The Role of the h-index and the Characteristic Scores and Scales in Testing the Tail Properties of Scientometric Distributions

Wolfgang Glänzel

Wolfgang.Glanzel@econ.kuleuven.be
K.U. Leuven, Centre for R&D Monitoring and Dept. MSI, Leuven (Belgium)

glanzw@iif.hu
Hungarian Academy of Sciences, IRPS, Budapest (Hungary)

Abstract

The tail properties of scientometric distributions are studied in the light of the h-index and the characteristic scores and scales. A statistical test for the h-core is presented and illustrated using the example of four selected authors. Finally, the mathematical relationship between the h-index and characteristic scores and scales is analysed. The results give new insights into important properties of rank-frequency and extreme-value statistics derived from scientometric and informetric processes.

Introduction

The statistic analysis of the tail of publication-activity and citation distributions has always been a challenge to the scientometric/informetric research community. On the one hand, these distributions share the "long-tail" property with other distribution resulting from social processes and, on the other hand, extreme values and ranking according to productivity and citation impact reflect important aspects of outstanding research performance. Recently, the introduction of the h-index by *Hirsch* (2005) has essentially stimulated the advancement of methodological research in this topic; the h-core (e.g., *Rousseau*, 2006, *Burrell*, 2007, *Jin*, 2007, or most recently, *Egghe* and *Rousseau*, 2008) defined as the set of those papers, that have received at least *h* citations, is certainly one of the promising new approaches to analyse the high-end of citation distributions. Another model for grouping ranked observations, the method of *characteristic scores and scales* (CSS), was introduced as early as in 1980s (*Glänzel* and *Schubert*, 1988). While the h-index was designed for the measuring the research performance of individual scientists, CSS was developed for gauging the performance of a subset against classes defined on the total. The latter method has been successfully applied to the level of journals and scientific disciplines (e.g., *Schubert* et al., 1987, 1989). Similarly, the idea of applying the h-index to journals or other levels of aggregation has brought interesting new results (e.g., *Schubert* and *Glänzel*, 2007). Nevertheless, in the present paper we will go back to the roots, and apply both methods to the level of individual authors. The reason is very simple; citation distributions of individuals are more flexible than and not always as skewed as their counterparts at higher levels of aggregation. In particular, the study will aim at solving the following three problems.

- Constructing robust statistics for testing the tail of bibliometric distributions
- Finding appropriate truncation points for the tail of the distribution
- Studying the mathematical relation between the h-index and CSS

In order to accomplish the above tasks, four individual authors have been chosen from Thomson Reuters' *Web of Science* representing a group of scientists with about 25 or more years of professional experience in three different subject areas, particularly, in *mathematics, chemistry* and *social sciences*. All individuals are treated anonymously and therefore they are denoted by A, B, C and D, respectively. It should be stressed that there is no concordance

Proceedings of ISSI 2009, pages 120–130.
Edited by B. Larsen and J. Leta.

between the alphabetical order of author codes and the order of the above-mentioned subject areas. Only *articles*, *letters*, *notes*, *reviews* and *proceeding papers* published in journals are taken into consideration. Data was retrieved from the Web of Science on 03 December 2008.

In this study we proceed from an earlier paper and a note by the author (*Glänzel*, 2008a, 2008b). These earlier results will be extended and applied to the four individual authors. First, an introduction into the statistics of ranked samples is given, where the question of what type of statistics might be best suited will be answered. In the subsequent section the test presented by *Glänzel* (2008a) is briefly described and applied to the above samples. These sections refer to the first research question. The following sections are devoted to the analyses of the relation between the h-index and the characteristic scores, the h-cores and the characteristic scales, respectively. This refers to the residual research questions.

Statistics of ranked samples

In this section the rudiments of the statistics of ranked samples will be briefly described. We will also discuss what kind of statistics is suited for the analysis of the tail of scientometric distributions.

Let X be a random variable. In the present case X represents the citation rate of a paper. Since citation distributions are assumed to be discrete we can define the probability mass function of X which is denoted by $p_k = P(X = k)$ for each $k \geq 0$. The distribution function is denoted by $F_k = P(X < k)$. Furthermore we put $G_k := 1 - F_k = P(X \geq k)$. In order to allow approximate solutions we will not restrict the further considerations to discrete distribution models.

