BACKGROUND

Jorge E. Hirsch (2005) proposed a new measure "to quantify an individual's scientific output".

He called his measure h-index:

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.

Egghe (2006) introduced the \( \textit{g-index} \) ("a second-moment \( h \)-index"). Besides inheriting all good properties of the \( h \)-index, it better takes into account the citation scores of the top articles.

Bihui et al. (2007) supplemented the \( h \)-index by their (A)R-index to improve its properties. A-index is the average citation rate of the Hirsch core and \( R=(A\cdot h)^{1/2} \).

First applications to research evaluation were made by Bormann & Daniel (2005, 2007). They demonstrated the use and validity of the \( h \)-index in comparison with peer review decisions.

Van Raan (2005) showed that the \( h \)-index correlates with other bibliometric indicators of ‘significance’ but he also stressed that scientific performance can hardly be expressed simply by one indicator alone.

Possible application of the \( h \)-index to the meso level, particularly to scientific journals was shown by Braun et al (2005) and Schubert & Glänzel (2007).
PROPERTIES OF THE H-INDEX

Glänzel (2006): “A scientist has index h if h is the largest number of his/her n papers having received at least h citations each”, that is,

\[ h := \max \{ r : R(r) \geq r \} \]

\[ = \max \{ r : \max \{ k : [1-F_k] \geq r/n \} \geq r \} \]

*R : Rank frequency, r : rank,

\( F \) : empirical frequency distribution and \( G_k = 1 - F_k \),

\( u_r = G^{-1}(r/n) \) is called Gumbel's \( r \)-th characteristic extreme value (Gumbel, 1958).

\( u_r \) is a theoretical value of the \( r \)-th empirical rank \( R(r) \).

For a generalised Pareto distribution (Lomax distribution) we have

\[ G(u) \sim N^\alpha u_r^\alpha = r/n, \text{ if } n >> r \]

Consequently, we have \( r \cdot u_r^\alpha = N^\alpha \cdot n \) and

\[ \zeta(r) := r^{1/(\alpha+1)} \cdot u_r^{\alpha/(\alpha+1)} = N^{\alpha/(\alpha+1)} \cdot n^{1/(\alpha+1)} \]

Since \( E(X) = N/(\alpha-1) \) for Lomax distributions, the Pareto property results in the following equation:

\[ \zeta(r) := r^A \cdot u_r^{(1-A)} \sim c(\alpha) \cdot n^A \cdot E(X)^{(1-A)} = y(A); \ A=1/(\alpha+1) \]

Since \( \zeta(h) = h \) by definition (i.e. \( u_h = h \)) and \( y \) does not depend on \( r \), we have \( \zeta(r) = h \) for all \( r << n \).
RELATIONSHIP WITH OTHER INDICATORS

In the case of scientific journals we have found the following relationship for 3-year citation windows with $\alpha = 2$:

$$h \sim c \cdot x^{2/3} \cdot n^{1/2}$$

with $c = O(1)$, where $n$ is the number of papers in the journal and $x$ the journal impact.

We found this result stable for small citation windows and it was independent of the subject field (see Schubert & Glänzel, 2007).

Correlation of the journal h-index with $n^{1/3}x^{2/3}$
(all science fields combined, PY=2002, cites=2002-2004)

$$y = 0.757x$$
$$R^2 = 0.956$$
Correlation of the journal h-index with $n^{1/3}x^{2/3}$
(chemistry, PY=2002, cites=2002-2004)

\[ y = 0.7315x, \quad R^2 = 0.9809 \]

Correlation of the journal h-index with $n^{1/3}x^{2/3}$
(virology, PY=2002, cites=2002-2004)

\[ y = 0.718x, \quad R^2 = 0.974 \]
More generally one obtains

\[ h \sim c^* x^{\alpha/(\alpha+1)} n^{1/(\alpha+1)} \]

with \( c^* \sim (\alpha-1)^{\alpha/(\alpha+1)} \).

With growing citation window \( \alpha \) (and \( c^* \)) decreases. This property is in line with observations by Vlachy (1976) and Pao (1986).

The \( c^* \) values are varying slowly, and are of order 1 for \( \alpha \) values ranging between 1.2 and 2.5.

A similar suggestion was made by Lindsey (1978) for journal impact. For his Corrected Quality Ratio (CQ) we have \( \text{CQR}^{0.4} = n^{0.4} x^{0.6} \) (i.e. \( \alpha = 1.5 \)).

---

**Correlation of the journal h-index with \( n^{1/3}x^{2/3} \)**

(all fields combined, PY=1980, cites=1980-1982)

\[ y = 0.784x \]
\[ R^2 = 0.957 \]
**THE z(r) STATISTICS AND THEIR PROPERTIES**

\( \zeta(r) \) is a function which, however, does not depend on \( r \), thus we have \( \zeta(r) \equiv h \) for all \( r \ll n \). Replacing the Gumbel extreme values by their empirical values we obtain the following statistics

\[
z(r) = r^h \cdot R(r)^{(1-A)}.
\]

\( z \) is not an unbiased estimator of \( h \) but the following property holds for all \( z(r) \).

