**Made available by Hasselt University Library in https://documentserver.uhasselt.be**

# Distributions of the h-index and the g-index

Non Peer-reviewed author version

EGGHE, Leo (2007) Distributions of the h-index and the g-index. In: Torres-Salinas, D. & Moed, H. (Ed.) Proceedings of the 11th International Society for Scientometrics and Informetrics. p. 245-253..

Handle: http://hdl.handle.net/1942/1791

# Distributions of the h-index and the g-index [1]

Leo Egghe

*leo.egghe@uhasselt.be*

Universiteit Hasselt (UHasselt), Campus Diepenbeek, Agoralaan, B-3590 Diepenbeek (Belgium)[2]

and Universiteit Antwerpen (UA), Campus Drie Eiken, Universiteitsplein 1, B-2610 Wilrijk (Belgium)

**Abstract**
In every scientific research area, each scientist has a unique h-index and g-index. This paper addresses the problem of determining the distribution of these indexes over the scientists. We apply aspects of linear three-dimensional Lotkaian informetrics to determine these distributions.

We show that, supposing the article – citation information production process (IPP) to be Lotkaian with exponent $\alpha$ and supposing the scientist-article IPP to be Lotkaian with exponent $\alpha^*$, we have that the scientist-h-index IPP and the scientist-g-index IPP are Lotkaian with exponent $\alpha\alpha^*$. This model is proved for discrete as well as continuous variables.

This shows that the size-frequency distributions of the h-index and the g-index are very skew in general due to the generally high value of $\alpha\alpha^*$. We also calculate the rank-frequency distributions of the h- and g-index, based on the size-frequency distributions in the continuous variables case. Examples are given.

**Keywords**
h-index; Hirsch index; g-index; distribution; Lotka

**Introduction**
In Hirsch (2005), the physicist Hirsch introduced his so-called Hirsch-index (or h-index) as follows (using our own terminology – cf. Egghe and Rousseau (2006)): if we rank an author's papers in decreasing order of the number of citations they have received (publication and citation periods are fixed but arbitrary) then this author's h-index is the largest rank $r = h$ such that all papers on ranks $1,...,h$ have at least h citations each.

Since its introduction, the h-index has received a lot of attention, also in the informetric community, see e.g. Ball (2005), Bornmann and Daniel (2005), Braun, Glänzel and Schubert (2005), Egghe (2007), Glänzel (2006a,b), van Raan (2005), where also advantages and disadvantages of the h-index are described. We do not go into this topic here, but we only indicate one disadvantage of the h-index, which lead Egghe (see Egghe (2006a,b,c)) to his definition of the g-index: a clear disadvantage of the

[1] The author is grateful to Prof. Dr. I.K. Ravichandra Rao (ISI, Bangalore, India) for mentioning the problem studied in this paper.
[2] Permanent address
Acknowledgement: The author is grateful to Prof. Dr. I.K. Ravichandra Rao (ISI, Bangalore, India) for mentioning the problem studied in this paper.

h-index is that, once an article is taken into account for the calculation of the h-index (i.e. once an article has a rank in the set $\{1,2,...,h\}$) its actual number of citations (above h of course), now and in the future, is not taken into account. Egghe finds this a clear disadvantage for an overall performance measure of a scientist. Egghe notes that the h-index satisfies that the first h articles have at least $h^2$ citations, together but h is not necessarily the largest value with this property.

Therefore, in Egghe (2006a,b,c), Egghe defines the g-index to be the largest rank $r = g$ (we use the same ranking as above) such that the papers on ranks $1,...,g$ have at least $g^2$ citations, together. Clearly, $g \geq h$ and in most cases, $g > h$. Examples in Egghe (2006b) show that the g-index has more "discriminative" power amongst scientists in a field but this topic is not further addressed in this paper. Also in this paper we suppose that the total number A of citations (to all papers of a scientist) is less than the square $T^2$ of the total number T of papers, so that g is always defined and also $g \leq T$. This property was always encountered in the examples in Egghe (2006b) but, when $A > T^2$, in Egghe (2006b), we also give a methodology to calculate g-indexes that are superior to T, by adding fictitious articles with 0 citations.

Having any scientific field, to be considered as a group of researchers, we can calculate, for each of them, a h-index and a g-index. One can then wonder what is the distribution of these h-indexes over the researchers and we can also ask the same question for the g-indexes.

