Journal of Informetrics 10 (2016) 863–874

Contents lists available at ScienceDirect

Journal of Informetrics

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

Citation count distributions for large monodisciplinary journals

Mike Thelwall

Statistical Cybermetrics Research Group, University of Wolverhampton, UK

ARTICLE INFO

Article history:
Received 23 February 2016
Received in revised form 14 July 2016
Accepted 14 July 2016

Keywords:
Citation distributions
Discretised lognormal distribution
Lognormal distribution
Hooked power law
Citation analysis
Shifted power law

ABSTRACT

Many different citation-based indicators are used by researchers and research evaluators to help evaluate the impact of scholarly outputs. Although the appropriateness of individual citation indicators depends in part on the statistical properties of citation counts, there is no universally agreed best-fitting statistical distribution against which to check them. The two current leading candidates are the discretised lognormal and the hooked or shifted power law. These have been mainly tested on sets of articles from a single field and year but these collections can include multiple specialisms that might dilute their properties. This article fits statistical distributions to 50 large subject-specific journals in the belief that individual journals can be purer than subject categories and may therefore give clearer findings. The results show that in most cases the discretised lognormal fits significantly better than the hooked power law, reversing previous findings for entire subcategories. This suggests that the discretised lognormal is the more appropriate distribution for modelling pure citation data. Thus, future analytical investigations of the properties of citation indicators can use the lognormal distribution to analyse their basic properties. This article also includes improved software for fitting the hooked power law.

1. Introduction

Journals, authors, departments and universities are sometimes evaluated with the aid of indicators derived from citation counts, such as the Journal Impact Factor (JIF) (Garfield, 2006), the h-index (Hirsch, 2005) or the Mean Normalized Citation Score (MNCS) (Waltman, van Eck, van Leeuwen, Visser, & van Raan, 2011). The appropriateness of any indicator depends upon the properties of the data on which it is based (Wang, Song, & Barabási, 2013). For example, the JIF is imprecise because sets of citation counts are highly skewed and its calculation uses the arithmetic mean, which is inappropriate for skewed data sets – the geometric mean is a better option (Thelwall & Fairclough, 2015; Zitt, 2012).

Knowledge about the statistical distribution that best fits citation data can also aid theoretical understanding of how citations accrue in order to give context to interpretations of scores. This is important because the straightforward explanation that citations reflect relevant contributions from prior work (Merton, 1973) is not the full truth. Citations are affected by factors that are apparently unrelated to the quality of the cited work, such as the number of prior citations (Merton, 1968) as well as the nationality of the authors in collaborations (Glänzel, 2001), and document-based properties, such as the readability of the abstract (Gazni, 2011). Identifying the influence of such factors requires, at least in part, a statistical approach in order to detect tendencies that may not be evident in individual articles (Didegah & Thelwall, 2013; Onodera

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

http://dx.doi.org/10.1016/j.joi.2016.07.006
1751-1577/

& Yoshikane, 2015). For this, identifying the most appropriate statistical distribution is essential because analyses that use incorrect distributions can reach unjustified conclusions (Thelwall, 2016a; Thelwall & Wilson, 2014b).

