A basic indicator for relative contributions and a remarkable difference between a ratio of averages and an average of ratios

Leo Egghe[1,2] and Ronald Rousseau[2,3,4]
[1]Universiteit Hasselt (UHasselt), Campus Diepenbeek, Agoralaan,
B-3590 Diepenbeek, BELGIUM
[2]Antwerp University (UA), IBW, Venusstraat 35, 2000 Antwerp, BELGIUM
[3]KHBO-VIVES, Faculty of Engineering Technology,
Zeedijk 101, 8400 Oostende, BELGIUM
[4] KU Leuven, Dept. Mathematics, 3000 Leuven, BELGIUM
e-mail: leo.egghe@uhasselt.be; ronald.rousseau@uantwerpen.be

ABSTRACT

This article has been inspired by the activity index and problems in its proper understanding. It studies a basic relative contribution index which is placed in the context of countries' publication contributions. Two versions are proposed: one being an average of ratios (AoR) and the other one a ratio of averages (RoA). A Lotkaian-Zipfian framework is used to model the two versions of the proposed indicator. A remarkable difference between the two approaches (RoA vs. AoR) is found when determining the fraction of units (countries) that have a value larger than one. This observation contributes to the understanding of the differences between these two approaches.

Keywords: Average of ratios (AoR); Ratio of averages (RoA); Relative contribution index; Independence.

INTRODUCTION: A BASIC RELATIVE INDICATOR

Comparing a value with an average or median value leads to a basic indicator which has been applied in many fields and many cases. Table 1 shows existing or potential examples in the context of scholarly communication, diffusion studies and educational assessment.

Table 1: Examples of Context and the Respective Indicator

Context	Focus	Indicator
Received citations during a given period (e.g. two years) by articles published in a given journal volume.	One article	Received citations of this article divided by the average number of received articles published in the same volume
Received citations during a given period (e.g. two years) by articles published in a given WoS subfield.	One article	Received citations of this article divided by the average number of received articles published in the same year and in the same WoS subfield
Articles published in a given field.	One country	Number of articles published by this country during a given period divided by the average number of articles in the same field, during the same period, by all countries
Diffusion studies: citing JCR categories of a journal volume	One article	Number of JCR categories that cite a given article divided by the average number of citing JCR categories for all articles in a given journal volume.
Scores in a given test (educational research)	One pupil	Score of this pupil in a given test divided by the average score of all pupils in her class.

As an example we focus on the number of publications of country C during a given period (e.g. one particular year) in a given field F. This number is denoted as s(C,F). For country C_k the number $s(C_k,F)$ is an absolute indicator which can easily be made into a relative one by dividing it by the average value of s(C,F), averaged over all countries(or regions) under consideration. We denote this indicator by $S_F(C_k)$:

$$S_F(C_k) = \frac{s(C_k, F)}{\frac{1}{m}\sum_{j=1}^{m} s(C_j, F)}$$

With the average-of-ratios versus ratio-of-averages (AoR-RoA) debate in mind (Larivière and Gingras 2011; Lundberg 2007; Opthof and Leydesdorff 2010) we also propose the following alternative:

$$S_F^1(C_k) = \frac{1}{m} \sum_{j=1}^{m} \frac{s(C_k, F)}{s(C_j, F)}$$

where the sum is taken over all countries for which $s(C_j,F) > 0$.The indicator S_F compares the production of a country in a given field with the average production of all countries in the same field, while S^1_F compares the production of this country with every country and determines the average value of this ratio. This is nothing but $s(C_k,F)$ divided by the harmonic mean of the $s(C_j,F)$.

Clearly it is possible that a country has an index (S_F or S^1_F) that is larger than one for all fields. Indeed, if country C_0 publishes more than half of all publications and this in every field, then $s(C_0,F)$ is strictly larger than the average of all $s(C_j,F)$ and this for every F. Similarly, $s(C_0,F)$ is larger than each $s(C_j,F)$ if $C_j \neq C_0$. Hence also S^1_F being an average of numbers that are larger than one, is also larger than one. Basically, although there is a comparison with reference values, S_F and S^1_F are "big is beautiful" indicators. Next we study the continuous form of these indicators in a Lotkaian-Zipfian framework, leading to a surprising result.