In the next section we study this problem in the discrete setting (researchers have 1, 2, 3,… cited articles) and in the third section the same problem will be studied in the continuous setting (with densities of articles and citations). In both sections we prove the same theorem (exact in the case of continuous variables and with good approximations in the case of discrete variables): suppose that articles are cited according to a Lotkaian size-frequency function (or distribution) with Lotka exponent $\alpha > 1$ and suppose that authors publish articles according to a Lotkaian size-frequency function with Lotka exponent $\alpha^* > 1$, then the distribution (size-frequency function) of the numbers of researchers with a certain h-index or g-index is Lotkaian with exponent $\alpha\alpha^*$. Concrete examples are given that make this observation clear but, of course, also exact mathematical proofs are given. So, in general, such distributions have large Lotkaian exponents showing that their size-frequency functions are very skew (take e.g. the most common Lotka-exponents $\alpha = \alpha^* = 2$, then the distributions of the h-index and of the g-index are Lotkaian with exponent $\alpha\alpha^* = 4$ which is very high and leads to very skew (concentrated – see Egghe (2005), Chapter III) size-frequency functions for the distribution of the h- and g-indexes.

Based on the size-frequency functions of the h- and g-index we also determine the rank-frequency functions of the h- and g-index.

**Size-frequency functions for the h- and g-index: discrete variables case**

So we have a situation where researchers in a certain field produce articles and that these articles receive citations (after their publication). We restrict ourselves to those articles that received at least one citation. Publication periods and citation periods are fixed but – for this model – are arbitrary.

We suppose the article-citation IPP to be Lotkaian, cf. Glänzel (2006b), Egghe and Rousseau (2006), Rousseau (1997), Redner (1998), but this model can also serve as a first, simple approximation of other decreasing models for the number $f(n)$ of papers with n citations, as in Burrell (2007) or Redner (2005): we suppose that the number of articles with n citations equals $(n = 1,2,3,...)$

$$f(n) = \frac{C}{n^{\alpha}} \qquad (1)$$

where $C > 0$ is a constant and $\alpha > 1$ is the Lotkaian exponent of this IPP.

Likewise, and even more classical – cf. Lotka (1926) – we can suppose the author-publication (articles) IPP to be Lotkaian: we suppose that the number of authors with T articles equals

$$\varphi(T) = \frac{D}{T^{\alpha^*}} \tag{2}$$

where $D > 0$ is a constant and $\alpha^* > 1$ is the Lotkaian exponent of this IPP.

While the number n of citations to articles can be arbitrarily large, we assume that authors produce a number of articles between 1 and $T_{max}$. We further, classically, suppose that we have only one author producing the maximum number of articles $(T_{max})$. Hence, using (2)

$$1 = \frac{D}{T_{max}^{\alpha^*}}$$

whence $D = T_{max}^{\alpha^*}$ and hence

$$\varphi(T) = \frac{T_{max}^{\alpha^*}}{T^{\alpha^*}} \tag{3}$$

*Size-frequency function for the h-index*
As proved in Glänzel (2006b) in the discrete case (approximately) and exactly in the continuous case in Egghe and Rousseau (2006), we have in case (1) when there are T articles in total (for a particular author):

$$T = \sum_{n=1}^{\infty} f(n) = \sum_{n=1}^{\infty} \frac{C}{n^{\alpha}} \tag{4}$$

that the h-index equals

$$h = T^{\frac{1}{\alpha}} \tag{5}$$

Combining (3) and (5) yields that, for each $T = 1, ..., T_{max}$, we have

$$\frac{T_{max}^{\alpha^*}}{T^{\alpha^*}} \text{ authors with } h = T^{\frac{1}{\alpha}} \tag{6}$$

(hereby supposing that $\alpha$ is fixed, independent of T). This proves that the size-frequency function for the h-index: $\varphi_1(h) =$ the number of authors with h-index h equals

$$\varphi_1(h) = \frac{T_{max}^{\alpha^*}}{h^{\alpha\alpha^*}} \tag{7}$$

Indeed: for each h-index $h = T^{\frac{1}{\alpha}}$, formula (7) gives a number of authors with this h-index equal to

$$\frac{T_{max}^{\alpha^*}}{T^{\frac{\alpha\alpha^*}{\alpha}}} = \frac{T_{max}^{\alpha^*}}{T^{\alpha^*}}$$

which is correct according to (3), since these authors have T articles.

