Information Processing & Management Vol. 22, No. 5, pp. 417-419, 1986
Printed in Great Britain.

0306-4573/86 $3.00 + .00

EMPIRICAL VALIDATION OF LOTKA'S LAW

PAUL TRAVIS NICHOLLS†
School of Library and Information Science, University of Western Ontario,
London, Canada N6G 1H1

Abstract—Two modifications to the Pao procedure for testing Lotka's law are proposed and applied to 15 samples drawn from the humanities, social sciences, and sciences.

Lotka's law is a discrete probability distribution function which describes author productivity. Originally proposed as an inverse-square model by Lotka[1], a more general inverse-power form is now recognized as Lotka's law:

$$y_x = kx^{-b} \quad , \quad x = 1,2, \ldots, x_{\max} \quad , \tag{1}$$

where y_x represents the probability that an author will make x published contributions to a subject, while k and b are parameters to be estimated from the data.

Recently, Pao[2] has proposed a standard testing procedure for Lotka's law. Although many previous studies testing the applicability of Lotka's law have been reported, the procedures have been inconsistent and the results therefore largely incomparable[3]. Some of the methods employed were also unsatisfactory. A sound and replicable testing methodology is a necessary prerequisite to the validation and generalization of Lotka's law.

THE PAO PROCEDURE

The main elements involved in "fitting" a bibliometric model are: measurement of the variable(s) and tabulation, form of the model, parameter estimation, and a criterion for goodness-of-fit. Pao recommends the following approach in the case of Lotka's law:

1. *Measurement and tabulation*: The number of senior authors y_x contributing x papers are organized into a size-frequency table of N x, y pairs.
2. *Model*: The generalized inverse-power model, $y_x = kx^{-b}$, is adopted.
3. *Estimation of slope b*: The ordinary linear least squares estimate of b in the transformed model,

$$\log y_x = \log k - b \log x \quad , \quad x = 1,2, \ldots, x_{\max} \tag{2}$$

   is calculated, but excluding that part of the data representing the more prolific authors. The "best" cut-off point is determined by visual inspection combined with trial-and-error, the aim being to optimize the linearity implied by eqn 2.
4. *Estimation of constant k*: The inverse of the Riemann Zeta function is used, such that

$$k^{-1} = \zeta(b) = \Sigma x^{-b} \quad , \quad x = 1,2, \ldots \tag{3}$$

   for which Pao provides a very accurate approximation formula.
5. *Test*: The Kolmogorov–Smirnov (K-S) goodness-of-fit test is applied to the full set of observed and expected values at the 0.01 significance level.

With the exception of the test, these procedures are modelled very closely on Lotka's own. Two modifications are proposed. First, if the data are to be truncated for estimation of b, the criterion should be formally defined. The point where ties for the produc-

tivity scores end (at the first $y_x = 1$) is suggested. This criterion has both a theoretical and empirical basis[4], and is objective. However, an alternative approach is maximum likelihood, which provides an estimate with optimal qualities (unbiased, consistent, sufficient) and obviates the necessity to truncate altogether. A maximum likelihood estimate of b will satisfy the equation

$$(\Sigma \log y_x / \Sigma y_x) = -\zeta'(b)/\zeta(b) \ , \tag{4}$$

which may be solved by numerical iterative methods or using tables of $-\zeta'(b)/\zeta(b)$ provided in Johnson and Kotz[5].

Secondly, the productivity measure should take account of all collaborating authors. Lotka counted only first authors because multiple authorship was less common at that time, and probably, because it was easier[3]. Today, inquiry in most fields is characterized by extensive and increasing collaboration which is reflected in multiple authorship; measures which are insensitive to this phenomenon are invalid, assuming that we are interested in the distribution of author productivity. The senior author measure cannot be considered to be a sampling strategy either, since the underlying processes are probably not random[6].

APPLICATION

The modified procedure was applied to 15 classic datasets previously reported in the literature†. Samples 1–4 are drawn from humanities disciplines, 5–9 from the social sciences, and 10–15 from the natural sciences. All collaborating authors were included in the counts. Parameters were estimated by both the truncated least squares (LS) and maximum likelihood (ML) methods proposed above, and the Kolmogorov–Smirnov test applied (Table 1). Significant values of the test statistic (D_{max}) are in bold type. The critical value at 0.01 significance is denoted by $D_{K\ S}$ and obtained by the asymptotic formula

$$1.63/(\Sigma y_x + (\Sigma y_x/10)^{1/2})^{1/2} \ . \tag{5}$$

It is clear that both estimators perform well; however, the ML estimator consistently yields a lower maximum deviation and retains the null hypothesis in two cases rejected using the LS estimate. Incidently, with the ML estimate $b = 1.95$ and $k = 0.5902$, Lotka's senior author data for *Chemical Abstracts* (A, B, and A&B) *do* conform to the generalized model according to the K–S criterion.

Pao observes that the critical value of the K–S statistic decreases rapidly with increasing sample size, requiring a very small deviation with large samples to reject the null hypothesis. The K–S test is therefore quite sensitive to sampling. The two samples which were finally rejected here are also the largest (1529 and 3162 respectively). Pao also notes that the test is conservative (although still valid) when applied to discrete variables. In fact, the exact critical values for the discrete case would often be about 1/3 of the continuous estimates normally employed[18], which would have resulted in rejecting the Lotka hypothesis in four more cases. Finally, the underlying assumption of a random sample has never been properly addressed in connection with the use of inferential tests such as X^2 or K–S on this type of data[4].

