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

December 2, 2003

The Journal
Cybermetrics News
Editorial Board
Guide for Authors
Issues Contents
The Seminars
The Source
Scientometrics
Tools
R&D Policy & Resources
World Sitation Report

VOLUME 4 (2000): ISSUE 1. CORRESPONDENCE

Comments to the article by Rousseau & Rousseau

Eric Archambault
Observatoire des Sciences et des Technologies
Institut National de la Recherche Scientifique
3465, rue Durocher
Montreal, Quebec
Canada H2X 2C6
eric.archambault@inrs-urb.uquebec.ca

This contribution aimed to inform on an additional way to calculate power law distributions. I believe that Brendan and Ronald Rousseau's contribution is useful since it uses a maximum likelihood approach. It would be interesting to compare the extent of the difference between this method compared to using least-square fits.
Many users have been using Excel to calculate regressions of hyperbolic (power-law) distributions a la Lotka. This can be performed in either of two ways.

Method 1

1) Plot the data on a XY (Scatter) graph
2) Select the data series on the graph and "Add Trendline..." in the "Chart" menu.
3) Select "Power" type of regression curve, in the Option Tab, select "Display equation on chart" as well as "Display R-squared value on chart"

Method 2

1) Select a two column by five rows area on the spreadsheet where you data is
2) Type "=Linest(log(Y:Yn);log(X:Xn);1;1)" where Y:Yn is the range of the Y-data (frequencies) and X:Xn is the range of the X data (number).
3) Press simultaneously Ctrl-Shift-Enter to create an array-formula. Read Excel's help to interpret the stats. To convert the b of the intercept, raise 10 to the power of b to obtain the constant of the power law (c=10^b).

The advantage of using these methods based on the least-square fit is to obtain the R-Value as well as, in the case of spreadsheet based method (Method 2) the F-statistics. The t-test can also be calculated from the results of the formula array.

Here you can download a template that calculate regressions from the original Lotka (1926) data in both the number-frequency (as in Lotka's paper) and rank-frequency (as in the form used by Auerbach long before Zipf) forms.

The rank-frequency form of Lotka falsify the assertion of Zipf that data following Lotka's law (not really a law since it fits only his data) would produce a rank-frequency distribution with a power of 1.

Discussion

I agree with Mark (**Newman's Comments in this issue**) on the danger of working with power law distributions. Those who have tried to replicate Lotka's work will have noticed that although he did no mention of it, Lotka excluded some data from his dataset. One has to transform some of the data through the use of outliers to alleviate the problems outlined by Mark, in the case of number-frequency distributions in scientometrics at least. As he so rightly pointed out, regressions will be overestimated or underestimated regardless of whether the data is number-frequency or rank-frequency. In number-frequency distributions, it is very difficult to calculate a valid regression whatever the method used. This is inherent to the data and not to the method. This applies more to problems in the social sciences than in the natural sciences, although I do not pretend it is absent in the latter. In social systems, for number-frequency distributions, it is very difficult to calculate a valid exponent without the use of outliers.

The use of rank-frequency only shifts the problem around. The solution that I have favored to calculate rank-frequency is to minimize this effect by binning the data, hence, using the mid-point for any given frequency, the data becomes a mean-rank - frequency distribution. Once this transformation is accomplished, I'm not certain that least-square fitting is so bad, hence my suggestion to test the difference obtained by maximum-likelihood and least-square methods, and why not the maximum-entropy method while we're at it.

In the end, the epistemological question remains of how to choose the best answer, and hence, the best method. If we do not know a priori the power coefficient of a distribution, and given the weakness of our theoretical knowledge on the why of power-law distributions in social systems (sorry for those drawing (weak) analogies between sand-piles, dinosaur extinction, and scientific publications, this is not a theory nor an explanation for what we observe in scientometric research) there is no foolproof method to determine which measure is the "real one".

*Updated version of a message originally submitted to **SIGMETRICS** listserv*

References

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

Rousseau , B. and Rousseau, R. (2000). **LOTKA: A program to fit a power law distribution to observed frequency data**. **Cybermetrics**, 4 (1), paper 4 <**http://www.cindoc.csic.es/cybermetrics/v4i1p4.htm**>

Received 24/January/2001
Updated 26/January/2001

MAIN PAPER	CORRESPONDENCE
	Comments to the article by Rousseau & Rousseau Eric Archambault

LOTKA: A program to fit a power law distribution to observed frequency data Brendan Rousseau, Ronald Rousseau	**Comments to the article by Rousseau & Rousseau** Mark Newman
	Software and Peer-Review: The Rousseau Case J. Sylvan Katz
	Rejoinder Brendan Rousseau, Ronald Rousseau