*Size-frequency function for the g-index*
When a particular author has T articles we have (proved exactly in Egghe (2006b) in the continuous case) that the g-index equals

$$g = \left(\frac{\alpha-1}{\alpha-2}\right)^{\frac{\alpha-1}{\alpha}} T^{\frac{1}{\alpha}} \tag{8}$$

Hence, using (3) and (8) we now have that, for each $T = 1,...,T_{max}$, we have

$$\frac{T_{max}^*}{T^*} \text{ authors with } g = \left(\frac{\alpha-1}{\alpha-2}\right)^{\frac{\alpha-1}{\alpha}} T^{\frac{1}{\alpha}} \tag{9}$$

This proves that the size-frequency function for the g-index: $\varphi_2(g) =$ the number of authors with g-index g equals

$$\varphi_2(g) = \frac{T_{max}^{\alpha^*}\left(\frac{\alpha-1}{\alpha-2}\right)^{(\alpha-1)\alpha^*}}{g^{\alpha\alpha^*}} \tag{10}$$

Indeed: for each g-index $g = \left(\frac{\alpha-1}{\alpha-2}\right)^{\frac{\alpha-1}{\alpha}} T^{\frac{1}{\alpha}}$, formula (10) gives a number of authors with this g-index equal to

$$\frac{T_{max}^{\alpha^*}\left(\frac{\alpha-1}{\alpha-2}\right)^{(\alpha-1)\alpha^*}}{\left(\frac{\alpha-1}{\alpha}\right)^{\frac{\alpha-1}{\alpha}\alpha\alpha^*} T^{\frac{\alpha\alpha^*}{\alpha}}} = \frac{T_{max}^{\alpha^*}}{T^{\alpha^*}}$$

which is correct according to (3) since these authors have T articles.

Both previous subsections show that the size-frequency functions of the h- and g-indexes are Lotkaian (with different constants in the numerator) with the same Lotkaian exponent $\alpha\alpha^*$, the product of the Lotkaian exponents of the article-citation IPP and of the author-article IPP.

This simple result also shows that the size-frequency functions of the h- and g-index are very skew or concentrated – see e.g. Egghe (2005), Chapter IV, Corollary IV.3.2.1.5, p. 204-205, since, usually, $\alpha\alpha^*$ is a large Lotka exponent: take e.g. the "classical" values $\alpha \approx \alpha^* \approx 2$ then $\alpha\alpha^* \approx 4$ which is extremally large (see the Lotka exponents described in the review subsection I.4 in Egghe (2005), p. 85-98).

The above results can also be proved – essentially – in the continuous setting. This will be done in the next section. The continuous results will also enable use to calculate the rank-frequency functions for the h- and g-index, which is not possible in the discrete variable setting.

**Size- and rank-frequency functions for the h-index and g-index: continuous variables case**

*Size-frequency function for the h-index*
We again have (2) but now for continuous variables T which we do not limit: $T \in [1, +\infty[$. Also, result (5) is exact, for every T, as proved in Egghe and Rousseau (2006). Hence we have that the density of authors with h-index h is proportional to (by (2) and (5)):

$$\varphi_1(h) \sim \frac{1}{h^{\alpha\alpha^*}} \tag{11}$$

We still have to normalise formula (11): we must have that

$$\int_1^\infty \varphi_1(h)dh = \int_1^\infty \varphi(T)dT \tag{12}$$

(= total number of authors).

Note that both h and T have 1 as minimal value. For T this is so because, with (2), we consider the number of articles as items, ranging from 1 (see Egghe (2005), Chapter II). Since $T = 1$ is the minimal value of cited articles, $h = 1$ is also the minimal value for h.

Defining the proportionality factor in (11) as E:

$$\varphi_1(h) = \frac{E}{h^{\alpha\alpha^*}} \tag{13}$$

we have (since $\alpha$, $\alpha^* > 1$) and by (2)

$$\frac{E}{\alpha\alpha^* - 1} = \frac{D}{\alpha^* - 1}$$

hence

$$E = D\frac{\alpha\alpha^* - 1}{\alpha^* - 1}$$

so that

$$\varphi_1(h) = \frac{D\frac{\alpha\alpha^* - 1}{\alpha^* - 1}}{h^{\alpha\alpha^*}} \tag{14}$$

Hence we again find Lotka's law with the exponent $\alpha\alpha^*$.

