Jointly published by Akadémiai Kiadó, Budapest
and Springer, Dordrecht

Scientometrics, Vol. 67, No. 2 (2006) 315–321
DOI: 10.1556/Scient.67.2006.2.12

Short communication

On the h-index – A mathematical approach to a new measure of publication activity and citation impact

WOLFGANG GLÄNZEL[a,b]

[a] *Steunpunt O&O Statistieken, K. U. Leuven, Leuven (Belgium)*
[b] *ISSRU/IRPS, Hungarian Academy of Sciences, Budapest (Hungary)*

Introduction

The traditional bibliometric toolset is based on simple statistical functions including means, relative frequencies and quantiles. Nonetheless, these publication- and citation based statistics proved to be robust and useful output measures of activity and performance of scientific research. In contrast, the evaluation at the micro level, above all, the assessment of the research performance of individual scientists remained most problematic. The reason is twofold, on one hand, a sufficiently large publication output produced in a relatively short time span is necessary to obtain statistically reliable indicators and, on the other hand, research productivity and citation impact are not necessarily correlated variables. That means, even if these statistical methods can be applied a set of different cases has still to be examined, namely how low/high publication activity relates to low/high citation impact. In order to overcome these shortcomings bibliometricians are faced with in micro-level studies, HIRSCH (2005) has recently suggested a new indicator for the assessment of the research performance of individual scientists. This measure – called *h-index* – is designed for application at the micro level, and measures both publication activity and citation impact. According to his definition, "a scientist has index h if h of his or her N_p papers have at least h citations each and the other $(N_p - h)$ papers have $\leq h$ citations each".

Hirsch's idea, which appears to be tracked back to Sir Arthur Eddington (EDWARDS, 2005), has immediately found interest in the public (BALL, 2005), and received positive

Received: December 1, 2005

Address for correspondence:
WOLFGANG GLÄNZEL
Steunpunt O&O Statistieken, K. U. Leuven, Dekenstraat 2, B-3000 Leuven, Belgium
E-mail: Wolfgang.Glanzel@econ.kuleuven.ac.be

0138–9130/US $ 20.00

reception both in the physics community (DINIZ BATISTA et al., 2005; POPOV, 2005) and the scientometrics literature (BORNMANN & DANIEL, 2005; BRAUN et al., 2005, GLÄNZEL, 2006). The latter two papers have shown that the h-index correlates with other bibliometric indicators of 'significance'. This could be confirmed by VAN RAAN (2005) as well, however he stressed that scientific performance can hardly be expressed simply by one indicator alone.

There is much uncertainty in the interpretation of the h-index. This is only in part a consequence of lacking experience with this indicator. Unlike the traditional bibliometric indicators, the mathematical-statistical background of the h-index has not yet been studied. In the case of means, percentages or quantiles simple but robust statistical tests can be used to answer the question of whether the indicator value of a sample significantly differs from that of another one or from a given reference standard. The Welch-test applied by SCHUBERT & GLÄNZEL (1983) to the comparison of citation means might serve just as an example. However, the h-index, as it is defined, heavily depends on the 'sample size' itself.

Hirsch himself attempted to analyse the basic properties of his measure. In particular, he applied a simple deterministic model with constant annual growth of both publications and citations, and developed some rather arbitrary reference standards for the evaluation of scientists in physics (cf., HIRSCH, 2005). He developed an *h'-index* ($h' = 1/(1/P+1/C)$) on basis of his linear growth model resulting in a derivative of the *harmonic mean* of the number of publications (P) and the number of citations (C) received by them. Thus he uses a simple composite of publication and citation counts as proxy for his h-index.