A study of S_F and S^1_F in a Lotkaian-Zipfian framework

We keep the field F fixed and assume that countries are ranked according to $s(C_r,F)$. In this framework the rank-parameter r is assumed to be a real-valued variable defined on the interval [0,T], where T denotes (the continuous analogue of) the total number of countries. Assuming a Zipfian model (Egghe 2005) we have:

$$s(C_r, F) = \frac{B}{r^\beta} = g(r), \qquad \text{with } B, \beta > 0$$

if μ denotes the average, we know (Egghe, 2005) that for $0 < \beta < 1$

$$\mu = \frac{1}{T}\int_0^T g(r)dr = \frac{1}{1-\beta}$$

In the corresponding Lotka framework (thus $\alpha > 2$) this leads to:

$$\mu = \frac{\alpha-1}{\alpha-2}, with\ B = (T)^{\frac{1}{\alpha-1}}$$

Hence,

$$S_F(C_r) = \frac{s(C_r,F)}{\mu} = \frac{\frac{B}{r^\beta}}{\frac{\alpha-1}{\alpha-2}} = \frac{\frac{(T)^{\frac{1}{\alpha-1}}}{r^{\frac{1}{\alpha-1}}}}{\frac{\alpha-1}{\alpha-2}} = \frac{\alpha-2}{\alpha-1}\left(\frac{T}{r}\right)^{\frac{1}{\alpha-1}}$$

Similarly, for $\alpha > 1$:

$$S^1_F(C_r) = \frac{1}{T}\int_0^T \frac{\frac{B}{r^\beta}}{\frac{B}{t^\beta}}dt = \frac{1}{T * r^\beta}\int_0^T t^\beta\, dt = \frac{1}{\beta+1}\left(\frac{T}{r}\right)^\beta = \frac{\alpha-1}{\alpha}\left(\frac{T}{r}\right)^{\frac{1}{\alpha-1}}$$

Now we investigate for which ranks r $S_F(C_r) > 1$, and similarly for $S^1_F(C_r)$.

$$S_F(C_r) > 1 \Leftrightarrow r < \left(\frac{\alpha-2}{\alpha-1}\right)^{\alpha-1} T = r_0$$

We note, as a check for our calculations, that $0 < r_0 < T$. The proportion of countries for which $S_F(C_r) > 1$ is:

$$0 < \frac{r}{T} = \left(\frac{\alpha-2}{\alpha-1}\right)^{\alpha-1} = \theta(\alpha)$$

We observe that θ(α) is an increasing function of α ($\alpha > 2$), that $\lim_{\alpha\to 2}\theta(\alpha) = 0$ and $\lim_{\alpha\to\infty}\theta(\alpha) = \frac{1}{e} \approx 0.368$.

Considering now $S^1_F(C_r)$ we have:

$$S^1_F(C_r) > 1 \Leftrightarrow r < \left(\frac{\alpha-1}{\alpha}\right)^{\alpha-1} T = r^1_0$$

Also here we see that $0 < r^1_0 < T$. The proportion of countries for which $S^1_F(C_r) > 1$ is:

$$0 < \frac{r}{T} = \left(\frac{\alpha - 1}{\alpha}\right)^{\alpha - 1} = \theta^1(\alpha)$$

We observe that now $\theta^1(\alpha)$ is a decreasing function of α, that $\theta^1(2) = 0.5$, $\lim_{\alpha \to 1} \theta^1(\alpha) = 1$and $\lim_{\alpha \to \infty} \theta^1(\alpha) = \frac{1}{e}$. Hence θ(α) and $\theta^1(\alpha)$ have the same limit for increasing values of α, but in the first case this limit is approached from below, while in the second case it is approached from above.

FURTHER OBSERVATIONS

The indicators S_F and S^1_F are independent (Bouyssou and Marchant 2011) in the sense that, if for countries C_k and C_n the relation $S_F(C_k) < S_F(C_n)$ holds and if we add the same number of publications, p> 0, in the field F, to the output of these two countries then still $S_F(C'_k) < S_F(C'_n)$, where the notations C'_k and C'_n refer to the same countries but with an increased number of publications in the field F.

