Journal of Informetrics 10 (2016) 336–346

Contents lists available at ScienceDirect

Journal of Informetrics

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

The discretised lognormal and hooked power law distributions for complete citation data: Best options for modelling and regression

Mike Thelwall *

Statistical Cybermetrics Research Group, School of Mathematics and Computer Science, University of Wolverhampton, Wulfruna Street, Wolverhampton WV1 1LY, UK

ARTICLE INFO

Article history:
Received 22 September 2015
Received in revised form
22 December 2015
Accepted 22 December 2015
Available online 3 March 2016

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

ABSTRACT

Identifying the statistical distribution that best fits citation data is important to allow robust and powerful quantitative analyses. Whilst previous studies have suggested that both the hooked power law and discretised lognormal distributions fit better than the power law and negative binomial distributions, no comparisons so far have covered all articles within a discipline, including those that are uncited. Based on an analysis of 26 different Scopus subject areas in seven different years, this article reports comparisons of the discretised lognormal and the hooked power law with citation data, adding 1 to citation counts in order to include zeros. The hooked power law fits better in two thirds of the subject/year combinations tested for journal articles that are at least three years old, including most medical, life and natural sciences, and for virtually all subject areas for younger articles. Conversely, the discretised lognormal tends to fit best for arts, humanities, social science and engineering fields. The difference between the fits of the distributions is mostly small, however, and so either could reasonably be used for modelling citation data. For regression analyses the best option is to use ordinary least squares regression applied to the natural logarithm of citation counts plus one, especially for sets of younger articles, because of the increased precision of the parameters.

1. Introduction

The citation impact of sets of articles from journals (Chandy & Williams, 1994), researchers (Meho & Yang, 2007), research groups (van Raan, 2006), departments (Oppenheim, 1995), universities (Charlton & Andras, 2007) or even countries (Braun, Glänzel, & Schubert, 1985) are often compared with quantitative indicators on the basis that citations tend to reflect scientific impact. In addition, sets of articles with different properties are also sometimes analysed with the aid of citation counts, such as to test whether open access articles tend to be more frequently cited (Harnad & Brody, 2004; McCabe & Snyder, 2015) or whether collaboration tends to increase citations (Gazni & Didegah, 2011; Glänzel, Schubert, & Czerwon, 1999). These comparisons often employ standard indicators, such as the *h*-index (Hirsch, 2005) or field normalised citation counts (Waltman, van Eck, van Leeuwen, Visser, & van Raan, 2011). If part of a formal evaluation, then these indicators may be used to inform qualitative judgements. For more theoretical reasons, citation counts are sometimes analysed using statistical

* Tel.: +44 1902 321470; fax: +44 1902 321478.
E-mail address: m.thelwall@wlv.ac.uk

http://dx.doi.org/10.1016/j.joi.2015.12.007
1751-1577/

regression, where the independent variables are factors to be tested for a relationship with research impact, such as the number or nationality of the authors (Didegah & Thelwall, 2013; Onodera & Yoshikane, 2015; Yu, Yu, Li, & Wang, 2014). For both of these purposes, it is essential to understand the broad properties of sets of citation counts so that the indicators developed, and the regression approaches used, can be as powerful and appropriate as possible. This is particularly important because citation counts are known to be highly skewed and so many statistical techniques, including the sample mean, are not appropriate for them.

There have been many studies of citation count distributions since the early realisation that they were highly skewed, with a small number of articles attracting very high citation counts (de Solla Price, 1965). This skewed nature was thought to be due to preferential attachment processes in science (the Matthew effect), with articles attracting citations at least partly because they had already been cited (de Solla Price, 1976; Merton, 1968). This process is possible because researchers can find cited articles from other articles' reference lists, being cited can grant prestige, and modern digital libraries, such as Google Scholar, tend to list more cited articles above less cited articles. Nevertheless, articles attract citations much more rapidly than accounted for by the publication lifecycle and so preferential attachment cannot fully explain the pattern of growth in citations because few authors can cite an article using knowledge about how may citations it will have attracted when their work is published. To investigate this, one study has found evidence from physics that interest in an article decays exponentially over time (Eom & Fortunato, 2011).

