Journal of Informetrics 10 (2016) 110–123

Contents lists available at ScienceDirect

Journal of Informetrics

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

The precision of the arithmetic mean, geometric mean and percentiles for citation data: An experimental simulation modelling approach

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 21 September 2015
Received in revised form 5 December 2015
Accepted 5 December 2015

Keywords:
Scientometrics
Citation analysis
Research evaluation
Geometric mean
Percentile indicators
MNCS

ABSTRACT

When comparing the citation impact of nations, departments or other groups of researchers within individual fields, three approaches have been proposed: arithmetic means, geometric means, and percentage in the top *X*%. This article compares the precision of these statistics using 97 trillion experimentally simulated citation counts from 6875 sets of different parameters (although all having the same scale parameter) based upon the discretised lognormal distribution with limits from 1000 repetitions for each parameter set. The results show that the geometric mean is the most precise, closely followed by the percentage of a country's articles in the top 50% most cited articles for a field, year and document type. Thus the geometric mean citation count is recommended for future citation-based comparisons between nations. The percentage of a country's articles in the top 1% most cited is a particularly imprecise indicator and is not recommended for international comparisons based on individual fields. Moreover, whereas standard confidence interval formulae for the geometric mean appear to be accurate, confidence interval formulae are less accurate and consistent for percentile indicators. These recommendations assume that the scale parameters of the samples are the same but the choice of indicator is complex and partly conceptual if they are not.

1. Introduction

The European Union and some countries and regions produce periodic reports that compare their scientific performance with the world average or with appropriate comparators (EC, 2007; Elsevier, 2013; NIFU, 2014; NISTEP, 2014; NSF, 2014; Salmi, 2015). One of the points of comparison is typically (but not always: NIFU, 2014) the citation impact of the research conducted (Aksnes, Schneider, & Gunnarsson, 2012; Albarrán, Perianes-Rodríguez, & Ruiz-Castillo, 2015; King, 2004) on the basis that this is a likely pointer to its average scientific quality or influence. Citation data is often reported in conjunction with a range of other indicators, such as expenditure, publication volumes, patenting and PhD completions. Monitoring such data over time may give insights into the success of a science system and perhaps also of individual large scale policy initiatives or restructuring. Sets of departments within a field are also sometimes evaluated with the aid of quantitative data, and other groups of researchers may also be compared for theoretical reasons, such as to contrast the impacts of collaborative and

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

http://dx.doi.org/10.1016/j.joi.2015.12.001
1751-1577/

non-collaborative research (e.g., Abramo & D'Angelo, 2015a). Whilst Mendeley readership counts have been proposed as an alternative to citation counts for articles published in recent years (Fairclough & Thelwall, 2015a), they have not yet been used in practice and disciplinary differences (Haustein, Larivière, Thelwall, Amyot, & Peters, 2014) make them unsuitable for some fields.

The typical statistic used for comparing citation impact is some form of field-normalised citation count, such as the new crown indicator, or Mean Normalized Citation Score (MNCS) (Waltman, van Eck, van Leeuwen, Visser, & van Raan, 2011a; Waltman, van Eck, van Leeuwen, Visser, & van Raan, 2011b). This approach has problems with robustness and interpretation (Leydesdorff & Opthof, 2011) but is still used for the convenience with which sets of publications can be compared. At the level of an individual field and year, this indicator is equivalent (other than a common scalar multiple) to the arithmetic mean of the citation counts of the articles from that field and year. For articles from multiple fields, the arithmetic mean is calculated only after field normalisation by dividing each article by the average citation count for its field, document type and year. The arithmetic mean is not ideal, however, due to the skewed nature of citation data (de Solla Price, 1976). The median (Rousseau, 2005) is suitable for skewed data but is probably to crude to be useful in many contexts. The geometric mean (Zitt, 2012) is also appropriate for skewed data and is fine grained enough for comparisons. Percentile ranks are an alternative to direct citation counting (Schreiber, 2013; Schubert & Braun, 1996; Tijssen, Visser, & van Leeuwen, 2002), as well as for individuals and research groups, and when multiple indicators are needed (Bornmann, Leydesdorff, & Mutz, 2013). For comparing the citation impact of countries, the proportion of a nation's share of the world's top X% of articles can be calculated. If this share is higher than X% then the nation is above the world average for the calculation. Different values of X suggest different interpretations of the results. For example, if $X = 50$ then the percentile statistic corresponds to the nation's share of above average research, whereas if $X = 1$ then the percentile statistic corresponds to the nation's share of the world's very high impact research.

Given the choice of (appropriately field/year/document type normalised) arithmetic means, geometric means and percentiles for citation impact comparisons, the latter two are preferable on the grounds of the skewness of citation data. Nevertheless, it is not clear whether one of these is better than the other, and whether there are theoretical grounds to prefer one particular percentile limit. In the absence of specific policy requirements or a need to report multiple statistics, a logical way to select an indicator is to choose the one that is best able to distinguish between different countries. This would mean that the best indicator is the one that is the most precise relative to the spread of likely values for different countries. Whilst the arithmetic mean should perform poorly in this regard, a previous study with empirical data found that the geometric mean was more precise than the percentage of a country's articles in the top 10% most cited (Fairclough & Thelwall, 2015a; Fairclough & Thelwall, 2015b), but it did not check that this was universally true and did not check other percentiles (e.g., 50%, 1%). This article addresses this issue using a different approach, experimental simulation modelling, by comparing the relative precision of the arithmetic mean, geometric mean and percentiles with a range of different parameters.

2. Modelling citation distributions

If the citation counts of all articles from a single field and year are examined, they typically exhibit a strong pattern that approximates a known statistical distribution. Several different distributions have been suggested as the most suitable.

2.1. Alternative citation distribution models

Articles from the same subject and year seem to fit the discretised lognormal distribution reasonably well (Evans, Kaube, & Hopkins, 2012; Thelwall & Wilson, 2014a) and better than most distributions tested so far. In particular, the discretised lognormal fits at least as well as the power law in almost all cases (Brzezinski, 2015) and even for the exceptions the power law only fits the tail of citation data well (i.e., ignoring articles with few citations), which excludes its use for modelling entire citation distributions.

Count data distributions are a more natural choice for citation counts because they directly model discrete data. Of these, the negative binomial distribution (Hilbe, 2011), or zero inflated, truncated or hurdle variants (Chen, 2012; Didegah & Thelwall, 2013), seem to fit citation data better than most alternatives tried, but the discretised lognormal fits citation data better than the negative binomial (Low, Wilson, & Thelwall, 2015) probably because the negative binomial does not model the very high values well. In other words, a heavy tailed distribution (Clauset, Shalizi, & Newman, 2009), such as the lognormal or power law, is needed to account for a small number of very high citation counts.

There is some evidence of a modified negative binomial stopped sum distribution fitting slightly better than the lognormal in some cases but this is impractical for use in citation analysis because of the difficulty in accurately estimating the distribution parameters (Low et al., 2015). The hooked (or shifted) power law also fits citation data approximately as well as the discretised lognormal (Eom & Fortunato, 2011; Thelwall & Wilson, 2014a) but also has problems with inaccuracy of parameter estimation (Thelwall & Wilson, 2014b). Whilst parameter estimation is not directly used in the modelling here, this difficulty suggests that it would be difficult to accurately model distributions for a predefined mean and standard deviation, as needed here.

Another exception is the Yule distribution, which is for discrete data and has been shown to fit citation data approximately as well as the discretised lognormal overall and slightly better for some sets of articles, although only after excluding articles with few citations (Brzezinski, 2015). For the current article, the option of excluding uncited articles or articles with few