Journal of Informetrics 10 (2016) 622–633

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Are there too many uncited articles? Zero inflated variants of the discretised lognormal and hooked power law distributions

Mike Thelwall

Statistical Cybermetrics Research Group, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1ST, UK

ARTICLE INFO

Article history:
Received 15 January 2016
Received in revised form 20 April 2016
Accepted 20 April 2016
Available online 6 May 2016

Keywords:
Discretised lognormal distribution
Hooked power law
Citation analysis
Uncited articles
Zero inflated discretised lognormal distribution
Zero inflated hooked power law

ABSTRACT

Although statistical models fit many citation data sets reasonably well with the best fitting models being the hooked power law and discretised lognormal distribution, the fits are rarely close. One possible reason is that there might be more uncited articles than would be predicted by any model if some articles are inherently uncitable. Using data from 23 different Scopus categories, this article tests the assumption that removing a proportion of uncited articles from a citation dataset allows statistical distributions to have much closer fits. It also introduces two new models, zero inflated discretised lognormal distribution and the zero inflated hooked power law distribution and algorithms to fit them. In all 23 cases, the zero inflated version of the discretised lognormal distribution was an improvement on the standard version and in 16 out of 23 cases the zero inflated version of the hooked power law was an improvement on the standard version. Without zero inflation the discretised lognormal models fit the data better than the hooked power law distribution 6 out of 23 times and with it, the discretised lognormal models fit the data better than the hooked power law distribution 9 out of 23 times. Apparently uncitable articles seem to occur due to the presence of academic-related magazines in Scopus categories. In conclusion, future citation analysis and research indicators should take into account uncitable articles, and the best fitting distribution for sets of citation counts from a single subject and year is either the zero inflated discretised lognormal or zero inflated hooked power law.

1. Introduction

Citation-based indicators are commonly used to help assess academic departments (Hicks, 2012; Wilsdon et al., 2015) and scholarly journals (Schweizer, 2010) as well as in some digital libraries to help rank search results (Lawrence, Giles, & Bollacker, 1999; Mayr & Walter, 2007). Citation analysis is still controversial, however, for general reasons as well as for specific criticisms of individual indicators (e.g. DORA, 2012; MacRoberts & MacRoberts, 1989). In order to construct the most suitable indicators and to evaluate the limitations of existing indicators, it is useful to know as much about citations as possible, such as why they are created, what their role is and the factors that influence their creation (Borgman & Furner, 2002; Moed, 2006; van Raan, 2005). Four ways to achieve this are to theorise about the role of citations in scholarly communication (Merton, 1973), to ask scholars why they cite (Brooks, 1985; Case & Higgins, 2000), to examine individual citations to ascertain their apparent purpose (Chubin & Moitra, 1975; Oppenheim & Renn, 1978) and to statistically analyse collections of articles

E-mail address: m.thelwall@wlv.ac.uk

http://dx.doi.org/10.1016/j.joi.2016.04.014
1751-1577/

to identify factors that associate with citation counts (Peters & van Raan, 1994; Zitt & Bassecoulard, 1998). For statistical analyses, it is important to identify the overall distribution of sets of citation counts so that appropriate regression (e.g. linear, negative binomial, or zero inflated; any transformations needed) or other techniques can be selected (Thelwall, 2016b). The most appropriate choice of indicator also depends upon the nature of citation distributions. For instance, impact factors for journals should be calculated with the geometric mean rather than the arithmetic mean because of the skewed nature of citation counts for journals (Thelwall & Fairclough, 2015; Zitt, 2012). The distribution of sets of citation counts is also needed to assess the precision (Thelwall, 2016a) and other properties of indicators generated from citation counts as well as to help identify when sets of citation counts are anomalous in some way. There have been many attempts to identify statistical distributions that are appropriate for sets of citation counts but there is no consensus about which is the best overall.

