Information Processing and Management 45 (2009) 484–489

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

An econometric property of the g-index

L. Egghe *

Universiteit Hasselt (UHasselt), Campus Diepenbeek, Agoralaan, B-3590 Diepenbeek, Belgium
Universiteit Antwerpen (UA), Stadscampus, Venusstraat 35, B-2000 Antwerpen, Belgium

ARTICLE INFO

Article history:
Received 19 December 2008
Received in revised form 31 March 2009
Accepted 2 April 2009
Available online 6 May 2009

Keywords:
g-Index
h-Index
Hirsch
R-index
Kosmulski
Lorenz curve

ABSTRACT

Let $X = (x_1, \ldots, x_N)$ and $Y = (y_1, \ldots, y_N)$ be two decreasing vectors with positive coordinates such that $\sum_{j=1}^N x_j = \sum_{j=1}^N y_j$ (representing e.g. citation data of articles of two authors or journals with the same number of publications and the same number of citations (in total)). It is remarked that if the Lorenz curve $L(X)$ of X is above the Lorenz curve $L(Y)$ of Y, then the g-index $g(X)$ of X is larger than or equal to the g-index $g(Y)$ of Y. We indicate that this is a good property for so-called impact measures which is not shared by other impact measures such as the h-index. If $L(X) = L(Y)$ and $\sum_{j=1}^N x_j > \sum_{j=1}^N y_j$ we prove that $g(X) \geqslant g(Y)$. We can even show that $g(X) > g(Y)$ in case of integer values x_i and y_i and we also investigate this property for other impact measures.

1. Introduction

First we will re-introduce Lorenz concentration theory and then discuss some well-known impact measures.

1.1. Concentration theory: discrete Lorenz curve

Lorenz concentration theory was invented by Lorenz in 1905 (Lorenz (1905)) and is used to measure the concentration or inequality between a set of positive numbers (e.g. the salaries of employees). Lorenz concentration theory has also found its way into informetrics, e.g. to measure the inequality in citations of papers of an author or to measure the inequality in productivity of authors (i.e. in the number of papers of these authors) – see basically Egghe (2005, chap. IV) and many references therein. The application of Lorenz concentration theory in informetrics is no surprise since – as in econometrics – many (if not all) source-item distributions are very skew: many sources have few items and few sources have many items – see Egghe (2005, chap. I and IV), where these inequalities are described via the laws of Lotka and Zipf (but we will not use these laws in this paper).

Let us, briefly, describe Lorenz concentration theory. Let $X = (x_1, \ldots, x_N)$ be a decreasing vector with positive coordinates x_i, $i = 1, \ldots, N$. The Lorenz curve $L(X)$ of X is the polygonal curve connecting $(0, 0)$ with the points $\left(\frac{i}{N}, \sum_{j=1}^{i} a_j\right)$, $i = 1, \ldots, N$, where

$$a_i = \frac{x_i}{\sum_{j=1}^N x_j} \tag{1}$$

Note that for $i = N$ we have $(1, 1)$ as end point of $L(X)$. Let X and $Y = (y_1, \ldots, y_N)$ be two such vectors. We say that X is more concentrated than Y if $L(X) > L(Y)$. We also say that the coordinates of X are more unequal than the ones of Y.

* Address: Universiteit Hasselt (UHasselt), Campus Diepenbeek, Agoralaan, B-3590 Diepenbeek, Belgium.
E-mail address: leo.egghe@uhasselt.be

0306-4573/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2009.04.001

That Lorenz curves are the right tool to measure concentration or inequality is seen by the result of Muirhead (1903) stating[1] that $L(X) > L(Y)$ if and only if X is constructed, starting from Y, by a finite applications of elementary transfers. An elementary transfer (e.g. on $Y = (y_1, \ldots, y_N)$) changes Y into the vector

$$(y_1, \ldots, y_i + h, \ldots, y_j - h, \ldots, y_N) \tag{2}$$

where $1 \leqslant i < j \leqslant N$ and $h > 0$. Since Y is decreasing, this means, in econometric terms that "we take away ($h > 0$) from the poor (j) and give it to the rich (i)" which indeed yields a more unequal (concentrated) situation, which is applied repeatedly to yield X out of Y.

