A "Citation Surplus" Should Be Added to the *h*-Index

Sergio Da Silva

Department of Economics, Federal University of Santa Catarina, Florianopolis, Brazil
Email: professorsergiodasilva@gmail.com

Abstract

The *h*-index is the largest number *h* such that *h* publications have at least *h* citations. The index reflects both the number of publications and the number of citations per publication. One unperceived deficiency of this metric is that it is Pareto-inefficient. A "citation surplus" would be absent and, thus, the *h*-index would be efficient for a researcher if all his *h* papers that are equal or above his *h*-index received exactly *h* citations. This inefficiency would not be of great concern if those *h* papers were normally distributed. However, the rank from top to bottom does not decay exponentially. The decay follows the power law known in the literature as Lotka's law. To remedy this deficiency, I suggest the *h*-index be supplemented by a researcher's citation surplus.

Subject Areas

Statistics

Keywords

h-Index, Scientific Productivity, Scientometrics, Pareto-Efficiency

---

The *h*-index (after Hirsch [1]) is the largest number *h* such that *h* publications have at least *h* citations. The index reflects both the number of publications and the number of citations per publication. One unperceived deficiency of this metric is that it is inefficient, in economic terms. This inefficiency relates to the very articles cited more than their rank, in particular, to those citations above the *h*-index.

To appreciate this, consider a researcher with *h*-index of 136, as in Figure 1. The citations received by his papers that overshoot 136 citations are left out from the *h*-index, although they are what economists call the producer surplus. Absence
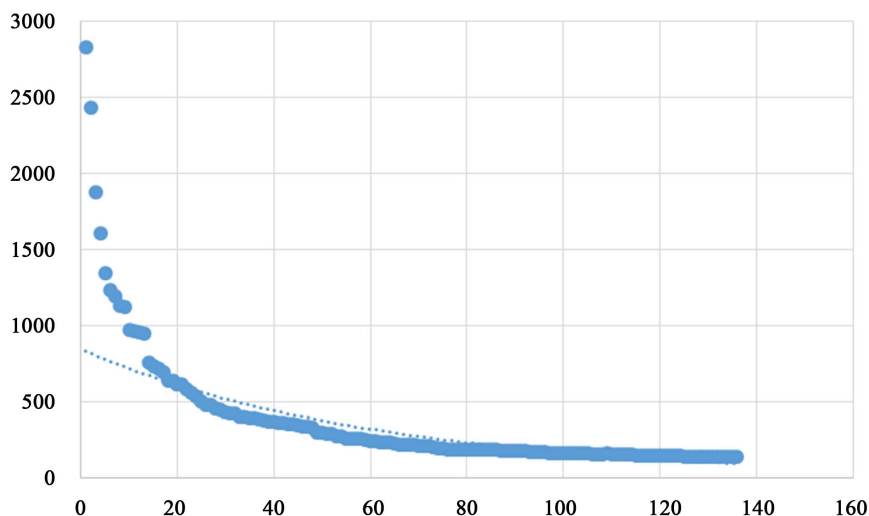
Figure 1. Citations received by Researcher One that fall above his *h*-index of 136. Vertical axis: number of citations; horizontal axis: rank. A power fit is appropriate, and the dotted line shows an exponential fit is unsatisfactory.

of surplus denotes what is called Pareto efficiency in economics. If the *h*-index was efficient for this researcher, all of his 136 papers that are equal or above the *h*-index of 136 would receive exactly 136 citations.

I collected the data in **Figure 1** from the website of one prominent physicist called Researcher One. For him, I computed a "citation surplus" of 51586 by summing all the citations that fall above his *h*-index of 136. His citation surplus represented 67.19 percent of his total citation of 76778, which is a figure I collected from Thomson ISI Web of Science. His most cited paper received 2832 citations, a figure hugely larger than his *h*-index of 136. The dataset is available at Figshare (https://doi.org/10.6084/m9.figshare.5404798.v1).

This inefficiency would not be of great concern if those very 136 papers of Researcher One were normally distributed. However, one established result is that citations do not distribute as a Gaussian. The rank from top to bottom does not decay exponentially, as shown by the dotted line in **Figure 1**. A power fit is more appropriate. Indeed, the decay follows the power law depicted as a straight line in the log-log plot of **Figure 2**. In the literature, this is known as Lotka's law [2].

I further collected data from two other researchers working on the same subject as that of Researcher One. The data was taken from Researcher ID of Thomson Reuters and is available at Figshare as well (https://doi.org/10.6084/m9.figshare.5404798.v1).

