Modelling some structural indicators in an h-index context: A shifted Zipf and a decreasing exponential model

Ronald Rousseau

ronald.rousseau@uantwerpen.be

Institute for Education and Information Sciences, IBW, University of Antwerp (UA), Venusstraat 35, Antwerp, B-2000, Belgium and
KU Leuven, Leuven, B-3000, Belgium

Abstract

We derive the values of the ratio between the e-area and the h-area and the average number of excess citations to papers in the h-core in a shifted Zipf model and in a decreasing exponential citation model.

Introduction

Consider a non-empty set of publications, Pub, and let r(p) denote the rank of publication p in Pub ordered according to the number of citations received during a given citation period. This number of citations is denoted as C(p). The symbol c(r) denotes the number of citations received by the publication ranked r. Hirsch (2005) originally defined the h-index as the highest rank such that the first h publications received each at least h citations. When all publications in Pub are uncited then h = 0. When discussing the h-index in a continuous context, it is characterized as the (real) number such that c(h) = h.

Set Pub can be subdivided into three sub sets: the first h publications, referred to as the h-core (Rousseau, 2006), the uncited publications and the other publications. The uncited publications and the other publications together form the h-tail. It is possible that the h-tail is empty and if h = 0 then the h-core is empty. In this contribution we do not pay special attention to the uncited publications, except for the fact that we assume that such articles may exist.

In the context of studies of the h-index the area under the citation curve, see Fig.1, can be divided into three sections. First, the area under the curve is divided into two sections: the section related to the citations received by the tail publications called the tail area and denoted as the t-area, and the section related to the citations received by the publications in the h-core. This second section can, in turn, be subdivided into two disjoint sections: a square corresponding to h^2 citations (see Fig.1), referred to as the h-area and the remaining part corresponding to the excess citations, denoted as the e-area. The core, excess citations, the tail and the e-t ratio have been studied e.g. in (Chen et al., 2013; Liu et al., 2013; Ye and Rousseau, 2010, Zhang, 2010, 2013a, 2013b). In this contribution we focus on the e/h, e^2/h^2 and the e^2/h ratios.

Let C_h denote the number of citations received by the publications in the h-core. Using the concept of excess citations, e^2 (Zhang, 2009), we have:

$$C_h = e^2 + h^2 \geq 0 \tag{1}$$

In a geometric interpretation (see Fig.1) we see that e^2/h^2 is the ratio between the e-area and the h-area. Moreover, e^2/h is the average number of excess citations to papers in the h-core. Together with e/h these three ratios can be considered as structural measures. Since e ≥ 0, we see that

$$C_h \geq h^2 \tag{2}$$

and

$$\sqrt{C_h} = R \geq h \tag{3}$$

where R denotes the R-index (Jin et al., 2007).

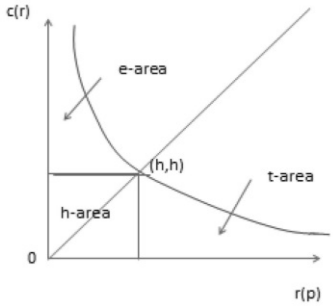

Figure 1: areas under the citation curve

Theoretical Modelling

In this section we calculate e/h, e²/h² and e²/h in the case of two basic theoretical models: a shifted Zipf model (Burrell, 2008; Egghe & Rousseau, 2012a, b; Glänzel, 2008, 2010) and a decreasing exponential model.

A. Shifted Zipf model

Consider first the framework of the shifted Zipf function, i.e. the rank-frequency function

$$g:]0,T] \to [1,+\infty[\, : r \to g(r) = \frac{B}{r^{\beta}} - 1$$

with B, β > 0. In (Egghe & Rousseau, 2012b) we showed that in this situation h is characterized by the relation $(h+1)h^{\beta} = B$. Then, by definition, we have that the number of citations (items) received by the publications (sources) in the h-core is:

$$\int_0^h \left(\frac{B}{r^{\beta}} - 1\right) dr = \frac{B}{1-\beta}\, h^{1-\beta} - h$$

valid for 0 < β < 1 (see e.g. Jin et al., 2007, p.858). This number is actually the square of the R-index.

The so-called excess citations (e²) are the citations received by the h-core publications in excess of the minimum number of citations in the h-core, namely h².

This leads to $e^2 = \frac{B}{1-\beta}\, h^{1-\beta} - h - h^2 = h^2 \left(\frac{B}{1-\beta}\, h^{-1-\beta} - \, h^{-1} - 1\right)$

We recall that in (Egghe & Rousseau, 2012a) it is shown that for a shifted system the average μ = $\frac{1}{\alpha-2}$ where α is the exponent in the equivalent shifted size-frequency system. As the relation between β and α is given by $\alpha = \frac{1+\beta}{\beta}$ we find that $\mu = \frac{1}{\alpha-2} = \frac{1}{\frac{1+\beta}{\beta}-2} = \frac{\beta}{1-\beta}$.