Consider now a given sample $\{X_i\}_{i = 1, \ldots, n}$ of size n. Assume that all elements are independent and identically distributed with F being the common distribution. Further assume that the sample elements X_i are ranked in decreasing order $X_1^* \geq X_2^* \geq \ldots \geq X_i^* \geq \ldots \geq X_n^*$. Although this can be readily obtained from an ordinary ordered sample by replacing index i by $(n–i+1)$ for all $i = 1, \ldots, n$, we will use the terms *rank statistics* of a statistical sample or simply *ranked sample* and denote the actual rank statistics by $R(r) := X_r^*$ in order to avoid any confusion with usual order statistics. We will actually use both notations, particularly, we use X_r^* whenever we would like to stress that the ranked observation originates from the sample $\{X_i\}_{i = 1, \ldots, n}$, otherwise, if we focus on the properties of the rank statistics, $R(r)$ is used.

In their comprehensive book on order statistics, *David* and *Nagaraja* (2003) present bounds and approximations for several statistical functions related to ordered samples, including moments, range and quantiles. As rank statistics can be obtained from order statistics, some of these methods can certainly be adopted to the present case as well. Nevertheless, most of these solutions are rather difficult and the tail of scientometrics distributions represents only a small part of the total. Robust approximations thus need not necessarily hold for the low-end of the distributions. We will therefore choose a different way. First, we discuss advantages and problems in using *medians* and the *expected values.*

The easiest way to obtain statistics and estimators for possible tests is using the median. In particular, one can readily determine a characteristic value which converges to the median of the corresponding ranked observation if the sample size tends to infinity. This idea is based on Gumbel's r-th characteristic extreme value (u_r).

$$u_r := G^{-1}(r/n) = \max \{k: G_k \geq r/n\} \quad (1)$$

n is a the size of a given sample with distribution F (see *Gumbel*, 1958). $R(r) := X_r^*$ can be considered an estimator of the corresponding r-th characteristic extreme value u_r. However, u_r is neither the median nor the expected value of $R(r)$. *Glänzel* and *Schubert* (1988) and *Schubert* and *Telcs* (1989) have shown that a minor correction results in an asymptotic solution for the median. In particular, we choose λ_r so that

$$P(X_r^* < G^{-1}(\lambda_r/n)) \sim P(X_r^* \geq G^{-1}(\lambda_r/n)) \sim 0.5 \quad \text{for} \quad n >> 1. \tag{2}$$

According to *Glänzel* and *Schubert* (1988) the *modified* Gumbel's r-th characteristic extreme values $u_r^* := G^{-1}(\lambda_r/n)$ with $\lambda_r \sim r-0.3$ asymptotically meet the median property and can readily be used for *multi-sample tests*, i.e. for testing whether the greatest, second, … r-th greatest observation of each sample is in line with the assumed joint model for all samples. Unfortunately, this does not work with single samples. Although the variables X_r^* ($r = 1, 2, ... k$; $k << n$) are apparently not independent, the particular events $\{X_r^* < u_r^*\}$ ($r = 1,2, ... k$; $k << n$) can still be considered independent. Nonetheless, the test statistics introduced in the above two articles cannot be used for statistical inference as the following example nicely illustrates. The h-core (that is, the set of papers that have received at least h citations) has been selected from the sample of author A. Although the citation rates $R(r)$ of nearly half the papers are below the corresponding median u_r^* and the statistic does not indicate significant deviation from the median, there is a clear trend showing that the corresponding hypothesis has yet to be rejected. This phenomenon is presented in Figure 1. The linear regression shows that the behaviour of tail elements is not in line with the assumed model, namely the first ten papers are below their median values while all other papers in the tail received more citations each than the corresponding median. (In this context, the underlying distribution model is not of particular interest since this example only serves to illustrate that a simple median test does not help decide whether the "top papers" of the sample are in line with the underlying model.)

