Journal of Informetrics 7 (2013) 774–783

Contents lists available at ScienceDirect

Journal of Informetrics

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

The *h*-index: A case of the tail wagging the dog?

Quentin L. Burrell*

Centre for R&D Monitoring (ECOOM), KU Leuven, Waaistraat 6, Leuven, Belgium

ARTICLE INFO

Article history:
Received 14 May 2013
Received in revised form 12 June 2013
Accepted 13 June 2013

Keywords:
Citation distribution
Publication-citation process
Geometric distribution
Non-linear equations
Quasi *h*-index
h-Index profile

ABSTRACT

From the way that it was initially defined (Hirsch, 2005), the *h*-index naturally encourages focus on the most highly cited publications of an author and this in turn has led to (predominantly) a rank-based approach to its investigation. However, Hirsch (2005) and Burrell (2007a) both adopted a frequency-based approach leading to general conjectures regarding the relationship between the *h*-index and the author's publication and citation rates as well as his/her career length. Here we apply the distributional results of Burrell (2007a, 2013b) to three published data sets to show that a good estimate of the *h*-index can often be obtained knowing only the number of publications and the number of citations. (Exceptions can occur when an author has one or more "outliers" in the upper tail of the citation distribution.) In other words, maybe the main body of the distribution determines the *h*-index, not the wild wagging of the tail. Furthermore, the simple geometric distribution turns out to be the key.

1. Introduction

When Jorge Hirsch first introduced the index that bears his name (Hirsch, 2005) he could hardly have foreseen the remarkable influence it would have in the world of informetrics and far beyond, even being included nowadays as one of the standard statistics for an author reported in, for instance, Web of Science (WoS®), Scopus® and Publish or Perish (PoP®). Its main attraction is the simplicity of its rationale – look at an author's most highly cited papers, with how many to be considered being determined by how many citations they attract. The formal definition is by now very well known.

Definition 1. If an author has N publications/papers, the h-index is the largest integer h such that h of the publications have at least h citations and the remaining $N-h$ have at most h citations.

The *h*-index is a purely empirical measure and, although Hirsch (2005) and Burrell (2007a) proposed, respectively, a deterministic and a stochastic model for the way the index develops in time, neither produced a theoretical formula for the index itself. The first proper formula – i.e. something that could be calculated from the model parameters – was given by Egghe and Rousseau (2006), later extended in Egghe and Rousseau (2012) and based on the assumption of the simple Pareto (continuous Lotka) model for the citation distribution. However, Burrell (2013a) has presented empirical examples that cast doubt on these formulae, suggesting that the problem is that the assumed so-called Lotkaian model for the entire distribution is not appropriate in this context. Here we will demonstrate that a formula based on a particularly simple case of the stochastic model of Burrell (2007a, 2013b) can often give surprisingly good results. Our approach is essentially the same as that of Egghe and Rousseau (2006, 2012) but based on a different distributional model.

* Correspondence address: 119 Friary Park, Ballabeg, Isle of Man IM9 4EX, via United Kingdom.
E-mail address: quentinburrell@manx.net

1751-1577/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.joi.2013.06.004

2. The empirical approach

The obvious way to determine the h-index is to rank the author's papers in decreasing order of number of citations and then identify the rank satisfying the definition. Graphically, one would plot rank on the horizontal scale and number of citations on the vertical scale – a rank–size plot – and look for the intersection with the "$x = y$" or equality line. See, for instance, Fig. 1 of Hirsch (2005) or Fig. 2.2 of Egghe (2010). In what follows, we shall instead adopt a traditional size–frequency approach, but focussing on the tail frequency distribution and plotting the number of citations on the horizontal scale and the number of papers with at least this number of citations on the vertical scale. A moments thought shows that, in comparison with the rank–size approach, this is just a matter of interchanging the major axes, or reflecting in the $x = y$ line, so that the h-index is given by the intersection of the tail frequency plot with the equality line. This equivalence between the standard frequency approach and the rank approach carries over, of course to theoretical models. Indeed this duality between the two approaches lies at the heart of much of the work in so-called Lotkaian informetrics, see in particular the standard text of Egghe (2005) and, in the context of the h-index, Egghe and Rousseau (2006, 2012) and Egghe (2010).

3. The probabilistic citation distribution

The stochastic model of Burrell (2007a, 2013b) is concerned with the development of an author's citation distribution over time. In all that follows, however, we will only be concerned with the citation distribution at some particular point in time so the time parameter will be suppressed. The probabilistic approach assumes that the numbers of citations to an author's publications at any time are equivalent to a random sample from some probability distribution. Thus if X = number of citations to a typical publication, then

$p_r = P(X = r)$ = probability that the publication has r citations, $r = 0, 1, 2, \ldots$ for some probability distribution on the non negative integers.

If we write $Z(r)$ = number of papers with exactly r citations, $r = 0, 1, 2, \ldots,$

then, given the number of papers that the author has published, say N, the conditional distribution of $Z(r)$ is binomial, corresponding to N "trials" (the number of papers) with probability $p_r = P(X = r)$ of "success" (getting r citations) at each trial.

It follows that

$$Z(r) \sim Bin(N, p_r) \quad \text{and} \quad \text{hence} \quad E[Z(r)] = Np_r = NP(X = r)$$

Thus the distribution of expected citation frequencies is just the underlying probability distribution scaled by the number of publications.

For the h-index we are interested in the number of papers with high numbers of citations so we define

$$N(r) = \text{Number of papers with at least} \quad r \quad \text{citations} = \sum_{k=r}^{\infty} Z(k), \quad r = 0, 1, 2, \ldots$$

Thus $N(0) = N$ and the empirical h-index is the integer h satisfying

$$h = \max\{n : n \le N(n)\}$$

When we turn to the theoretical framework, we do not have the observed number of papers, only the expected number. Thus we consider

$$E[N(r)] = E\left[\sum_{k=r}^{\infty} Z(k)\right] = N\sum_{k=r}^{\infty} P(X = k) = N\phi(r)$$

where we have written $\phi(r) = P(X \ge r)$ for what is usually called the tail distribution function but in some contexts is referred to as the survivor or reliability function. Note that this is just the underlying tail distribution scaled up by a factor N.

By analogy to the empirical definition we have that the theoretical h-index is the integer h satisfying

$$h = \max\{n : n \le E[N(n)]\} = \max\{n : n \le N\phi(n)\}$$

and note that this is well defined since $E[N(n)]$ decreases with increasing n and $E[N(0)] = N$.

Remark. In the empirical case we have that h satisfies $h = N(h)$ where $N(n)$ is necessarily an integer. Although the theoretical index is also an integer, since expected frequencies are not necessarily integers, we can have $h < E[N(h)]$ so that h would naturally be defined as the integer part of $E[N(h)]$.

In essence, then, to determine the theoretical h-index we need to solve the equation

$$h = N\phi(h) \quad \text{or} \quad \phi(h) = \frac{h}{N} \tag{1}$$

When we come to solving Eq. (1) in a theoretical framework, we are treating h as a real variable, not necessarily an integer. This leads to the following: