Characteristic scores and scales in a Lotkaian framework

Peer-reviewed author version

EGGHE, Leo (2010) Characteristic scores and scales in a Lotkaian framework. In: SCIENTOMETRICS, 83(2). p. 455-462.

DOI: 10.1007/s11192-009-0009-y
Handle: http://hdl.handle.net/1942/9685

Characteristic scores and scales in a Lotkaian framework

by

L. Egghe

Universiteit Hasselt (UHasselt), Campus Diepenbeek, Agoralaan, B-3590 Diepenbeek, Belgium
leo.egghe@uhasselt.be

ABSTRACT

The characteristic scores and scales (CSS), introduced by Glänzel and Schubert [Journal of Information Science 14, 123-127, 1988] and further studied in subsequent papers of Glänzel, can be calculated exactly in a Lotkaian framework. We prove that these CSS are simple exponents of the average number of items per source in general IPPs. The proofs are given using size-frequency functions as well as using rank-frequency functions.

We note that CSS do not necessarily have to be defined as averages but that medians can be used as well. Also for these CSS we present exact formulae in the Lotkaian framework and both types of CSS are compared. We also link these formulae with the h-index.

I. Introduction

The authors of the paper Glänzel and Schubert (1988) were far ahead of their time, when they introduced "characteristic scores and scales" (CSS). CSS subdivide an author's ranked set of articles (ranked in decreasing order of number of received citations) according to a certain

Keywords and phrases: characteristic scores and scales, CSS, average, median, Lotka

device of papers having a "minimum" number of citations (we will be more specific in the sequel) and in this sense such a sequence has a more "dynamic" value (than one indicator such as, e.g., the h-index (Hirsch (2005)).

CSS are defined in Glänzel and Schubert (1988) and redefined in Glänzel (2008, 2009) as follows. Let $\beta_0 = 0$ (if one wishes this can be deleted) and let $\beta_1 = \mu$ = the average number of citations to the author's papers. Now we discard all papers cited less than the mean $\beta_1 = \mu$. The average number of citations to the remaining papers is denoted by β_2 and, evidently, $\beta_2 > \beta_1$. Again those papers cited less than β_2 are removed. The average number of citations to the remaining papers is denoted β_3 and again we have $\beta_3 > \beta_2$.

In practise, this procedure ends at a certain point $\beta_{k_{\max}}$ but in theoretical models (see further) we can obtain an infinite strictly increasing sequence $\beta_0, \beta_1, ..., \beta_k, ...$ In Glänzel and Schubert (1988) one uses a discrete rank-frequency model as well as a discrete size-frequency model to estimate the values of β_k (in Glänzel (2008, 2009) only the discrete rank-frequency model is used). The fact that one uses a discrete model has the disadvantage that only approximations of the discrete sums are obtained.

In the next section we will define the CSS in a continuous infinite Lotkaian framework (both in the size-frequency model and in the rank-frequency model). There we will show that, simply, $\beta_k = \mu^k$, for all $k = 1, 2, ...$, were $\mu = \beta_1$ is as above. Some comparisons with the single h-index are made.

In the third section we will define other CSS, now based on the median number of citations, rather than the average number of citations as explained above. Let us denote this CSS by β_k^*, then we prove, again in the continuous infinite Lotkaian framework, that $\beta_k^* = 2^{k/(\alpha-2)}$ for all $k = 1, 2, ...$, where α is the Lotka-exponent. Relations with β_k and the h-index are given.

The paper closes with conclusions and some suggestions for further research.

II. The Glänzel-Schubert CSS in a continuous Lotkaian framework

We will use the simple continuous Lotkaian model:

$$f(j) = \frac{C}{j^{\alpha}} \tag{1}$$

, $C > 0$, $j \geq 1$, $\alpha > 1$, for the density of the sources with item-density j, (see Egghe (2005) for more details).

Based on the description of the CSS in Glänzel and Schubert (1988) and Glänzel (2008, 2009), we have the following defining equation for the CSS β_k, $k = 1, 2, \ldots$.

$$\beta_k = \frac{\int_{\beta_{k-1}}^{\infty} jf(j)dj}{\int_{\beta_{k-1}}^{\infty} f(j)dj} \tag{2}$$

where we define $\beta_0 = 1$ (for calculatory reasons we start in $j \geq 1$, hence $\beta_0 = 1$). Take now $\alpha > 2$. Then it is clear that (see also Egghe (2005), p.115)

$$\int_1^{\infty} jf(j)dj = \frac{C}{\alpha - 2} = A \tag{3}$$

$$\int_1^{\infty} f(j)dj = \frac{C}{\alpha - 1} = T \tag{4}$$

, the total number of items (A), respectively the total number of sources (T). Hence $\beta_1 = \frac{A}{T} = \mu$, the average number of items per source. We have the following Proposition concerning the CSS β_k, $k = 1, 2, \ldots$.