This example might illustrate that we need a real test on the distribution of the tail elements, which includes that we have to decide where the tail actually begins, that is, where we have to cut off the tail from the rest of the distribution. A first approach has been presented in the paper by *Glänzel* and *Schubert* (1988) using the expected values of statistics derived from sample ranks. Unlike in the case of the median, here we have to assume a particular distribution of the underlying sample elements X_k ($k = 1, 2, ..., n$). Simple solutions are obtained if the sample elements have uniform or exponential distribution. *Glänzel* et al., (1984) have shown that $r \cdot \ln(X_r^*/X_{r+1}^*) = r \cdot \ln[R(r)/R(r+1)]$ are asymptotically independent, identically distributed random variables with joint exponential distribution if the elements of the original sample $\{X_i\}_{i = 1, ..., n}$ have a *Paretian distribution*. This property results in a simple test for the tail of the distribution (see *Glänzel* and *Schubert*, 1988). However, two important problems could not be solved. Firstly, where should the tail be cut off and secondly, the proposed statistics proved very sensitive to ties. In practice, rank statistics of integer-valued discrete distributions often include ties (i.e. $R(r) = R(r+1)$ for some $r = 1, 2, …$) resulting in $r \cdot \ln[R(r)/R(r+1)] = 0$. These ties can heavily distort the fit of the exponential distribution and the applied goodness-of-fit tests. In the following section we attempt to solve both problems.

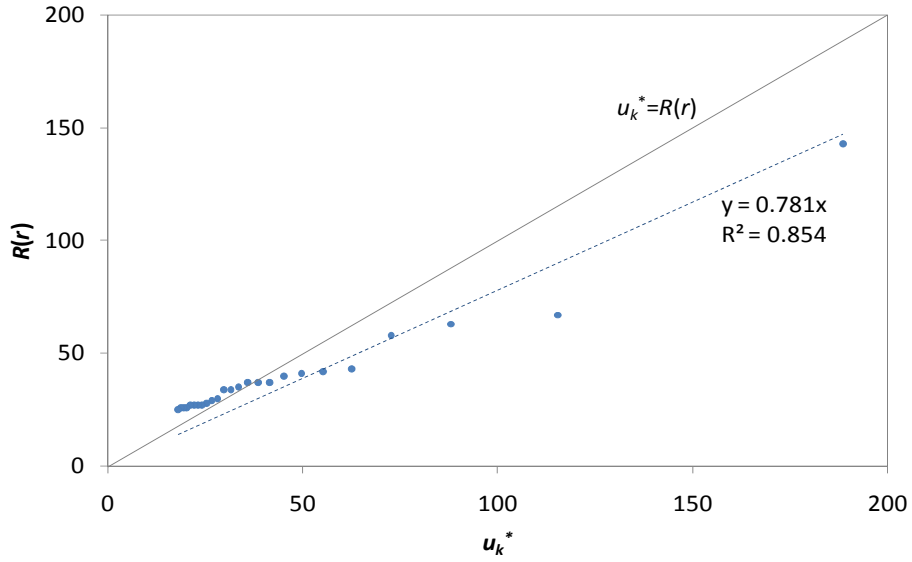

Figure 1 Example for invalid conclusions drawn from the median test

However, before we proceed we modify the underlying distribution model in order to facilitate the calculation of the estimators of the rank statistics. We assume that the citation distribution under study can be approximated by a non-negative continuous distribution. In the case of continuous distributions we will write $F(x)$ and $G(x)$ instead of F_x and G_x, respectively. Furthermore, we assume that the underlying citation rates follow a Pareto distribution of the second kind. This general form of the Pareto distribution, also referred to as *Lomax* distribution, can be obtained from the infinite beta distribution if one of the parameters is chosen 1 (e.g., *Johnson* et al., 1994). In particular, we say that the non-negative random variable X has a Lomax distribution if

$$G(x) = P(X \geq x) = N^{\alpha}/(N+x)^{\alpha}\ , \text{ for all } x \geq 0 \tag{3}$$

If x is large with respect to N, the parameter in the denominator can be neglected and we have

$$G(x) \sim N^{\alpha}/x^{\alpha}\ , \text{ for } x >> N. \tag{4}$$

Assuming a statistical sample with Lomax distribution and size n we obtain

$$G(u_r) \sim N^{\alpha}/u_r^{\alpha} = r/n, \text{ if } n >> r. \tag{5}$$

The assumption of a Lomax distribution instead of the Pareto distribution does not results in any essential restriction or change of the model; if statistics are based on $R(r)/R(r+1)$ ratios, the parameter N disappears.

