International Journal of Scientometrics, Informetrics and Bibliometrics
ISSN 1137-5019

December 2, 2003

VOLUME 4 (2000): ISSUE 1. PAPER 4

LOTKA: A program to fit a power law distribution to observed frequency data

Brendan Rousseau and Ronald Rousseau
The Oakies, Blauwvoetstraat 10.
B-8400 Oostende, Belgium
oak@pandora.be

Abstract

LOTKA, a computer program for fitting a power law distribution such as Lotka's is presented. It basically follows Nicholl's methodology : using a maximum likelihood approach to estimate parameters, and a Kolmogorov-Smirnov test for goodness-of-fit. When input data are converted (from rank-frequency to size-frequency) this program can also be used to test Zipf's law. It can be downloaded **here**.

Keywords

computer program, power law distribution, Lotka's law, maximum likelihood estimation, Nicholls' methodology, Kolmogorov-Smirnov test, Zipf

Introduction

We are interested in fitting observed frequency data to the power law distribution

$$f(k)=\frac{C}{k^{\beta}} \quad (1)$$

with k = 1, 2, …. This power law distribution has two parameters: **ß** and **C**. These parameters are, however, not independent but are related through the requirement that

$$\sum_{k=1}^{\infty}\frac{C}{k^{\beta}}=1 \quad (2)$$

Power laws of this form abound in the scientific literature. Author production

(Lotka, 1926), interactions, i.e. links and sitations on the WWW (Barabási & Albert, 1999; Rousseau, 1997), distribution of family names (Miyazima et al., 1999), even the number of gold records in the popular music industry (Cox et al., 1995) are all said to follow a power law distribution.

The program we present follows Nicholls' methodology: specification of the model (here Lotka's power law), organization of the data (here in a size-frequency form, using all data, i.e. no truncation has been performed), estimation (using the maximum likelihood approach) and testing (using Kolmogorov-Smirnov) (Nicholls, 1986, 1987, 1989). Note also that if data are given in rank-frequency form (Zipf form) they can always be converted into a size-frequency form, and hence this program can be used.

Method

Conventionally the parameters ß and C, or at least the parameter C, have been estimated by the least squares method (Pao, 1985, 1986; Kawamura et al. 1999). This approach has, however, several drawbacks, the most important one being the fact that the linear least squares method gives only acceptable results when data are truncated. Nicholls (1987) convincingly showed that the maximum likelihood estimator for ß is by far the best method for estimating ß. Once the parameter ß has been estimated, C follows from equation (2). In practice, both the maximum likelihood estimator and the corresponding value for C are not easy to determine and one needs a computer program. Our program is simply based on tables taken from Nicholls (1987), Egghe & Rousseau (1990) and Rousseau (1993).

The program also applies a Kolmogorov-Smirnov goodness-of-fit test. This test is based on the maximum absolute deviation between the observed and the theoretical distribution functions. It has the advantage that it is non-parametric and requires no pooling of categories. It is, however, well known, that strictly speaking its use is unwarranted. One of the main problems being the fact that the data to which we want to apply the test are certainly not random data. Anyway, even if this were the case, the test is (very) conservative. Yet, we defend its use if only for comparative and descriptive use. Moreover, regularities observed in the social sciences are not the same as the precise laws of physics: they only describe a general trend. We accept that for this purpose the Kolmogorov-Smirnov test is adequate. Of course, those who do not agree with us may ignore the fitting part of the program.

An example

In Informetrics 87/88 and Informetrics 89/90, the conference proceedings of the first two international conferences on bibliometrics, scientometrics and informetrics there was

1 journal (namely JASIS) cited 76 times

1 journal (namely Scientometrics) cited 62 times

1 journal (namely Journal of Documentation) cited 58 times

and so on, leading to the following frequency distribution (taken from Rousseau (1997).

Table 1. Frequency distribution of periodicals cited in Informetrics 87/88 and Informetrics 89/90

Number of sources (i.e. journals)	Production (i.e. number of citations)
1	76
1	62
1	58
1	15
1	14
2	11
1	9
4	6
1	5
7	4
11	3
20	2
80	1

In the program one finds the same two columns: one headed *sources*, and headed *production*. Here one has to fill in the data. It does not matter in which order this is done: starting with the most productive source (or sources) as is done in Table 1, or starting with all the sources that have produced one item, or even in any other order. It is even allowed to write '0 sources with production 70'. Leaving a row blank or adding blanks after the numbers leads to an error message. It is at any time possible to correct mistakes. Once one is sure that the input contains no error one clicks one the 'analyse' button, and obtains the best fitting C and ß values. For this example C = 0.6244 and ß = 2.0489. At the same time a Kolmogorov-Smirnov test has been performed. We see (see Fig.1) that for this example the maximum (absolute) deviation is 0.0232, leading to the acceptance of the power law,

$$f(k) = \frac{0.6244}{k^{2.049}}$$

whichever of the critical values one prefers.

