3 + \mu_4 P_4 = 0$
⋮	⋮
n	$(\lambda_{n-1} P_{n-1} - \mu_n P_n) - \lambda_n P_n + \mu_{n+1} P_{n+1} = 0$

3 Derivation of Surname Distribution as a Branching Process

The Galton–Watson branching process has been used to study the survival of the progeny of a mutant gene and for many ecological and biological problems. Consul & Shoukri (1988) used it to obtain the probability distribution of the number of infected individuals upto the generation when the epidemic dies out and when the process started with a random number of ancestors.

The population settlement in any particular area is generally started by a single individual (i.e. a single surname) who decides that it is a good location for his business. This individual (surname) forms the 0th generation of the branching process so that $X_0 = 1$. Then this individual persuades his wife and children and hires some other persons to assist him in his business and thus gets the first generation of surnames X_1, which is a random variable. Since all these new persons have come on account of their past associations with the single individual ($X_0 = 1$), it may not be unreasonable to assume that X_1 has a negative binomial distribution (a contagion model) whose p.g.f. is $g(t) = (1-\theta)^{\beta-1}(1-\theta t)^{1-\beta}$, $\beta > 1$. As the conditions get stabilized and the people become more optimistic about the future of the area and realize that there are more available opportunities, they give birth to children and they encourage their friends and relations to move into the area. Thus, each surname in X_1 produces the second generation X_2 of surnames under the same conditions as stated above and so on in successive generations of surnames. Our objective is to find the probability distribution of the number of surnames in this area.

Let $X_0 = 1, X_1, X_2, \ldots, X_n, \ldots$ denote the number of surnames in the 0th, 1st, 2nd, ... generations. One basic assumption here is that the probability distribution of the number of surnames generated by each individual surname remains unaltered over all generations. Let $g_n(s) = E[s^{X_n}]$. It is clear that $g_0(s) = s$ and $g_1(s) = g(s)$. Also, for $n = 2, 3, 4, \ldots,$

$$\begin{aligned} g_{n+1}(s) &= \sum_{k=0}^{\infty} P_r(X_{n+1} = k) \cdot s^k \\ &= \sum_{k=0}^{\infty} s^k \sum_{j=0}^{\infty} P_r(X_{n+1} = k \mid X_n = j) P_r(X_n = j) \\ &= \sum_{j=0}^{\infty} P_r(X_n = j)(g(s))^j = g_n(g(s)), \end{aligned} \tag{3.1}$$

and

$$g_2(s) = g_1(g(s)) = g(g(s)) = g(g_1(s)).$$

Similarly,

$$g_{n+1}(s) = g(g_n(s)). \tag{3.2}$$

Assume that the process of increase in surnames reaches a steady state (i.e. the growth stops) after the nth generation. Let $Y_n = X_0 + X_1 + X_2 + \ldots + X_n$ (where $X_0 = 1$) and let $G_n(s)$ be the p.g.f. of $Z_n = X_1 + X_2 + \ldots + X_n$. Then, $G_1(s) = g_1(s) = g(s)$. Let the p.g.f. of Y_n be $R_n(s)$. It can be easily shown that

$$R_n(s) = sG_n(s) = sg(R_{n-1}(s))$$

which gives the limiting form (as n increases)

$$R(s) = sg(R(s)). \tag{3.3}$$

Putting $R(s) = t$ in the above, we obtain the Lagrange transformation

$$t = sg(t). \tag{3.4}$$

Thus the Lagrange expansion of t, under the transformation (3.4) as a function of s can be easily obtained (Consul & Shenton, 1972) by putting $f(t) = t$ and by the formula

$$f(t) = t = f(0) + \sum_{x=1}^{\infty} \frac{s^x}{x!} \frac{d^{x-1}}{dt^{x-1}} [(g(t))^x f'(t)]_{t=0}. \tag{3.5}$$

The above gives the probability distribution of the total number of surnames (starting with one surname) as

$$P(Y = j) = \frac{(\beta j - j)_{(j-1)}}{j!} \theta^{j-1}(1-\theta)^{\beta j - j}, \quad (j = 1, 2, 3, \ldots), \tag{3.6}$$

which is the Geeta distribution (Consul, 1990a).