Constructing a statistical test using the h-index

According to *Glänzel* (2006), the theoretical h-index (h) can be defined as

$$h := \max\ \{r\colon u_r \geq r\} = \max\ \{r\colon \max\ \{k\colon G_k \geq r/n\} \geq r\}\ . \tag{6}$$

Obviously, if there is such index r so that $u_r = r$ then we can write $h := u_h$.

The following important result (*Glänzel*, 2008a) can be directly obtained from Eqs (5) and (6).

$$\zeta(r) := r^{1/(\alpha+1)} \cdot u_r^{\alpha/(\alpha+1)} = N^{\alpha/(\alpha+1)} \cdot n^{1/(\alpha+1)}. \tag{7}$$

Since the right-hand side does not depend on the particular rank r, Eq. (7) immediately results in $\zeta(r)$:=const. Furthermore, we have $\zeta(h) = h$ by definition (i.e., $\zeta(h) = h^{1/(\alpha+1)} \cdot h^{\alpha/(\alpha+1)} = h$). The first step in constructing a test is replacing the Gumbel extreme values in Eq. (7) by the corresponding statistics, i.e., by the corresponding elements of the ranked sample. Furthermore, we define the following statistics

$$z(r) := r^{\frac{1}{\alpha+1}} \cdot R(r)^{\frac{\alpha}{\alpha+1}} \tag{8}$$

where $z(r)$ is an estimator of $\zeta(r) \equiv h$ for each $r << n$.
Although z itself is not an unbiased estimator of h, it has be shown (*Glänzel*, 2008a) that the statistics $r \cdot \ln(z(r)/z(r+1))$ are independent, exponentially distributed random variables with expectation

$$\mathrm{E}\{r \cdot \ln[z(r)/z(r+1)]\} = \{r \cdot \ln[r/(r+1)] + 1\}/(\alpha+1), \quad r < n. \tag{9}$$

The statistics $r \cdot \ln[z(r)/z(r+1)$ have the following properties.

Theorem (*Glänzel*, 2008): The expected value of $\mathrm{mean}_m(r \cdot \ln(z(r)/z(r+1)))$ tends to 0 as $m, n \to \infty$ and for its standard deviation we have

$$\mathrm{D}[\mathrm{mean}_m(r \cdot \ln[z(r)/z(r+1)])] = Z(m)/(\alpha+1) = m^{-1/2}/(\alpha+1), \tag{10}$$

with

$$Z(m) := (\alpha+1) \cdot \mathrm{E}\left[\frac{1}{m} \cdot \sum_{r=1}^{m} r \cdot \ln \frac{z(r)}{z(r+1)}\right] \quad \text{for all } m < n. \tag{11}$$

Furthermore, Eq. (11) immediately results the following approximation.

$$Z(h) \sim \frac{0.5 \ln(h+1) - 0.081}{h} \quad \square$$

This means that $Z(h)$ is close to 0 if h is large enough. Examples for Z values as a function of h are presented in Table 1.

Table 1 *h* and *Z*(*h*) values for different orders of magnitude for *h*

h	10	25	50	100	500	1000
***Z*(*h*)**	0.113	0.062	0.037	0.022	0.006	0.003

As a consequence of the above results, now a simple Welch-test can be applied to the above mean values. As described in the study by *Glänzel* (2008a), the statistic

$$w = (\alpha+1)\cdot h^{1/2}\{\text{mean}_h(r\cdot\ln[z(r)/z(r+1)])-Z(h)/(\alpha+1)\}$$

has approximately a standard normal distribution.

