Mathematical derivation of the impact factor distribution

Peer-reviewed author version

EGGHE, Leo (2009) Mathematical derivation of the impact factor distribution. In:
JOURNAL OF INFORMETRICS, 3(4). p. 290-295.

DOI: 10.1016/j.joi.2009.01.004
Handle: http://hdl.handle.net/1942/9681

Mathematical derivation of the impact factor distribution

by

L. Egghe

Universiteit Hasselt (UHasselt), Campus Diepenbeek, Agoralaan, B-3590 Diepenbeek, Belgium[1]
and
Universiteit Antwerpen (UA), Stadscampus, Venusstraat 35, B-2000 Antwerpen, Belgium
leo.egghe@uhasselt.be

ABSTRACT

Experimental data in Mansilla, Köppen, Cocho and Miramontes [Journal of Informetrics 1(2), 155-160, 2007] reveal that, if one ranks a set of journals (e.g. in a field) in decreasing order of their impact factors, the rank distribution of the logarithm of these impact factors has a typical S-shape: first a convex decrease, followed by a concave decrease. In this paper we give a mathematical formula for this distribution and explain the S-shape. Also the experimentally found smaller convex part and larger concave part is explained. If one studies the rank distribution of the impact factors themselves we now prove that we have the same S-shape but with inflection point in μ, the average of the impact factors. These distributions are valid for any type of impact factor (any publication period and any citation period). They are even valid for any sample average rank distribution.

[1] Permanent address

Key words and phrases: impact factor, rank distribution, S-shape, Central Limit Theorem, average.
Acknowledgement: The author is grateful to Prof. Dr. R. Rousseau for his advise of using the Central Limit Theorem in the solution of this problem.

I. Introduction

Impact factors (IF) show a typical S-shape if one draws their decreasing rank-order distribution. Fig. 1 shows such a graph, which appeared in Mansilla, Köppen, Cocho and Miramontes (2007).

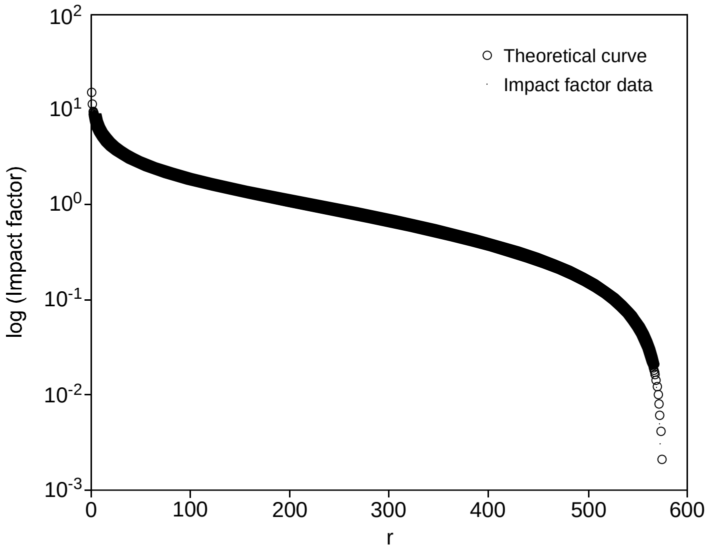

Fig. 1 Rank-order distribution of ln(IF). Reprinted from Mansilla, Köppen, Cocho and Miramontes (2007) with kind permission from Elsevier.

If one orders a set of journals in decreasing order of their IF and if one draws the graph of ln(IF) versus the rank r one obtains a rank-order distribution as in Fig. 1: convex decrease followed by a concave decrease. Here one deals with physics journals. In Mansilla, Köppen, Cocho and Miramontes (2007), however, the same shape is found in every other set of journals that they consider (mathematics, environmental sciences).

Such a shape cannot be a coincidence and hence needs an explanation. In Mansilla, Köppen, Cocho and Miramontes (2007) only a statistical fitting is given. Although their fittings are very good (outperforming earlier work of Lavalette (1996) – see also Popescu (2003)), statistical fittings do not yield a mathematical explanation.