Several studies have shown that citation counts tend to follow a power law distribution (or variants: van Raan, 2001) quite well, at least if articles with few citations are excluded (Clauset, Shalizi, & Newman, 2009; Garanina & Romanovsky, 2015; Redner, 1998). This is sometimes described as fitting the tail of the distribution. The hooked/shifted power has been show to fit better than the power law and about as well as the discretised lognormal distribution for citations to papers from 12 American Physics Society journals if articles with few citations are excluded (Eom & Fortunato, 2011). Some regression analyses have used the negative binomial distribution instead (e.g., Didegah & Thelwall, 2013; Hanssen & Jørgensen, 2015; Onodera & Yoshikane, 2015), on the basis that it is for discrete data and can cope with highly skewed data. It does not fit citation distributions as well as the discretised lognormal (Low, Thelwall, & Wilson, 2015), because of the heavy tailed nature of sets of citation counts (i.e., relatively many very high values within the data). Conversely, the Yule–Simon distribution, which is essentially a discrete version of the power law based upon assumptions about preferential attachment, seems to fit the tail of citation count distributions well (Brzezinski, 2015). Nevertheless, it unlikely to fit citation distributions well if zeros are included and it is shifted by 1 to allow zeros, because it is a strictly decreasing function and in some fields the mode is not zero (e.g., Developmental Biology: Radicchi, Fortunato, & Castellano, 2008).

For articles from a single subject and year, if uncited articles (only) are excluded, then the discretised lognormal (Evans, Hopkins, & Kaube, 2012; Radicchi et al., 2008) and hooked power law (Pennock, Flake, Lawrence, Glover, & Giles, 2002) (see below for descriptions of the distributions) fit substantially better than the power law distribution (Thelwall & Wilson, 2014a) and there do not seem to be any serious alternatives (excluding those with unstable parameters: Low et al., 2015). Uncited articles are typically removed when fitting most distributions because some of them, including the power law and discretised lognormal, are usually implemented in a way that excludes zeros, although logarithmic binning is a way of avoiding this problem (Evans et al., 2012). The omission of uncited articles is a problem since they are important for any full analysis of groups of articles. Hence, approaches are also needed to model the full range of citation counts.

One article has previously addressed this issue by comparing negative binomial and lognormal regression models for citation count data in a way that includes uncited articles, using 1337 journal articles published between 2001 and 2010 matching a Scopus title search for "knowledge management", and using as independent variables the number of years since publication and the number of references in the article. It also analysed a data set of articles from the online Information Research journal between 2001 and 2011, and using as independent variables the number of website views, Mendeley readers, and years since publication (Ajiferuke & Famoye, 2015). The negative binomial regression model was found to fit better than the discretised lognormal model but in both cases the data sets are relatively small, and the use of the publication year as an independent variable for a data set with multiple years is problematic because the relationship between publication year and citation counts is not simple (Adams, 2005; Eom & Fortunato, 2011) and hence may not be modelled well by regression. A previous study using simulations had shown that negative binomial regression had a tendency to identify non-existent relationships at a rate above the significance level set, showing that conclusions drawn from negative binomial regression are unsafe (Thelwall & Wilson, 2014b). Whilst this conclusion was not confirmed by the analysis of Information Research articles and knowledge management articles (Ajiferuke & Famoye, 2015), the number of dependant variable tested was too small and the nature of the datasets tested too restricted to give convincing evidence and so the use of negative binomial regression for citation data remains problematic.

This article uses a simple approach to model uncited articles with distributions that do not allow zeros: adding 1 to all citation counts before fitting a model. This simple transformation, which is a common way of dealing with zeros in a dataset that needs a log transformation (O'Hara & Kotze, 2010), allows the discretised lognormal distribution to be fitted to the full range of data and allows it to be compared against the main current alternative, the hooked power law. This transformation could perhaps be justified on the theoretical grounds that each article announces itself by its existence and is therefore a kind of self-citation. If data naturally fits the negative binomial distribution, however, then it is preferable to use negative binomial regression than to log transform the data before using regression (O'Hara & Kotze, 2010). This article compares