Simple models and the corresponding h- and g-index

Ronald Rousseau
KHBO (Association K.U.Leuven), Industrial Sciences and Technology, Zeedijk 101, 8400 Oostende, Belgium, & Hasselt University, Agoralaan, 3590 Diepenbeek, Belgium & Antwerp University, IBW, Universiteitsplein 1, 2610 Wilrijk, Belgium
Email: ronald.rousseau@khbo.be

Abstract

The relation between the Hirsch index and Egghe's g-index is determined for some simple models such as the uniform model, the point model, the linear model and Zipf's model.

Introduction

Recently the Hirsch-index, in short: h-index, has attracted a lot of attention in the scientific community (Ball, 2005; Bornmann & Daniel, 2005; Liang, 2006; Egghe, 2006c; Egghe & Rousseau, 2006; Rousseau, 2007). This index, introduced by J.E. Hirsch (2005) is calculated as follows. Consider the list of publications [co-]authored by scientist S, ranked according to the number of citations each of them has received over a given period. Then S' h-index is m if the first m publications received each at least m citations, while the publication ranked m+1 received strictly less than m+1 citations.

Clearly, this definition can also be applied to some other source-item pairs, besides a scientist's publications and citations (Braun et al. 2005; Egghe & Rousseau, 2006; Rousseau, 2006). In general we will denote the production of the source ranked r, as P(r), and its piecewise linear interpolation as P(x), this is: the function connecting the points (r, P(r)), where r denotes the rank (r = 1, 2, ...).

A slight generalization of Hirsch' original definition is obtained by defining the h-index as the abscissa of the intersection of the lines y = x and the observed function P(x). The original h-index is always a strictly positive integer, while this generalization, denoted as h_r, is a real number. Note that h_r is an index derived from observed data. If h_r is known than the corresponding h-value is equal to $\lfloor h_r \rfloor$. This is the floor function of h_r, or the largest natural number smaller than or equal to h_r. Note that using a real-valued Hirsch index is the natural thing to do when, e.g., citations are counted fractionally. Of course the h-index may also be modelled in a continuous context (Egghe, 2006c) but then this index is not anymore derived from observed data.

The g-index

The g-index has been introduced by my colleague Leo Egghe (2006a,b,d). It is calculated as follows: one draws the same list as for the h-index, but now the g-index is the highest rank such that the cumulative sum of the number of citations received is larger than or equal to the square of this rank. Clearly h ≤ g.

The g-index too can be calculated as a real number. It is then defined as the abscissa of the intersection of the curves $y = x^2$ and $y = C(x)$, where $C(x)$ is the function connecting the points $C(r) = \sum_{k=1}^{r} P(k)$. Similar to the notation h_r, this index is denoted as g_r.

Calculation of the h-index and the g-index for some simple models

We first consider two simple extreme cases, and will then consider a linear model. Also the Zipf model and the exponential model are briefly considered.

Model 1. The uniform model

In this case the production function is constant, say equal to $c \in \mathbf{N}$. Then clearly h = g = c.

Model 2. The point model

In this case P(1) = c > 0, while all other P(r) are zero. Then clearly h = 1 and $g_r = \sqrt{c}$.

Model 3. The linear model

Here we assume that P(r) = a – br, with a, b > 0. The h-index is determined by the requirement that a – bh = h. Solving this equation actually yields not h but $h_r = \frac{a}{1+b}$. As h, and hence also h_r must at least be equal to one, we must require that b ≤ a-1 (and hence certainly a > 1).

In this model the g-index is determined by:

$$\sum_{r=1}^{g}\left(a-br\right)=g^2$$
$$\Leftrightarrow ag-b\frac{g(g+1)}{2}=g^2$$
$$\Leftrightarrow\left(1+\frac{b}{2}\right)g^2+\left(\frac{b}{2}-a\right)g=0$$
$$\Leftrightarrow g=\frac{2a-b}{2+b}\quad\left(as\ g\neq 0\right)$$

Again this value is actually g_r. Note that we have to require that $g_r > 0$, or $2a > b$ (one may also want to require that $g_r \geq 1$, or $a \geq 1+b$), and that $P(g_r) > 0$, or $2a - ab+b^2 > 0$. Hence, not every decreasing linear function can be used as a model for a production function.

Model 4. The Zipf model

We assume that $P(r)=\frac{A}{r^{\beta}}$, with $A > 0$, $\beta \geq 1$. Then $h_r = A^{\frac{1}{\beta+1}}$. For the special case $\beta = 1$, $h_r = \sqrt{A}$. These values can also be found in (Egghe & Rousseau, 2006).

The corresponding g-index is determined by: $\sum_{r=1}^{g}\frac{A}{r^{\beta}}=g^2$. This equation can only be solved numerically. Some examples for the integer-valued g-index are given in Table 1

Table 1. Calculated integer-valued g-index for some values of A and β.

		A			
		10	50	100	200
β	1	4	12	18	28
	1.5	4	9	14	20
	2	3	8	12	17
	3	3	7	10	15

For β = 1, the g-index can also be found (approximately) as follows:

$\sum_{r=1}^{g}\frac{A}{r}=g^2 \Rightarrow A\left(\ln(g)+\gamma\right)=g^2$, where γ is Euler's constant (≈ 0.5772). Also this equation must be solved numerically. For A = 10, 50, 100, 200 this yields: g_r = 4.48, 12.45, 18.73, 27.96. The corresponding g-values are 4, 12, 18 and 28 (rounded), as shown in Table 1.

Model 5: The exponential model

Now P(r) = K a^{-r}, with K, a > 1. The h-index is determined as h = K a^{-h}. This leads to ln(K) – h. ln(a) – ln(h) = 0, an equation which can only be solved numerically. See Table 2.