A mathematical explanation is necessary, first of all, for the simple reason of developing informetrics into a science in which mathematical reasonings explain informetric regularities.

Secondly, explaining the shape of a graph as in Fig. 1, shows what is, informetrically, behind it. For instance, in Egghe (2009) we present a rationale for the Hirsch-index (Hirsch (2005)) rank-order distribution. The shape of this distribution is explained based on the theory of the law of Lotka (see Egghe (2005)) and is different from the shape of the impact factor distribution, studied here: here the Central Limit Theorem (CLT) is involved and hence the “bell curve” of Gauss. The latter is a symmetric function while Lotka’s law is very skew, explaining the different aspects that ly beneath these two distributions.

This problem is the topic of this paper. In the next section we will develop the theory that leads to the model. We remark that all IFs are averages, being the average number of citations per paper in a journal. So, all journals in a field yield a sample of averages (IFs), i.e. the average number of citations to the journal’s papers, hence, by the Central Limit Theorem (CLT), are normally distributed, i.e. according to a Gaussian distribution (bell curve). This distribution is used to derive a mathematical formula for IF(r): the rank-order distribution of IF in function of the rank r of the journal (in decreasing order of their IFs). We prove that

$$\mathrm{IF}(r) = F^{-1}(T - r), \tag{1}$$

where T is the total number of journals and where F^{-1} is the inverse of the injective function

$$F(x) = \int_0^x \varphi(y)\,dy \tag{2}$$

where

$$\varphi(y) = Ae^{-\frac{(y-\mu)^2}{2\sigma^2}} \tag{3}$$

(μ = average IF, σ^2 = variance of the IFs) where A is a parameter such that

$$\int_0^\infty \varphi(y)\,dy = T, \tag{4}$$

the total number of journals.

The function in Fig. 1 is then nothing else than $\ln\left(F^{-1}(T-r)\right)$. We then calculate the first and second derivative of this function, hereby explaining the shape of the graph in Fig. 1. We also explain why the convex part is shorter than the concave part, hence giving a full mathematical explanation of the rank-order distribution $\ln\left(IF(r)\right)$.

We are surprised that, in Mansilla, Köppen, Cocho and Miramontes (2007), the evident IF(r) (rank-order distribution of the IFs themselves) is not presented. This function, being given by formula (1) is also studied here. We show that also this function has the same S-shape as $\ln\left(IF(r)\right)$ but their inflection point (the ordinate) is at $IF(r)=\mu$, the average of the IFs (for $\ln\left(IF(r)\right)$ we have the inflection point in a value $IF(r)>\mu$).

The paper ends by making a remark on the function $\ln r \rightarrow \ln IF(r)$ (i.e. the function $x \rightarrow \ln IF\left(e^x\right)$) of which the graph is also presented in Mansilla, Köppen, Cocho and Miramontes (2007) and also in Taylor, Perakakis and Trachana (2008). We also present some open problems.

II. Mathematical model for the rank-order distribution of IF

The Central Limit Theorem says that sample averages are distributed (approximately) according to the Gaussian bell curve. If all values in $\mathbb{R}$ are allowed we have that this distribution equals ($\overline{x}$ = sample average)

$$\varphi(\overline{x})=\frac{1}{\sqrt{2\pi\sigma}}e^{-\frac{(\overline{x}-\mu)^2}{2\sigma^2}} \tag{5}$$

where μ and σ are the overall average and standard deviation (i.e. of the population).

Note that $\sigma^2 = \frac{s^2}{N}$ where N is the sample size and s is the standard deviation of the citation distribution itself. We assume σ to be fixed for the time being (see further for a remark on the N-dependency) and that N is large enough in order to have the Gaussian bell curve approximation as predicted by the Central Limit Theorem (CLT) – see e.g. Blalock (1987) or Altman (1991).

