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

December 2, 2003

> Homepage > The Journal > Issues Contents > Vol. 4 (2000) > Correspondence

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

M. E. J. Newman
Santa Fe Institute
Santa Fe, New Mexico
USA
mark@santafe.edu

As has been pointed out by many people before, performing a least-squares fit to the logarithm of a histogram in order to fit a power law is fraught with danger. The principal objection to the method is that the statistical fluctuations in the logarithms of the data are greater in the downward direction than in the upward one. This effect is more pronounced in the tail of the power law, and this has the result that there is a systematic tendency for least-squares fits to overestimate the slope of the power law.

How much they overestimate depends on the size of the statistical fluctuations. Two common methods are used to circumvent this problem, neither of which is perfect:

> 1) One calculates a backward cumulated histogram of one's data (also called a rank/frequency plot). The slope of such a rank/frequency plot, on logarithmic scales, is one less than the slope that the original histogram would have. Using the rank/frequency plot much improves the statistical fluctuations, but has the undesirable property that successive data points become correlated, making the simple statistical estimate of error on the fit invalid.
>
> 2) One performs logarithmic binning of the data, i.e., binning where the widths of adjacent bins are a constant ratio, and normalizes by bin width. This reduces the effects of the fluctuations, but for power laws with slope steeper than -1 it does not eliminate them altogether.

Ronald Rousseau proposes a further method based on maximization of likelihood. This is also a good method to use, but is also not perfect since, like all maximum likelihood methods, it implicitly assumes that the probability of the model given the data is equal to the probability of the data given the model, which is only strictly true if the prior probability of the model is uniform in the parameter space used, which in general it is not.

The ultimate correct way of performing the fit is to use maximization of entropy, given the correct prior on the model. The trouble is, we rarely know what the correct prior is, which is why maximum likelihood is popular.

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 26/January/2001

MAIN PAPER	CORRESPONDENCE
LOTKA: A program to fit a power law distribution to observed frequency data Brendan Rousseau, Ronald Rousseau	**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

Copyright information | **Editor** | **Webmaster** | **Updated: 11/27/2003** TOP