International Journal of Pure and Applied Mathematics

Volume 42 No. 3 2008, 353-359

A THEORY OF PROBABILITY DISTRIBUTIONS

Gabriele Stoppa

Faculty of Economics
University of Trento
5, Via Vigilio Inama, Trento, I-38100, Italy
e-mail: gabriele.stoppa@economia.unitn.it

Abstract: The aim of this paper is to develope a new system of distributions based on *elasticity* function. A huge variety of size models, including the best known continous univariate one supported on positive half line, are placed in a very natural order. The system here suggested contains 1940 and 4943 families with no more than four or five parameters respectively.

AMS Subject Classification: 60E05, 62E10
Key Words: generating system, linear regression, non parametric estimation, empiric distribution, generalized least squares and robust stepwise regression method

1. Introduction

The continuous improvement in information systems and data collection techniques allows the automatic generation of sizeable data-based collection, calling for sophisticated statistical investigations, aimed both at initially uncovering structure and proposing models in different application contexts. It is of particular interest to have, apart from classical generating systems (Pearson [10], Burr [1], Johnson [5], D'Addario [2] and Dagum [3]), a classification system that yields, starting from a few basic principles, a framework for models that could be useful not only for the main estimators, but also for the modelling of size phenomena (income, assets, actuarial loss, etc.). The classical approach upon the problem of deciding probabilities and frequencies has been made through the use of the density function **f** (Pearson [10]) or through the distribution function (Burr [1]). The *elasticity* function with respect to **F**, $\boldsymbol{\epsilon_F}$, would seems to be theoretical much better qualified for this problem. The *elasticity* of **F**,

Received: August 17, 2007

defined as $\frac{xdF}{FdX}$, gives the dynamic of relative expected number of observations less than a given value. Hence the expected relative frequencies in any given range are found simply by multiply this *elasticity* with the relative range. The aim of this paper is to suggest the direct use of the *elasticity* so as to use this theoretical advantage. The problem of fitting the elasticity is considered and new *elasticity* functions possessing practical aspects, useful to a large variety of observed fenomena, will be discussed.

2. The Method

It is possible to use the advantage of the *elasticity* i) finding suitable *elasticities* and ii) suggesting a method of fitting $\boldsymbol{\epsilon_F}$ directly. These points are now discussed. The method is based on following observations: 1) $\boldsymbol{\epsilon_F}$ is more informative than $\mathbf{F}$ because, by definition, it describes the dynamicity of $\mathbf{F}$; 2) The expressions of $\boldsymbol{\epsilon_F}$ may be very complex, but they may be easier if they are considered as functions of suitable *filters* (operators, transformators); 3) $\boldsymbol{\epsilon_F}$, equalized to a set of operators, $\boldsymbol{O_i, 0 \leq i \leq h}$, may be written as $\boldsymbol{g(O_0, ..., O_h) = 0}$; 4) The *filters* here considered are: $\boldsymbol{O_0 = e}$, $\boldsymbol{O_1 = F}$, $\boldsymbol{O_2 = 1 - F}$, $\boldsymbol{O_3 = x}$, $\boldsymbol{O_4 = x^{\log x}}$, $\boldsymbol{O_5 = \log x}$, $\boldsymbol{O_6 = e^F}$, $\boldsymbol{O_7 = e^x}$, $\boldsymbol{O_8 = e^{x^2}}$, $\boldsymbol{O_9 = (1 - x)}$, $\boldsymbol{O_{10} = (1 + x)}$, $\boldsymbol{O_{11} = (1 - x^2)}$, $\boldsymbol{O_{12} = (1 + x^2)}$, (the last four have limited support contest), $\boldsymbol{O_{13} = (e^x - 1)}$ and $\boldsymbol{O_{14} = (e^x + 1)}$ for quantities measured with great precision; 5) The sequence of *filters* appears in $\boldsymbol{g}$ with exponential coefficients $\boldsymbol{b_0}$, $\boldsymbol{b_1}$,...,$\boldsymbol{b_{14}}$ (see Appendix). The *filter* order is convenient but is undeniable a logic priority of $\mathbf{F}$ with respect to $\mathbf{x}$ because $\boldsymbol{\epsilon_F}$, by definition is made respect to $\mathbf{F}$. It is also reasonable to assign priority to with respect to $\boldsymbol{e^F}$ and to $\mathbf{x}$ with respect to $\boldsymbol{e^x}$. A huge variety of *elasticity*, supported on positive half-line, has been placed in an order that develops in a very natural way. Having understood what *filters* are in play, it is interesting and instructive to reverse the approach, that is: i) fixing $\boldsymbol{\epsilon_F}$ as function of above *filters* or of all subset of them (see Appendix, a corresponding table of distributions, quoted in Kotz and Kleiber [6], p. 53, is in Stoppa [11]); ii) discovering that $\boldsymbol{\epsilon_F}$ is, respect to above *filters*, of multiplicative type and it is very useful from the estimation point of view; iii) using, instead of generalized least squares (Hocking [4]; Thompson [13]), the robust stepwise regression method (Stoppa [12]) we give non-parametric (suggested directly from the data) *elasticity* estimate, using the empiric versions (sample counterparts) of *filters* and of $\boldsymbol{\epsilon_F}$.

