Central Indexes to the Citation Distribution: A complement to the h-index

Pablo Dorta-González[1] and María-Isabel Dorta-González[2]

[1] *pdorta@dmc.ulpgc.es*
Universidad de Las Palmas de Gran Canaria (Spain)

[2] *isadorta@ull.es*
Universidad de La Laguna (Spain)

Introduction

The citation distribution of a researcher shows the impact of their production and determines the success of their scientific career. Some bibliometric indexes that try to synthesize in a numerical value the principal characteristics of this distribution have been proposed recently.

In contrast with other bibliometric measures, the biases that the distribution tails provoke, are reduced by the h-index. A researcher has an h-index when h of its publications have received at least h citations, and the rest have h or less citations (Hirsch, 2005). This index has been extensively studied (see reviews by Bornmann & Daniel, 2007; Alonso et al., 2009).

Based on the definition of the Hirsch core, several authors have proposed new variants. The g-index (Egghe, 2006) considers all citations of the g most cited articles, and represents an average citation of these g publications. In fact, the h-index and the g-index are special cases of a family of Hirsch index variants (Schreiber, 2010). Similarly, the A-index and the AR-index (Jin et al., 2007) have the particularity of taking into account the citations of the Hirsch core.

In this work, a new complement to the h-index that increases the accuracy giving information about the shape of the citation distribution is given. Variants which have been proposed previously are a function of all citations included in a core of most cited papers. By contrast, we establish an upper limit to the maximum number of citations considered for each publication in order to reduce the effect that isolated successes and/or large collaborations may have on the final result (as was pointed out by Hirsh). This upper limit can be modified without further changing the radius of the central index.

Central indexes

In this work, two original sets of indexes, the central area indexes and the central interval indexes, that complement the h-index to include the central shape of the citation distribution, are proposed.

Given the published articles of an author in decreasing order of citations, let be c_i the number of citations received by the publication i, $i=1,\ldots,N_p$.

$H = h^2$ is a lower bound for the number of citations of those papers in the Hirsch core. The upper tail U is the excess citations received by the Hirsch core over the lower bound. The lower tail L is the number of citations received by those papers outside the Hirsch core.

It seems reasonable to measure part of U and L in order to complement the h-index with the area around H. Thus, the discrimination capacity is increased.

Let E (F) be the upper (lower) area next to H, that is, the part of the upper tail U (lower tail L) in the citation distribution closest to H. The lower area corresponds to those articles that will likely contribute to increasing the value of h in the future, since they are closer to the Hirsch core. The upper area includes those citations that

will form part of H at the time the h-index increases its value.

The central area index of radius j is defined as the citations of the h+j most cited papers limited to the number of citations of paper h-j. That is, the citations of those papers in the Hirsch core, restricted by the citations of paper h-j, jointly with the citations of papers from h+1 to h+j. The arithmetic definition is the following:

$$A_j = (h-j)\cdot c_{h-j} + \sum_{i=h-j+1}^{h+j} c_i,\ j = 1,\ldots,h-1.$$

Note that A_{h-1} includes the total upper tail U.

The central interval index of radius j is defined as the aggregated citations of the articles from h-j to h+j:

$$I_j = \sum_{i=h-j}^{h+j} c_i,\ j = 1,\ldots,h-1.$$

That is, the citations at the interval [h-j,h+j]. Note that $I_{h-1} = A_{h-1}$. The geometrical representation is shown in Figure 1.

Materials and Methods

The publications and citations for those scientists who have received the Price Medal are obtained from the ISI Web of Science database in February 2010. In order not to distort further analysis, especially regression analysis, only the 15 existing and currently productive scientists were considered.

Empirical application

The behavior of central indexes is analyzed in this section. To estimate the predictive capacity of indexes for five and ten years ahead, and their comparison with the h-index, the cited articles and the number of citations obtained in 1999, 2004, and 2009, have been considered.

Production-impact scatter plots are obtained. Linear correlation between the number of articles and the number of citations exists. Authors located above the regression line show a more selective behavior than those below this line.

Linear correlation coefficients among indexes for 5 and 10 years are obtained. The area indexes for 1999 are strongly correlated with those for 2004, so they look like good estimators for 5 years. In all cases, correlations are higher than 0.94. Most of them are greater than the correlation between h-indexes 0.977. Best correlations correspond to j=7, which is half the average h-index of the sample.

Area indexes for 1999 also show high correlations with 2009, although slightly lower than those mentioned in the previous paragraph, making them also good estimators for 10 years. All of the coefficients are greater than the correlation between the h-indexes 0.812. Finally, the area indexes for 2004 also present correlations with the year 2009. Most of the elements are higher than the correlation between h-indexes 0.889.

With respect to interval indexes, something very similar occurs. Correlations are also high in all cases. However, the correlations for the area index are greater than for the interval index.

Conclusions

The h-index is a bibliometric indicator that attempts to measure the success of a researcher with just a part of the total amount of publications and citations. Due to not considering all production and impact, this index corrects biases of mass collaborations and punctual successes, which may not be significant in the researcher's career as a whole. However, different citation distributions, like those of a selective researcher and a large producer, may cause similar h-indexes, and in these cases, it is not possible to distinguish between these researchers using the h-index exclusively.

In this paper two complements to the h-index, the area and the interval indexes, have been proposed with the aim of increasing the capacity of discrimination among researchers with similar h, and improving the prediction of future successes. These indicators consider some

areas that are larger for selective authors than for large producers. Thus, a problem described in the literature about the h-index, which penalizes selective researchers compared to large producers, is corrected.

Both central indexes are good estimators and correlations are generally higher for the area index than for the interval index. Moreover, a radius that well describes the shape of the citation distribution has been estimated empirically. This radius is about half the average h-index of researchers being evaluated.

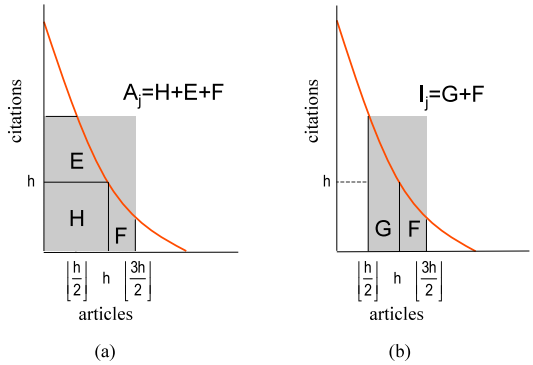

Figure 1. Central area index (a) and central interval index (b) of radius j=h/2 (integer part).

References

Alonso, S., Cabrerizo, F. J., Herrera-Viedma, E. & Herrera, F. (2009). h-Index: A review focused in its variants, computation and standardization for different scientific fields. *Journal of Informetrics*, 3(4), 273–289.

Bornmann, L. & Daniel, H. D. (2007). What do we know about the h-index? *Journal of the American Society for Information Science and Technology*, 58(9), 1381–1385.

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

Hirsch, J. E. (2005). An index to quantify an individual's scientific research output. *Proceedings of the National Academy of Sciences*, 102(46), 16569–16572.

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.

Schreiber, M. (2010). A new family of old Hirsch index variants. *Journal of Informetrics*, 4(4), 647–651.