Similarly, if $S^1_F(C_k) < S^1_F(C_n)$ and if, as above, we add the same number of publications, p, in the field F to the two countries then still $S^1_F(C'_k) < S^1_F(C'_n)$, where C'_k and C'_n have the same meaning as above.

Proof of this proposition for the indicator S_F.

If $S_F(C_k) < S_F(C_n)$ then clearly $s(C_k,F) < s(C_n,F)$. Adding p publications in the field F to the publication output of the two countries yields $s(C_k,F) + p < s(C_n,F) + p$, and as $S_F(C'_k)$ and $S_F(C'_n)$ have the same denominator the required inequality is trivially satisfied.

Proof of the proposition for the indicator S^1_F.

If $S^1_F(C_k) < S^1_F(C_n)$ then again $s(C_k,F) < s(C_n,F)$ (as $S^1_F(.)$ can be written as s(.,F) times a factor which is the same for each country). If now p publications are added to C_k and C_n then $s(C'_k,F) = s(C_k,F) + p < s(C'_n,F) = s(C_n,F) + p$, and this factor is multiplied by a factor which is still the same for the two countries (be it changed with respect the original situation) leading to the required inequality: $S^1_F(C'_k) < S^1_F(C'_n)$.

If countries C_k and C_n have no publications in common and they are considered to be one country (denoted as $C_k \cup C_n$) what is then the relation between $S_F(C_k) + S_F(C_n)$ and $S_F(C_k \cup C_n)$, and similarly for S^1_F ? The answer to the first question is that $S_F(C_k \cup C_n) < S_F(C_k) + S_F(C_n)$. Indeed: if countries C_k and C_n have no publications in common then $s(C_k,F) + s(C_n,F) = s(C_k \cup C_n,F)$. This operation does not alter the total number of publications involved, but after taking the union of two countries the average number of publications per country has increased. This implies that $S_F(C_k \cup C_n) < S_F(C_k) + S_F(C_n)$.

An illustration: if m = 4 and $s(C_1,F) = s(C_2,F) = s(C_4,F) = 10$, $s(C_3,F) = 100$and C_2 and C_3 are brought together leading to country $C_{2,3}$ then $S_F(C_{2,3}) = S_F(C_2 \cup C_3) = \frac{110}{\frac{1}{3}(10+110+10)} \approx 2.538$ while $S_F(C_2) + S_F(C_3) = \frac{10+100}{\frac{1}{4}(10+10+100+10)} \approx 3.385$. Generally: $S_F(C_k \cup C_n) = (m-1)/m\ (S_F(C_k) + S_F(C_n))$, where m denotes the number of different countries.
Remarkably the corresponding inequality: $S^1_F(C_k \cup C_n) < S^1_F(C_k) + S^1_F(C_n)$ is not generally valid. Indeed, let m = 3, $s(C_1,F) = s(C_2,F) = 100$ and $s(C_3,F) = 1$ and let k =1 and n = 2. Then

$S^1_F(C_k \cup C_n) = S^1_F(C_1 \cup C_2) = \frac{1}{2}\left(\frac{200}{200} + \frac{200}{1}\right) \approx 100.5$ while $S^1_F(C_k) + S^1_F(C_n) = S^1_F(C_1) + S^1_F(C_2) = \frac{2}{3}\left(\frac{100}{100} + \frac{100}{100} + \frac{100}{1}\right) = 68$. Hence this inequality is not valid. If however m = 2 or if all $s(C_j,F)$ are equal then the inequality holds. Indeed if m = 2, then $S^1_F(C_1 \cup C_2) = 1$, while $S^1_F(C_1) + S^1_F(C_2) = \frac{1}{2}\left(1 + \frac{s(C_1)}{s(C_2)}\right) + \frac{1}{2}\left(\frac{s(C_2)}{s(C_1)} + 1\right) > 1$. Further, if all $s(C_j,F)$ are equal, then $S^1_F(C_k \cup C_n) = \frac{1}{m-1} * \left((m-2) * 2 + 1\right) = \frac{2m-3}{m-1}$ while $S^1_F(C_k) + S^1_F(C_n) = 2$, showing that also in this special case the inequality holds.