Researcher Two had an *h*-index of 27, as of September 2017. For him, I computed a citation surplus of 3541, which makes up 95.55 percent of his total citation of 3868. His most cited paper received 731 citations, which is 27 times greater than his *h*-index of 27. **Figure 3** shows the inadequacy of an exponential fit (dotted line) for the tail of the citation distribution of Researcher Two. Actually, the tail decays as a power law (**Figure 4**).
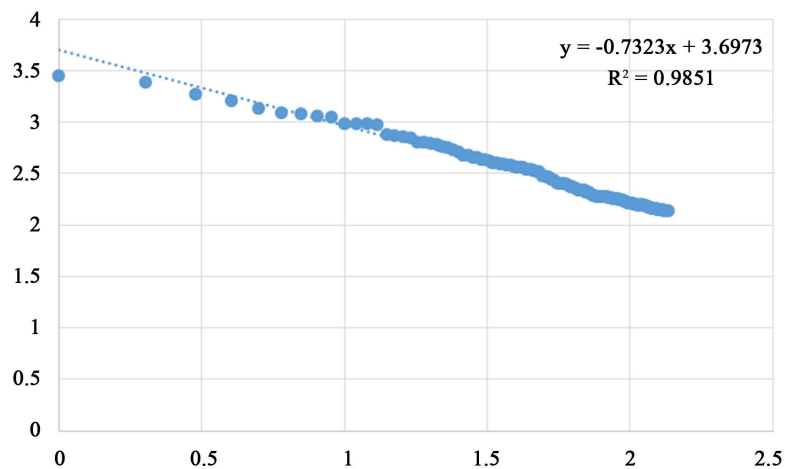
Figure 2. Log of rank versus log of citations received by Researcher One that fall above his *h*-index of 136. A power law describes well the tail of his citation distribution.
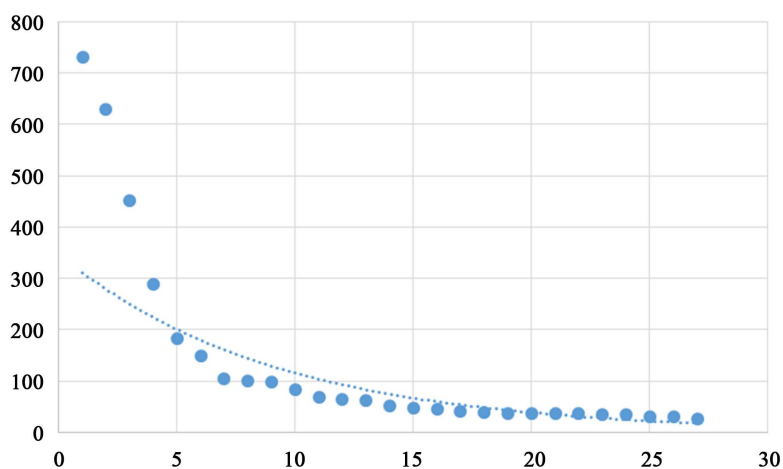

Figure 3. Citations received by Researcher Two that fall above his *h*-index of 27. Vertical axis: number of citations; horizontal axis: rank. The dotted line shows the poor exponential fit for the tail of the distribution of citations.
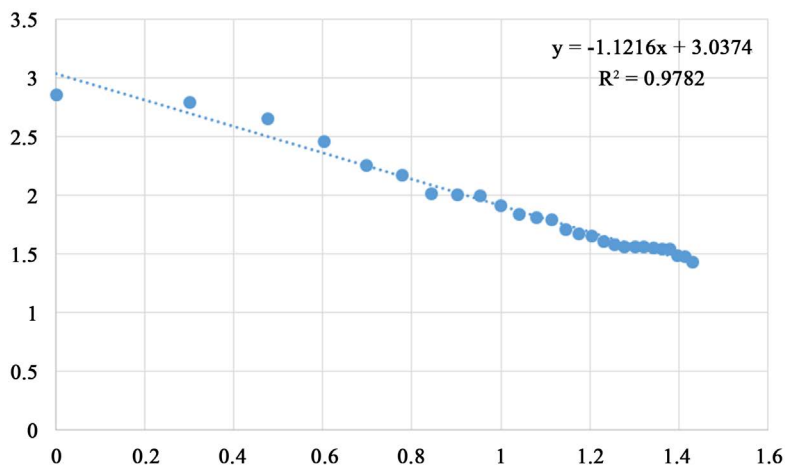

Figure 4. Log of rank versus log of citations received by Researcher Two that fall above his *h*-index of 27. A power law describes well the tail of his citation distribution.

In turn, Researcher Three had an *h*-index of 8 (as of September 2017). I computed his citation surplus as 95, which makes up 57.23 percent of the 166 citations he received from all of his papers. A pattern similar to those for Researchers One and Two is seen (Figure 5 and Figure 6).