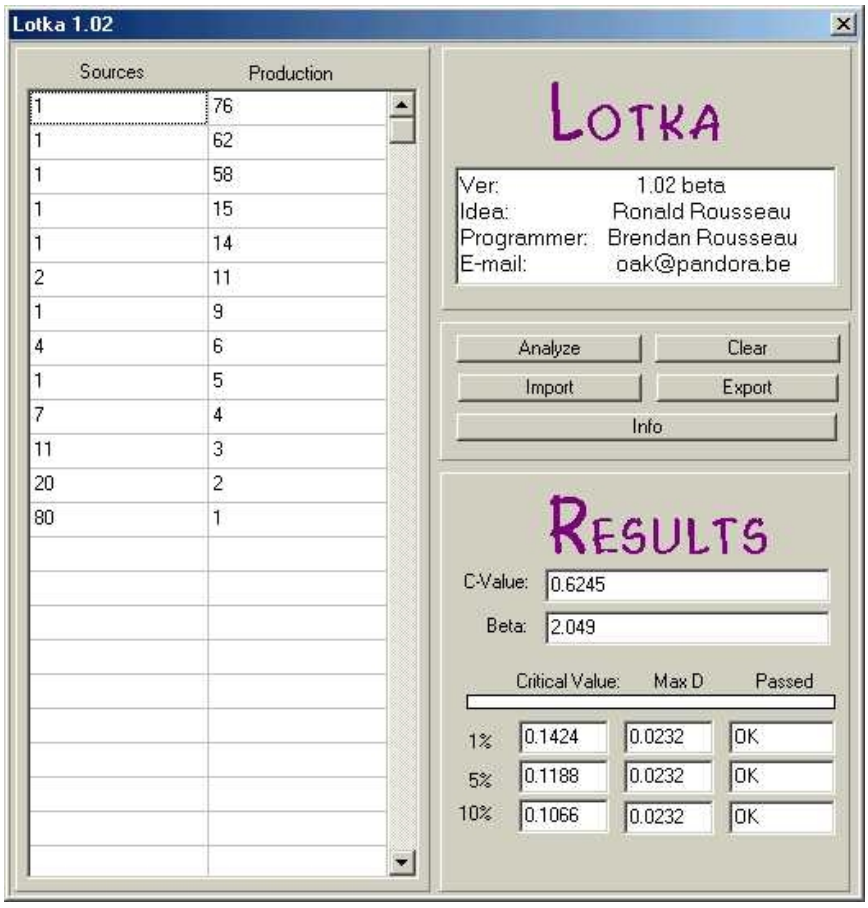

Fig.1. An example of a LOTKA screen

If one wants to use these results in another program (e.g. inserting these data and results in a text) one has to click on the 'export' button. Clicking on the 'clear' button allows one to continue with another set of data.

Just as a test we used our program to analyse Cox's data on golden records (a list at that time topped by *the Beatles*) (Cox et al.,1995). We found a ß-value of 1.863 and a corresponding C-value of 0.557. The corresponding power law, however, was rejected by a K-S test, with a maximum absolute deviation of 0.0717. Cox et al. found, using a least squares procedure, a best ß-value of 1.578. They use the observed value for C, but the correct C-value (otherwise there is no distribution in a statistical sense, unless they specify the interval, which they don't) is 0.426. Testing (K-S) with these values yields a maximum absolute deviation of 0.135, which is far worse than the ML-estimate. Note, however, that their least squares fit yields an R^2 value of 0.9966, from which they conclude that the fit is excellent. This, again, shows the difference not only between different estimation methods, but also between criteria to judge if the obtained fit is acceptable.

Technical restrictions

The program yields only ß-values between 1.26 and 3.49. Note that, according to Egghe's formal theory on IPPs (Egghe,1990) a ß-value larger than 3 cannot occur. Further, the percolation model (Bogaert et al., 2000) always seems to yield ß-values larger than 1.7. So allowing ß-values between 1.26 and 3.49 seems to be on the safe side for most applications. The maximum production is 10,000. Finally, all calculated values are shown as truncated (not rounded) numbers.