Same Family. In the same family we have distributions with $\boldsymbol{\epsilon_F}$ involving the same *filters*; so we may use the term of maximum degree of closeness. The mathematical and statistical properties of the distributions of the same family have strong analogies and the same interpretative structure. The first family of the system has $\boldsymbol{O_0}$ *filter*, in the sense that $\boldsymbol{\epsilon_F}$ is constant, and so hold one coefficient alone: $\boldsymbol{b_0}$. Than we have families with couple of *filters*, and so on. The system here suggested is extremely rich because, for example with $h = 5$, it contains $\sum_{j=0}^{5} \binom{15}{j} = 4943$ families. In terms of known distributions the following families are very important examples ("G-" stands for "generalized"):

system number	**present name**	**system number**	**present name**
16	G-Log-Power	250	G-Euler
17	G-Log-Logistic	251	G-Normal
18	G-Frechet	252	G-Beta
43	G-Log-Pareto	253	G-Beta II
44	G-Pareto	254	G-Pearson
45	G-Log-Normal	255	G-Chauchy
68	G-Logistic	256	G-Bose-Einstain
224	G-Weibull	257	G-Fermi-Dirac

Same Period. We say that a set of families forms a period when the *elasticity* has a set of configurations with the same number of *filters* but of different nature and so the families are characterized by distinct properties. The first period (stratum) numbered from 1 to 15, includes families having an $\boldsymbol{\epsilon_F}$ depending on a *filter* alone. It is a set of elementary families, in the sense that the families have an *elasticity* of simple structure with only one operator. The number of parameters determines the position and the step with respect to another model and indicates the degree of generalization of the $\boldsymbol{\epsilon_F}$. Composing groups and periods we have the system, holding an increasing sequence. From one period to another we expect families with very different properties. Appendix includes an outline of the *elasticities* of the first 258 families of the system.

3. Estimation

The exploratory data analysis is concerned with the extraction from the observations of all informations that may facilitate the determination of a statistical

procedure suitable to the analysis of data themselves; in general the latter depends on a random variable **X** that follows an unknown distribution **F**. A non parametric estimate of the process, that generates the data themselves, appears to be the first-choice candidate for an analysis of this kind. Let $\boldsymbol{X_1, ..., X_n}$ be a set of random variables distributed like the random variable **X** defined on positive half line of unknown shape **F** and suppose we want to estimate the process $\boldsymbol{\epsilon_F}$ that can generate **F**. It should be remember that, under the above-mentioned conditions, the maximum likelihood method is not applicable. Moreover, the question can be faced from a general point of view if one undertakes to figure **F** through the estimate of *elasticity* $\boldsymbol{\epsilon_F}$ that generates it. Among others, this would allow also the consideration of those distributions that do not have an explicit form for **F**. For reasons of estimation we want to construct a regression equation:

$$\widehat{\log \epsilon_F} = b_0 + b_1 \log O_1 + ... + b_{14} \, , \log O_{14} \, , \tag{1}$$

where the estimated predicted value is $\boldsymbol{\widehat{\log \epsilon_F}}$ and the regressors are $\boldsymbol{\log O_i}$, $\boldsymbol{0 \leq i \leq 14}$ and where we use the sample distribution, $\boldsymbol{F_n}$, in the empiric versions of *filters* of $\boldsymbol{\epsilon_F}$. There are many useful diagnostic tools on regression literature. Then we choose a smallest set predictor *filters* by a robust stepwise regression procedure and $\boldsymbol{R^2}$ criterion. The set of predictor *filters* choosed indicates the family involved.

4. An Application

Analysing data on duration of UK strikes, broken down by industry and causes of dispute, while Lancaster [7] suggests inverse Gaussian distribution, Lawrence [8] uses Lognormal (system number 45) and Morrison and Schmittlein [9] indicate Weibull one (system number 43), applying the non parametric approach referred above we found Fréchet distribution (system number 18).