CONCLUSIONS

The Pao procedure is a rationalization of the best methods employed by previous investigators. The modifications proposed here are refinements which have shown promise in a limited trial; and the procedure remains simple to apply using a statistical package

†1: Rao [7, p. 119], 2: Schorr [8, p. 206], 3: Pao [9, p. 105], 4: Munch-Petersen [10, p. 9,'A'], 5: Rao [7, p. 120], 6: Schorr [11, p. 32], 7: Rao [7, p. 116], 8: Coile [12, p. 134, 1966], 9: Frohmann [13, p. 116], 10: Dufrenoy [14, p. 207, 1932], 11: Radhakrishnan and Kernizan [15, p. 51, CACM], 12: *ibid.*, p. 51, JACM, 13: Subramanyam [16], 14: Subramanyam [17], 15: Rao [7, p. 114].

Table 1. Kolmogorov-Smirnov test for two estimation methods

	b_{LS}	k_{LS}	D_{max}	b_{ML}	k_{ML}	D_{max}	$D_{K\text{-}S}$
1.	2.1047	.6424	.0936	2.7500	.7935	.0288	.1235
2.	2.6652	.7784	.0190	2.6500	.7756	.0162	.0510
3.	1.9789	.6006	.0521	2.2000	.6709	.0232	.0723
4.	1.7264	.4990	.0727	1.8500	.5519	.0477	.0878
5.	2.6513	.7758	.0243	2.6500	.7756	.0241	.0895
6.	3.1773	.8544	.0599	3.6500	.9001	.0147	.1113
7.	2.1471	.6554	.0314	2.1500	.6563	.0311	.0487
8.	3.4507	.8830	.0062	3.5000	.8875	.0053	.0453
9.	2.3512	.7111	.0206	2.4500	.7344	.0122	.0521
10.	3.0120	.8335	**.1239**	2.5500	.7560	**.0464**	.0415
11.	3.4880	.8864	.0558	3.0500	.8386	.0103	.0931
12.	3.4880	.8864	.0367	3.4000	.8782	.0285	.0662
13.	2.0032	.6090	**.1596**	2.3500	.7108	.0306	.0790
14.	2.3619	.7137	**.1622**	2.0500	.6248	**.0733**	.0289
15.	1.7020	.4877	**.0599**	1.8500	.5519	.0464	.0564

or dedicated program (a FORTRAN program is available from the author on request). A measure which takes account of co-authors is preferred because it is valid where productivity is collaborative (and remains so where it is not). The maximum likelihood estimator of b is preferred because of its optimal qualities and close relation to Pao's estimator for k. There are some misgivings with respect to the K-S test; however, it does avoid pooling categories, a correction for discreteness is available [18], and the D_{max} statistic retains a comparative descriptive value even outside the context of hypothesis testing[4].

Acknowledgement—The author wishes to thank Professors Craven, Kinnucan, and Tague of the School of Library and Information Science, University of Western Ontario, for their critical comments on this note.

REFERENCES

1. Lotka, A. J., The frequency distribution of scientific productivity, *Journal of the Washington Academy of Science*, **16**(12): 317-323; 1926.
2. Pao, M. L., Lotka's law: A testing procedure. *Information Processing and Management*, **21**(4): 305-320; 1985.
3. Potter, W. G., Lotka's law revisited. *Library Trends*, **31**(2): 21-39; 1981.
4. Nelson, M. J., and Tague J. M., Split size-rank models for the distribution of index terms. *Journal of the American Society for Information Science*, **36**(5): 283-296; 1985.
5. Johnson, N. L., and Kotz, S., *Discrete Distributions*. Houghton Mifflin, Boston; 1969.
6. Lindsey, D., Production and citation measures in the sociology of science—the problem of multiple authorship. *Social Studies of Science*, **10**(2): 145-162; 1980.
7. Rao, I. K. R., The distribution of scientific productivity and social change. *Journal of the American Society for Information Science*, **31**(2): 111-122; 1980.
8. Schorr, A. E., Lotka's law and the history of legal medicine. *Research in Librarianship*, **30**: 205-209; 1975.
9. Pao, M. L., Bibliometrics and computational musicology. *Collection Management*, **3**(1): 97-109; 1979.
10. Munch-Petersen, E., Bibliometrics and fiction. *Libri*, **31**(1): 1-21; 1981.
11. Schorr, A. E., Lotka's law and library science. *Reference Quarterly*, **14**(1): 32-33; 1974.
12. Coile, R. C., Letter. *Journal of the American Society for Information Science*, **26**(2): 133-134; 1975.
13. Frohmann, B., A bibliometric analysis of the literature of cataloguing and classification. *Library Research*, **4**(4): 355-373; 1982.
14. Dufrenoy, J., The publishing behaviour of biologists. *Quarterly Review of Biology*, **13**: 207-210; 1938.
15. Radhakrishnan, T., and Kernizan, R., Lotka's law and computer science literature. *Journal of the American Society for Information Science*, **30**(1): 51-54; 1979.
16. Subramanyam, K., Research productivity and breadth of interest of computer scientists. *Journal of the American Society for Information Science*, **35**(6): 369-371, 1984.
17. Subramanyam, K., Lotka's law and the literature of computer science. *IEEE Transactions on Professional Communication*, **22**(4): 187-189, 1979.
18. Conover, W. J., *Practical Nonparametric Statistics*. Wiley, New York; 1980.