The program is an update and extension of a program R.R. wrote in 1995, and which has been available from him since then. Version 1.02 differs from

version 1.01 by the fact that it is now possible to import data in comma-delimited form. This form is, for example, one of the output formats of an Excel spreadsheet. Note, however, that data must be stored (e.g. in Excel with .csv extension) in exactly the same way as in our program: two columns, the first relating to the number of sources, and the second one relating to the number of produced items. No comments or other textual features may be present. In case you notice that the Excel comma-delimited format is actually not comma delimited (but e.g. semi-colon delimited - this may occur, depending on Windows' regional settings) you have to do the following in Windows: go to *settings*, then *control panel*, then *regional settings*, *number* and finally make sure that the *list separator* is a comma. This should work.

The output screen of version 1.01 showed the beta-value in the C-box and vice versa. This has been corrected.

References

Barabási, A.-L. and Albert, R. (1999). "Emergence of scaling in random networks". **Science**, 286 (1999), 509-512.

Bogaert, J., Rousseau, R. and Van Hecke, P. (2000). "Percolation as a model for informetric distributions: fragment size distribution characterised by Bradford curves". **Scientometrics**, 47 (2000), 195-206.

Cox, R.A.K., Felton, J.M. and Chung, K.C. (1995). "The concentration of commercial success in popular music: an analysis of the distribution of gold records". **Journal of Cultural Economics**, 19 (1995), 333-340.

Egghe, L. (1990). "The duality of informetric systems with applications to the empirical laws". **Journal of Information Science**, 16 (1990), 17-27.

Egghe, L. and Rousseau, R. (1990). **Introduction to Informetrics. Quantitative methods in library, documentation and information science**. Elsevier, Amsterdam.

Kawamura, M., Thomas, C.D.L., Kawaguchi, Y. and Sasahara, H. (1999). "Lotka's law and the pattern of scientific productivity in the dental science literature". **Medical Informatics & The Internet in Medicine**, 24 (1999), 309-315.

Lotka, A.J. (1926). "The frequency distribution of scientific productivity". **Journal of the Washington Academy of Sciences**, 16 (1926), 317-323.

Miyazima, S., Lee, Y., Nagamine, T. and Miyajima, H. (1999). "Family name distribution in Japanese societies". **Journal of the Physical Society of Japan**, 68 (1999), 3244-3247.

Nicholls, P.T. (1986). "Empirical validation of Lotka's law". **Information Processing and Management**, 22 (1986), 417-419.

Nicholls, P.T. (1988). "Estimation of Zipf parameters". **Journal of the American Society of Information Science**, 38 (1987), 443-445. + Erratum, JASIS, 39 (1988), p. 287.

Nicholls, P.T. (1989). "Bibliometric modeling processes and the

empirical validity of Lotka's law". **Journal of the American Society for Information Science**, 40 (1989), 379-385.

Pao, M.L. (1985). "Lotka's law: a testing procedure". **Information Processing and Management**, 21 (1985), 305-320.

Pao, M.L. (1985). "An empirical examination of Lotka's law". **Journal of the American Society for Information Science**, 37 (1986), 26-33.

Rousseau, R. (1993). "A table for estimating the exponent in Lotka's law". **Journal of Documentation**, 49 (1993), 409-412.

Rousseau, R. (1997). "The proceedings of the first and second international conferences on bibliometrics, scientometrics and informetrics: a data analysis". In: B. Peritz & L. Egghe (eds.), **Proceedings of the Sixth Conference of the International Society for Scientometrics and Informetrics**, Hebrew University, Jerusalem, 1997, 371-380.

Rousseau, R. (1997). **Sitations: an exploratory study**. **Cybermetrics**, 1(1), paper 1. <http://www.cindoc.csic.es/cybermetrics/v1i1p1.htm>

Received 6/August/2000
Accepted 3/December/2000
Updated 22/January/2001

CORRESPONDENCE		
Comments to the article by Rousseau & Rousseau Eric Archambault	**Comments to the article by Rousseau & Rousseau** Mark Newman	**Software and Peer-Review: The Rousseau Case** J. Sylvan Katz
Rejoinder Brendan Rousseau, Ronald Rousseau		