For further use, we also note the following. If $L(X) = L(Y)$ we have that $X = aY$ for a certain value $a > 0$, namely (use (1))

$$a = \frac{\sum_{j=1}^{N} x_j}{\sum_{j=1}^{N} y_j} \tag{3}$$

Indeed, denote a_i for X as in (1) and denote

$$b_i = \frac{y_i}{\sum_{j=1}^{N} y_j} \tag{4}$$

$i = 1, \ldots, N$ for Y. Since $L(X) = L(Y)$ we have $a_1 = b_1$, $a_1 + a_2 = b_1 + b_2, \ldots, a_1 + \cdots + a_N = b_1 + \cdots + b_N$ (=1) so that $a_i = b_i$ for all $i = 1, \ldots, N$ from which $X = aY$ with a as in (3) follows.

1.2. Impact measures

This theory will now be linked with a set of new impact measures, defined only since 2005 onwards. Impact measures are defined on the same type of vectors X and Y as described above. Usually we now interpret the coordinates as the number of citations to N papers of an author or a journal X or Y, but this is not really necessary. In the above interpretation, impact measures then measure the overall visibility, impact,... of a journal or of an author's career. Let us briefly re-introduce the impact measures that we will use in this paper.

It all started with the introduction of the Hirsch index (or h-index): Hirsch (2005). Let the vector $X = (x_1, \ldots, x_N)$ be as above: a decreasing sequence of N positive numbers. Then X has h-index h if $r = h$ is the largest rank such that each paper on rank $1, \ldots, h$ has h or more citations. As mentioned in many papers, h is a unique index that combines quantity (number of papers) with quality (or rather visibility) (number of citations to these papers) and it is a robust measure in the sense that it is not influenced by a set of lowly cited papers nor by the exact number of citations to the first h papers in the ranking in X (the so-called h-core) (see Braun, Glänzel, and Schubert (2006), Egghe (2006)). However, the latter property is considered as a disadvantage of the h-index: once a paper is in the h-core, it does not matter how many citations (above h) it received or will receive: this does not influence the value of h. We agree with a measure that does not take into account some (or several) lowly cited papers as long as it takes into account the number of citations to the highly cited papers. Therefore, Egghe introduced in 2006, see Egghe (2006), an improvement of the h-index: the g-index.

Note that the papers in the h-core, together, have at least h^2 citations. Now the g-index is the largest rank $r = g$ such that all papers on rank $1, \ldots, g$, together, have at least g^2 citations. Obviously $g \geqslant h$ but that is not an important issue here. It has been recognized that the g-index has more discriminatory power than the h-index (Schreiber, 2008a, 2008b; Tol, 2008).

The R-index, introduced in Jin, Liang, Rousseau, and Egghe (2007), serves the same goal as the g-index on the improvement of the h-index although it uses the h-index in its definition:

$$R = \sqrt{\sum_{i=1}^{h} x_i} \tag{5}$$

where $X = (x_1, \ldots, x_N)$ is as above and h is the h-index of X. Note again, as in the case of the g-index, that the actual x_i-values (the highest ones) are effectively used.Kosmulski's $h^{(2)}$-index is similar to the h-index but now one requires $r = h^{(2)}$ to be the largest rank such that each paper on rank $1, \ldots, h^{(2)}$ has $(h^{(2)})^2$ or more citations – see Kosmulski (2006). It was introduced to save time in calculating impact measures: $h^{(2)}$ is much smaller than h since one requires (at least) the square of the rank as the number of citations (see below for the impact measure values of this author).

Since $g \geqslant h$ it might be interesting to apply Kosmulski's idea also to the g-index. Note that the first $h^{(2)}$ papers, together, have, at least $(h^{(2)})^3$ citations. We now define $g^{(2)}$ as the highest rank such that the first $g^{(2)}$ articles, together, have, at least $(g^{(2)})^3$ citations. The $g^{(2)}$ impact measure is new and is introduced here for the first time.

Table 1 gives the citation data of this author, based on the Web of Science on July 24, 2008. We only present the first 23 papers since we do not need higher ranks. As needed for the calculation of the g-index, we also present the cumulative scores and the squares of the ranks.

[1] Muirhead's theorem was published in 1903, two years before Lorenz introduced the Lorenz curve (Lorenz (1905)). Muirhead's theorem hence did not use the Lorenz terminology but a combinatorial variant of it. Here we present the Lorenz variant of Muirhead's theorem. Muirhead's theorem can also be found in Hardy, Littlewood, and Pólya (1952) and in Egghe and Rousseau (1991).