Journal of Informetrics 10 (2016) 43–47

Contents lists available at ScienceDirect

Journal of Informetrics

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

Expected number of citations and the crown indicator

Lawrence Smolinsky*

Department of Mathematics, Louisiana State University, Baton Rouge, LA 70803, USA

ARTICLE INFO

Article history:
Received 30 July 2015
Accepted 28 October 2015

Keywords:
Bibliometric indicator
Crown indicator
Field normalization
Research performance assessment

ABSTRACT

The mean normalized citation score or crown indicator is a much studied bibliometric indicator that normalizes citation counts across fields. We examine the theoretical basis of the normalization method and, in particular, the determination of the expected number of citations. We observe a theoretical bias that raises the expected number of citations for low citation fields and lowers the expected number of citations for high citation fields when interdisciplinary publications are included.

1. Introduction

The mean normalized citation score (MNCS) is a bibliometric indicator of research performance developed at The Centre for Science and Technology Studies, Leiden University. Bibliometric indicators are important as they are used to compare research performance for individuals, research groups, institutions, and countries. A recent survey of bibliometric indicators is included in Waltman (2015). The MNCS bibliometric indicator normalizes citation counts for differences between fields while keeping year and document type fixed. For a set of n publications, it is defined as

$$\text{MNCS} = \frac{1}{n}\sum_{i=1}^{n}\frac{c_i}{e_i} \tag{1}$$

where c_i is the number of citations to the ith publication and e_i is the expected number of citations to the ith publication.

The MNCS is an indicator that normalizes by dividing by an expected value. The determination of these normalizing variables has been heavily discussed for both the MNCS indicator and its predecessor, the mean field citation score/citations per publication (CPP/FCSm). Leydesdorff and Opthof (2011) criticize the MNCS while Lundberg (2007) and Opthof and Leydesdorff (2010) included a discussion in the context of the CPP/FCSm. There was a response by van Raan, van Leeuwen, Visser, van Eck, and Waltman, 2010 and a proposal in Waltman, van Eck, van Leeuwen, Visser, and van Raan (2011). Much of the discussion deals with the issues of validity of field classification and a proper reference set for the normalization. Although there was a suggestion to ignore fields and use reference counts, citations counts were generally accepted. Ultimately the discussion by Waltman et al. of the MNCS indicator used field classification and citation counts.

The citation counts c_i in Eq. (1) are well-defined as variables by choosing to use the values from the Web of Science or Scopus. This is not the case for the expected number of citations e_i. Waltman et al. (2011) write, "We also determine for each publication its expected number of citations. The expected number of citations of a publication equals the average number of citations of all publications of the same document type (i.e., article, letter, or review) published in the same field and in the same year." Translating this intent into mathematics is not straightforward. This note addresses the expected number

* Tel.: +1 225 578 1570.

http://dx.doi.org/10.1016/j.joi.2015.10.007
1751-1577/

of citations from a theoretical perspective and, in particular, critiques the methodology of Waltman et al., when there are overlapping fields. Their method introduces a bias that hurts researchers in low citation fields and benefits researchers in high citation fields. We also discuss that altering the normalizing variables does not violate the uniqueness statement of Waltman et al.'s Theorem 1.

2. The expected value and properties of the MNCS

Waltman et al. show that MNCS is a bibliometric indicator of average performance of a set of publications that satisfies two properties: consistency and homogeneous normalization. Both the properties of consistency and homogeneous normalization involve comparisons between sets of publications of the same size, say n, at a time. The property of consistency is also known as Independence (Bouyssou & Marchant, 2011).

We use the notation from Waltman et al., where a publication is represented as an ordered pair of numbers (c,e), c is the number of citations to the publication and e the expected number of citations. A collection of n publications is then represented as a set of n ordered pairs and a bibliometric indicator is a nonnegative function on the sets of ordered pairs.

A bibliometric indicator of average performance is said to have the property of consistency if $f(S_1) \geq f(S_2) \Leftrightarrow f\left(S_1 \cup \left\{(c,e)\right\}\right) \geq f\left(S_2 \cup \left\{(c,e)\right\}\right)$ for all sets of publications S_1 and S_2 of the same size and all publications (c,e) in the complement of $S_1 \cup S_2$. In other words, the relative ranking of two sets of publications does not change by the addition of the same publication to both sets.

It is important that the sets are of equal size. For example, consider Eq. (1) for two sets of publications all with the same e_i value e. A set S_1 of one publication with 1 citation will be evaluated with a lower value than a set S_2 of 10 publications where one publication has two citations and the nine others have one citation. However, the addition of a publication c with 20 citations to both sets will switch the evaluation having a more dramatic effect on the smaller group, i.e., $f(S_1) < f(S_2)$ but $f(S_1 \cup \{(c,e)\}) > f(S_2 \cup \{(c,e)\})$.

The second property of homogeneous normalization precisely defines the indicator in the case that every publication in the set has the same e coordinate. If $S = \{(c_1,e), (c_2,e), \ldots, (c_n,e)\}$, then homogeneous normalization requires

$$f(S) = \frac{1}{e}\sum_{i=1}^{n}\frac{c_i}{n}$$

The set of publications is considered homogeneous since they all have the same expected number of citations. The indicator is required to average the citations and then weights them by dividing by e.

Waltman et al. furthermore state (Theorem 1) that MNCS is the unique bibliometric indicator of average performance to satisfy these conditions. However, a little caution is required. The proof of Theorem 1 uses the notion 'bibliometric indicator' simply as collection of functions (parameterized by n) of two variables without meaning assigned to the two variables. There is no requirement of the meaning of each variable or a formula for computing a variable. In other words, the statement of Theorem might more properly be stated as: Eq. (1) is the unique function f: $(\mathbf{N}_0 \times \mathbf{R}_+)^n \to \mathbf{R}$ satisfying the properties of homogeneous normalization and consistency of the average performance.

The c_i's may be taken as the number of citations to an article on the Web of Science or Scopus. However, the e_i's may have a number of variations each of which define a function that satisfies Theorem 1. In fact, if the e_i's are arbitrary positive numbers with $e_i = e_j$ whenever publications i and j are in the same field, then the collection of e_i's will allow for an MNCS indicator that satisfies the properties of homogeneous normalization and consistency. For example, if e_i is set equal to 1 for all i, then the indicator is just the average number of citations for articles in the collection and satisfies homogeneous normalization and consistency. Similarly, one may replace the expected number of citations with the median number of citations for each field and obtain another indicator satisfying homogeneous normalization and consistency.

In addition to the properties of homogeneous normalization and consistency, Waltman et al. discuss one further property: "A nice property that we would like the MNCS indicator to have is that the indicator has a value of one when calculated for the set of all publications published in all fields." This condition will be included in later computations, but was not a requirement of Theorem 1. We will call it the unity property.

3. Computing the e_i's in the example from Waltman et al. (2011)

Waltman et al. point out that a reasonable definition of e_i is straightforward if each article in the determining set has a uniquely defined field, i.e., just the average. We look at their example in their Section 6 (*How to handle overlapping fields*) when articles do not have a single classification and determination of e_i is not straightforward. We reproduce their Table 9 (Overview for each publication of the field in which it has been published and the number of citations it has received) as our Table 1.

The expected number of citations is determined by the number of citations in the field as determined by the chart. In any method we consider, $e_4 = 6$. Publication 4 is the only Publication in the table in field Z. Also note $e_1 = e_2$ since both these publications are only in field X.

We look at three methods for computing e_i's for potential MNCS indicators. In each method the MNCS indicator does have a value of one for the entire collection. Before proceeding to the methods, we give a word about the harmonic mean