Journal of Informetrics 10 (2016) 454–470

Contents lists available at ScienceDirect

Journal of Informetrics

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

Are the discretised lognormal and hooked power law distributions plausible for citation data?

Mike Thelwall *

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

ARTICLE INFO

Article history:
Received 9 January 2016
Received in revised form 12 March 2016
Accepted 12 March 2016
Available online 7 April 2016

Keywords:
Citation distributions
Hooked power law
Discretised lognormal distribution

ABSTRACT

There is no agreement over which statistical distribution is most appropriate for modelling citation count data. This is important because if one distribution is accepted then the relative merits of different citation-based indicators, such as percentiles, arithmetic means and geometric means, can be more fully assessed. In response, this article investigates the plausibility of the discretised lognormal and hooked power law distributions for modelling the full range of citation counts, with an offset of 1. The citation counts from 23 Scopus subcategories were fitted to hooked power law and discretised lognormal distributions but both distributions failed a Kolmogorov–Smirnov goodness of fit test in over three quarters of cases. The discretised lognormal distribution also seems to have the wrong shape for citation distributions, with too few zeros and not enough medium values for all subjects. The cause of poor fits could be the impurity of the subject subcategories or the presence of interdisciplinary research. Although it is possible to test for subject subcategory purity indirectly through a goodness of fit test in theory with large enough sample sizes, it is probably not possible in practice. Hence it seems difficult to get conclusive evidence about the theoretically most appropriate statistical distribution.

1. Introduction

Citation-based indicators are sometimes used to support formal evaluations of groups of academics and for self-evaluations. The Journal Impact Factor (JIF) is the most widely used, in the belief that it may tend to correspond to human judgements of journal quality in some subjects (e.g., Gordon, 1982). Moreover, field normalised indicators, such as the Mean Normalised Citation Score (MNCS) (Waltman, van Eck, van Leeuwen, Visser, & van Raan, 2011a,b) are commonly used for evaluations and reports (e.g., EC, 2007; Elsevier, 2013; NIFU, 2014; NSF, 2014). Most citation indicators do not cope well with the highly skewed nature of sets of citation counts, however, and this has led to a shift toward percentile-based indicators, such as the proportion of a department's outputs in the most highly cited 10% in its field. This approach has been criticised in turn on the basis that percentile-based indicators are imprecise and that geometric mean field normalised citation counts is more precise and therefore more likely to detect genuine differences and less likely to indicate spurious differences (Thelwall, 2016a; for a similar JIF argument see also: Zitt, 2012). This conclusion was drawn based upon a set of limiting assumptions, however, and is not therefore definitive. In order to give a more conclusive statement about which indicator is the most suitable for comparing the impacts of sets of articles, theoretical insights are needed into the mathematical properties of sets

* Corresponding author. Tel.: +44 01902321000.
E-mail address: m.thelwall@wlv.ac.uk

http://dx.doi.org/10.1016/j.joi.2016.03.001
1751-1577/

of citation counts so that randomness within citation counts can be effectively separated from any underlying patterns. This is particularly important because in some cases the decision about which indicator to use affects importance of a set of articles, not because of imprecision in the indicators but because different indicators can give fundamentally different answers (Thelwall, 2016a). For example, an indicator to compare the citation impact of nations (e.g., Albarrán, Perianes-Rodríguez, & Ruiz-Castillo, 2015) would not be useful if it could be shown to be an imprecise side-effect of a more fundamental and more precisely measurable property of the citation distribution. In some cases the choice of indicator may be driven by policy decisions but even these need to be informed by an understanding about the properties of the different indicators.

Two attractive distributions for citation data are the power law (Clauset, Shalizi, & Newman, 2009; Egghe, 2005), with probability density function $f(x) = Ax^{-\alpha}$ for $x > c > 0$ (for some $c > 0$, and where A is chosen to ensure that the probabilities sum to 1, so the distribution has only one free parameter, α) and the Yule process (Brzezinski, 2015), with its limiting equilibrium form having probability mass function $f(x) = \rho B(x, \rho + 1)$ for $x = 1, 2, \ldots$, where B is the Euler beta function. Both can be generated by cumulative advantage processes in which the probability that an article attracts citations is related to the number of citations that it already has, so that highly cited articles naturally attract an increasingly large share of new citations (de Solla Price, 1976). This is intuitively reasonable because articles can be found through citations from other articles and by searches in digital libraries that use citation counts within their search results ranking mechanisms (Lawrence, Giles, & Bollacker, 1999). Both of these distributions fit citation data well asymptotically (Brzezinski, 2015) but poorly overall (Thelwall & Wilson, 2014a; Thelwall & Wilson, in press). The poor overall fit is due to both having monotonic decreasing (point mass) functions, whereas many subject areas have citation count modes of at least 1 (Thelwall, 2016a), indicating that they are increasing for some of their values. These distributions only model positive values but can model the full range of citation counts if 1 is added first. This could be thought of as adding a citation to all papers to reflect an implicit self-citation or the basic value that an article has just from being published (see also: de Solla Price, 1976).