*Size-frequency function for the g-index*
Now (2) is still valid and also (8) is an exact result. Hence we have that the density of authors with g-index g is proportional to (by (2) and (8)):

$$\varphi_2(g) \sim \frac{1}{g^{\alpha\alpha^*}} \tag{15}$$

(we can omit D as well as $\left(\frac{\alpha-1}{\alpha-2}\right)^{\frac{\alpha-1}{\alpha}}$ in (15) since both are constants and since the normalising constant in (15) still must be determined). This goes as follows: as in (12), T starts in 1 but now, since h (or T) starts in 1 and by (8), we have that g starts in $\left(\frac{\alpha-1}{\alpha-2}\right)^{\frac{\alpha-1}{\alpha}}$. Hence we have the requirement:

Total number of authors

$$= \int_{\left(\frac{\alpha-1}{\alpha-2}\right)^{\frac{\alpha-1}{\alpha}}}^{\infty} \varphi_2(g)\,dg = \int_1^{\infty} \varphi(T)\,dT \tag{16}$$

Defining the proportionality factor in (15) as F:

$$\varphi_2(g) = \frac{F}{g^{\alpha\alpha^*}} \tag{17}$$

we have (since $\alpha$, $\alpha^* > 1$) and by (2):

$$\frac{F}{1-\alpha\alpha^*}\left[g^{1-\alpha\alpha^*}\right]_{\left(\frac{\alpha-1}{\alpha-2}\right)^{\frac{\alpha-1}{\alpha}}}^{\infty} = \frac{D}{1-\alpha^*}\left[T^{1-\alpha^*}\right]_1^{\infty}$$

$$= \frac{D}{\alpha^*-1}$$

Or

$$\frac{F}{\alpha\alpha^*-1}\left(\frac{\alpha-1}{\alpha-2}\right)^{\frac{\alpha-1}{\alpha}\left(1-\alpha\alpha^*\right)} = \frac{D}{\alpha^*-1}$$

Hence

$$F = D\frac{\alpha\alpha^*-1}{\alpha^*-1}\left(\frac{\alpha-1}{\alpha-2}\right)^{\frac{\alpha-1}{\alpha}\left(\alpha\alpha^*-1\right)} \tag{18}$$

so that the size-frequency function for the g-index is

$$\varphi_2(g) = \frac{D\frac{\alpha\alpha^*-1}{\alpha^*-1}\left(\frac{\alpha-1}{\alpha-2}\right)^{\frac{\alpha-1}{\alpha}\left(\alpha\alpha^*-1\right)}}{g^{\alpha\alpha^*}} \tag{19}$$

Formulae (7), (10), (14) and (19) are evidence for the following Theorem.

*Theorem*:

If the article-citation IPP is Lotkaian with exponent $\alpha$ and if the author-article IPP is Lotkaian with exponent $\alpha^*$, then both the author-h-index IPP and the author-g-index IPP are Lotkaian (i.e. their size-frequency functions are Lotkaian) with exponent $\alpha\alpha^*$.

This represents a case where one regularly finds power laws with high exponents $\alpha\alpha^*$.

*Example*:
Let us take the "most classical" case that $\alpha = \alpha^* = 2$ (see Egghe (2005) for a treatment of this special case of Lotka exponents equal to 2). Let us take $\left(\alpha^* = 2\right)$ (cf. (2))

$$\varphi(T) = \frac{100}{T^2}$$

(author-article size-frequency function). Since also the article-citation IPP has Lotka exponent $\alpha = 2$ we have here that (Glänzel (2006b), Egghe and Rousseau (2006)) $h = \sqrt{T}$ for every production T. We have

for T = 1 : 100 researchers have h = 1
T = 2 : 25 researchers have $h = \sqrt{2}$
.
.
.
T = 10 : 1 researcher has $h = \sqrt{10}$

Hence we have (cf. (7))

$$\varphi_1(h) = \frac{100}{h^4}$$

as size-frequency distribution. Indeed: $h = 1$ occurs with 100 researchers, $h = \sqrt{2}$ occurs with $\frac{100}{h^4} = 25$ researchers, …, $h = \sqrt{10}$ occurs with $\frac{100}{h^4} = 1$ researcher.

Redner (1998) even reports on Lotkaian exponents $\alpha \approx 3$ for the article-citation IPP making $\alpha\alpha^*$ (most likely) to be even larger than four (probably around 6 if $\alpha^* \approx 2$)!

Similar examples can be given for the g-index (for $\alpha > 2$ now) based on (10).

Based on (14) and (19) we can also determine the rank-frequency functions of the h- and g-index, i.e. the functions $h(r)$ and $g(r)$ being the h-index (g-index respectively) at rank r.

