Journal of Informetrics 7 (2013) 907–913

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

A simulation study to investigate the accuracy of approximating averages of ratios using ratios of averages

J.M. van Zyl

Department of Mathematical Statistics and Actuarial Science, University of the Free State, Bloemfontein, South Africa

ARTICLE INFO

Article history:
Received 7 June 2013
Received in revised form 22 August 2013
Accepted 29 August 2013

Keywords:
Averages of ratios
Ratio of averages
Power-law
Citations

ABSTRACT

For a number of researchers a number of publications for each author is simulated using the zeta distribution and then for each publication a number of citations per publication simulated. Bootstrap confidence intervals indicate that the difference between the average of ratios and the ratio of averages are not significant. It was found that the log–logistic distribution which is a general form for the ratio of two correlated Pareto random variables, give a good fit to the estimated ratios.

1. Introduction

This study is concerned with a common problem in bibliometrics, whether the ratio of averages (or totals) can be used as a proxy for the average of the ratios of the individuals. An application is where totals of publications and citations are available for each of various scientific fields and one wish to compare the average number of citations per publication over scientific fields. The question is if the ratio of the total number of citations to the total number of publications can be used as a proxy for the average ratio, citations per publication, calculated over the individual researchers. The ratio citations per publication is an integral aspect of this problem and important when only the total number of publications and citations of a group, and not the results of individual researchers are available. Two of the important references on the problem of comparing different scientific or subject fields are those by Waltman, van Eck, van Leeuwen, Visser, and van Raan (2011) and Opthof and Leydesdorff (2011).

Summary citation data are often available but not results for the individual researchers. An example is where Scopus provides a huge database of research output of countries in terms of totals per subject field (SJR—SCImago Journal & Country Rank, 2007). Van Zyl and van der Merwe (2012) used the Scopus totals to find an approximate ranking of the average number of citations per publication over subject fields.

Egghe (2012) derived mathematical results concerning relationships between averages of ratios (AoR) and ratios of averages (RoA). He proved that the mean AoR and RoA are equal if the correlation between the ratios and the denominator used to calculate the ratio is zero. If the data of individuals are available, this result can be used, but in practice often only totals are available and a correlation cannot be calculated. In the simulation study the correlation between the ratio, citations/publications, and publications were found to be approximately zero and the variation of the correlation coefficient is decreasing as the sample size increases. In small samples, say less than 15 observations, the approximation of AoR by using RoA might not be advisable since the observed correlation have a large variation and the correlation can be as large as for example 0.6 in some samples, and it is mostly positive in these cases.

E-mail address: wwjvz@ufs.ac.za

1751-1577/$ – see front matter © 2013 Elsevier Ltd. All rights reserved.
http://dx.doi.org/10.1016/j.joi.2013.08.008

Larivière and Gingras (2011) conducted a thorough study and tested equality of the medians of AoR and RoA using the Wilcoxon signed-rank test (Gibbons & Chakraborti, 2011). For symmetric distributions medians and averages are equal, but not in general for skewed distributions. The distribution of differences of the AoR's and RoA's and of the logs of AoR and RoA are not symmetric in the cases investigated in this study, which implies that the medians and means may differ. Conclusions based on tests for medians may thus not be valid for the means.

In large samples, testing for equality of means or medians, will be statistically significant even if the difference is negligible from a practical viewpoint. For example consider two average citations per publication calculated from large samples, say 10.5 and 10.6. This difference can be statistically significant since the standard statistical test is for equality, but in a practical problem such a difference is negligible when comparing average number of citations per publication.

In order to investigate results where citation statistics are involved, the simulation should be according to the distributional laws involved in citation data and two aspects which must to be considered is the distribution of the number of publications per author and also the distribution of citations per publication. The number of publications per researcher was simulated used a parametric approach, the zeta density, and the number of citations for each publication simulated using a nonparametric approach by simulating from observed probabilities in a large sample of real data.

1.1. *Distribution of number of publications per author*

In the simulation study it will be assumed that the output or number of publications per researcher is generated according to Lotka's Law (Lotka, 1926). There is still much research and development of models in this field, but the discrete Pareto distribution or zeta distribution to model the number of papers generated by individual authors is often used.

The zeta density or pure power-law discrete density is

$$p(k) = \frac{k^{-\gamma}}{\zeta(\gamma)}, \quad k = 0, 1, 2, 3\ldots, \tag{1}$$

where $\zeta(\gamma)$ is the Riemann zeta function which is finite for $\gamma > 1.0$ and is defined as $\zeta(\gamma) = \sum_{k=1}^{\infty} k^{-\gamma}$.

The zeta density or discrete Pareto distribution is reviewed in the book by Johnson, Kemp, and Kotz (2005). Estimated values of γ in the region of three for the output of researchers are often found when using real data. Applications of this distribution can be found in the papers by Goldstein, Morris, and Yen (2004) and Redner (1998). In the work of Goldstein, Morris, and Yen (2004), the zeta distribution was fitted to the number of publications of 1354 authors, comprising 900 papers, in the field of complex networks, and a good fit was found with $\gamma = 2.544$.

