

## A Probability Model Involving the Use of the Zero-Truncated Yule Distribution for Analysing Surname Data

J. PANARETOS

*University of Patras, Greece*

[Received 17 August 1988]

### 1. Introduction

IN AN attempt to interpret the mechanism that governs the occurrence of surnames, Fox & Lasker (1983) considered the discrete Pareto distribution (also known as the zeta distribution or Zipf's law). This distribution appeared in connection with some empirical data on the frequency of appearance of surnames in the area of Reading, UK (Lasker *et al.*, 1979). All of the observed distributions were characterized by a J-shape and very long upper tails. Moreover, as Fox & Lasker observed, the logarithm of the proportion of names with  $x$  occurrences was linearly related to the logarithm of  $x$  with a negative slope less than  $-1$ . They therefore concluded that the observed surname distribution could be approximated by a function of the form

$$x^{-(c+1)} / \sum x^{-(c+1)} \quad (x = 1, 2, \dots), \quad (1.1)$$

which is precisely the probability density function of the discrete Pareto distribution. However, as noted by these authors, as well as by Zei *et al.* (1983), it seems difficult to give a theoretical justification as to why the discrete Pareto distribution may provide a satisfactory representation of the mechanism that generated the observed distribution of surnames. Such a justification can probably be found in a biological context, the reason being that in most societies surnames are transmitted by the male line and hence can be thought of as behaving like genes. Moreover, the distribution of surnames can serve as a valuable source of information concerning the genetic structure of the population, for data on surnames are easier to collect than those on genes.

Within such a framework, this paper considers an alternative approximation of the true surname distribution provided by another J-shaped discrete distribution which can also be highly skewed with a long upper tail where it coincides with (1.1). This distribution is the zero-truncated version of Yule's (1924) distribution

with probability function

$$P(X = x) = \frac{cx!}{(c+2)_{(x)}} \quad (x = 1, 2, \dots; c > 0), \quad (1.2)$$

where  $a_{(b)} = \Gamma(a+b)/\Gamma(a)$  ( $a > 0, b \in \mathbb{R}$ ).

As  $x$  increases, it is obvious that  $x!/(c+2)_{(x)}$  becomes approximately proportional to  $x^{-(c+1)}$ . In fact, (1.2) is considered to be the discrete analogue of the continuous Pareto distribution (Kendall, 1961). In addition, (1.2) seems to be an appropriate choice because it can arise in the context of surname frequency analysis through a probability model described in Section 2. Finally, the truncated Yule distribution is shown to provide a satisfactory fit to the data of Lasker *et al.* (1979).

## 2. The probability model

Assuming that surnames (genes) are transmitted through the male line, let  $X$  be the number of males with a given surname from a pool of size  $\lambda$  ( $\lambda > 0$ ). We can reasonably assume that  $X$  has a Poisson distribution and let its parameter be  $\lambda\vartheta$  ( $\vartheta > 0$ ), where  $\vartheta$  is characteristic of the commonality of the particular name. (The commonality of names can be regarded as corresponding to the richness of a gene pool.) Assume further that the degree of commonality  $\vartheta$  varies from name to name according to an exponential distribution with probability density function

$$f(\vartheta) = e^{-\vartheta} \quad (\vartheta > 0). \quad (2.1)$$

(So surnames behave like alleles of  $Y$  chromosomes or genes.)

Then, for the given name (i.e. for given  $\lambda$ ), the probability of its occurrence is defined by the probability generating function (pgf)

$$\begin{aligned} G_\lambda(s) &= \int_0^\infty e^{\lambda\vartheta(s-1)} e^{-\vartheta} d\vartheta = \int_0^\infty e^{-\vartheta[1+\lambda(1-s)]} d\vartheta \\ &= [1 + \lambda(1-s)]^{-1}. \end{aligned} \quad (2.2)$$

Let us now suppose that the pool size varies from name to name according to a distribution with probability density function given by

$$c(1+\lambda)^{-(c+1)} \quad (\lambda > 0, c > 0). \quad (2.3)$$

Therefore, the resulting surname distribution will have pgf

$$G(s) = c \int_0^\infty (1+\lambda)^{-(c+1)} [1 + \lambda(1-s)]^{-1} d\lambda = \left(\frac{c}{c+1}\right) \sum_{r=0}^\infty \frac{r!}{(c+2)_{(r)}}.$$

This implies that the number  $Y$  of occurrences of a surname has a distribution with probability function

$$P(Y = y) = \frac{cr!}{(c+1)_{(r+1)}} \quad (r = 0, 1, 2, \dots).$$

The above form leads to (1.2) in the case of unobserved-zero-frequency data.

It becomes obvious from the above hypotheses that the distribution of surname occurrences is considered to be a Poisson mixture. This is not an unreasonable requirement as one would not expect the frequency of surnames to be governed solely by pure chance. As to the hypothesized specific forms of the distributions of  $\lambda$  and  $\partial$ , they may at first appear to be not a well-justified choice, thus making the model seem to be not adequately explanatory. However, there is some supporting logic that can be found in the following results.

