Pergamon

Information Processing & Management, Vol. 31, No. 1, pp. 133–137, 1995

0306-4573/95 $9.50 + .00

0306-4573(94)00036-0

BRIEF COMMUNICATION

A MODIFICATION OF LOTKA'S FUNCTION FOR SCIENTIFIC PRODUCTIVITY

P. H. FANG
F. S. Lab, 156 Common Street, Belmont, MA 02178, U.S.A.

and

JOHN M. FANG
Evaluation and Analysis Laboratory, Hughes Aircraft Co.,
P. O. Box 902, El Segundo, CA 90245, U.S.A.

(*Received* 1 *March* 1993; *accepted in final form* 23 *May* 1994)

Abstract—In this paper, we have examined the validity of Lotka's function for a measurement of scientific productivity. A systematic deviation of this function from numerical examples has been found by introducing a different least square formulation instead of a logarithmic linearization. Therefore, a modification on the measurement of the publication frequencies is proposed.

I. INTRODUCTION

In 1926, Lotka published a canonical relation between the number of authors and their productivity measured by the number of publications [1]. This relation, referred to as Lotka's law, has been extensively investigated formalistically [2]. In its application, there is an analytical difficulty, which is traced to an adaptation of a logarithmic linearization procedure. Therefore, in II, a different procedure of a least square minimization is presented, which is free from the mathematical deficiency. The extended result obtained from this consistent analysis manifests a systematic deviation from Lotka's law and, in Part III, a modification is proposed by introducing a *virtual* number of publication frequency through which the deviation is corrected.

II. A DIRECT MINIMIZATION FOR A SOLUTION OF LOTKA'S PARAMETERS

Following Lotka's original notations,

$$x^n y = c, \tag{1}$$

where y is the frequency of persons making x publications, n and c are constant parameters. This is Lotka's law. In eqn (1), x runs numerically from 1 to an arbitrary large number, for example, 346 in one of Lotka's examples. In the case of y, there are some large numbers, but also zeroes for certain x, especially in the regions of large x. The objective of the numerical analysis of statistical data is to find the value of the parameters n and c.

A conventional approach to find n and c is by a least square procedure to minimize the following expression, which is a logarithmic representation of eqn (1):

$$n \ln x + \ln y = \ln c. \tag{2}$$

Since eqn (2) is a linear equation of n and b, a simple linear least square method can be applied to find the values of n and c. However, this procedure has two defects: (1) A frequency occurring number is $y = 1$. Since $\ln 1 = 0$, this would have a zero statistical weight; and (2) for $y = 0$, $\ln y$ would become a negative infinity and cannot be accommodated in a finite analysis. The $y = 0$ case has been ignored in a vague name of data fluctuations [1]. This practice is not objective and also not justifiable, because the very nature of the statistical analysis is its ability to accommodate the fluctual occurrence of the data. In one publication, $\ln y$ for $y = 0$ is treated as zero [3]. This is simply an elementary mathematical error. In still another paper [4], results are given for data including $y = 0$, for example, in Auerbach data [1] for $N = 19$, 20, 23, and 26. These results obviously cannot be obtained from eqn (2) of that reference. We conclude that the least square of the logarithmic representation of Lotka's law, simple in its appearance, is not applicable in practice.

As far as we know, the pathological consequence of a logarithmic linearization has not been treated in the literature. Neither is the origin of this linearization known, but judging by the work, Lotka may well have used this procedure himself. In fact, as will be demonstrated in the present paper, we can solve the problem directly without a linearization.