Proposition 1: Let $\alpha > 2$. The CSS β_k have the values, for $k = 1, 2, \ldots$

$$\beta_k = \beta_1^k = \mu^k \tag{5}$$

Proof: The proof is by complete induction. By the above, the Proposition is true for $k = 1$ (and even for $k = 0$: $\beta_0 = \mu^0 = 1$ which shows that $\beta_0 = 1$ is natural in this framework). Suppose now that $\beta_{k-1} = \mu^{k-1}$ for a certain $k \in \mathbb{N}$. By (2) we have, using (1) with $\alpha > 2$

$$\beta_k = \frac{\frac{C}{\alpha-2}\frac{1}{\beta_{k-1}^{\alpha-2}}}{\frac{C}{\alpha-1}\frac{1}{\beta_{k-1}^{\alpha-1}}} \tag{6}$$

$$\beta_k = \frac{\alpha-1}{\alpha-2}\beta_{k-1}$$

$$\beta_k = \beta_1\beta_{k-1} \tag{7}$$

since $\beta_1 = \frac{A}{T} = \mu$ and by (3) and (4). Formula (7) proves that

$$\beta_k = \mu^k$$

, finishing the induction argument. □

Glänzel and Schubert (1988) use both the size-frequency and rank-frequency approach while Glänzel (2008, 2009) only uses the rank-frequency approach. The above argument is the size-frequency approach, using the size-frequency function (1).

The rank-frequency approach, equivalent to (1), is given by, for $r \in]0,T]$,

$$g(r) = \frac{\beta}{r^{\beta}} \tag{8}$$

where

$$B = T^{\frac{1}{\alpha-1}} \tag{9}$$

and where

$$\beta = \frac{1}{\alpha-1} \tag{10}$$

, cf. Exercise II.2.2.6 in Egghe (2005), p.134 (a proof is presented in Egghe and Rousseau (2006)). Here $g(r)$ is the rank-frequency function of the item-density $g(r)$ at source-density $r \in]0,T]$. Formula (8) is known as Zipf's law.

The CSS of Glänzel and Schubert (1988) are now defined as (continuous framework)

$$\beta_k = \frac{\int_0^{\nu_{k-1}} g(r)dr}{\nu_{k-1}} \tag{11}$$

and

$$g(\nu_k) = \beta_k \tag{12}$$

for $k = 1, 2, \ldots$ and $\nu_0 = T$, the total number of sources. Evidently, also this framework yields Proposition 1. Indeed, trivially, $\beta_1 = \mu$, by definition of $g(r)$ and T. By complete induction, suppose $\beta_{k-1} = \mu^{k-1}$ for a certain $k \in \mathbb{N}$. Then, by (11) and (8),

$$\beta_k = \frac{B}{\nu_{k-1}^{\beta}(1-\beta)} \tag{13}$$

But by (8) and (12) we see that $\nu_{k-1}^{\beta} = \dfrac{B}{\beta_{k-1}}$ hence, by this and (13) we have

$$\beta_k = \frac{\beta_{k-1}}{1-\beta} \tag{14}$$

But, by (10),

$$\frac{1}{1-\beta} = \frac{\alpha-1}{\alpha-2} = \mu = \beta_1$$

By (3), (4) and (10) and so (14) yields

$$\beta_k = \beta_1 \beta_{k-1} = \mu^k$$

, concluding this complete induction argument.

The function $\mu = \dfrac{\alpha-1}{\alpha-2}$ is decreasing in $\alpha > 2$, hence increasing in $L(g)$, the Lorenz-curve of the size-frequency function $g(.)$ (see Egghe (2005), p.205 for this result and the definition of the Lorenz-curve). In words, this means that, the more unequal the source-item table values are (sources in decreasing order of their number of items), the higher the CSS values β_k.