In the present study we will analyse the basic properties of the h-index on basis of a probability distribution model which is in fact the most important distribution family used in bibliometrics, namely the Paretian distributions. In order to facilitate the application of this model to the h-index, we have slightly modified the definition by HIRSCH (2005). "A scientist has index h if h is the largest number of his/her n papers having received at least h citations each". If we translate this into mathematical parlance, we can represent the individual citation rates of the n papers by a given author by an ordered citation-rate sample. In particular, we consider a set of n papers, of which the i-th one has received exactly X_i citations. The elements of the ordered sample $\{X_i^*\}_{i=1}^n$ are then formed by the ranked observations, namely $X_1^* \geq X_2^* \geq \ldots \geq X_i^* \geq \ldots \geq X_n^*$, $i = 1, 2, \ldots , n$ with X_1^* (X_n^*) denoting the number of citations received by the most (least) cited paper. According to the original definition by Hirsch the *h-index* is defined as $h \leq X_h^*$, provided $h > X_{h+1}^*$. In order to solve the problem $X_{h+1}^* = X_h^* = h$ we apply the modified definition which reads in mathematical formulation as follows. $h = \max \{j : X_j^* \geq j\}$. The h-index can obviously be applied to any set of papers. Since $h \leq n$, we can even define $h = 0$ for completely inactive authors, i.e., for $n = 0$. Now the question arises of how can the empirical h-index be linked to a model, that is, how can Hirsch's problem be solved *theoretically.* In the following section we will give a solution for a special type of bibliometric distributions, namely for the family of Paretian distributions.

The theory of h-index: A mathematical etude

The most convenient way is using Gumbel's characteristic extreme values. Before we introduce these statistics we briefly summarise the basic concepts and notations.

Let X be a random variable. In our case X represents the citation rate of a paper. The probability distribution of X is denoted by $p_k = P(X = k)$ for every $k \geq 0$ and the cumulative distribution function is denoted by $F(k) = P(X < k)$. Put $G_k = G(k) := 1–F(k) = P(X \geq k)$. Gumbel's r-th characteristic extreme value (u_r) is then defined as $u_r := G^{-1}(r/n) = \max\ \{k\colon G(k) \geq r/n\}$, where n is a given sample with distribution F. The theoretical h-index (H) can consequently be defined as $H := \max\ \{r\colon u_r \geq r\} = \max\ \{r\colon \max\ \{k\colon G(k) \geq r/n\} \geq r\}$. If there exists such index r so that $u_r = r$ then we have obviously $H := r$ and we can write $H := u_H$.

In the following, we will study two special cases, namely (1) discrete Paretian distributions with finite expectation and (2) the Price distribution. These two distributions cover most distributions used for modelling publication activity and citation processes.

Discrete Paretian distributions with finite expectation

We say, that the distribution of a random variable X is Paretian if it asymptotically obeys Zipf's law, i.e, if $\lim_{k\to\infty} G_k \cdot k^{-\alpha}$ = constant. Asymptotically Pareto distributed random variables obviously meet this definition since $p_k = P(X = k) \approx d(N+k)^{-(\alpha+1)}$ if $k >>1$; $\alpha > 1$, where N and d are positive constants. In what follows we will deal with this family of distributions. For $k >>N$ we obtain $p_k = P(X = k) \approx d\cdot k^{-(\alpha+1)}$ and $G_k = P(X \geq k) \approx d_1\cdot k^{-\alpha}$ where d_1 is a positive constant. Hence we have,

$$E(X) = \sum_{k=0}^{\infty} kp_k = \sum_{k=0}^{\infty} G_k < \infty \text{ if } \alpha > 1.$$

By elementary manipulation of the cumulative distribution function we obtain the following approximation from the above definition of Gumbel's r-th characteristic extremes.

$$u_r \approx c_1\cdot(n/r)^{1/\alpha},$$

where c_1 is a positive constant. Applying the Hirsch condition to this approximation results in the property $H = u_H \approx c_1\cdot(n/H)^{1/\alpha}$, if $n >>1$. Hence we have $H \approx c_2\cdot n^{1/(\alpha+1)}$, if $n >>1$, where $c_2 = c_1^{\alpha/(\alpha+1)}$ is a positive constant. In verbal terms, the h-index is approximately proportional to the (α+1)-th root of the number of publications. Further properties for Paretian distributions can be derived from the analysis of Gumbel's characteristic extreme values. The first property is straightforward from the definition of the Gumbel's extreme values. We have $u_1/u_r \approx r^{1/\alpha}$ for any finite r, if $n >>1$. Thus the value of each observation is inversely proportional to the 1/α-th power of its rank (Zipf property). If we apply the Zipf property to the h-index, we immediately obtain $u_1 \approx H^{(1+1/\alpha)} \approx c_1 \cdot n^{1/\alpha}$, if $n >>1$. The highest citation rate is thus approximately equal to the (1+1/α)-th power of the h-index if the number of underlying papers is very large.