**THEOREM 2.1** (Bondesson, 1979) *Let  $\{N(t), t \geq 0\}$  be a homogeneous Poisson process with parameter 1 and let  $Y$  be a non-negative random variable independent of  $\{N(t), t \geq 0\}$ . Then the distribution of  $N(Y)$  is a generalized negative binomial convolution if and only if the distribution of  $Y$  is a generalized gamma convolution, i.e. if and only if the probability density function of  $Y$  is of the form*

$$f_Y(y) = \int_0^\infty \frac{\lambda^\beta}{\Gamma(\beta)} y^{\beta-1} e^{-\lambda y} dF(\lambda) \quad (y > 0, \beta > 0), \quad (2.4)$$

where  $F$  is a proper distribution function.

**THEOREM 2.2** (Xekalaki & Panaretos, 1988) *Let  $\{N(t), t \geq 0\}$  and  $Y$  be defined as in Theorem 2.1. Suppose that the probability density function of  $Y$  is of the form (2.4). Then the distribution of  $N(Y)$  is the Yule distribution if and only if  $F$  is the distribution function of the Pareto distribution.*

The above theorems combined with the fact that the Yule distribution belongs to the family of generalized negative binomial convolutions (being a mixture on  $p$  of the geometric distribution) provide the needed theoretical justification. Indeed, they lead to the conclusion that a satisfactory fit of the observed data by the Yule distribution necessarily implies that the distribution of  $\lambda\partial$  is of the form (2.4) for  $\beta = 1$  (Theorem 2.1), where  $F$  should be defined by the probability density function in (2.3) (Theorem 2.2). This in turn implies that the distribution of  $\partial$  is as defined by (2.1).

### 3. Interpreting actual surname data

Having postulated a mechanism that describes the occurrence of surnames, one would be interested in examining how well the resulting distribution fits actual data. For this purpose, the Yule distribution as defined by (1.2) was fitted to the surname data of Lasker *et al.* (1979) with reference to their division into eight districts. The results are summarized in Table 1. The upper entries of the table refer to actual observed frequencies, while lower entries are the expected Yule frequencies.

The fit is quite satisfactory in seven of the eight districts. Inspecting Table 1, one can see that the description of the data provided by the discrete Pareto distribution of Fox & Lasker (1983) has the same degree of adequacy as that of the description provided by the truncated Yule distribution. District 1 seems to be an exception: the fit of (1.2) is poor, possibly owing to the fact that the corresponding observed frequency of surnames behaves somewhat erratically in the tail.

TABLE 1  
*Observed and expected frequencies of surnames using the zero-truncated Yule distribution*

<i>x</i>	District							
	5	4	6	3	8	7	2	1
1	234	243	281	292	282	349	329	832
	231.295	239.613	273.966	286.965	275.188	343.249	324.392	786.508
2	19	17	23	28	34	30	43	151
	23.485	22.646	33.507	34.462	42.203	38.819	47.948	183.584
3	5	4	9	6	11	7	11	39
	3.404	3.066	5.793	5.856	9.017	6.233	9.899	57.559
>4	1	2	2	3	3	4	3	33†
	0.815	0.675	1.735	1.716	3.592	1.698	3.76	41.347
<i>c</i>	16.697	18.161	13.353	13.654	10.041	14.684	10.531	5.568
$\chi^2$	1.639	2.819	5.291	2.265	2.296	5.316	0.852	18.217
df	1	1	2	2	2	2	2	4

† 20 at 4, 11 at 5, 2 at 6, 4 at 7, 5 at 8, 1 at 10, 2 at 12, 1 at 13, and 1 at 24.

Therefore the fit provided by the Yule distribution does not show any appreciable difference from that provided by the discrete Pareto distribution. This can be thought of as reflecting the fact that, in the context of the problem considered in this paper, the discrete Pareto distribution represents an excellent approximation to the Yule distribution, while the latter seems to be theoretically more satisfactory than the former. The Yule distribution is of direct significance for theoretical interpretation, since it has been derived from a specific, reasonably realistic, evolutionary model, so that it can lead to information useful from a genetic point of view.

#### REFERENCES

- BONDESSON, L. 1979 On generalized gamma and generalized negative binomial convolutions. *Scand. Actuarial J.*, 125–46 (Part I), 147–66 (Part II).
- FOX, W. R., & LASKER, G. W. 1983 The distribution of surname frequencies. *Int. Statist. Rev.* **51**, 81–87.
- KENDALL, M. G. 1961 Natural law in the social sciences. *J. R. Statist. Soc. A* **124**, 1–16.
- 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**, 445–60.
- XEKALAKI, E., & PANARETOS, J. 1988 On the association of the Pareto and the Yule distributions. *Teor. Ver. Prim.* **33**, 206–10.
- 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.
- ZEI, B., MATESSI, R. G., SIRI, E., MORONI, A., & CAVILLI-SFORZA, L. 1983 Surnames in Sardinia. I. Fit of frequency distributions for neutral alleles and genetic population structure. *Ann. Human Genet.* **47**, 329–52.