Table 1 sums up the three cases. The degree of inefficiency of the *h*-index looms larger as one considers a researcher with more citations, as evidenced by growing citation surpluses. Because the tails of the citation distributions do not decay exponentially, the problem of the inefficiency of the *h*-index is consequential. The most cited paper impacts a discipline disproportionally more—a fact that cannot be captured by the *h*-index metric alone. Therefore, it has to be supplemented by the citation surplus, as suggested here.
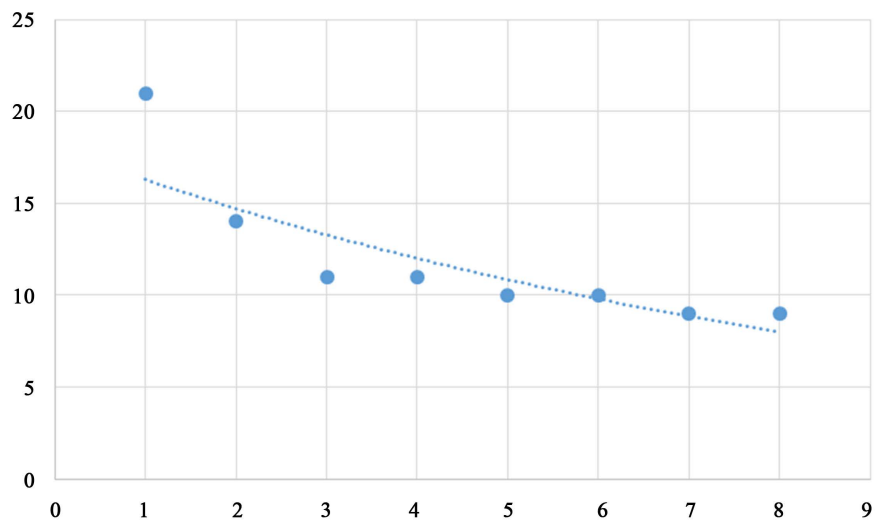

Figure 5. Citations received by Researcher Three that fall above his *h*-index of 8. Vertical axis: number of citations; horizontal axis: rank. The dotted line shows the poor exponential fit for the tail.
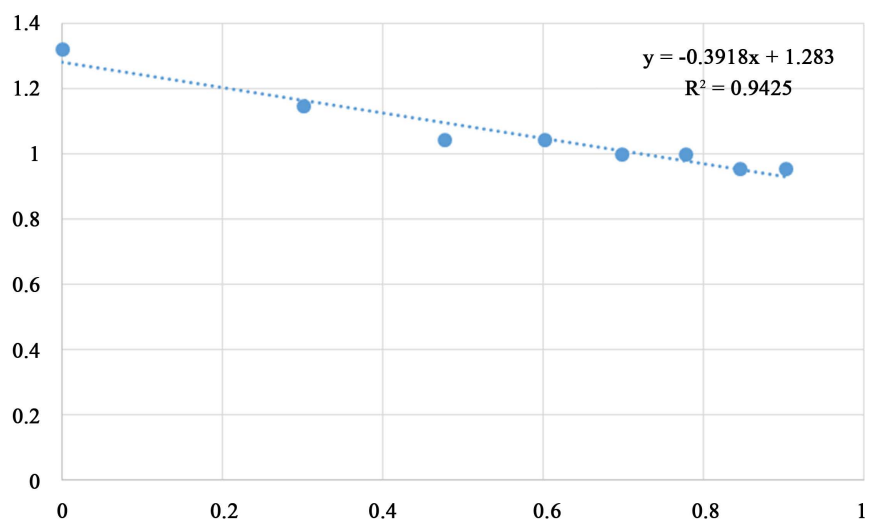

$$y = -0.3918x + 1.283$$
$$R^2 = 0.9425$$

Figure 6. Log-log plot of rank versus citations received by Researcher Three that fall above his *h*-index of 8. The tail decay follows a power law.

Table 1. Summary of the illustrations employed.

	h-index	citation surplus	citations received by the most cited paper
Researcher One	136	51586	2832
Researcher Two	27	3541	731
Researcher Three	8	95	21

My suggestion echoes what statisticians do when supplementing averages with measures of variance. For example, the median indicates only the value in the middle but ignores how much larger the larger numbers are and how much smaller the smaller numbers are. In our terms, the median is Pareto-inefficient. One could supplement the h-index by similar measures of variation—for example, by counting the total number of citations beyond h of all papers with more than h citations, as suggested here. Computing this citation surplus is equivalent to determining the value of the cumulative distribution function beyond the mean, which might differ from 50 percent for nonsymmetric distributions.

Googling for "h-index" produces various suggested extensions to the h-index. Despite that, the h-index continues to be widely employed because it seems parsimonious enough and to convey only the critical relevant information. Simply counting the total number of citations ignores the distribution of citations across papers, and the h-index provides the optimal information regarding this distribution by focusing on the number of papers h that have been cited at least h times. Here, I am not suggesting another extension to the h-index. What I do is simply supplement it with a straightforward measure that tracks what is ignored by the h-index—the information about how many times the papers with more than h citations have been cited. It is a variance to supplement an average for statisticians, and it is a citation surplus for economists.

References

[1] Hirsch, J.E. (2005) An Index to Quantify an Individual's Scientific Research Output. *PNAS*, **102**, 16569-16572. https://doi.org/10.1073/pnas.0507655102

[2] Lotka, A.J. (1926) The Frequency Distribution of Scientific Productivity. *Journal of the Washington Academy of Sciences*, **16**, 317-323.