The Distribution of the Uncitedness Factor and its Functional Relation with the Impact Factor

Leo Egghe

leo.egghe@uhasselt.be
Universiteit Hasselt (UHasselt), Campus Diepenbeek, Agoralaan, B-3590 Diepenbeek (Belgium)

Abstract

The uncitedness factor of a journal is its fraction of uncited articles. Given a set of journals (e.g. in a field) we can determine the rank-order distribution of these uncitedness factors. Hereby we use the Central Limit Theorem which is valid for uncitedness factors since it are fractions, hence averages.

A similar result was proved earlier for the impact factors of a set of journals. Here we combine the two rank-order distributions, hereby eliminating the rank, yielding the functional relation between the impact factor and the uncitedness factor. It is proved that the decreasing relation has an S-shape: first convex, then concave and that the inflection point is in the point (μ',μ) where μ is the average of the impact factors and μ' is the average of the uncitedness factors.

Introduction

In Egghe (2009) we studied the rank-order distribution of impact factors (IF) say of a set of journals (e.g. in a field). Remark that IFs are averages (average number of citations per article in a journal). Hence we can use the Central Limit Theorem (CLT) for the distribution of IFs over these journals: the normal (or Gaussian) distribution

$$\varphi(x) = Ae^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{1}$$

where $x = \mathrm{IF} \geq 0$ and where the constant A is such that

$$\int_0^\infty \varphi(x)\,dx = T\,, \tag{2}$$

the total number of journals.

Ranking the journals in decreasing order of their IFs we argued in Egghe (2009) that

$$r = \int_x^\infty \varphi(y)\,dy \tag{3}$$

with $x = \mathrm{IF}(r)$, the IF of the journal at rank r (continuous argument). Denoting

$$F(x) = \int_0^x \varphi(y)\,dy \tag{4}$$

, the cumulative normal distribution, we have by (3) and (2)

Key words and phrases: impact factor, uncitedness factor, rank distribution, rank-order distribution, S-shape, Central Limit Theorem.

Acknowledgement: The author is grateful to Prof. Dr. R. Rousseau for his advise of using the Central Limit Theorem in the treatment of this problem.

Proceedings of ISSI 2009, pages 22–30.
Edited by B. Larsen and J. Leta.

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

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

Hence, denoting by F^{-1} the inverse of the injective function F,

$$IF(r) = x = F^{-1}(T - r) \tag{5}$$

yielding the desired distribution. In Egghe (2009) it is then shown that this function has an S-shape: first decreasing convexly followed by a concave decrease. Furthermore we proved in Egghe (2009) that the inflection point of (5) is in a point $(r, IF(r))$ where $IF(r) = \mu$, the average of distribution (1).

In Egghe (2009), the theoretical results where compared with the shapes of the experimental curves in Mansilla, Köppen, Cocho and Miramontes (2007) and, this way, the theoretical results were confirmed.

This paper addresses two problems. First we can develop, in the same way as above, the rank-order distribution of the uncitedness factors (U) of these journals. This is done in the next section.

This result and the above mentioned result in Egghe (2009) on IF are combined (by eliminating the ranks) to yield a functional expression of IF in function of U. It is a rather intricate function, involving two different normal (Gaussian) distributions. We are able to prove that $IF(U)$ also has an S-shape: first convexly decreasing followed by a concave decrease. We will show that the inflection point is in (μ', μ) where μ is the average of the IFs, appearing in formula (1) and where μ' is the average of the Us.

Both $IF(U)$ and $\ln(IF(U))$ are studied and the latter compared with the experimental curve, obtained in van Leeuwen and Moed (2005), hence explaining this typical S-shape. This paper considerably improves the result on $IF(U)$, obtained in Egghe (2008) where, for practical reasons, a simple - perhaps too simple – size-frequency function was used.