4 Comparison of Geeta Model with Yule Model and Discrete Pareto Model

All the three models (1.1), (1.2) and (3.6) are J-shaped, long-tailed and are defined over the domain $x = 1, 2, 3, \ldots$, and their bar diagrams look very similar for various values of their parameters. However, they are not identical and their parameters are such that the probability P_n for one model cannot be expressed as a function or limiting value of the corresponding probability for the other model. In view of the above the best way of showing the differences between them is to draw the graphs between their means μ and variances σ^2 for various values of their parameters.

Consul (1990a) has shown that the relation of variance σ^2 to the mean μ for the Geeta distribution is

$$\sigma^2 = \mu(\mu - 1)(\beta\mu - 1)/(\beta - 1), \tag{4.1}$$

which decreases monotonically with β. Thus the minimum value of σ^2 becomes $\mu^2(\mu - 1)$, which implies that for all values of μ and β the full domain for Geeta models is $\sigma^2 \geqslant \mu^2(\mu - 1)$. This covers more than 50% of the quadrant given by $\sigma^2 \geqslant 0$, $\mu \geqslant 1$ as indicated by the shaded area in Fig. 2.

Values of μ and σ^2 were computed for 20 values of c from 0·4 to 20 for the Yule distribution (1.2), applied by Panaretos (1989). Twelve of these points are (1·665, 4·39), (1·50, 2·16), (1·40, 1·297), (1·333, 0·888), (1·286, 0·660), (1·25, 0·521), (1·2, 0·36), (1·167, 0·271), (1·143, 0·218), (1·125, 0·181), (1·111, 0·154) and (1·050, 0·067) which are marked by small dots on Fig. 2. It is clear from these points that for all values of c, the points (μ, σ^2) for the Yule distribution are on a curve within the (μ, σ^2) domain of the Geeta distribution.

Since the value of σ^2 does not exist for $c \leqslant 2$ for the discrete Pareto distribution (1.1), the values of μ and σ^2 were computed for numerous values of $c \geqslant 2{\cdot}2$. The (μ, σ^2) points for $c = 2{\cdot}2$, 2·3, 2·4, 2·5, 2·6, 2·7, 2·8, 2·9, 3·0, 3·5, 4, 4·5 for the discrete Pareto model are

(1·029, 0·040), (1·044, 0·069), (1·068, 0·131), (1·111, 0·281), (1·123, 0·336),
(1·136, 0·408), (1·152, 0·498), (1·170, 0·617), (1·190, 0·777), (1·215, 0·988),
(1·243, 1·278), (1·276, 1·680), (1·316, 2·238).

The above thirteen points are marked in Fig. 2 by thick black dots. Their location clearly indicates that for various values of c in the discrete Pareto distribution the points (μ, σ^2) lie on a curve which is above the similar curve for the Yule distribution; however, this curve is also in the domain of (μ, σ^2) for Geeta distribution.

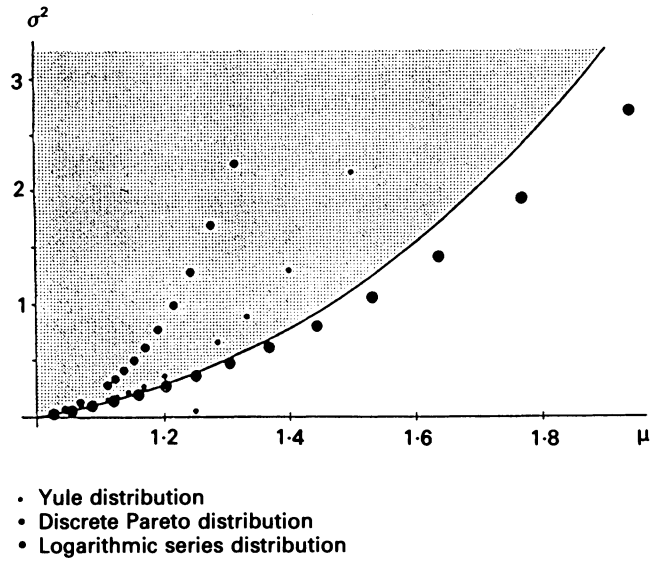

Figure 2. *(μ, σ^2) curves for discrete Pareto, Yule and Logarithmic Series Distributions and the (μ, σ^2) domain for Geeta distribution.*

Thus, for any given value of μ the discrete Pareto distribution has a larger variance than the Yule distribution but the Geeta distribution is such that it can have a larger variance as well as a smaller variance as needed by the conditions.