Table 2 *h*-related statistics of four authors active in the fields of chemistry, mathematics and social sciences

A			B			C			D		
r	***R*(*r*)**	***z*(*r*)**	***r***	***R*(*r*)**	***z*(*r*)**	***r***	***R*(*r*)**	***z*(*r*)**	***r***	***R*(*r*)**	***z*(*r*)**
17	28	23.7	8	28	18.4	24	37	32.0	13	32	23.7
18	27	23.6	9	25	17.8	25	35	31.3	14	30	23.3
19	27	24.0	10	22	16.9	26	35	31.7	15	29	23.3
20	27	24.4	11	21	16.9	27	34	31.5	16	29	23.8
21	27	24.8	12	20	16.9	28	34	31.9	17	28	23.7
22	26	24.6	13	17	15.5	29	33	31.6	18	27	23.6
23	26	25.0	14	17	15.9	30	33	32.0	19	26	23.4
24	26	25.3	15	16	15.7	31	32	31.7	20	26	23.8
25	25	25.0	16	16	16.0	32	32	32.0	21	23	22.3
26	25	25.3	17	15	15.6	33	31	31.7	22	21	21.3
h	$\hat{\boldsymbol{z}}$	***w***	***h***	$\hat{\boldsymbol{z}}$	***w***	***h***	$\hat{\boldsymbol{z}}$	***w***	***h***	$\hat{\boldsymbol{z}}$	***w***
25	23.9	-1.16	16	17.3	1.24	32	31.0	-0.69	21	27.9	3.48

The mean values $\hat{z}(h)$ and the means of the *w* statistics for the four authors are given in Table 2. A joint parameter of $\alpha = 2$ is assumed for all authors. This assumption is in line, for instance, with the observations by *Schubert* and *Glänzel* (2007). The mean values of the first three authors do not exceed their critical values belonging to the significance level of 0.95. Thus the statistics based on the author samples A, B and C reflect the *h*-property of the tail of citation distributions described by the right-hand side of Eq. (7) sufficiently well. The result of the test for the fourth author clearly suggests rejecting the hypothesis. His h-index is not in line with the model represented by authors A, B and C. At the same time we can conclude that *h* can be used as an appropriate truncation point for the tail of a distribution.

The h-index and the method of 'characteristic scores and scales'

Authors in many areas of the sciences and social have a individual h-index of 15, 20, 30 or even more according to the Web of Science. This is in part due to the recent extension of the database towards proceedings literature, but also a consequence of what can be called *inflationary bibliometric values* (*Persson* et al, 2004). A solution for shorting the tail is to subdivide distributions into several zones according to performance criteria. Such solution has been suggested, among others, by *Zitt* et al. (2007). Another method to split up distributions into performance classes and to delimit the tail of a distribution is the method of *characteristic scores and scales* introduced by *Glänzel* and *Schubert* (1988). This method was original developed to gauge subsamples against the standard set by the entire population, for instance, to compare journal citation impact with a field standard. Nonetheless, CSS can also be used to gauge sample elements against the own individual standard. Before we apply this method to the four selected authors, we give a concise description of the procedure.

In verbal terms, this method can be summarised as originated from iteratively truncating samples at their mean value and recalculating the mean of the truncated sample until the

procedure is stopped or no new scores are obtained. In the present study, we proceed from the citation distribution of the papers by a given author.
Using a more mathematical approach, we first put $\beta_0 = 0$ and $\nu_0 = n$ to obtain the *characteristic scores and scales* of an underlying citation distribution. β_1 is then defined as the sample mean

$$\beta_1 = \sum_{i=1}^{n} \frac{X_i}{n} = \sum_{i=1}^{n} \frac{X_i^*}{v_o} \ , \tag{12}$$

where the citations $\{X_i\}_{i=1}^{n}$ received by each paper by the author in question are then ranked in descending order $X_1^* \geq X_2^* \geq \ldots \geq X_n^*$ according to Section entitled "Statistics of ranked samples".
The value ν_1 is defined by the following inequality

$$X_{\nu_1}^* \geq \beta_1 \quad \text{and} \quad X_{\nu_1+1}^* < \beta_1 \, . \tag{13}$$