DISCUSSION

The study of this indicator was inspired by the activity index. Recall (Frame 1977; Schubert and Braun 1986; Schubert, Glänzel and Braun 1988) that the activity index (AI) of country C with respect to a given field F (and with respect to the world, W, which in practice means with respect to the used database) in a given year Y (or, more generally, period P) is defined as:

$$AI(C,F,W,Y) = \frac{the\ country's\ \ share\ in\ the\ publication\ output\ in\ field\ F}{the\ country's\ share\ in\ the\ world's\ \ publication\ output\ in\ all\ fields}$$

Denoting by s(C,F) the publication output of country C in field F; by t(C) the total publication output of country C; by v(F) the total number of publications published in field F; and by w the total number of publications in the world over all fields, we see that:

$$AI(C,F) = \frac{s(C,F) * w}{v(F) * t(C)}$$

Yet it is not straightforward to correctly understand the meaning of this indicator and its analogues such as the attractivity and the Balassa index (Rousseau and Yang 2012; Rousseau 2012). Moreover its definition implies that a country cannot have an activity index strictly larger than one (or strictly smaller than one) in all fields. Indeed: consider country C and assume that for each j = 1,…, n $AI(C,F_j) \geq 1$, with at least one strict inequality. Then we have for each j = 1, …,n:

$$\frac{\frac{s(C,F_j)}{t(C)}}{\frac{v(F_j)}{w}} \geq 1 \text{ and hence: } s\left(C,F_j\right) * w \geq t(C) * v(F_j)$$

As this inequality holds for each j = 1,…,n and there is at least one strict inequality we obtain:

$$\sum_{j=1}^{n} s\left(C,F_j\right) * w > t(C) * \sum_{j=1}^{n} v(F_j)$$

leading to t(C)*w > t(C)*w which clearly is a contradiction. This simple proof has already appeared in Rousseau (2012) but was included here as it was found independently by the first author.

These facts inspired us to consider the simpler indicator studied above.

CONCLUSION

In this short note we collected some observations related to the simple indicator constructed by comparing a value associated with a unit with its average over a set of comparable units. Depending on the RoA or the AoR approach, we found – at least in a Lotkaian framework - a remarkable difference when determining the fraction of units (countries) that have a value larger than one.

ACKNOWLEDGEMENT.

This research is supported by NFSC grant no. 71173154.

REFERENCES

Bouyssou, D. and Marchant, T. 2011 Ranking scientists and departments in a consistent manner. *Journal of the American Society for Information Science and Technology*, Vol. 62, no. 9: 1761-1769.

Egghe, L. 2005. *Power laws in the information production process: Lotkaian informetrics*. Elsevier: Amsterdam.

Frame, J.D. 1977. Mainstream research in Latin America and the Caribbean. *Interciencia*, Vol. 2, no. 3, 143-148.

Larivière, V. and Gingras, Y. 2011. Averages of ratios vs. ratios of averages: an empirical analysis of four levels of aggregation. *Journal of Informetrics*, Vol. 5, no. 3: 392-399.

Lundberg, J. 2007. Lifting the crown – citation z-score. *Journal of Informetrics*, Vol. 1, no. 2: 145-154.

Opthof, T. and Leydesdorff, L. 2010. Caveats for the journal and field normalizations of the CWTS ("Leiden") evaluations of research performance. *Journal of Informetrics*, Vol. 4, no. 3: 423-430.

Rousseau, R. 2012.Thoughts about the activity index and its formal analogues. *ISSI Newsletter*, Vol. 8, no. 4: 73-75.

Rousseau, R. and Yang, L.Y. 2012. Reflections on the activity index and related indicators. *Journal of Informetrics*, Vol. 6: 413-421.

Schubert, A. and Braun, T. 1986. Relative indicators and relational charts for comparative assessment of publication output and citation impact. *Scientometrics*, Vol. 9, no. 5-6: 281-291.

Schubert, A., Glänzel, W. and Braun, T. 1988. Against absolute methods: relative scientometric indicators and relational charts as evaluation tools, In: A.F.J van Raan (ed.) *Handbook of Quantitative Studies of Science and Technology* (pp. 137-175). Amsterdam: Elsevier.