5 Fitting the Model to Actual Data

In this section the Geeta distribution, derived in §§ 2 and 3 and defined by (2.8), has been fitted to the observed frequency distributions of surnames (Fox & Lasker, 1983) in all nine non-overlapping districts by the maximum likelihood method. Here $f(x)$ denotes the number of surnames occurring x times.

The results are summarized in Table 2. For each value of x in every district the upper entry is the observed frequency and the lower entry is the expected frequency for the Geeta distribution. The last six rows of the table contain n, the total number of surnames, values of $\bar{x}$ and estimated values of β, value of chi-square goodness of fit statistic, degrees of freedom ν on which χ^2 is based and the p-value. The occurrences x for districts 1 and 9 were more than $x = 13$ and the values of $\bar{x}$ for those two districts are based on the actual values of x in the original data.

It is clear from Table 2 that the expected frequencies of Geeta distribution are very close to the observed frequencies and that the fits are quite reasonable. Even for all districts, where the number of classes is quite large and the degrees of freedom is 10, the p-value of 0·69 is very high. The least value of p is 0·16, for district 6, which is quite reasonable. These values of p have been obtained when the expected frequencies for χ^2 have been taken upto 2·5. Of course one reason for a better fit of the Geeta distribution may also be that it has two parameters. Thus, the Geeta distribution is a very plausible model for the distributions of surnames.

Though the p-value for district #1 is 0·31 there seems to be a substantial difference in the observed frequencies and expected frequencies. One reason for this is that the observed data are somewhat abnormal as there are three modes. The frequency drops from 11 to 2, then rises to 5 and then drops to 0 and then again rises to 2. Such abnormalities are

Table 2

Frequencies $f(x)$, observed (upper entries) and expected (lower entries), of the occurrences x of surnames in nine districts of Reading using Geeta distribution.

Occurrence	District								
x	5	4	6	3	8	7	2	1	all
1	234	243	281	292	282	349	329	832	1925
	234·03	243·01	280·98	292·09	282·02	348·95	329·34	823·78	1927·91
2	19	17	23	28	34	30	43	151	347
	18·51	17·49	23·79	27·07	34·51	30·47	41·33	142·65	327·89
3	5	4	9	6	11	7	11	39	140
	4·79	3·85	6·40	6·59	8·87	7·19	10·31	51·45	140·65
4	0	2	1	2	2	3	1	20	77
	1·61	1·10	2·24	2·06	2·87	2·18	3·21	23·32	77·28
5	1	0	0	0	0	1	0	11	50
	0·61	0·35	0·89	0·73	1·04	0·75	1·12	11·86	47·90
6	1	0	0	0	0	0	1	2	26
	0·25	0·12	0·38	0·28	0·40	0·28	0·42	6·42	31·91
7	0	0	0	1	0	0	0	4	19
	0·11	0·05	0·17	0·11	0·17	0·11	0·16	3·69	22·31
8	0	0	1	0	1	0	1	5	13
	0·05	0·02	0·08	0·05	0·07	0·04	0·07	2·18	16·14
9	0	0	0	0	0	0	0	0	9
	0·02	0·007	0·04	0·02	0·03	0·02	0·03	1·32	11·98
10	0	0	0	0	0	0	0	1	6
	0·01	0·003	0·02	0	0·01	0·01	0·01	0·82	9·08
11	0	0	0	0	0	0	0	0	4
	0·01	0	0·01	0	0·01	0	0·01	0·51	6·99
12	0	0	0	0	0	0	0	2	5
	0	0	0	0	0	0	0	0·30	5·45
⩾13	0	0	0	0	0	0	0	2	28
	0	0	0	0	0	0	0	0·66	23·51
n	260	266	315	329	330	390	386	1069	2649
$\bar{x}$	1·14615	1·11654	1·16190	1·15806	1·20909	1·14615	1·20725	1·43779	1·80974
$\hat{\beta}$	1·28237	1·32117	1·29757	1·44394	1·83204	1·41563	2·02681	1·85764	1·48919
χ^2	0·1903	0·0937	1·9568	0·1039	0·0755	1·1220	0·9269	4·7973	7·3706
d.f.	1	1	1	1	1	1	1	4	10
p	0·66	0·76	0·16	0·76	0·30	0·73	0·34	0·31	0·69