The rank-order distribution of the uncitedness factor U

Similar as what we did for the IF, cf. (1) and (2), we suppose that U is distributed according to a normal (Gaussian) distribution:

$$\psi(y) = Be^{-\frac{(y-\mu')^2}{2\sigma'^2}} \tag{6}$$

where B is such that

$$\int_0^1 \psi(y)\,dy = T \tag{7}$$

(note that $y \in [0,1]$ being an uncitedness factor). Indeed, also for U we can apply the CLT since U is a fraction: the fraction of the papers in the journal which are uncited. Fractions are indeed averages (so that the CLT applies): give each uncited paper in the journal a value 1 and each cited paper a value 0 then U is nothing else than the average of all these 0s and 1s.

Now we rank the journals in <u>increasing</u> order r of their uncitedness factors U. The defining relation for $U(r)$ is now, evidently,

$$r = \int_0^y \psi(y')dy' \tag{8}$$

where $y = U(r)$. (8) is a function of y which we denote by $G(y)$ (it is the cumulative normal distribution). Hence we have

$$r = G(U)$$

hence

$$U(r) = G^{-1}(r), \tag{9}$$

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

Let us study the shape of (9)

$$U'(r) = \frac{1}{G'\left(G^{-1}(r)\right)} > 0$$

since $G' = \psi > 0$. Hence U indeed increases strictly (by construction). Further

$$U''(r) = -\ \frac{1}{\left(G'\left(G^{-1}(r)\right)\right)^2}\ \frac{G''\left(G^{-1}(r)\right)}{G'\left(G^{-1}(r)\right)}$$

$$U''(r) = -\ \frac{G''\left(G^{-1}(r)\right)}{\left(G'\left(G^{-1}(r)\right)\right)^3} \tag{10}$$

Since $G' = \psi$ and since this implies, using (6), that

$$G''(y) = \psi(y)\left(-\ \frac{y-\mu'}{\sigma'^2}\right) \tag{11}$$

we have that the sign of (10) is equal to the sign of

$$\frac{U-\mu'}{\sigma'^2}$$

hence <0 for $U<\mu'$ and >0 for $U>\mu'$: first concavely increasing followed by a convex increase. The inflection point is in $(r, U(r))$ where $U(r)=\mu'$.

This ends the study of the rank-order distribution of the uncitedness factor $U(r)$. Now we will combine both results on IF and U in order to obtain the functional relationship $IF(U)$ between IF and U.

The functional relation between the impact factor IF and the uncitedness factor U.

In this section we make one more assumption: the rank r occurring in (5) is the same as the rank r in (9). This is model-theoretically acceptable since this assumption is equivalent by supposing that IF decreases with U (note that $IF(r)$ decreases in r and $U(r)$ increases in r).

From (5) we have

$$r = T - F(x) \tag{12}$$

with $x = IF(r)$ and from (9) we have

$$r = G(y) \tag{13}$$

with $y = U(r)$. (12) and (13) yield

$$T - F(x) = G(y)$$

$$F(x) = T - G(y)$$

$$x = F^{-1}(T - G(y))$$

So,

$$IF(U) = IF = F^{-1}(T - G(U)) \tag{14}$$

being the desired relation between IF and U. Now

$$IF'(U) = \frac{-G'(U)}{F'\left(F^{-1}\left(T - G(U)\right)\right)} \tag{15}$$

$$= \frac{-\psi(U)}{\varphi(IF(U))} < 0 \tag{16}$$

by (8) (definition of G) and by (4) (definition of F). Hence we confirm that IF is a strictly decreasing function of U. We also see that $U=0$ implies (by (14) and the definition of F and G) $IF = F^{-1}(T) = +\infty$ and that $U=1$ implies $IF = F^{-1}(0) = 0$. The fact that $IF(0) = +\infty$

implies that $\mathrm{IF}(\mathrm{U})$ starts decreasing convexly. We now prove that $\mathrm{IF}(\mathrm{U})$ has an inflection point. By (15)

$$\mathrm{IF}''(\mathrm{U}) = -\frac{1}{\left(\mathrm{F}'\left(\mathrm{F}^{-1}\left(\mathrm{T}-\mathrm{G}(\mathrm{U})\right)\right)\right)^2}\frac{\mathrm{F}''\left(\mathrm{F}^{-1}\left(\mathrm{T}-\mathrm{G}(\mathrm{U})\right)\right)}{\mathrm{F}'\left(\mathrm{F}^{-1}\left(\mathrm{T}-\mathrm{G}(\mathrm{U})\right)\right)}\left(-\mathrm{G}'(\mathrm{U})\right)\left(-\mathrm{G}'(\mathrm{U})\right) - \frac{1}{\mathrm{F}'\left(\mathrm{F}^{-1}\left(\mathrm{T}-\mathrm{G}(\mathrm{U})\right)\right)}\mathrm{G}''(\mathrm{U})$$