*Rank-frequency function for the h-index*
From every size-frequency function $\varphi(j)$ one can derive the corresponding rank-frequency function $\psi(r)$ using the following Lemma (Exercise II.2.2.6, p. 134 in Egghe (2005) or Appendix in Egghe and Rousseau (2006) where also a proof is given)

*Lemma*:
The following assertions are equivalent:

(i) $\varphi(j)=\frac{C}{j^{\alpha}}$ with $C>0$, $\alpha>1$ (constants) and $j\in[1,+\infty[$ (size-frequency function)

(ii) $\psi(r)=\frac{B}{r^{\beta}}$ with $B$, $\beta>0$ (constants) and $r\in]0,T]$ (rank-frequency function), where T denotes the total number of sources. Moreover, the relation between the parameters are:

$$B=\left(\frac{C}{\alpha-1}\right)^{\frac{1}{\alpha-1}} \tag{20}$$

$$\beta=\frac{1}{\alpha-1} \tag{21}$$

When we apply the above Lemma to (14) (as $\varphi$) we find as rank-frequency function for h

$$h(r)=\psi_1(r)=\frac{D^{\frac{1}{\alpha\alpha^*-1}}\left(\frac{\alpha\alpha^*-1}{\alpha^*-1}\right)^{\frac{1}{\alpha\alpha^*-1}}}{\left(r\left(\alpha\alpha^*-1\right)\right)^{\frac{1}{\alpha\alpha^*-1}}} \tag{22}$$

*Rank-frequency function for the g-index*
Similarly, based on the above Lemma and (19) we have the following rank-frequency function for g

$$g(r)=\psi_2(r)=\frac{D^{\frac{1}{\alpha\alpha^*-1}}\left(\frac{\alpha\alpha^*-1}{\alpha^*-1}\right)^{\frac{1}{\alpha\alpha^*-1}}}{\left(r\left(\alpha\alpha^*-1\right)\right)^{\frac{1}{\alpha\alpha^*-1}}}\left(\frac{\alpha-1}{\alpha-2}\right)^{\frac{\alpha-1}{\alpha}} \tag{23}$$

Note that it follows from (22), (23), (5) and (8) that the source-rankings in (22) are the same as the source-rankings in (23) since the source on rank r in (22) has $h(r)$ as h-index and hence has

$$h(r)\left(\frac{\alpha-1}{\alpha-2}\right)^{\frac{\alpha-1}{\alpha}}$$

$$=g(r)$$

as g-index. This is logical since, for any two sources A and B: $h_A<h_B \Leftrightarrow g_A<g_B$, hence the rankings must be the same. So formula (23) also serves as a control for the correctness of our models.

**Conclusions**
We showed that the size-frequency functions (or distributions) of the h-index as well as the g-index (with respect to authors) is Lotkaian if we suppose the same for the size-frequency functions of citations (with respect to articles) and of articles (with respect to authors). Moreover the Lotka exponent for the h- and g-index distribution is the product of the respective Lotka distributions of citations and articles.

This also shows that we encounter here, in most cases, large exponents making the exponent 4 the "classical" value for the h- and g-index distributions (since exponents 2 are "classical" in the underlying cases).

**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.

Burrell, Q.L. (2007). Hirsch's h-index: a stochastic model. Journal of Informetrics, 1(1), to appear.

Egghe, L. (2005). Power Laws in the Informetric Production Process: Lotkaian Informetrics. Elsevier, Oxford, UK.

Egghe, L. (2006a). An improvement of the h-index: the g-index. ISSI Newsletter, 2(1), 8-9.

Egghe, L. (2006b). Theory and practise of the g-index. Scientometrics, 69(1), 131-152.

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

Egghe, L. (2007). Dynamic h-index: The Hirsch index in function of time. Journal of the American Society for Information Science and Technology, 58(3), 452-454.

Egghe, L. & Rousseau, R. (2006). An informetric model for the Hirsch-index. Scientometrics, 69(1), 121-129.

Glänzel, W. (2006a). On the opportunities and limitations of the h-index. Science Focus, 1(1), 10-11 (in Chinese).

Glänzel, W. (2006b). On the h-index – a mathematical approach to a new measure of publication activity and citation impact. Scientometrics, 67(2), 315-321.

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, 16569-16572.

Lotka, A.J. (1926). The frequency distribution of scientific productivity. Journal of the Washington Academy of Sciences, 16(12), 317-324.

Redner, S. (1998). How popular is your paper ? An empirical study of the citation distribution. The European Physical Journal, B4(2), 131-134.

Redner, S. (2005). Citation statistics from 110 years of Physical Review. Physics Today, 58(6), 49-54.

Rousseau, R. (1997). Sitations: an exploratory study. Cybermetrics, 1(1), paper 1. http://www.cindoc.csic.es/cybermetrics/articles/v1i1p1.html

van Raan, A.F.J. (2006). Comparison of the Hirsch-index with standard bibliometric indicators and with peer judgement for 147 chemistry research groups. Scientometrics, 67(1), 491-502.