5. Conclusive Remarks

The problem of closeness between probability distributions is here considered with respect to the composition of *elasticity*. Two families can be defined close to each other when their *elasticity* depends on the same number of *filters* and on the same type. The closeness will be less when depends on a different number of *filters*. From this point of view we may say that the distributions logistic

and inverse-logistic have similar characteristics (system number 56); they have an $\boldsymbol{\epsilon_F}$ with direct or inverse proportionality with respect to $\mathbf{x}$. The system proposed has: 15 families with a number of parameters ≤ 1, 120 with a number of parameters ≤ 2, 575 ≤ 3, 1940 with ≤ 4 and 4943 with ≤ 5. At the moment, many families do not have explicit form with respect to $\mathbf{x}$: for example 7, 21, 46, exc., but these families have also a known and stimable $\boldsymbol{\epsilon_F}$. In 44 we found Pareto, Inverse Pareto, Dagum and Weibul-Exponential distribution; 68 holds Logistic and Inverse Logistic models; 224 contains Weibull, Inverse Weibull, Exponential, Gordon-Normal, IV-Pearson, VI-Pearson, II-Pareto, IV-Pareto and XII-Burr; in 250 there are Eulero, Generalized Exponential (G-Exponential), Chi-Square, Herlag, Helley and Richard; in 251 there are Normal, Maxwel-Boltzman, Ferreri, X-Burr, Generalized Gamma, Chi, Helmert and 342-Kendall; in 252 we found Beta I, Inverse-Beta I, VI-Burr, Bradford, Simpson and 282-Kendall; 253 holds Beta II, VIII-Pearson and IX-Pearson; 254 counts II-Pearson, Lambert, Lagrange, 21-Kendall, 27-Kendall and 49-Kendall; 255 have Chauchy, VII-Pearson, T-Student and Agnesi distribution; 256 have Bose-Einstain and Plank; 257 coniugates the Fermi-Dirac model. The estimation question is here faced from a general point of view undertaking to figure $\mathbf{x}$ through the estimate of *elasticity* $\boldsymbol{\epsilon_F}$ that generates it. Among others, this would allow also the consideration of those distributions that do not have an explicit form for $\mathbf{F}$. Generalized Least Squares and Robust Stepwise regression estimators of the *elasticity* have explicit form and are efficient in the sense of Markov. It is possible to extend the sequence of operators using special *filters* and it is also possible to obtain *elasticity* estimations, generally with procedures of iterative type. The most promising additional *filters* seem to be: $\mathbf{\log F}$, $\mathbf{\log (1 - F)}$.

References

[1] I. Burr, Cumulative frequency functions, *Annals of Mathematical Statistics* (1942), 215-32.

[2] R. DÁddario, Ricerche sulla curva dei redditi, *Giornale degli Economisti e Annali di Economia* (1949), 91-114.

[3] C. Dagum, A new model of personal income distribution: specification and estimation, *Economia Applicata* (1977), 413-37.

[4] R.R. Hocking, The analysis and selection, *Biometrics* (1976), 32, 1-40.

[5] N.I. Johnson, System of frequency curves, *Biometrika* (1949), 149-76.

[6] S. Kotz, C. Kleiber, *Statistical Size Distributions in Economic and Actuarial Sciences*, Wiley, New York (2003).

[7] T. Lancaster, A stochastic model for the duration of a strikes, *Journal of the Royal Statistical Society, A*, **135** (1972), 257-71.

[8] R.J. Lawrence, The lognormal distribution of duration of strikes, *Journal of the Royal Statistical Society, A*, **147**, No. 3 (1984), 464-83.

[9] D.G. Morrison, D.C. Schmittlein, Jobs, strikes and wars: Probability model for duration, *Org. Behav. and Hum. Performance*, **25** (1980), 224-251.

[10] K. Pearson, *Early Statistical Papers*, Cambridge University Press (1948).

[11] G. Stoppa, Una tavola per modelli di probabilita, *Metron*, **LI**, No. 3-4 (1993), 99-117.

[12] G. Stoppa, Scelta robusta delle variabili per via grafica, *Statistica Applicata*, **9**, No. 2 (1997), 247-64.

[13] M.L. Thompson, Selection of variables, *International Stastistical Review*, **46**, (1978), 129-46.