By definition of F and G, by (11) and the similar formula for F and by (14) we have

$$\mathrm{IF}''(\mathrm{U}) = -\frac{1}{\varphi(\mathrm{IF})^3}\varphi(\mathrm{IF})\left(-\frac{\mathrm{IF}-\mu}{\sigma^2}\right)\psi^2(\mathrm{U}) - \frac{1}{\varphi(\mathrm{IF})}\psi(\mathrm{U})\left(-\frac{\mathrm{U}-\mu'}{\sigma'^2}\right) \tag{17}$$

It is hard to determine the sign of $\mathrm{IF}''(\mathrm{U})$ but we can prove the following: since F and G are cumulative normal distributions and by (2) and (8) we have that $\mathrm{G}(\mu') = \frac{\mathrm{T}}{2}$ and $\mathrm{F}(\mu) = \frac{\mathrm{T}}{2}$. Now (14) yields

$$\mathrm{IF}(\mu') = \mu \tag{18}$$

, implying that the point (μ', μ) is on the curve (14). This is also the inflection point of $\mathrm{IF}(\mathrm{U})$: indeed (18) implies that, in this point, $\mathrm{IF}''(\mathrm{U}) = 0$. We now have that $\mathrm{IF}(\mathrm{U})$, first decreases convexly up to $\mathrm{U} = \mu'$ and then decreases concavely. We hence have a graph as in Fig. 1

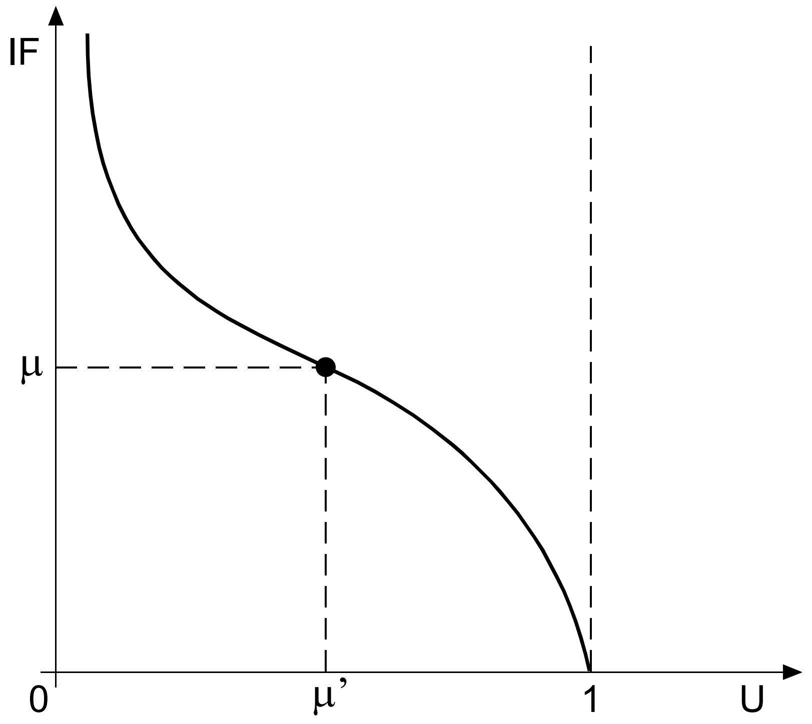

Figure 1. Graph of $\mathrm{IF}(\mathrm{U})$ and inflection point in (μ', μ).