The problem is to minimize the expression,

$$\sum_{i=1}^{N} (x_i^n y_i - c)^2, \tag{3}$$

where i is the index of (x_i, y_i) pairs and N is the largest number of publications. The minimization is reached by setting the first partial derivative of expression (3) with respect to n and c to zero. The results are,

$$\Sigma x^{2n} y^2 \ln x - c \Sigma x^n y \ln x = 0, \tag{4}$$

$$\Sigma x^n y - Nc = 0. \tag{5}$$

In the above, and henceforth, the cumbersome indices' notations are omitted; thus,

$$\sum x^n y \ln x \text{ represents } \sum_{i=1}^{N} x_i^n y_i \ln x_i.$$

A consequence of this formulation is that due to an absence of $\ln y$, the terms in eqns (4) and (5) are all finite and, therefore, one can analyze general data to including $y = 0$.

From the system of eqns (4) and (5), n and c can be solved. This we have done and is illustrated in Table 1. The data are from the Auerbach Table in Lotka's paper [1].

A strong deviation of n is manifested in Table 1. Furthermore, this deviation is not a random, but a systematic deviation: When the number of data points N is increased, from $N = 17$ to 48, n increases from 0.91 to 3.22, $N = 48$ being the largest data point. In this whole range, n increases monotonically and an existence of a limiting or a stable value is not evident.

The N dependence of n has been pointed out already by Pao [4] based on various N, up to $N = 27$. By extending to $N = 48$ in our work, the variation is even more exagger-

Table 1. Values of the exponents for Lotka's data (chemical abstracts)

N	n	N	n
17	0.91	25	1.95
18	1.11	30	2.33
19	1.27	35	2.64
20	1.42	40	2.89
		48	3.22

ated. In view of this N-dependency in the values of n, it seems superfluous to carry out an elaborated Komogorov-Smirnov analysis, as applied in [3] and discussed in [4] and [5], based on a single n value.

III. A MODIFICATION OF LOTKA'S FUNCTION

We propose the following modified Lotka's function:

$$x^n(y + u) = c, \tag{6}$$

where the additional parameter u has the following interpretations:

1. Although the values of y derived from formal publications run from 0 to large numbers, in fact, we suggest y is not the only measurement of the productivity of an author. There are many other reasons that y does not have zero in reality. Some examples are (a) the author also has unpublished papers or contributions which have only a limited circulation, but their importance is substantial, (b) due to circumstances, the results are not published, but these results form the basis for future development of contributions, (c) the editors or referees are not immune from a misjudgment and the manuscripts are rejected, (d) an idea could constitute part of an incubative process for later formal publications, etc. It would not be simple to trace these cases and we denote all these factors by a collective parameter u. Conceptually, u would be positive and not necessarily an integer.

2. Statistically, u can be described as a fuzzy number. Utilistically, u has a mathematical consequence in that now the term involving $y = 0$ is no longer singular, nor $y = 1$ gives a zero contribution. This effect alone should be a justification for the introduction of u in order to remedy Lotka's function.

3. When writing eqn (6) in a different form with $1/n = m$,

$$x(y + u)^m = c', \tag{7}$$

where $\ln c' = m \ln c$. Equation (7) now formalistically becomes Mandelbrot function [5] or generalized Zipf function [6]; however, our formulation is in a different context.

With these credentials for u, we can take advantage of the simplicity obtained through a logarithmic linearization in the numerical analysis, without the menace of a logarithmic divergence. The system of simultaneous equations obtained through the least square procedure are,

$$\sum (\ln x)^2 \cdot n + \sum \ln(y + u) \ln x - \sum \ln x \cdot b = 0, \tag{8}$$

$$\sum \ln xn + \sum \ln (y + u) - N \cdot b = 0, \tag{9}$$

$$\sum \ln x \ln(y + u) \cdot n + \sum \ln(y + u)/(y + u) - \sum 1/(y + u) \cdot b = 0, \tag{10}$$

Table 2. Analysis of Lotka's data (Auerbach) according to eqn (6)

N	u	n	b
5	42	1.491	6.666
10	12.4	1.613	6.578
15	5.32	1.756	6.558
17	3.04	1.676	6.526
18	3.49	1.731	6.541
19	4.31	1.685	6.509
25	5.96	1.438	6.206
30	4.67	1.437	6.136
35	4.08	1.405	6.039
40	4.29	1.310	5.837