Table 2. Calculated h_r-index for some values of a and K (rounded to one decimal).

		K			
		10	50	100	200
a	1.1	5.8	13.6	18.0	22.8
	1.2	4.4	9.3	11.7	14.4
	1.5	3.0	5.5	6.7	8.0
	2.0	2.2	3.7	4.5	5.3

The g-index is obtained as the solution of $\sum_{r=1}^{g} Ka^{-r} = g^2$ or $K\frac{1-a^{-g}}{a-1}=g^2$, which is again an equation that can only be solved numerically. See Table 3.

Table 3. Calculated g_r-index for some values of a and K (rounded to one decimal).

		K			
		10	50	100	200
a	1.1	7.0	20.8	30.8	44.4
	1.2	5.7	15.3	22.2	31.6
	1.5	4.0	9.9	14.1	20.0
	2.0	3.0	7.0	10.0	14.2

Construction of h for a given g

In this section we consider the following problem. Given a g-value and a particular model, find the parameters of the model yielding this g-value and find the corresponding h-value. We restrict ourselves to the two extreme cases and the linear model.

Problem 1: the uniform model

Given the natural number g = $g_0 \geq 1$, then clearly P(1) = …= P(g_0) = g_0 = h.

Problem 2: a point model

Given the natural number g = $g_0 \geq 1$, then P(1) = g_0^2 and P(2) = … = P(g_0) = 0. For the point model the corresponding h-index is 1.

Problem 3: a linear production model

Let a value of the g-index, $g_0 > 1$, be given. Again g_0 is assumed to be a natural number. Then we want to determine a linear production function P(x) such that its g-index is equal to the given value g_0. We will also determine the corresponding h-index.

Put P(x) = a + bx and assume further that P(g_0) = c (c $\in$ **N**). From this requirement we see that c = a + bg_0, hence $b = -\frac{a-c}{g_0}$. As P(x) must be a decreasing function, b must be negative, and hence the problem has only a solution if c < a. From $\sum_{r=1}^{g_0} P(r) = g_0^2$ we obtain:

$$a.g_0 - \frac{a-c}{g_0}.\frac{g_0(g_0+1)}{2} = g_0^2$$

$$\Leftrightarrow \; a\left(2g_0 - g_0 - 1\right) + c\left(g_0+1\right) = 2g_0^2$$

$$\Leftrightarrow a = \frac{2g_0^2 - c(g_0+1)}{g_0-1}$$

$$\Leftrightarrow \; a - c = \frac{2g_0^2 - c(g_0+1) - c(g_0-1)}{g_0-1} = \frac{2g_0^2 - 2c.g_0}{g_0-1} = \frac{2g_0\left(g_0-c\right)}{g_0-1}$$

We conclude that P(x) = $\frac{2g_0^2 - c(g_0+1)}{g_0-1} - \frac{2(g_0-c)}{g_0-1}x$.

In particular, P(1) = $\frac{2g_0^2 - c(g_0+1)}{g_0-1} - \frac{2(g_0-c)}{g_0-1} = 2g_0 - c$.

If c = 1 then P(x) = $2g_0+1-2x$ and P(1) = 2g_0 -1.

The requirement c < a becomes: $c<\frac{2g_0^2-c(g_0+1)}{g_0-1}$ or $c<g_0$.

The corresponding h_r-index is the solution of. Hence $\frac{2g_0^2-c(g_0+1)}{g_0-1}-\frac{2(g_0-c)}{g_0-1}h_r=h_r$

$$\frac{2g_0^2-c(g_0+1)}{g_0-1}=h_r\left(1+\frac{2(g_0-c)}{g_0-1}\right)$$
$$or$$
$$h_r=\frac{2g_0^2-c(g_0+1)}{3g_0-1-2c}$$

Note that because c < g_0 and g_0 > 1, the denominator is always strictly positive. For the same reason the numerator is also strictly positive, so that h_r is a positive number. If g_0 is large we see that $h_r\approx\frac{2g_0}{3}$. This shows that in this model the g-index is about 50% larger than the h-index.

If c = 1, then $h_r=\frac{2g_0+1}{3}$.

Conclusion

The relation between the h- and the g-index is determined for some simple models such as the uniform model, the point model, the linear model, Zipf's model and the exponential model.

Acknowledgements

Research for this article was performed while the author was visiting Dalian University of Technology and the National Library of Sciences CAS (Beijing). He would like to thank professor Liu Zeyang, professor Jin Bihui, and their colleagues and students for their hospitality. Research for this article was supported through NSFC grant 70373055.

References

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

Bornmann, L. & Daniel, H.-D. (2005). Does the h-index for ranking of scientists really work? *Scientometrics*, 65, 391-392.

Braun T., Glänzel, W. ,& Schubert A., & (2005). A Hirsch-type index for journals. *The Scientist*, 19(22), p.8.

Egghe, L. (2006a). How to improve the h-index. *The Scientist*, 20(3), p. 14.

Egghe L. (2006b). An improvement of the H-index: the G-index. *ISSI Newsletter*, 2(1), 8-9.

L. Egghe (2006c). Dynamic h-index: the Hirsch index in function of time. *Journal of the American Society for Information Science and Technology* (to appear).

L. Egghe (2006d). Theory and practice of the g-index. *Scientometrics* (to appear).

L. Egghe and R. Rousseau (2006). An informetric model for the Hirsch index. *Scientometrics* (to appear).

J.E. Hirsch (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences of the USA*, 102(46), 16569-16572.

Liang, L. (2006). H-index sequence and h-index matrix: construction and applications. *Scientometrics* (to appear)

Rousseau R. (2006). A case study: evolution of JASIS' h-index. E-LIS: ID-code 5430.

Rousseau R. (2007). The influence of missing publications on the h-index. *Journal of Informetrics* (submitted).