This ends the study of the function $\mathrm{IF}(\mathrm{U})$. In van Leeuwen and Moed (2005), however, one gives graphs of $\ln\left(\mathrm{IF}(\mathrm{U})\right)$, the semi-logarithmic version of $\mathrm{IF}(\mathrm{U})$. In order to be able to compare our model with these experimental results, we will now study the function

$$\mathrm{h}(\mathrm{U}) =: \ln\left(\mathrm{IF}(\mathrm{U})\right) = \ln\left(\mathrm{F}^{-1}\left(\mathrm{T}-\mathrm{G}(\mathrm{U})\right)\right) \tag{19}$$

We have

$$\mathrm{h}'(\mathrm{U}) = \frac{1}{\mathrm{F}^{-1}\left(\mathrm{T}-\mathrm{G}(\mathrm{U})\right)} \frac{-\mathrm{G}'(\mathrm{U})}{\mathrm{F}'\left(\mathrm{F}^{-1}\left(\mathrm{T}-\mathrm{G}(\mathrm{U})\right)\right)} \tag{20}$$

Of course, also this function is strictly decreasing. The second derivative is complicated but can, nevertheless, be used further on

$$\mathrm{h}''(\mathrm{U}) = -\frac{1}{\left(\mathrm{F}^{-1}\left(\mathrm{T}-\mathrm{G}(\mathrm{U})\right)\right)^2} \frac{\left(-\mathrm{G}'(\mathrm{U})\right)\left(-\mathrm{G}'(\mathrm{U})\right)}{\left(\mathrm{F}'\left(\mathrm{F}^{-1}\left(\mathrm{T}-\mathrm{G}(\mathrm{U})\right)\right)\right)^2}$$

$$+\frac{1}{\mathrm{F}^{-1}\left(\mathrm{T}-\mathrm{G}(\mathrm{U})\right)}\left(-\frac{1}{\left(\mathrm{F}'\left(\mathrm{F}^{-1}\left(\mathrm{T}-\mathrm{G}(\mathrm{U})\right)\right)\right)^2} \frac{\mathrm{F}''\left(\mathrm{F}^{-1}\left(\mathrm{T}-\mathrm{G}(\mathrm{U})\right)\right)}{\mathrm{F}'\left(\mathrm{F}^{-1}\left(\mathrm{T}-\mathrm{G}(\mathrm{U})\right)\right)}\left(-\mathrm{G}'(\mathrm{U})\right)\left(-\mathrm{G}'(\mathrm{U})\right)\right)$$

$$+\frac{1}{\mathrm{F}^{-1}\left(\mathrm{T}-\mathrm{G}(\mathrm{U})\right)} \frac{-\mathrm{G}''(\mathrm{U})}{\mathrm{F}'\left(\mathrm{F}^{-1}\left(\mathrm{T}-\mathrm{G}(\mathrm{U})\right)\right)}$$

Using the definition of F and G, by (11) and the similar one for F and by (14) we have

$$\mathrm{h}''(\mathrm{U}) = \frac{-\psi^2(\mathrm{U})}{\left(\mathrm{IF}(\mathrm{U})\right)^2\left(\varphi\left(\mathrm{IF}(\mathrm{U})\right)\right)^2} - \frac{\psi^2(\mathrm{U})\varphi\left(\mathrm{IF}(\mathrm{U})\right)\left(-\dfrac{\mathrm{IF}(\mathrm{U})-\mu}{\sigma^2}\right)}{\mathrm{IF}(\mathrm{U})\left(\varphi\left(\mathrm{IF}(\mathrm{U})\right)\right)^3} + \frac{\psi(\mathrm{U})\left(\dfrac{\mathrm{U}-\mu'}{\sigma'^2}\right)}{\mathrm{IF}(\mathrm{U})\left(\varphi\left(\mathrm{IF}(\mathrm{U})\right)\right)}$$

which has the same sign as

$$-\frac{\psi(\mathrm{U})}{\mathrm{IF}(\mathrm{U})\varphi\left(\mathrm{IF}(\mathrm{U})\right)} + \frac{\psi(\mathrm{U})\left(\dfrac{\mathrm{IF}(\mathrm{U})-\mu}{\sigma^2}\right)}{\varphi\left(\mathrm{IF}(\mathrm{U})\right)} + \frac{\mathrm{U}-\mu'}{\sigma'^2}$$

We now see that in the point $(\mu', \ln\mu)$ (i.e. for $\mathrm{U}=\mu'$ and $\mathrm{IF}=\mu$) that $\mathrm{h}''(\mathrm{U})$ has the sign of

$$\frac{-\psi(\mathrm{U})}{\mathrm{IF}(\mathrm{U})\varphi\left(\mathrm{IF}(\mathrm{U})\right)} < 0$$

hence strictly negative.