For a fixed topic or field (e.g. physics as in Fig. 1), each journal in this field can be considered as a random sample in the total population of all articles in the field and each journal's IF can be considered as the sample average $\bar{x}$ of the number of citations per paper in this journal. These sample averages are, according to the CLT, distributed according to a Gaussian bell curve with population average μ being the average number of citations per paper in the field. Different with situation (5) is here that $\mathrm{IF} \geq 0$ necessarily. So another normalization constant (other than $\frac{1}{\sqrt{2\pi\sigma}}$) is necessary. Since I also want $\int_0^\infty \varphi(y)dy$ to be T (the total number of journals in this field) instead of 1, let A be this constant such that, with $(y \geq 0)$

$$\varphi(y) = Ae^{-\frac{(y-\mu)^2}{2\sigma^2}} \qquad (6)$$

we have that

$$\int_0^\infty \varphi(y)dy = T \qquad (7)$$

(note that y goes over $[0, +\infty[$, the theoretical range of the IFs).

We have the following easy but basic theorem.

Theorem 1: The rank-order impact factor distribution, denoted IF(r) ($r \in [0, T]$) is equal to

$$\mathrm{IF}(r) = F^{-1}(T - r) \qquad (8)$$

where F^{-1} is the inverse of the injective function

$$F(x) = \int_0^x \varphi(y)dy \tag{9}$$

with φ as in (6), hence $F(x)$ is the cumulative Gaussian distribution (multiplied by T, using (7)).

Proof: If we define

$$r = \int_x^\infty \varphi(y)dy \tag{10}$$

then $x = IF(r)$. Indeed the journals on ranks $\leq r$ have an $IF \geq x$, hence on rank r we have exactly $IF = x$, hence $x = IF(r)$ (continuous argument). Now (10) yields, by (7)

$$r = T - \int_0^x \varphi(y)dy$$

$$r = T - F(x)$$

Hence

$$F(x) = T - r$$

$$IF(r) = x = F^{-1}(T - r)$$

hence (8) is proved. □

We now start studying Fig. 1, hence the function $\ln(IF(r))$.

Theorem 2: The function

$$f(r) = \ln\left(IF(r)\right) = \ln\left(F^{-1}(T-r)\right) \tag{11}$$

is strictly decreasing, first, on an interval $[0, r_0]$ convexly and on the interval $[r_0, T]$ concavely. Furthermore we have that $IF(r_0) > \mu$ in the inflection point $\left(r_0, \ln\left(IF(r_0)\right)\right)$.

Proof: By (9) and (6) we have

$$F'(x) = \varphi(x) \tag{12}$$

$$F''(x) = -\frac{x-\mu}{\sigma^2}\varphi(x) \tag{13}$$

We have, by (11), for all $r \in [0,T]$

$$f'(r) = \frac{1}{F^{-1}(T-r)} . \frac{1}{F'\left(F^{-1}(T-r)\right)}(-1) \tag{14}$$

, hence f is strictly decreasing by (12), (9) and (6).

$$f''(r) = -\frac{1}{\left(F^{-1}(T-r)\right)^2}\frac{1}{F'\left(F^{-1}(T-r)\right)}(-1)\frac{1}{F'\left(F^{-1}(T-r)\right)}(-1)$$

$$+\frac{1}{F^{-1}(T-r)}\left[\frac{1}{F'^2\left(F^{-1}(T-r)\right)}F''\left(F^{-1}(T-r)\right)\frac{1}{F'\left(F^{-1}(T-r)\right)}(-1)\right](-1)$$

$$f''(r) = \frac{-1}{\left(F^{-1}(T-r)\right)^2 F'^2\left(F^{-1}(T-r)\right)} - \frac{1}{F^{-1}(T-r)F'^3\left(F^{-1}(T-r)\right)}F''\left(F^{-1}(T-r)\right)$$

Since $F'^2 > 0$, $F^{-1} > 0$ (by (9)) we have that the sign of $f''(r)$ is equal to the sign of

$$-\frac{1}{F^{-1}(T-r)}-\frac{F''\left(F^{-1}(T-r)\right)}{F'\left(F^{-1}(T-r)\right)}=-\frac{1}{IF(r)}-\frac{\varphi\left(IF(r)\right)\left(-\frac{IF(r)-\mu}{\sigma^{2}}\right)}{\varphi\left(IF(r)\right)}$$

by (12), (13) and (8).