The most appropriate sets of articles to examine from the perspective of citation count distributions are sets of journal articles because journal articles are the primary scholarly record in most areas of scholarship (excluding the arts and humanities and some social sciences). These sets should also be homogeneous in the sense of being from the same subject and year (or at least a common citation window: Abramo, Cicero, & D'Angelo, 2011) because different fields attract citations at different rates and so should not be mixed, if possible, and citations accumulate over time. For a typical homogeneous set of articles, the citation counts are highly skewed because few articles are highly cited and many remain uncited (Seglen, 1992). If articles with few citations are excluded, then sets of citation counts fit a power law or a discretised lognormal distribution quite well (Garanina & Romanovsky, 2015; Redner, 1998; van Raan, 2001), but if all articles are included then the power law is a very poor fit for almost all subjects (exception: physics) and the discretised lognormal distribution and hooked power law (described below) tend to fit much better (Eom & Fortunato, 2011; Radicchi, Fortunato, & Castellano, 2008; Thelwall & Wilson, 2014). Of these two, the hooked power law seems to fit better for a majority of subjects, although the discretised lognormal fits better for a minority (Thelwall & Wilson, 2014). A discrete version of the power law, the Yule–Simon process, also fits citation data reasonably well (Brzezinski, 2015) but cannot cope with subjects that have modes (most common values) greater than 0. A range of other distributions have also been proposed but all have problems. Since citation data sets are a type of count (integer) data, a range of count data models have been tested. The negative binomial distribution (Hilbe, 2011) model is a count model for skewed data sets but does not fit citation counts as well as the other two models (Low, Wilson, & Thelwall, 2015). Stopped sum models (Neyman, 1939) seem to fit slightly better but have parameter estimation issues that make them impractical to use (Low et al., 2015).

Although the discretised lognormal and hooked power law seem to be the distributions that fit sets of citation counts best, at least in practice, neither are perfect fits for most Scopus subjects and have particular problems in estimating the number of uncited articles (Thelwall, 2016c). This article assesses whether the main problem is that there are too many uncited articles published in journals, in the sense that if uncited articles are given special treatment then the remaining articles fit citation distributions much better. This could occur, for example, if some articles are almost uncitable because they make a valid contribution but either close off an area of research, address a highly niche topic, or are of interest only to students, practitioners or policy makers. This issue is investigated with a purely quantitative approach by systematically removing uncited articles in order to ascertain whether statistical distributions fit better afterwards. In addition, this article introduces two new distributions, the zero inflated discretised lognormal (ZIDL) and the zero inflated hooked power law (ZIHP) to deal with this situation, as well as software to fit them.

2. The zero inflated discretised lognormal and hooked power law distributions

The hooked power law, also known as the shifted power law (Pennock, Flake, Lawrence, Glover, & Giles, 2002), has probability mass function $f(n) = A(B+n)^{-\alpha}$, where α and B are the parameters of the distribution and A is determined by the choice of α and B because the sum of $f(n)$ for all theoretically possible values of n must be 1. The parameter α is also found in the power law and primarily determines how high the citation counts are likely to be, and the shift parameter B primarily affects the extent to which very low values occur, including the value of the mode.

The continuous lognormal distribution $\ln\mathcal{N}\left(\mu, \sigma^2\right)$ has probability density function $f(x) = 1/x\sigma\sqrt{2\pi}e^{-(\ln(x)-\mu)^2/2\sigma^2}$ with location parameter μ and scale parameter σ that are the mean and standard deviations of the natural log of the data (Limpert, Stahel, & Abbt, 2001). This can be converted into the discretised lognormal distribution $\ln\overset{\dots}{\mathcal{N}}\left(\mu, \sigma^2\right)$ by integrating a unit interval around each positive integer, giving probability mass function $f(n) = \frac{1}{A}\int_{(n-0.5)}^{(n+0.5)} f(x)\mathrm{d}x$, where $A = \int_{0.5}^{\infty} f(x)\,\mathrm{d}x$ compensates for the missing interval (0,0.5] from the continuous distribution.

The discretised lognormal distribution is only defined for positive integers and not zero, so cannot accommodate uncited articles. The standard way to circumvent this issue, and the one used in this article, is to add 1 to all citation counts before analysing them. This is not necessary for the hooked power law but, if done, makes no difference except for decreasing the value of B by 1.

3. Zero inflation calculations

A zero inflated variant of a distribution incorporates a procedure to remove some zeros from a dataset before applying the main distribution to the remaining data. This can be theorised in terms of some of the zeros being "natural" or predetermined