It is impossible to logically or empirically *prove* that any given statistical distribution fits citation counts perfectly, which is a generic issue with mathematical models of real data (e.g., Burnham & Anderson, 2002, p. 20). Nevertheless, researchers can assess whether a distribution fits citation counts reasonably well and can also compare two or more distributions to check which fits best. Although some such attempts exclude articles with few citations and have found that the remaining articles fit single parameter distributions well, such as the power law and the Yule-Simon process (Brzezinski, 2015; Clauset, Shalizi, & Newman, 2009), this does not help citation analysis in practice because uncited and low-cited articles are rarely completely ignored by citation-based indicators (the h-index is an exception). When including low cited articles and uncited articles, the shifted/hooked power law (for background see: Pennock, Flake, Lawrence, Glover, & Giles, 2002) and discretised lognormal distributions (for continuous lognormal background see: Limpert, Stahel, & Abbt, 2001) fit substantially better (Eom & Fortunato, 2011; Evans, Kaube, & Hopkins, 2012; Radicchi, Fortunato, & Castellano, 2008; Thelwall, 2016a; Thelwall & Wilson, 2014a) and the lognormal distribution seems to have the wrong shape for subject categories (Thelwall, 2016b). The negative binomial distribution has also been suggested but does not fit as well as the hooked power law and discretised lognormal distributions (Low, Thelwall, & Wilson, 2015). Stopped sum models have been found to fit better on some data sets but have parameter estimation problems (Low et al., 2015), as does the hooked power law in a minority of cases. Models have also been proposed for predicting the growth of citations over time (Yao, Peng, Zhang, & Xu, 2014; Wu, Fu, & Chiu, 2014), with one suggesting that the lognormal may not be appropriate for individual articles with a long term total citation count above 8.5 (Wang, Song, & Barabási, 2013).

Although the hooked power law and discretised lognormal distribution seem to be the best distributions found so far for citation analysis, in terms of their fit to citation data and (relative) robustness of parameter estimation, each is preferable to the other in some subject areas but not in others. If uncited articles are excluded, then the hooked power law fits better than the discretised lognormal for 15 of out 20 varied Scopus categories for journal articles from 2004 (Thelwall & Wilson, 2014a). If no articles are excluded then a similar conclusion holds: the hooked power law is a better fit than the discretised lognormal for 22 out of 26 varied Scopus categories for journal articles from 2009, although the discretised lognormal fits better than the hooked power law for a larger percentage of categories for more recent articles (Thelwall, 2016a). The hooked power law has been found to fit better than the discretised lognormal for a set of ten physics journals, using different subsets of articles from 1950 to 2008 (Eom & Fortunato, 2011). The (not discretised) lognormal has also been shown to fit articles from 20 different Web of Science subject categories reasonably well (Radicchi, Fortunato, & Castellano, 2008). A limitation of the first two studies is that Scopus subject categories can include journals with very different specialisms within a field and if any of these specialisms have different citation properties then the overall subject category citation distribution will be impure. The third study investigated only one subject area, physics, and the fourth did not compare the hooked power law with the lognormal distribution.

A logical way around the problem of impure subject categories is to select single journals rather than entire subject categories. Non-general journals often target a specific field and hence should have a narrower focus than collections of journals within a subject. Some studies have adopted this strategy (e.g., using *Physical Review D:* Redner, 1998), but none have compared the discretised lognormal with the hooked power law without excluding low cited articles. Moreover, larger scale systematic studies across disciplines (e.g., not restricted to physics) are needed to make general conclusions possible. This study fills this gap by analysing a set of 50 different large non-general journals to see whether there is evidence that one of the two models tends to fit these purer distributions better than the other. This would give evidence that the better fitting distribution is the pure distribution whereas the other may only fit subject categories that are impure. Large amounts of data are needed to get accurate fits of statistical models and so the 50 non-general journals with the most articles indexed in Scopus were selected. The research question is therefore the following.

- RQ: Which out of the hooked power law and the discretised lognormal distribution is the best fitting for sets of citation counts from articles published in large non-general journals?

This article uses a similar main strategy to a previous paper (Thelwall, 2016b) but uses a new and different type of data set (journals rather than subject categories), has an improved method for fitting the hooked power law, and reaches different conclusions.

2. Methods

2.1. Data

To identify the journals with the most articles in Scopus, the query PUBYEAR IS 2006 AND DOCTYPE(ar) was run to match all journal articles from 2006. The year 2006 was chosen to give a decade to attract citations so that the citation distribution should be mature and there should not be a substantial difference between articles published early in the year compared with articles published late in the year. Scopus was selected in preference to the Web of Science for its larger coverage of academic literature (Li, Burnham, Lemley, & Britton, 2010; López-Illescas, de Moya-Anegón & Moed, 2008; Moed & Visser, 2008). The Refine option was then used to identify the 50 titles with the most matching articles. One conference proceedings (IEEE