This implies, when compared to $\mathrm{IF}(\mathrm{U})$ (where in this point one had an inflection point), that $\ln(\mathrm{IF}(\mathrm{U}))$ is already concavely decreasing in this point. Since again $\ln(\mathrm{IF}(0)) = +\infty$ (by (19)), the curve $\ln(\mathrm{IF}(\mathrm{U}))$ starts decreasing convexly. In conclusion: the inflection point of the curve $\ln(\mathrm{IF}(\mathrm{U}))$ occurs at an abscissa $< \mu'$, hence the "convex part" is relatively smaller than the "concave part", certainly in comparison with the graph of $\mathrm{IF}(\mathrm{U})$ (Fig. 1). Note also that $\ln(\mathrm{IF}(1)) = -\infty$ (by (19)) so that we reached a graph as in Fig. 2.

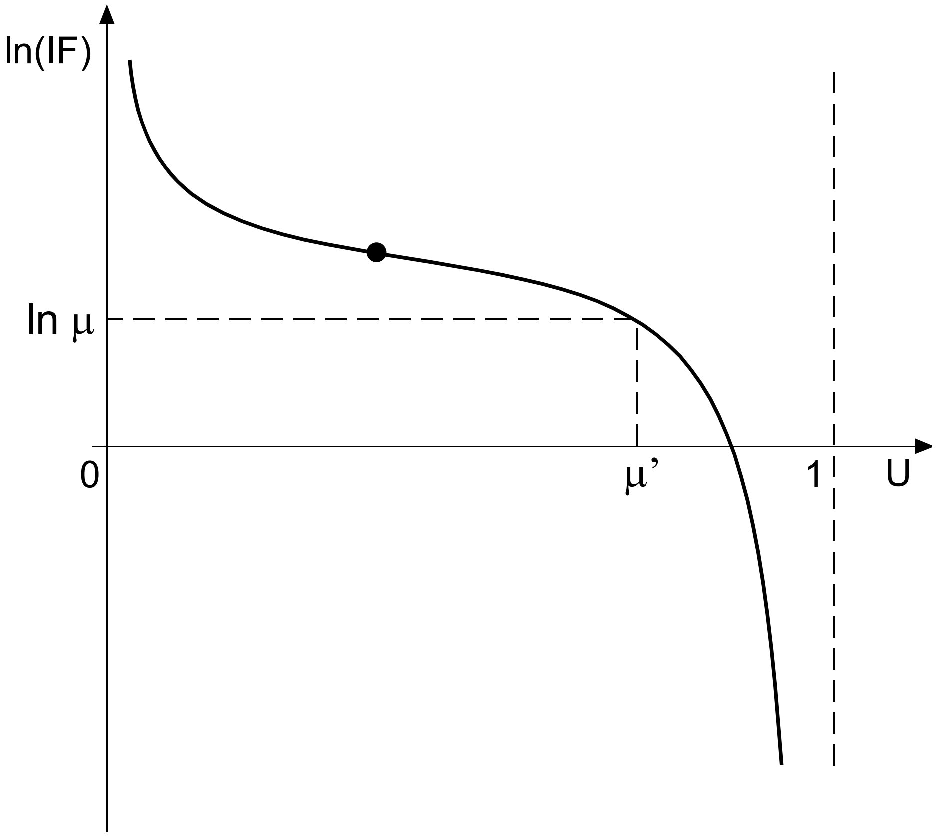

Figure 2. Graph of $\ln(\mathrm{IF}(\mathrm{U}))$ and inflection point at an abscissa $< \mu'$

It is clear that this graph has the same shape as the one in van Leeuwen and Moed (2005) which we reproduce here (Fig. 3). Note also that the smaller convex part, compared to the concave part, is explained (note that the ordinate values below 1 are actually negative since they express the $\ln(\mathrm{IF})$ values but the numbers in van Leeuwen and Moed (2005) are actually IF-values instead of $\ln(\mathrm{IF})$-values).