Hence the sign of $f''(r)$ is equal to the sign of

$$-\frac{1}{IF(r)}+\frac{IF(r)-\mu}{\sigma^{2}}=\frac{IF^{2}(r)-\mu IF(r)-\sigma^{2}}{IF(r)\sigma^{2}}$$

Since the denominator is positive we have that the sign of $f''(r)$ is equal to the sign of (replace $IF(r)$ by x)

$$x^{2}-\mu x-\sigma^{2} \qquad (15)$$

Putting (15) equal to 0 we have that the two roots of this equation are given by

$$x_{1}=\frac{\mu-\sqrt{\mu^{2}+4\sigma^{2}}}{2}<0$$

and

$$x_{2}=\frac{\mu+\sqrt{\mu^{2}+4\sigma^{2}}}{2}>\mu$$

Note that $x_1 < 0$ is not a real IF but one of the roots of equation (15): IF is restricted to $IF \geq 0$. Hence, since $x_1 < 0$ we have that on the interval $[0, x_2[$ the sign of (15), hence of $f''(r)$, is negative, hence $f(r) = \ln\left(IF(r)\right)$ is concavely decreasing in an interval $[0, x_2]$ comprising μ. On the interval $]x_2, +\infty[$ we have that the sign of (15), hence $f''(r)$, is positive, hence $f(r)$ is convexly decreasing. The inflection point is $x_2 > \mu$ and hence there is a rank r_0 such that $x_2 = IF(r_0) > \mu$ and, since f decreases, we have that f convexly decreases on $[0, r_0]$ and concavely decreases on $[r_0, T]$. This explains the “shorter” convex part of the graph in Fig. 1 and the “longer” concave part. □

Now we will study the function $IF(r)$ itself.

Theorem 3: The function

$$g(r) = IF(r) = F^{-1}(T-r) \tag{16}$$

is strictly decreasing, first on an interval $[0, r_1]$ convexly and on the interval $[r_1, T]$ concavely. Furthermore we have that $IF(r_1) = \mu$ in the inflection point $(r_1, IF(r_1))$.

Proof: For all $r \in [0,T]$

$$g'(r) = \frac{1}{F'\left(F^{-1}(T-r)\right)}(-1) < 0$$

hence g is strictly decreasing since $F' = \varphi > 0$ by (12) and (6).

$$g''(r) = -\frac{1}{\left(F'\left(F^{-1}(T-r)\right)\right)^2}\frac{F''\left(F^{-1}(T-r)\right)}{F'\left(F^{-1}(T-r)\right)}(-1)(-1)$$

$$g''(r) = -\frac{F''\left(F^{-1}(T-r)\right)}{F'\left(F^{-1}(T-r)\right)^3}$$

$$g''(r) = -\frac{\varphi(IF(r))\left(-\frac{IF(r)-\mu}{\sigma^2}\right)}{\varphi(IF(r))^3}$$

which has the same sign as

$$\frac{IF(r)-\mu}{\sigma^2} \tag{17}$$

Now (17) is < 0 for $IF(r) < \mu$ and (17) is > 0 for $IF(r) > \mu$. Hence $IF(r)$ starts decreasing convexly and then continues decreasing concavely. The inflection point is for this rank $r = r_1$ such that $IF(r_1) = \mu$, by (17). □

<u>Note</u>: As said above we disregarded the sample size (N)-dependency. Yet we could explain the shapes of the $\ln(IF(r))$ and $IF(r)$-curves. As N varies we have “sheaves” of these curves, all of the same shape and yielding a graph with the same shape as proved above.