Table 3. The u dependence of the results of Table 2

u	L	n	b
0.1	−24	1.92	8.35
1	−16	1.88	8.31
2	−10	1.85	8.27
5	−.06	1.76	8.17
5.03	−.001	1.76	8.17
5.04	+.019	1.76	8.17
7	+3.0	1.72	8.12
10	+5.2	1.65	8.05

where $b = \ln c'$. In the calculation, we solve n and b from eqns (8) and (9) first in terms of u and known values of (x,y) pairs, and substituting into eqn (10) to solve for u. The value of u that gives a value closest to zero on the left-hand side of eqn (10) is accepted as the solution. Based on this u value, n and b are determined.

We now reanalyze the data used in Lotka's classical work. The objective of this calculation is primarily an illustration of the methodology so that this approach can be adapted by other researchers. The calculation is made on a PC computer with software [8]. The same calculation would not be practically adaptable at Lotka's time, but can be a pedagogical routine today. Once the program is made, calculation can be completed in a matter of seconds. The results are shown in Table 2 for Auerbach data.

Some significant results are listed below:

1. While there is a considerable variation in the value of u, the values of n and b are fairly constant, around 1.4 to 1.7 for n and around 6 for b. This is a considerable stability improvement from the result of Table 1.

2. Except for $u = 0$, which is tantamount to a return to the deficient Lotka's law, the values of n and b are quire insensitive to the actual value of u. Table 3 illustrates this point. In this table, L denotes the numerical value of the left-hand and of eqn (10) for the corresponding u value. For a good solution, L should be near zero. This demonstration implies the existence of u and the utility of u to seek an unbiased, stable solution of n and b. In this respect, the new value of n and b, based on our modified Lotka's law given by eqn (7), should be more meaningful than that derived from Lotka's original law, given by eqn (1).

3. While Table 3 proves the stability of the values of n and b, quite independent of the exact value of u, the results also prove that u value can be determined quite precisely given by the values boxed in the table. With this credential, we hope informetricists would offer a concrete interpretation of u beyond the generalities we have suggested in this paper.

4. A final remark we would like to make is that the motivation and the formulation of the present paper are from mathematical consistency and logical necessity. Although in our establishment we have used a minimization procedure that has also been used in statistical analysis, the validity and the significance of our work are not dependent on statistical criteria.

Acknowledgements—We express our appreciation to Josephine M. Fang, Professor Emerita of the Graduate School of Library and Information Science of Simmons College, Boston, Massachusetts, for her introduction of ref. [3] leading to the present investigation, and her help in the preparation of this paper. We thank Linda Watkins, Librarian of the same School, for providing us with some important reference materials.

REFERENCES

1. Lotka, A.J. The frequency distribution of scientific productivity. *Journal of the Washington Academy of Science*, *16*(12):317-323; 1926.

2. See, for example, Egghe, L., and Rousseau, R. *Introduction to informetrics, quantitative methods in library, documentation and information science*. Amsterdam: Elsevier; 1990, and references therein.
3. Ho, J. Quantitative analysis of sample articles on bibliometrics (in Chinese). *Journal of Library and Information Science* (National Taiwan Normal University), *18*(1):48–82; 1992.
4. Pao, M.L. Lotka's law: A testing procedure. *Information Processing & Management*, *21*(4):305–320; 1985.
5. Nicholls, P.T. Empirical validation of Lotka's law. *Information Processing & Management*, *22*:417–419; 1986.
6. Mandelbrot, B. *The fractal geometry of nature*. New York: Freeman; 1977.
7. Tague, J., and Nicholls, P.T. The maximum value of Zipf size variable: Sample properties and relationship to other parameters. *Information Processing and Management*, *23*:155–170; 1987.
8. The software we used is Math CAD from MathSoft Inc., Cambridge, MA 02139.