Now we introduce the terms *h-papers* (H_p) and *h-citations* (H_c). H-papers are papers meeting the Hirsch criterion, that is, papers that have received at least H citations each. H-citations are the total of citations received by h-papers. Finally, we will give an approximation for h-citations. We have

$$H_c = n\sum_{k=0}^{\infty} kp_k \approx n\sum_{k=0}^{\infty} ck^{-\alpha} \quad \text{if } k >>1,\ \alpha >1.$$

Hence,

$$H_c \approx cn \int_{x=H}^{\infty} x^{-\alpha} = \frac{cnH^{1-\alpha}}{\alpha-1} = c^{*} n \cdot n^{\frac{1-\alpha}{1+\alpha}} = c^{*} n^{\frac{2}{\alpha+1}} = c^{**} H^2 .$$

The number of h-papers is thus proportional to the square of the h-index if $\alpha > 1$ and k is large.

The Price distribution

The Price distribution which is a special Paretian distribution has been introduced by GLÄNZEL & SCHUBERT (1985). Actually, it is a special case of the Waring distribution with $\alpha = 1$. The general form of the Waring distribution is

$$p_k = P(X = k) = \frac{\alpha}{N+\alpha} \cdot \frac{N}{N+\alpha+1} \cdot \ldots \cdot \frac{N+k-1}{N+\alpha+k},\ k \geq 0,$$

where α and N are positive parameter. For $\alpha = 1$ we obtain the Price distribution

$$p_k = P(X = k) = \frac{N}{(N+k)(N+k+1)} = N\left(\frac{1}{N+k} - \frac{1}{N+k+1}\right),\ k \geq 0 .$$

Obviously, $G_k = N/(N + k)$ for $k \geq 0$. Hence we have $E(X) = \sum_{k=0}^{\infty} kp_k = \sum_{k=0}^{\infty} G_k = \infty$. The property $G_k \sim (N + k)^{-1}$ substantiated that this distribution is Paretian with $\alpha = 1$.

We just mention in passing that the Price distribution is asymptotically Lotka, but, in contrast to the Lotka distribution, the event $k = 0$, i.e, uncitedness, is defined.

From the definition of Gumbel's r-th extreme value we have immediately

$$u_r = \mathrm{Int}[N\cdot(n - r)/r],\ r = 1, 2, \ldots , n,$$

where Int [·] denotes the integer part of the argument, or without loss in rigor, in the following we can simply use

$$u_r = N(n - r)/r.$$

In verbal terms, this property means that u_r is proportional to the ratio of ranked sample elements below and above their rank. Hence we can readily derive the h-index for the case of the Price distribution. In particular we have

$$H = u_H = N(n - H)/H.$$

This results in a quadratic equation the solution of which is

$$H = -\frac{N}{2} + \sqrt{\frac{N^2}{4} + nN} \approx (nN)^{½}, \text{ if } n >>1.$$

The h-index is thus proportional to the square root of the number of publications (cf. Price' square root law, GLÄNZEL & SCHUBERT, 1985).

Finally, we will have a look at the Zipf property in the case of the Price distribution.

$$\lim_{n\to\infty} u_1/u_r = \lim_{n\to\infty} r(n-1)/(n-r) = r \quad \text{for any finite } r = 1, 2, \ldots\,.$$

The value of each observation is inversely proportional to its rank (cf., GLÄNZEL & SCHUBERT, 1985). If we apply the Zipf property to the h-index we obtain the following important property

$$u_1 \approx H^2 \approx nN\text{, if } n >>1$$

The highest citation rate is approximately equal to the square of the h-index if the number of underlying papers is very large. The above properties of the Price distribution can be obtained from the corresponding properties of the Paretian distribution for $\alpha = 1$ (see the previous subsection). Finally, we will give an approximation for the number of citations received by h-papers, that is, the papers meeting the Hirsch criterion by having received at least h citations each. According to GLÄNZEL & SCHUBERT (1985), we have $\sum_{k=0}^{u_r} kp_k \sim \ln(u_r + N) + \gamma(N)$, where $\gamma(N)$ is a constant depending only on N. In particular, for $N = 0$, $\gamma(0) \approx 0.577215665$ is the well-known Euler-Mascheroni constant. Hence,

$$\left(\sum_{k=0}^{u_1} kp_k - \sum_{k=0}^{u_r} kp_k \right) \approx \{\ln(n) - \ln(n/r)\} = \ln(r)$$

and

$$H_c \approx n \cdot \ln(H) \approx H^2 \cdot \ln(H)/N \approx ½n \cdot \ln(n/N)$$

In verbal terms, this property means that the number of h-citations is proportional to the square of the h-index multiplied with the natural logarithm of H.