In Mansilla, Köppen, Cocho and Miramontes (2007) as well as in Taylor, Perakakis and Trachana (2008) one also studies the log-log variant of the rank order impact factor function, i.e. the function $\ln r \rightarrow \ln(IF(r))$, hence the function

$$h(x) = \ln\left(IF\left(e^x\right)\right) \tag{18}$$

$$h(x) = \ln\left(F^{-1}\left(T - e^x\right)\right) \tag{19}$$

by (8). It is easy to prove that h decreases strictly but we are not in a position to determine (as above) where h is concave and where h is convex (the second derivative h'' becomes too complex in this case). But h is only a mathematical variant of the functions f and g, which have been explained mathematically (informetrically). In Taylor, Perakakis and Trachana (2008) there is an indication that the same S-shape is there for h but with even a smaller convex part than in the case of f (there called the “King Effect”). Although the fitting in Fig. 1 in Mansilla, Köppen, Cocho and Miramontes (2007) seems to indicate that this function is concavely decreasing, a closer look at the data points indicate that indeed the same S-shape is valid but with a very small convex part. We cannot conclude this from (19), nor can we prove this in general from the analogous property of f and g. Indeed below is an example of a concavely decreasing h and with the S-shape as in f:

$$h(x) = \sqrt{a - x}$$

$0 £ x £ a$: concavely decreasing while

$$f(r) = h(\ln r)$$

has the S-shape as studied here (easy verification). So an S-shape of f does not always imply an S-shape of h.

III. Conclusions

The CLT can be applied to prove the shape of the rank order distribution of $IF(r)$ and $\ln(IF(r))$. This double application of the CLT is necessary since a direct general proof that the S-shape of a general function f implies the S-shape of $\ln(f)$ or vice-versa is not possible. The application of the CLT to IFs is possible since IFs are averages (average number of citations per paper).

It is now clear that any sample average $\bar{x}$ ordered decreasingly has such an S-shape, which I think is a new probabilistical-statistical result.

We leave open the explanation of the curve $\ln IF(e^x)$: it does not follow directly from the shape of $IF(r)$ or $\ln(IF(r))$ and an argument with the CLT leads to intricate formulae. But, essentially, a mathematical proof of one of these three curves suffices to understand the informetric behavior of IFs and in this paper we could even explain two of these distributions: $IF(r)$ and $\ln(IF(r))$.

In Egghe (2008), using simplifications, we could prove a similar S-shape relationship between the IF of a journal and its uncitedness factor U (an experimental curve of this S-shape can be found in van Leeuwen and Moed (2005). Can this relationship also be explained (in its full generality) using the CLT ?

References

D.G. Altman (1991). Practical Statistics for medical Research. Chapman and Hall, London (UK).

H.M. Blalock, Jr. (1987). Social Statistics. Revised second edition. Mc Graw-Hill, Singapore.

L. Egghe (2005). Power Laws is the Information Production Process: Lotkaian Informetrics. Elsevier, Oxford, UK.

L. Egghe (2008). The mathematical relation between the impact factor and the uncitedness factor. Scientometrics 76(1), 117-123.

L. Egghe (2009). A rationale for the Hirsch-index rank-order distribution and a comparison with the impact factor rank-order distribution. Preprint.

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

D. Lavalette (1996). Facteur d'impact: impartialité ou impuissance ? Internal Report, INSERM U350, Institut Curie, Paris.

R. Mansilla, E. Köppen, G. Cocho and P. Miramontes (2007). On the behavior of journal impact factor rank-order distribution. Journal of Informetrics 1(2), 155-160.

I. Popescu (2003). On a Zipf's law extension to impact factors. Glottometrics 6, 83-93.

M. Taylor, P. Perakakis and V. Trachana (2008). The siege of science. Ethics in Science and environmental Politics 8, 17-40.

T.N. van Leeuwen and H.F. Moed (2005). Characterisics of journal impact factors: the effects of uncitedness and citation distribution on the understanding of journal impact factors. Scientometrics 63(2), 357-371.