In terms of the h-index we can remark that we proved in Egghe and Rousseau (2006), in this framework, that (for $\alpha > 1$)

$$h = T^{\frac{1}{\alpha}} \tag{15}$$

and hence, since $\mu = \dfrac{A}{T}$ and by (3) and (4), for $\alpha > 2$

$$h = \left(\frac{\alpha-2}{\alpha-1} A\right)^{\frac{1}{\alpha}} \tag{16}$$

This yields, using (3), (4) and (5):

$$h = \left(\frac{A}{\beta_k^{1/k}}\right)^{\frac{1}{\alpha}} \tag{17}$$

(h in function of the CSS values β_k) or inversely

$$\beta_k = \left(\frac{A}{h^\alpha}\right)^k \tag{18}$$

for all values $k = 1, 2, \ldots$. This shows that the CSS values β_k can be derived from the h-index (and vice-versa), given A, the total number of items.

In Glänzel (2008), it is argued that an interesting property of the CSS is that it can yield sets of important articles of an author that contain less articles than the h-articles in the h-core. In the above model, this can always be obtained. It suffices to require that at the CSS value β_k one has less than h sources with item density β_k or higher:

$$\int_{\beta_k}^{\infty} f(j)dj < h \tag{19}$$

yielding, using (1) and (15)

$$\frac{C}{\alpha - 1}\frac{1}{\beta_k^{\alpha-1}} < h = T^{\frac{1}{\alpha}} \tag{20}$$

But by (4) we have that (20) is equivalent with

$$\beta_k > h \tag{21}$$

(in fact (21) could have been used directly since, by $g(h) = h$, (21) implies that there are less articles with item-densities β_k or higher than in the h-core). Formulae (21) and (18) imply the condition

$$h < A^{\frac{k}{\alpha k+1}} \tag{22}$$

which can always be reached for k sufficiently high, since

$$\lim_{k\to\infty} A^{\frac{k}{\alpha k+1}} = A^{\frac{1}{\alpha}} > \left(\frac{\alpha-2}{\alpha-1}A\right)^{\frac{1}{\alpha}} > h$$

by (16). That (21) can be reached is also clear from (5) and the fact that $\mu > 1$: this implies

$$\lim_{k\to\infty} \beta_k = \infty > h$$

and hence there exists a $k_0 \in \mathbb{N}$ such that, for all $k \geq k_0$, $\beta_k > h$, i.e. condition (21). This is reached for the natural number $k_0 = [k_1] + 1$, where

$$\mu^{k_1} = h$$

or

$$k_1 = \frac{\ln h}{\ln \mu} \tag{23}$$

($[k_1]$ is the largest natural number, smaller than or equal to k_1).

In the next section, we will study another set of CSS, now based on the median, instead of the average as used above.

III. CSS based on the median in a continuous Lotkaian framework

Similar as in (11) and (12) we define now a new set of CSS, now based on medians. Hence we define $v_0^* = T$ and for every $k = 1, 2, ...$

$$\int_0^{v_k^*} g(r)dr = \frac{1}{2}\int_0^{v_{k-1}^*} g(r)dr \tag{24}$$

$$g(v_k^*) = \beta_k^* \tag{25}$$

where $g(.)$ is as in (9).

We have the following Proposition concerning the CSS β_k^*, $k = 1, 2, ...$

Proposition 2: Let $\alpha > 2$. The β_k^* have the values, for $k = 1, 2, ...$

$$\beta_k^* = 2^{\frac{k}{\alpha - 2}} \tag{26}$$

Proof: The proof is by complete induction. For $k = 1$ we have, since $v_0^* = T$

$$\int_0^{v_1^*} g(r)dr = \frac{1}{2}\int_0^{T} g(r)dr$$

By (9) we find (since $\beta < 1$, since $\alpha > 2$ and by (10))

$$v_1^* = \left(\frac{1}{2}\right)^{\frac{1}{1-\beta}} T \tag{27}$$

By (10) we have that

$$\frac{1}{1-\beta} = \frac{\alpha-1}{\alpha-2} = \mu \tag{28}$$

by (3) and (4), so that (27) can also be written as

$$v_1^* = \left(\frac{1}{2}\right)^{\mu} T \tag{29}$$

By (8) and (25) we find

$$\beta_1 = \frac{B}{v_1^{\beta}}$$

This formula, (9), (10) and (27) yield

$$\beta_1 = \frac{T^{\beta}}{\left(\frac{1}{2}\right)^{\frac{\beta}{1-\beta}} T^{\beta}}$$

$$\beta_1 = 2^{\frac{1}{\alpha-2}} \tag{30}$$

by (10).