Conclusions

We have found strong relation between both number of papers published and citations received with h-index for practically all Paretian distributions. In particular,

- The h-index is proportional to the (α+1)-th root of the number of publications, in the case of Price distributions this results in a square-root law.
- The highest citation impact is also a power function of the h-index.
- The property H_c/H^2 = constant for Paretian distributions with α>1 confirms one important finding by HIRSCH (2005). Note that this constant is dependant of the parameter α; different distributions thus result in different ratios H_c/H^2.
- The number of h-citations is a function of the square of the h-index and a constant dependent of the Pareto exponent and in the Price case this coefficient becomes a logarithmic function of the number of publications.
- The relatively low number of papers of most individuals does, however, not allow reliable statistical analysis of extreme values (cf. GLÄNZEL & SCHUBERT, 1988).

Summarising these properties as well as the pros and cons discussed in the outset, we can conclude that the h-index is certainly an interesting indicator with interesting mathematical properties. The strength of this index lies in the potential application to the assessment of small paper sets were other, traditional bibliometric indicators often fail or at least were their application proved usually problematic. The fact that the list of the cons somewhat exceeds that of the pros, does not necessarily mean that the disadvantages predominate. It just means that problems might arise in several applications. The main problem is perhaps that the h-index crashes the multi-dimensional space of bibliometrics into one single dimension. Besides additional theoretical work on the mathematical background, systematic research in the application of this measure is also necessary to reveal more perspectives and/or further limitations.

Nonetheless, the h-index is a useful supplement to the bibliometric toolset but it is certainly not suited to substitute the advanced indicators which have long ago become standard in bibliometric work.

*

The author would like to express his gratitude to Dr. András Schubert for his constructive comments and suggestions on an earlier version of the paper.

References

BALL, P. (2005), Index aims for fair ranking of scientists, *Nature,* 436 : 900.

BORNMANN, L., DANIEL, H.-D. (2005), Does the h-index for ranking of scientists really work? *Scientometrics*, 65 (3) : 391–392.

BRAUN, T., GLÄNZEL, W., SCHUBERT, A. (2005), A Hirsch-type index for journals, *The Scientist*, 19 (22) : 8.

DINIZ BATISTA, P., GUIMARAES CAMPITELI, M., KINOUCHI, O., SOUTO MARTINEZ, A. (2005), A complementary index to quantify an individual's scientific research output, arXiv:physics/0509048, accessible via http://arxiv.org/abs/physics/0509048.

EDWARDS, A. W. F. (2005), System to rank scientists was pedalled by Jeffreys, *Nature*, 437 : 951.

GLÄNZEL, W., SCHUBERT, A. (1985), Price distribution. An exact formulation of Price's "Square Root Law". *Scientometrics*, 7 (3–6) : 211–219.

GLÄNZEL, W., SCHUBERT, A. (1988), Theoretical and empirical studies of the tail of scientometric distributions. In: L. EGGHE, R. ROUSSEAU (Eds), *Informetrics 87/88*, Elsevier Science Publisher B. V., pp. 75–83.

GLÄNZEL, W. (2006), 也谈*h* 指数 的机会和局限性 (On the opportunities and limitations of the h-index, in Chinese); *Science Focus*, 1 (1) 10–11.

GUMBEL, E. J. (1958), *Statistics of Extremes*, Columbia University Press, New York.

HIRSCH, J. E. (2005), An index to quantify an individual's scientific research output, *Proceedings of the National Academy of Sciences of the United States of America*, 102 (46) : 16569–16572. (also available at: arXiv:physics/0508025, accessible via http://arxiv.org/abs/physics/0508025).

POPOV, S. B. (2005), A parameter to quantify dynamics of a researcher's scientific activity, arXiv:physics/0508113, accessible via http://arxiv.org/abs/physics/0508113.

SCHUBERT, A., GLÄNZEL, W. (1983), Statistical reliability of comparisons based on the citation impact of scientific publications, *Scientometrics*, 5 (1) : 59–74.

VAN RAAN, A. F. J. (2005), Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgment for 147 chemistry research groups, arXiv:physics/0511206, accessible via http://arxiv.org/abs/physics/0511206.