Journal of Informetrics 11 (2017) 530–540

Contents lists available at ScienceDirect

Journal of Informetrics

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

Regular article

The accuracy of confidence intervals for field normalised indicators

Mike Thelwall*, Ruth Fairclough

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

ARTICLE INFO

Article history:
Received 8 February 2017
Received in revised form 10 March 2017
Accepted 11 March 2017

Keywords:
Citation analysis
Field normalised citation indicators
Confidence intervals

ABSTRACT

When comparing the average citation impact of research groups, universities and countries, field normalisation reduces the influence of discipline and time. Confidence intervals for these indicators can help with attempts to infer whether differences between sets of publications are due to chance factors. Although both bootstrapping and formulae have been proposed for these, their accuracy is unknown. In response, this article uses simulated data to systematically compare the accuracy of confidence limits in the simplest possible case, a single field and year. The results suggest that the MNLCS (Mean Normalised Log-transformed Citation Score) confidence interval formula is conservative for large groups but almost always safe, whereas bootstrap MNLCS confidence intervals tend to be accurate but can be unsafe for smaller world or group sample sizes. In contrast, bootstrap MNCS (Mean Normalised Citation Score) confidence intervals can be very unsafe, although their accuracy increases with sample sizes.

1. Introduction

Citation indicators that estimate the average citation rate of articles produced by a group are widely used in research assessment and for ranking universities, countries and departments (Aksnes, Schneider, & Gunnarsson, 2012; Albarrán, Perianes-Rodríguez, & Ruiz-Castillo, 2015; Braun, Glänzel, & Grupp, 1995; Elsevier, 2013; Fairclough & Thelwall, 2015). For example, in the U.K., they have been proposed for the national Research Excellence Framework (REF) to cross-check peer review judgements (Stern, 2016). If average citation indicators are to be used in such a role, then they must be calculated in a fair way and accompanied with an estimate of statistical variability so that strong conclusions are not drawn from small or biased differences.

Field normalised citation impact indicators adjust average citation counts for the field and year of publication to allow fair comparisons of citation impact between sets of articles that were published in different combinations of fields and years. For example, if group A published 100 medical humanities articles in 2014 with an average of 4 citations each but group B published 100 oncology articles in 2013 with an average of 30 citations each then it is not clear which had generated the most impactful research. Group B has two advantages: its articles are older, with longer to attract citations, and it publishes in an area where citations accrue rapidly. A field normalised indicator may divide by the average number of citations for the field and year so that the normalised counts are 1 if the average citation impact is equal to the world average. After this,

* Corresponding author.
E-mail addresses: m.thelwall@wlv.ac.uk (M. Thelwall), r.fairclough@wlv.ac.uk (R. Fairclough).

http://dx.doi.org/10.1016/j.joi.2017.03.004
1751-1577/

it would be reasonable to compare the field normalised values of A and B. Nevertheless, confidence intervals or statistical hypothesis tests are needed to be able to judge whether the difference between A and B is likely to reflect an underlying trend rather than a random fluctuation of the data.

The use of statistical inference or confidence intervals to compare the average citation impact is uncommon within scientometrics and there are arguments against it, such as a lack of clarity about what exactly is being sampled (Waltman, 2016). Statistical inference is typically used when data is available about a sample whereas in scientometrics, relatively complete sets of publications are normally analysed and so there is no necessity to infer population properties from a sample, at least in the obvious sense. Nevertheless, research is a social process and therefore each citation is the product of activities that are affected by processes that can be thought of as random in the sense of not predictable in advance (Williams & Bornmann, 2016). The exact citation count of an article is therefore partly a result of chance factors rather than just the quality or value of an article. For example, if two essentially identical papers are published at the same time then one may become more highly cited than the other for spurious reasons, such as the prestige of the publishing journal (Larivière & Gingras, 2010), or the extent to which the citing literature is covered by the database used for the counts (Harzing & Alakangas, 2016; Table 3 in: Kousha & Thelwall, 2008). Thus, it seems impossible to regard citation counting as precisely measuring the impact of publications and it seems better to regard it instead as an inaccurate estimate (see the similar argument in: Waltman & Traag, 2017). Moreover, the purpose of research evaluation is often to make decisions about future funding allocations or strategies based on past performance. In this context, the exact citation count of a paper is less important than the underlying capacity of a group to produce impactful research. Each article produced by a group can also be thought of as the product of both the underlying research power of the group and chance factors that affect the value of each paper produced. These chance factors include creativity-related factors that are internal to the researchers (Simonton, 2004) as well as external factors that are partly outside of their control, such as whether external technical or social developments turn their topic into one of societal importance (e.g., the recent rise in the importance of Arabic natural language processing and Middle Eastern studies). Thus, for example, Nobel Prize winners may occasionally produce rarely-cited research even if most of their output has high impact. In both contexts, statistical inference is reasonable and aligns with the standard social sciences practice of treating the situation as having an apparent population of plausible outcomes from the known parameters (Berk, Western, & Weiss, 1995; Bollen, 1995).