Appendix

Number	Name	Number	Name
1	e^{b_0}	148...	$e^{b_0}e^{b_6 F}e^{b_7 x}$
2	F^{b_1}	...155	$e^{b_0}e^{b_6 F}(e^x+1)^{b_{14}}$
3	$(1-F)^{b_2}$	156...	$e^{b_7 x}e^{b_8 x^2}$
4	x^{b_3}	...162	$e^{b_7 x}(e^x+1)^{b_{14}}$
5	$x^{b_4 \log x}$	163...	$e^{b_0}e^{b_7 x}e^{b_8 x^2}$
6	$(\log x)^{b_5}$	...169	$e^{b_0}e^{b_7 x}(e^x+1)^{b_{14}}$
7	$e^{b_6 F}$	170...	$e^{b_8 x^2}(1-x)^{b_9}$
8	$e^{b_7 x}$	...175	$e^{b_8 x^2}(e^x+1)^{b_{14}}$
9	$e^{b_8 x^2}$	176...	$e^{b_0 e^{b_8 x^2}}(1-x)^{b_9}$
10	$(1-x)^{b_9}$	...181	$e^{b_0}e^{b_8 x^2}(1+x)^{b_{14}}$
11	$(1+x)^{b_{10}}$	182...	$(1-x)^{b_9}(1+x)^{b_{10}}$
12	$(1-x^2)^{b_{11}}$	...186	$(1-x)^{b_9}(e^x+1)^{b_{14}}$
13	$(1+x^2)^{b_{12}}$	187...	$e^{b_0}(1-x)^{b_9}(1+x)^{b_{10}}$
14	$(e^x-1)^{b_{13}}$	...191	$e^{b_0}(1-x)^{b_9}(e^x+1)^{b_{14}}$
15	$(e^x+1)^{b_{14}}$	192...	$(1+x)^{b_{10}}(1-x^2)^{b_{11}}$
16...	$e^{b_0}F^{b_1}$	...195	$(1+x)^{b_{10}}(e^x+1)^{b_{14}}$
...29	$e^{b_0}(e^x+1)^{b_{14}}$	196...	$e^{b_0}(1+x)^{b_{10}}(1-x^2)^{b_{11}}$
30...	$F^{b_1}(1-F)^{b_2}$	...199	$e^{b_0}(1+x)^{b_{10}}(e^x+1)^{b_{14}}$
...42	$F^{b_1}(e^x+1)^{b_{14}}$	200...	$(1-x^2)^{b_{11}}(1+x^2)^{b_{12}}$
43...	$e^{b_0}F^{b_1}(1-F)^{b_2}$	...202	$(1-x^2)^{b_{11}}(e^x+1)^{b_{14}}$
...55	$e^{b_0}F^{b_1}(e^x+1)^{b_{14}}$	203...	$e^{b_0}(1-x^2)^{b_{11}}(1+x^2)^{b_{12}}$
56...	$(1-F)^{b_2}x^{b_3}$	...205	$e^{b_0}(1-x^2)^{b_{11}}(e^x+1)^{b_{14}}$
...67	$(1-F)^{b_2}(e^x+1))^{b_{14}}$	206	$(1+x^2)^{b_{12}}(e^x-1)^{b_{13}}$
68...	$e^{b_0}(1-F)^{b_2}x^{b_3}$	207	$(1+x^2)^{b_{12}}(e^x+1)^{b_{14}}$
...79	$e^{b_0}(1-F)^{b_2}(e^x+1)^{b_{14}}$	208	$e^{b_0}(1+x^2)^{b_{12}}(e^x-1)^{b_{13}}$
80...	$x^{b_3}x^{b_4 \log x}$	209	$e^{b_0}(1+x^2)^{b_{12}}(e^x+1)^{b_{14}}$
...90	$x^{b_3}(e^x+1)^{b_{14}}$	210	$(e^x-1)^{b_{13}}(e^x+1)^{b_{14}}$
91...	$e^{b_0}x^{b_3}x^{b_4 \log x}$	211	$e^{b_0}(e^x-1)^{b_{13}}(e^x+1)^{b_{14}}$
...101	$e^{b_0}x^{b_3}(e^{x+1})^{b_{14}}$	212...	$F^{b_1}(1-F)^{b_2}x^{b_3}$
102...	$x^{b_4 \log x}(\log x)^{b_5}$	...223	$F^{b_1}(1-F)^{b_2}(e^x+1)^{b_{14}}$
...111	$x^{b_4 \log x}(e^x+1)^{b_{14}}$	224...	$e^{b_0}F^{b_1}(1-F)^{b_2}x^{b_3}$
112...	$e^{b_0}(x^{b_4 \log x})(\log x)^{b_5}$	...235	$e^{b_0}F^{b_1}(1-F)^{b_2}(e^x+1)^{b_{14}}$
...121	$e^{b_0}x^{b_4 \log x}(e^x+1)^{b_{14}}$	236...	$F^{b_1}x^{b_3}x^{b_4 \log x}$
122...	$(\log x)^{b_5}e^{b_6 F}$	...246	$F^{b_1}x^{b_3}(e^x+1)^{b_{14}}$
...130	$(\log x)^{b_5}(e^x+1)^{b_{14}}$	247...	$e^{b_0}F^{b_1}x^{b_3}x^{b_4 \log x}$
131...	$e^{b_0}(\log x)^{b_5}e^{b_6 F}$	...257	$e^{b_0}F^{b_1}x^{b_3}x^{b_4 \log x}$
...139	$e^{b_0}(\log x)^{b_5}(e^x+1)^{b_{14}}$	258...	$F^{b_1}x^{b_3}(\log x)^{b_5}$
140...	$e^{b_6 F}e^{b_7 x}$	...	...
...147	$e^{b_6 F}(e^x+1)^{b_{14}}$	...	...