This procedure is repeated recurrently, particularly,

$$\beta_k = \sum_{i=1}^{\nu_{k-1}} \frac{X_i^*}{v_{k-1}} \ , \tag{14}$$

and ν_k is chosen so that

$$X_{\nu_k}^* \geq \beta_k \quad \text{and} \quad X_{\nu_k+1}^* < \beta_k \, , \ k \geq 2 \, . \tag{15}$$

The properties $\beta_0 \leq \beta_1 \leq \ldots$ and $\nu_0 \geq \nu_1 \geq \ldots$ are obvious from the definition. The interval $[\beta_k, \infty)$ then contains ν_k papers by definition. This procedure is repeated until no new scores are obtained (see Glänzel and Schubert, 1988). By practical reasons, we stop the procedure already at k if $\beta_{k+1} = X_1^*$. We denote the greatest such index k by k_{max}. Now we can define the following zones or classes. $[0, \beta_1)$ is the class of 'poorly cited' papers, $[\beta_1, \beta_2)$ contains 'fairly cited' papers, $[\beta_2, \beta_3)$ and $[\beta_3, \infty)$ are the two classes of highly cited papers called 'remarkably cited' for $k = 2$ and 'outstandingly cited' papers for $k = 3$, respectively. If $k > k_{max}$ then the corresponding class is empty by definition. For our purpose we choose the classes of *remarkably cited* and *outstandingly cited* papers with $k = 2$ and $k = 3$, respectively (cf. *Glänzel* and *Schubert*, 1988).

Citation ranks for the 15 most cited papers are presented in Table 3. In addition, sample size, h-index and the mean citation rate are given.

The interval $[\beta_0, \infty)$ contains all papers (ν_0), $[\beta_1, \infty)$ all papers with attribute *fairly, remarkably* and *outstandingly* cited (ν_1), $[\beta_2, \infty)$ papers that are *remarkably* and *outstandingly* cited (ν_2) and, finally, $[\beta_3, \infty)$ forms the group of *outstandingly* cited papers (ν_3). Note that these attributes refer to the *authors' own standards* here. Zones above score β_3 usually contain very few papers, or are even empty. The scores β_k for the three authors and the number of papers in the corresponding zones ν_k are presented in Table 4. The score and scale for $k = 2$ is emphasised since this might, from the viewpoint of the analysis of the tail of scientometric distributions, form an alternative to the h-core.

Table 3. The 15 most cited papers of three selected authors A, B and C (*n*: number of publications, *h*: h-index, *x*: mean citation rate)

Rank	Citations			
	A	**B**	**C**	**D**
1	143	86	127	793
2	67	61	104	120
3	63	45	102	101
4	58	36	80	71
5	43	34	69	65
6	42	29	66	53
7	41	28	59	49
8	40	28	58	48
9	37	25	53	43
10	37	22	51	42
11	37	21	51	42
12	35	20	50	33
13	34	17	48	32
14	34	17	48	30
15	30	16	47	29
…	…	…	…	…
n	110	73	161	58
h	25	16	32	21
x	16.3	9.6	19.5	33.9

Table 4. Thresholds β_k and number of papers in $[\beta_k, \infty)$ for the authors A, B, C and D ($k = 3$ corresponds to the set of outstandingly cited authors)

k	β_k				Papers ν_k			
	A	**B**	**C**	**D**	**A**	**B**	**C**	**D**
1	16.3	9.6	19.5	33.9	43	22	59	11
2	**32.1**	**26.2**	**39.5**	**129.7**	**14**	**8**	**22**	**1**
3	50.8	43.4	59.5	:	4	3	6	:
4	82.8	64.0	91.3	:	1	1	3	:
5	:	:	111.0	:	:	:	1	:

It should be noticed that the method introduced above provides acceptable results if the distribution is skewed enough. This is usually the case if $x \leq h$. The citation distributions of authors A, B and C meet this criterion. Otherwise, if $x > h$, the identification of the tail containing the top papers on basis of the suggested method becomes rather difficult. For author D, for instance, we obtain an almost degenerate tail (cf. Table 4). He has 58 papers that received 1954 citations in total. The mean citation rate amounts to 33.9 and the h-index is 21. In this case $[\beta_3, \infty)$ is already empty and $[\beta_2, \infty)$ contains only one paper, namely the most cited one. In this context, we just mention in passing that the share of *remarkably* and *outstandingly* cited papers is surprisingly stable for authors A, B and C. ν_2/ν_0 ranges between 11% and 14% and ν_3/ν_0 amounts to about 4%. Only author D considerably deviates from this scheme. This example raises the question of how the relation between *characteristic scores and scales* and *h-related indexes* determine the high end of citation distributions. This question will be answered in the following.