sometimes possible though in such cases it is very desirable that the observations may be repeated. The expected frequencies for the Yule distribution, given by Panaretos (1989), for district 1 do not seem to be correct as the value of $c = 3{\cdot}284$ is the same as for district 5 and accordingly the expected frequencies for district 1 should be proportional to the expected frequencies for district 5.

Table 3

Values of the X^2-test statistic for the Yule and Geeta models

Model		District								
		5	4	6	3	8	7	2	1	all
Yule	χ^2	1·35	1·35	3·45	0·35	1·51	1·25	1·00	6·85	12·64
	d.f.	1	1	2	1	2	1	2	5	11
Geeta	χ^2	0·19	0·09	1·96	0·10	1·08	0·12	0·93	4·80	7·37
	d.f.	1	1	1	1	1	1	1	1	10

It may not be out of place to mention that we cumulated the frequencies for each value of x for all the districts which provided good frequencies for all values of x. The Geeta distribution fitted this cumulative data extremely well with a p-value of 0·69.

The goodness of fit test statistic χ^2 values for the Yule and Geeta distributions are given together with the degrees of freedom in Table 3. It is clear from these values that the Geeta model provides a better fit than the Yule model. The χ^2 values for the Pareto model are much larger than those for the Yule model.

6 Some Remarks

We have provided two theoretical justifications in §§ 2 and 3 for the use of Geeta distribution as a surname distribution. Possibly, there can be some other justification as well. From the discussion in § 4 it seems that the Geeta model will not only fit all those sets of data which fit the discrete Pareto model and the Yule model but also many more sets of data which cannot be fitted by these two models.

Acknowledgement

I thank Ms. Karen Browne for providing the French translation of the Summary and Ms. Maria Dourado for typing the manuscript.

References

Consul, P.C. (1990a). Geeta distribution and its properties. *Comm. Statist.* A **19,** 3051–3068.
Consul, P.C. (1990b). Two stochastic models for Geeta distribution. *Comm. Statist.* A **19,** 3699–3706.
Consul, P.C. & Shenton, L.R. (1972). Use of Lagrange expansion for generating discrete generalized probability distributions. *SIAM J. Appl. Math.* **23,** 239–248.
Consul, P. C. & Shoukri, M. M. (1988). Some chance mechanisms related to a generalized Poisson probability model. *Am. J. Math. Manag. Sci.* **8,** 181–202.
Fox, W.R. & Lasker, G.W. (1983). The distribution of surname frequencies. *Int. Statist. Rev.* **51,** 81–87.
Panaretos, J. (1989). On the evolution of surnames. *Int. Statist. Rev.* **57,** 161–167.

Résumé

Plusieurs auteurs ont considéré le problème de déterminer un modèle de probabilité convenable pour décrire la distribution des noms propres dans des régions diverses. Dans cet article un modèle de processus de naissance et de mort ainsi qu'un modèle de processus de branchement sont proposés afin d'expliquer l'évolution des noms propres. Les deux modèles donnent lieu à la distribution de Geeta qui fut adaptée à des données réelles et l'adaption est comparée à celles de la distribution discrète de Pareto et de la distribution de Yule. Les trois distributions sont aussi comparées par rapport á leurs domaines d'espérance et de variance.

[*Received June* 1990, *accepted April* 1991]