There are two alternative reasonable strategies to generate confidence limits. The parametric strategy assumes that the data follows a specific statistical distribution and then derives confidence limit formulae from an analysis of this distribution. The bootstrapping strategy resamples from the existing data, with replacement, and then calculates confidence limits in order that 95% (say) of the resampled indicator values fall within them (Efron & Tibshirani, 1986). Neither approach is perfect. The parametric strategy is reliant upon the distribution assumption and may also involve additional assumptions, such as that the distribution of a discretised distribution is like the continuous distribution that it was derived from. Bootstrapping is also unreliable for many data distributions and tasks (Hall, 1992; Hillis & Bull, 1993) and seems to be particularly unsuited to highly skewed data sets, such as those based on untransformed citation counts. In this context, it is not clear whether bootstrapping or parametric formulae are preferable for any given indicator and whether the optimal choice depends on basic properties of the data.

This article assesses the accuracy of bootstrapping for the calculation of confidence intervals for two field normalised average citation indicators. The Mean Normalised Citation Score (MNCS) (Waltman, van Eck, van Leeuwen, Visser, & van Raan, 2011a, Waltman, van Eck, van Leeuwen, Visser, & van Raan, 2011b), is used in the Leiden university ranking (Waltman et al., 2012), and the Mean Normalised Log-transformed Citation Score (MNLCS) (Thelwall, 2017) is a more recent variant. This study focuses on a single field and year for pragmatic reasons: to allow an exploration of the impact of the mean and standard deviation without generating unmanageably many results from experiments with multiple fields and/or years. Confidence interval formulae have been proposed for the MNLCS and so these are also assessed for accuracy at the same time. Although there are many other field normalised indicators, these represent two of the main variants, with MNCS being well known and MNLCS being designed as a logical extension to deal with skewing in citation count data. One recent quite different indicator is the Relative Citation Ratio (RCR) (Hutchins, Yuan, Anderson, & Santangelo, 2016) but this is not included because it is not clear that it is relevant outside of biomedical science and its design makes bootstrapping highly complex because a paper's citations and the impact factors of the publishing journals for their references need to be modelled.

2. Background

The parametric strategy in statistics requires an assumption about the distribution of a citation data set. It has been known for a long time that citation counts diverge substantially from the normal distribution (de Solla Price, 1965) and that the power law is a much better fit if articles with few citations are ignored (Clauset, Shalizi, & Newman, 2009). Since field normalised indicators do not omit rarely cited articles and these often form the clear majority within a collection, the power law is an inappropriate distribution (Thelwall & Wilson, 2014a). Instead, both the discretised lognormal distribution and the hooked power law are reasonable fits for most sets of articles from a single field (or large monodisciplinary journal) and year (Radicchi & Castellano, 2012; Thelwall, 2016a, 2016b). Many alternative distributions and approaches have also been tested on the full range of citation counts, but none are clearly better than the discretised lognormal or hooked power law and most are worse, when fully tested. Appropriate stopped sum models have been found to fit citation data reasonably, but there is limited evidence of this and their parameters are too unstable to be useful in practice (Low, Wilson, & Thelwall,