In order to analyse the relation between β_k and ν_k values and h-related indexes one has only to modify the initial condition in Eq. (12). Instead of $\beta_1=\sum X_i^*/\nu_1$, choosing the initial condition $\beta_1 = \nu_1$ immediately results in the definition of the h-index since $X^*_{\nu_1} \geq \nu_1$ and $X^*_{\nu_1+1} < \nu_1$ is a version of its definition (cf. *Glänzel*, 2006). Consequently, Eq. (12) yields Jin's A-index (Jin, 2007), which is actually the average citation rate of the h-core. In particular, one obtains

$$A = \beta_2 = \sum_{i=1}^{\nu_1} \frac{X_i^*}{\nu_1} \text{ with } \beta_1 = \nu_1 = h. \qquad (12^*)$$

This is an interesting result that illustrates that both self-adjusting approaches to the statistics of distribution tails, i.e., Hirsch's approach and the *characteristic scores and scales*, result in consistent solutions. Therefore the question arises of both solutions could be also made truly congruent. In order to analyse this, one has to distinguish the following three cases.

(i) $h = x$,

(ii) $h > x$,

(iii) $h < x$,

where h is the h-index and x is the mean citation rate.

Case (i): Since h is always an integer while x can be an arbitrary rational number, we assume here $h = \lceil x \rceil$, where $\lceil \cdot \rceil$ is called the *ceiling function* giving the smallest integer $\geq x$. In this case both the Hirsch and the CSS approach coincide and yield the same results, i.e., $\beta_1=h$ and $\beta_2=A$.

Case (ii): The question arises of whether the h-core could be extended by removing uncited and poorly cited articles so that $\beta_1=h=\sum X_i^*/\nu'_0$ is obtained, where $\nu'_0<n$ is a number that can be used to truncate the original ranked sample to be in line with *case (i)*. This means that the elements $X_{\nu'_0+1}{}^* \geq \ldots \geq X_n{}^*$ have to be removed. Since $h > x$, there is always an approximate solution, that is, $\beta_1 \sim h$. There for we call this extension of the h-core *regular* and the value $\theta:=n-\nu'_0$ indicates the extent of modification.

Case (iii): In this case the mean exceeds the h-index. Regular core extension is obviously not possible; removing elements from the low end of the ranked sample would further increase the mean value. Therefore we can sequentially add uncited papers until we reach the solution $\beta_1=h=\sum X_i^*/\nu'_0$. Here $\nu'_0>n$ and $X_{n+1}{}^* = \ldots = X_{\nu'_0}{}^* = 0$. Again, there is always an approximate solution with $\beta_1 \sim h$. This extension is not regular; this is indicated by negative values of $\theta = n-\nu'_0$.

In the above examples the following θ values are obtained. For author A index θ amounts to 45. In particular, we have

$$h \sim \beta_1 = \sum_{i=1}^{\nu'_0} \frac{X_i^*}{\nu'_0} = \sum_{i=1}^{65} \frac{X_i^*}{65} = 25.0 \quad \text{and} \quad A \sim \beta_2 = \sum_{i=1}^{\nu_1} \frac{X_i^*}{\nu_1} = \sum_{i=1}^{25} \frac{X_i^*}{25} = 40.4\,.$$

Similarly, we have θ=31 for author B with $h \sim 16.0$ and $A = 31.3$, and for author C the extension index θ even amounts to 75. Consequently, we have $h \sim 31.9$ and $A = 51.6$. In the case of the citation distribution of author D we obtain a non-regular extension with θ=–36. This means that we have to add 36 uncited items in this case. Hence we obtain $h \sim 20.9$ and $A = 81.4$. The ν_1 values, of course, coincide with the corresponding h-indexes, ν_2 amounts to 7, 5, 9 and 3 for author A, B, C and D, respectively. These results also suggest that both the h-index and alternatively whether the characteristic score β_2 or Jin's A-index can be used as appropriate truncation points for the tail of the distribution.