Hence:

$$\frac{e^2}{h^2} = \frac{B}{1-\beta}\, h^{-1-\beta} - h^{-1} - 1 = h^{-1}\left(\frac{h+1}{1-\beta} - 1\right) - 1 = h^{-1}\left(\frac{h+\beta}{1-\beta}\right) - 1 = \frac{\beta}{1-\beta}\frac{h+1}{h} = \mu\frac{h+1}{h}$$

and $\frac{e}{h} = \sqrt{\mu\frac{h+1}{h}}$

For large h, such as in the case of countries, universities and even famous scientists we may say that $\frac{e^2}{h^2} \approx \mu$. Indeed even if h is only 25, the ratio (h+1)/h =1.04, which is very close to 1.

Finally, $\frac{e^2}{h} = \frac{e^2}{h^2}.h = \ \mu\,(1+h)$.

B. Decreasing exponential model

Next we consider the exponential model. In this model we have an exponential size-frequency function

$$f: [0, +\infty[\rightarrow \ \mathbb{R}^{+}: x \rightarrow C. e^{-ax}$$

with a > 0. Then T, the total number of sources (articles) is:

$$T = \int_{0}^{+\infty} C. e^{-ax} dx = \frac{C}{a}$$

and the total number of items (citations) is:

$$A = \int_{0}^{+\infty} x. C. e^{-ax} dx = \frac{C}{a^2}$$

Hence, the average μ = 1/a. If 0 < a < 1 then this average is larger than 1, if a > 1, this average is smaller than 1.

Next we determine the equivalent rank-size form, g(r):

r = g^{-1}(j) = $\int_{j}^{+\infty} C. e^{-ax} dx = \frac{C}{a}\ e^{-aj}$

From this we find that j = g(r) = -(1/a). ln(r.a/C)

$$g:]0, T] \rightarrow [1, +\infty[\ : r \rightarrow g(r) = -\frac{1}{a} \ln\left(\frac{a}{C} r\right)$$

with a > 0. Now the h-index is characterized by the equation:

$$-\frac{1}{a}\ ln\left(\frac{a}{C}\ h\right) = h$$

or: $-\frac{1}{a}\ ln\left(\frac{h}{C}\right) - \frac{1}{a}\ln(a) = h$

Now the core contains

$$\int_{0}^{h} -\frac{1}{a}\ ln\left(\frac{a}{C}\ r\right) dr = \ -\frac{1}{a}\left(\ln(\frac{h}{C})h + h.(\ln(a) - 1)\right)$$

$$= \ -\frac{1}{a}(-ah^2 - \ln(a)\,h + h.\ln(a) - h)$$

= h^2+ h/a items (citations).

Hence e^2 = h/a and $(e/h)^2$ = 1/(ah) = μ/h and e^2/h = μ.

Hence, remarkably, in the exponential model the average number of excess citations to papers in the h-core is equal to the average number of citations per paper in the whole set of papers.

Conclusion

We have determined the ratios the e/h, e^2/h^2 and the e^2/h in the case of a shifted Zipf and decreasing exponential model. We note that we have tried to apply these results on a set of country-related data but these data did not yield an acceptable fit (Yuan et al., 2013). Of course, this just shows that these data could not be described by a shifted Zipf or a decreasing exponential model. When data are distributed according to one of these models then our theoretical calculations will prove to be valid.

Acknowledgements

The author thanks Leo Egghe and Fred Y. Ye for helpful discussions and Weina Hua and Xiaoyuan Yuan for encouragements.

References

Burrell, Q.L. (2008). Extending Lotkaian informetrics. *Information Processing and Management* 44(5), 1794-1807.

Chen, D-Z; Huang, M-H & Ye, F. Y. (2013). A probe into dynamic measures for h-core and h-tail. *Journal of Informetrics*, 7(1), 129-137

Egghe, L. and Rousseau, R. (2012a). Theory and practice of the shifted Lotka function. *Scientometrics*, 91(1), 295-301.

Egghe, L. and Rousseau, R. (2012b). The Hirsch index of a shifted Lotka function and its relation with the impact factor. *Journal of the American Society for Information Science and Technology,* 63(5), 1048-1053.

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

Glänzel, W. (2010). The role of the h-index and the characteristic scores and scales in testing the tail properties of scientometric distributions. *Scientometrics*, 83(3), 697-709.

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

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

Liu, J.Q., Rousseau, R., Wang, M.S. & Ye, F.Y. (2013). Ratios of h-cores, h-tails and uncited sources in sets of scientific papers and technical patents. *Journal of Informetrics*, 7(1), 190-197.

Rousseau, R. (2006). New developments related to the Hirsch index. *Science Focus*, 1, 23–25 (in Chinese). An English translation is available online at http://eprints.rclis.org/6376/.

Ye, F. Y., & Rousseau, R. (2010). Probing the h-core: An investigation of the tail-core ratio for rank distributions. *Scientometrics*, 84, 431–439.

Yuan, XY., Hua, W., Rousseau, R, & Ye, F.Y. (2013). A preliminary study of the relationship between the h-index and excess citations. Preprint.

Zhang, CT. (2009). The e-index, complementing the h-index for excess citations. *PLoS ONE*, 4, e5429.

Zhang, CT. (2010). Relationship of the h-index, g-index, and e-index. *Journal of the American Society for Information Science and Technology*, 61(3), 625–628.

Zhang, C.T. (2013a). A novel triangle mapping technique to study the h-index based citation distribution. *Scientific Reports*, 3: 1023, 1-5.

Zhang, C.T. (2013b). The h’ index, effectively improving the h-index based on the citation distribution. *PLoS ONE*, 8, e59912.