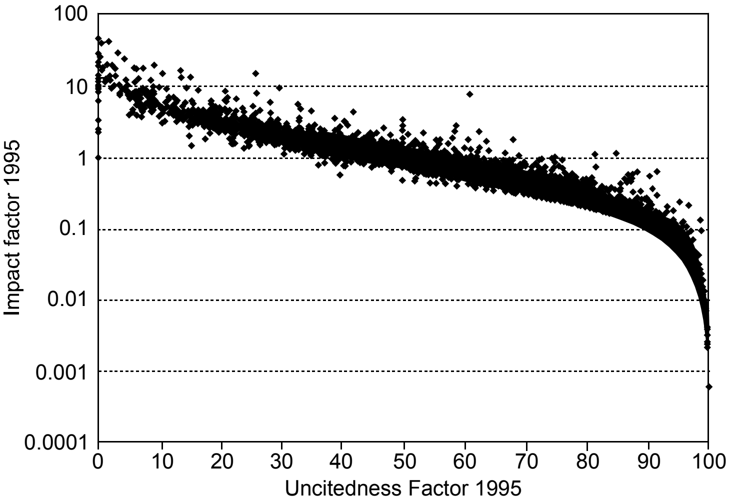

Figure 3. $\ln(\mathrm{IF}(\mathrm{U}))$, experimentally: SCI (all fields), 1995, van Leeuwen and Moed (2005), reprinted with kind permission of Springer.

Conclusions and open problems

Based on the fact that, for a set of journals, IF (any impact factor, irrespective of the publication or citation period) as well as U (the uncitedness factor, i.e. the fraction of the papers in a journal that are not cited) are averages, the CLT (Central Limit Theorem) could be applied to reach the following results

(i) Modelling the shape of the rank-order distribution $\mathrm{IF}(\mathrm{r})$ where journals are ranked in decreasing order of their IF. This was realized in Egghe (2009) explaining experimental results in Mansilla, Köppen, Cocho and Miramontes (2007).

(ii) Modelling the shape of the rank-order distribution $\mathrm{U}(\mathrm{r})$ where journals are ranked in decreasing order of their U.

(iii) Based on (i) and (ii): modelling the shape of the functional relation between U and $\mathrm{IF}(\mathrm{U})$. We proved that this function has an S-shape: starting convexly decreasing and followed by a concave decrease. The inflection point is in (μ',μ) where μ' is the average of the U-values and μ is the average of the IF-values.

(iv) Since we only have the van Leeuwen and Moed (2005) graph we also studied the relation between U and $\ln(\mathrm{IF}(\mathrm{U}))$. Now the graph starts convexly decreasing followed by a concave decrease but the inflection point has an abscissa $<\mu'$ which makes the convex part smaller than the concave part, which is also confirmed by the van Leeuwen and Moed graph (Fig. 3).

It would be interesting to calculate (exactly or numerically) the functions $\mathrm{IF}(\mathrm{U})$ and $\ln(\mathrm{IF}(\mathrm{U}))$ for some concrete values of μ, σ, μ' and σ'. Also, the “thicker” part of the graph in Fig. 3 should be explained, i.e. why we have there a relation and not exactly a function between $\ln(\mathrm{IF})$ and U. Indeed: the rank-order distributions $\ln(\mathrm{IF}(\mathrm{r}))$ in Mansilla, Köppen, Cocho and Miramontes (2007) are clearer functions than the graph of Fig. 3 in van Leeuwen and Moed (2005).

References

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

Egghe, L. (2009). Mathematical derivation of the impact factor distribution. Journal of Informetrics, to appear.

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

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