Several alternative distributions have been proposed to solve the problem of the ability of pure cumulative advantage distributions to account for the full range of citation counts. The hooked power law (Pennock, Flake, Lawrence, Glover, & Giles, 2002), with probability mass function $f(x) = A(B + x)^{-\alpha}$ (where A is chosen to ensure that the probabilities sum to 1, so the distribution has two free parameters, B and α). It is a discrete version of the Pareto type II distribution (Burrell, 2008) or, more precisely, the Lomax distribution (Lomax, 1954). The Lomax distribution is a Pareto type I distribution shifted to start at zero rather than 1. In the continuous case (i.e., the Lomax distribution) the additional parameter B is a scale parameter (larger values indicate a more spread out distribution) whilst α is a shape parameter. It asymptotically converges to a power law and fits sets of citation counts substantially better than the power law (Eom & Fortunato, 2011; Thelwall & Wilson, 2014a; Thelwall & Wilson, in press).

The lognormal distribution $\ln\mathcal{N}(\mu, \sigma^2)$ (Limpert, Stahel, & Abbt, 2001) with probability density function $f(x) = \left(1/x\sigma\sqrt{2\pi}\right) e^{-\left((\ln(x)-\mu)^2/2\sigma^2\right)}$ has a fundamentally different basis than the previous three distributions because it is not based on cumulative advantage processes, but is based on the related idea of data being multiplicative rather than additive (Limpert et al., 2001). More specifically, a continuous lognormal distribution can arise in the limit when positive, independent identically distributed random variables are multiplied together, although it is not clear how this relates to citation distributions. Like the hooked power law, it has two parameters. Its location and scale parameters are the mean and standard deviation, respectively, of the distribution of the natural logarithm of the variable. In its discretised version, the probability of a discrete value x is the integral of the above lognormal probability density function in the unit interval around x, $\left(1/B\right)\int_{x-0.5}^{x+0.5} f(x)\,\mathrm{d}x$. Here the constant $B = \int_{0.5}^{\infty} f(x)\,\mathrm{d}x$ compensates for the interval (0,0.5] that is not used for any integer value of x. This arbitrary removal of the interval (0,0.5] is aesthetically undesirable and it may be that alternative discretisation solutions that are almost equivalent in practice could be found that do not involve removing any of the original distribution. The location and scale terminology is used here for the corresponding discretised lognormal parameters, even though these do not fit the classical definitions of location and scale after discretisation. The discretised lognormal distribution will be denoted $\ln\overset{\cdots}{\mathcal{N}}(\mu, \sigma^2)$ and fits sets of citation counts from a single field and year much better than the power law and about as well as the hooked power law, depending on the particular field and year (Eom & Fortunato, 2011; Thelwall & Wilson, 2014a; Thelwall & Wilson, in press). It also fits the distribution of citation counts for articles from a single academic journal and year, for almost all Web of Science journals (Stringer, Sales-Pardo, & Amaral, 2010). Overall, the discretised lognormal distribution fits citation counts (with one added) for individual years less well than does the hooked power law for at least two thirds of Scopus fields (Thelwall, 2016b).

Although following standard practice in the literature, the discretisation of the lognormal distribution uses integration, the discretisation for the hooked power law uses an alternative approach of using the probability density function as a point mass function, with a correction constant. This approach was used for convenience of calculation but it seems likely these strategies have little difference in practice and do not affect the results.

There has been a claim that citation counts for sets of articles from a single subject and year may follow a discretised lognormal distribution with a variable location parameter but with a common scale parameter $\sigma = \sqrt{1.3} \cong 1.14$ (Evans, Kaube, & Hopkins, 2012; Radicchi, Fortunato, & Castellano, 2008; see also: Perianes-Rodríguez & Ruiz-Castillo, in press). This is not true for all fields, as can be shown by an analysis of related percentile indicators (Waltman, van Eck, & van Raan, 2012). It is also undermined by datasets with scale parameters varying from 1 to 1.3 (Radicchi et al., 2008) and from 1.0 to