**Proposition 1:** \( r \cdot \ln(z(r)/z(r+1)) \) are independent (non identically) exponentially distributed r.v.s with

\[
\mathbb{E}\{r \cdot \ln[z(r)/z(r+1)]\} = \{r \cdot \ln[r(r+1)] + 1\}/(\alpha+1)
\]

which tends to 0 if \( r \to \infty \).

**Corollary**: For the mean of the *Hirsch core* we have

$$Z(h) := (\alpha + 1) \left\{ \sum r \ln(z(r)/z(r+1)) \right\}/h \sim 0$$

provided $h$ is large enough.

**Proof**: $E(Z(h)) = \ln((h+1)!/(h+1)^{h+1})+1 \sim$

$$\sim \{0.5 \ln(h+1)/h - 0.081/h\}$$

The rest of the proof is straightforward. + + +

<table>
<thead>
<tr>
<th>$h$</th>
<th>10</th>
<th>25</th>
<th>50</th>
<th>100</th>
<th>500</th>
<th>1000</th>
</tr>
</thead>
<tbody>
<tr>
<td>$Z(h)$</td>
<td>0.113</td>
<td>0.062</td>
<td>0.037</td>
<td>0.022</td>
<td>0.006</td>
<td>0.003</td>
</tr>
</tbody>
</table>

The deviation for the individual $r$ values is large but the median of the $z(r)$ values proved a strikingly robust estimator of $h$.

For the *Hirsch core* we obtain $y = \text{median}(z(r)) \sim h$ as estimator for the h-index (with $\alpha = 2$ for a tree-year citation window).

Also the $\text{mean}_r\{r \ln[z(r)/z(r+1)]\}$ values around zero substantiate the applicability of the model. (Note that $\text{mean}_r\{r \ln[z(r)/z(r+1)]\} = Z(h)/(\alpha+1)$.)
Proposition 2: The expected value of \( m \cdot \ln(z(r)/z(r+1)) \) tends to 0 as \( m,n \to \infty \) and for its standard deviation we have \( D[\text{mean}_m(r \cdot \ln(z(r)/z(r+1)))] = m^{-\frac{3}{2}}((\alpha+1)) \) for all \( m < n \).

We can apply a Welsch-test if \( m \geq 25 \) (cf. Schubert and Glänzel, 1983).

Examples

Assuming an h-index of 25 and a small citation window with \( \alpha=2 \), we can accept the h-property for the z statistics if

\[
\text{mean}_h(r \cdot \ln(z(r)/z(r+1))) \in (-0.100, 0.162).
\]

For \( h=100 \), we have

\[
\text{mean}_h(r \cdot \ln(z(r)/z(r+1))) \in (-0.058, 0.072).
\]
Hirsch-type indexes for journals (PY=2002)

<table>
<thead>
<tr>
<th>Journal</th>
<th>h-index</th>
<th>$y = \text{median}(z(r))$</th>
<th>$\text{mean}_r(z(r)/z(r+1))$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Science</td>
<td>106</td>
<td>104.5</td>
<td>-0.036</td>
</tr>
<tr>
<td>Nature</td>
<td>108</td>
<td>110.0</td>
<td>+0.021</td>
</tr>
<tr>
<td>Cell</td>
<td>78</td>
<td>78.0</td>
<td>-0.038</td>
</tr>
<tr>
<td>Blood</td>
<td>50</td>
<td>49.5</td>
<td>-0.053</td>
</tr>
<tr>
<td>Angewandte Chemie</td>
<td>43</td>
<td>43.6</td>
<td>+0.017</td>
</tr>
<tr>
<td>Astrophysical Journal</td>
<td>49</td>
<td>47.1</td>
<td>-0.056</td>
</tr>
<tr>
<td>Analytical Chemistry</td>
<td>32</td>
<td>32.0</td>
<td>-0.021</td>
</tr>
<tr>
<td>Trends in Neurosciences</td>
<td>24</td>
<td>24.5</td>
<td>-0.026</td>
</tr>
</tbody>
</table>

CONCLUSIONS

Theoretical considerations and empirical analysis lead to the conclusion that the h-index strongly correlates with an indicator combining publication output and mean citation rate.

However, the composite indicator is not designed to substitute the h-index.

Furthermore, Hirsch-related statistics can be used for the analysis of the tails of bibliometric distributions at any level of aggregation. In particular, the h-index is useful as truncation point for rank frequency analysis (e.g. Hirsch core analysis and Z statistics).