Assume now that (26) is valid for k replaced by $k-1$ and where

$$v_{k-1}^* = \left(\frac{1}{2}\right)^{\frac{k-1}{1-\beta}} T \tag{30}$$

Since v_k^* is defined as

$$\int_0^{v_k^*} g(r)dr = \frac{1}{2}\int_0^{v_{k-1}^*} g(r)dr$$

we have, by (8) and again since $\beta < 1$, that

$$v_k^{*1-\beta} = \frac{1}{2} v_{k-1}^{*1-\beta}$$

$$v_k^{*1-\beta} = \frac{1}{2}\left(\frac{1}{2}\right)^{k-1} T$$

, by (30). Hence

$$v_k^* = \left(\frac{1}{2}\right)^{\frac{k}{1-\beta}} T \tag{31}$$

So the induction argument is finished for ν_k^* (by (27), (30) and (31)). Using (8) and (25) we now have, using again (9) and (10), that

$$\beta_k^* = \frac{T^\beta}{\left(\frac{1}{2}\right)^{k\frac{\beta}{1-\beta}} T^\beta}$$

$$\beta_k^* = 2^{\frac{k}{\alpha-2}}$$

by (10)), ending the induction hypothesis. □

There is a direct relation between the β_k and β_k^* for all k. By (5), (3), (4) and (26) we have

$$\ln \beta_k^* = \frac{k}{\alpha-2} \ln 2$$

$$\alpha - 2 = \frac{k \ln 2}{\ln \beta_k^*}$$

$$\beta_k = \mu^k$$

$$\beta_k = \left(\frac{\alpha-1}{\alpha-2}\right)^k$$

$$\beta_k = \left(\frac{\frac{k \ln 2}{\ln \beta_k^*} + 1}{\frac{k \ln 2}{\ln \beta_k^*}}\right)^k$$

$$\beta_k = \left(\frac{\ln\left(2^k \beta_k^*\right)}{\ln\left(2^k\right)}\right)^k \tag{32}$$

This can also be rewritten as

$$\beta_k = \left(\frac{k + \log_2 \beta_k^*}{k}\right)^k \tag{33}$$

, using $\log_2$.

We can also relate β_k^* ($k = 1, 2, \ldots$) with β_1:

$$\beta_k^* = \left(2^{\frac{1}{\alpha-2}}\right)^k \tag{34}$$

Since $\beta_1 = \mu = \dfrac{\alpha - 1}{\alpha - 2}$ (by (3), (4) and (5)) we see that

$$\alpha = \frac{2\beta_1 - 1}{\beta_1 - 1}$$

and this yields in (33), after some calculation

$$\beta_k^* = 2^{(\beta_1 - 1)k} \tag{35}$$

This relation yields a relation between β_k^* and h: from (35) we have

$$\beta_1 = \frac{\ln_2 \beta_k^*}{k} + 1 \tag{36}$$

Now (3), (4) and (5), (36) and (16) yield

$$h = \left(\frac{A}{\dfrac{\ln_2 \beta_k^*}{k} + 1} \right)^{\frac{1}{2}}$$

$$h = \left(\frac{Ak}{\ln_2 \beta_k^* + 1} \right)^{\frac{1}{\alpha}}$$

for every $k = 1, 2, \ldots$

IV. Conclusions and suggestions for further research

Exact formulae for the Glänzel – Schubert CSS are given in the continuous infinite Lotkaian framework and their relation with the h-index is presented.

Then we define new CSS based on the median (instead of the average in the Glänzel – Schubert case) and formulae are presented. We also show the relation of these CSS with the Glänzel – Schubert one and with the h-index.

It would be interesting to define other CSS and compare them with the ones defined here.

Is it possible to construct CSS using h-type indices (instead of the average or the median)? What are their properties? This will be studied in a forthcoming paper. Can one define "natural" properties that CSS should have in order to be "good" CSS?

References

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

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

W. Glänzel (2008). What are your best papers? ISSI Newsletter 4(4), 64-67.

W. Glänzel (2009). The role of the h-index and the characteristic scores and scales in testing the tail properties of scientometric distributions. Preprint.

W. Glänzel and A. Schubert (1988). Characteristic scores and scales in assessing citation impact. Journal of Information Science 14, 123-127.

J. E. Hirsch (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(6), 16569-16572.