Conclusions

This study has shown how appropriate tests can be used to test h-index conformity and to identify cases when the distribution tail is not in line with the commonly assumed distribution models. This can also be considered a step towards making h-indexes in a sense comparable. Furthermore, the relation between characteristic scores and the h- and A-index was shown and analysed. Both methods provide excellent tools for studying the tail behaviour of bibliometric distributions.

The 'extension' of the h-core by adding uncited papers or removing uncited and possibly poorly cited papers in order to "synchronise" the Hirsch approach with the method of *characteristic scores and scales* is an interesting option since neither the h-index nor the A-index changes if the low-end of the distribution is modified. Modification of the sample below the h-index merely results in adjusting the characteristic scores and scales to become conform with the Hirsch approach. This also means that statistical tools developed for the analysis of h-index related questions can be adopted for the CSS approach and vice versa.
Both h-index related indicators and characteristic scores proved useful as truncation point for rank frequency analysis as well, for instance, in the context of truncating the ranked sample for testing tail properties or for selecting 'top publications'.

References

Burrell, Q. L. (2007), On the h-index, the size of the Hirsch core and Jin's A-index, *Journal of Informetrics*, 1 (2) 170–177.

David, H. A., Nagaraja, H. N. (2003), *Order Statistics* (3rd Edition). Wiley, New Jersey.

Egghe, L., Rousseau, R. (2008), An h-index weighted by citation impact. *Information Processing & Management*, 44(2), 770–780.

Glänzel, W., Schubert, A., Telcs, A. (1984), *Goodness of Fit Test for the Tail of Distributions*. Bolyai Colloquium on Goodness of fit (Debrecen, Hungary, June 25-28, 1984).

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, 75–83.

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

Glänzel, W. (2008a), On some new bibliometric applications of statistics related to the h-index. *Scientometrics*, 77(1), 187–196.

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

Gumbel, E. J. (1958). *Statistics of extremes*. New York: Columbia University Press.

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).

Jin, B.H., Liang, L.M., Rousseau., R., Egghe, L. (2007), The R- and AR-indices: Complementing the h-index. *Chinese Science Bulletin*, 52 (6) 855–863.

Johnson, N. L., Kotz, S., Balakrishnan, N. (1994), *Continuous univariate distributions*. Volume 1, 2nd Edition, John Wiley & Sons, Ney York.

Persson, O., Glänzel, W., Danell, R. (2004), Inflationary bibliometric values: the role of scientific collaboration and the need for relative indicators in evaluative studies, *Scientometrics*, 60 (3), 421–432.

Rousseau, R. (2006), New developments related to the Hirsch index. *Science Focus*, 1(4), 23–25 (in Chinese); English version accessible at http://eprints.rclis.org/archive/00006376/.

Schubert, A., Glänzel, W., Braun, T., (1987), Subject field characteristic citation scores and scales for assessing research performance. *Scientometrics,* 12(5-6), 267-291.

Schubert, A., Glänzel, W., Braun, T., (1989), Scientometric datafiles. A comprehensive set of indicators on 2649 journals and 96 countries in all major fields and subfields 1981-1985. *Scientometrics,* 16(1-6), 3–478.

Schubert, A. Glänzel, W. (2007), A systematic analysis of Hirsch-type indices for journals. *Journal of Informetrics*, 1 (3), 179–184.

Schubert, A., Telcs, A. (1989), Estimation of the publication potential in 50 U.S. states and in the District of Columbia based on the frequency distribution of scientific productivity. *Journal of the American Society for Information Science*, 40(4), 291–297.

Zitt M, Bassecoulard E, Filliatreau G, Ramanana-Rahary S. (2007), *Revisiting country and institution indicators from citation distributions: Profile performance measures*. Proceedings of ISSI 2007: 11th International Conference of the International Society for Scientometrics and Informetrics, 797–802.