1.2. *Distribution of the number of citations per publication*

The number of citations per publication was simulated using a non-parametric approach without making an assumption about a parametric distribution of citations. The probabilities were calculated directly from a set of data. A comprehensive and typical set of citation data is that compiled by H. Small and D. Pendlebury of the data of the Institute for Scientific Information (ISI) which covers all publications from ISI catalogued journals that were published in 1981 and cited during the period January 1981–June 1997. The frequency for each number of citations is also given, thus an empirical density estimate for the distribution of citations. This set of data is available on the website of Sidney Redner and also described in the papers of Redner (1998) and Peterson, Pressè, and Dill (2010). It is a very large sample and comprises the number of citations of 344,589 papers and a total of 783,339 citations.

There were 368,110 papers with zero citations in the sample, and the probability of zero citations is estimated as 368,110/783,339 = 0.4699. There were 70,836 publications of the 783,339 publications with one citation, and the probability for a publication to have one citation is estimated as 0.0904 = 70,636/783,336, 44,127 with 2 citations and the probability of 2 citations for a paper 44,127/783,336 = 0.0563 and so on. These probabilities were used to simulate the number of citations for each publication.

Various researchers used a parametric approach to model the distribution of citations and found good results when fitting parametric discrete and also continuous distributions to the number of citations. Peterson, Pressè, and Dill (2010) developed a discrete probability model, Redner (2005), Radicchi, Fortunato, and Castellano (2008), and Limbert, Stahel and Abt (2001) favored the lognormal distribution. Newman (2005) suggested the Yule distribution as an alternative to the zeta distribution. In this work the nonparametric approach was used with making any distributional assumptions.

The average over citations per publication per researcher was calculated and also the average of the ratio of total number of citations to total number of publications for this group of researchers. This was repeated $m = 1000$ times, leading for a specific number of authors, n and γ to 1000 estimated pairs of AoR's and RoA's. Confidence intervals for the differences between RoA and AoR were constructed using the sample of 1000.

2. Distribution of the ratios

A possible distribution would be the ratio of two correlated Pareto random variables. This was tested and found to give good results when fitted to observed ratio, RoA's and AoR's. There are more than one version of a bivariate Pareto distribution

(Mardia, 1962), Arnold (1983). Let X and Y be two dependent Pareto distributed random variables and consider the bivariate distribution of the two variables

$$f(x,y)=\left[\frac{\alpha(1+\alpha)}{k_1k_2}\right]\left(\frac{1+x}{k_1}+\frac{y}{k_2}\right)^{-\alpha-2},\quad k_1,k_2,\alpha>0. \tag{2}$$

It can be shown, see for example Markovich (2009), that the ratio R = X/Y has the cdf $F_R(x)=1-k_1/(xk_2+1)$, which leads to a density of the form

$$\begin{aligned} f_R(x) &= \frac{k_1^2}{(xk_1+k_2)^2} \\ &= \frac{k_1^2/k_2^2}{(1+x(k_1/k_2))^2} \end{aligned} \tag{3}$$

It can be seen that this is a generalized Pareto distribution (GPD) with index equal to 1. This specific case of the GPD is a special case of the log–logistic density (Kleiber & Kotz, 2003), which is more heavy-tailed that the log–normal. Various densities were tested on different sets of the simulated ratios and for example the log–normal yields good results in some cases, but the log–logistic density gave the most consistent best results when fitted to the ratios. There is a theoretical motivation as explained above, since it is a general form of the ratio of two correlated Pareto random variables. Theoretically the mean of the log–logistic is finite for $\alpha>0$, and is equal to $E(x)=\pi\beta/\alpha\ \sin(\pi/\alpha)$.

Let $\beta=k_2/k_1$ and $\alpha>0$, then the log–logistic density with scale parameter β and shape parameter $\alpha>0$ is

$$f(x)=\frac{(\alpha/\beta)(x/\beta)^{\alpha-1}}{(1+(x/\beta)^{\alpha})^2}. \tag{4}$$

The distribution function is

$$F(x)=\frac{(x/\beta)^{\alpha}}{(1+(x/\beta)^{\alpha})},\quad x\geq 0.$$

Some authors and software use a notation which is in terms of the corresponding logistic density, which is the distribution of $\log(x)$ and of the form:

$$f(x)=\frac{\alpha e^{\alpha(x-\log(\beta))}}{(1+e^{\alpha(x-\log(\beta))})^{-2}},\quad -\infty<x<\infty, \tag{5}$$

and for example Matlab give estimates for $\log(\beta)$, $1/\alpha$. When tested on the ratios for example on simulated samples with $\gamma=3.5$, estimates of α in the region of 4.5 were found. In Figs. 1 and 2 in Section 3 examples are shown of fitted and observed observations.

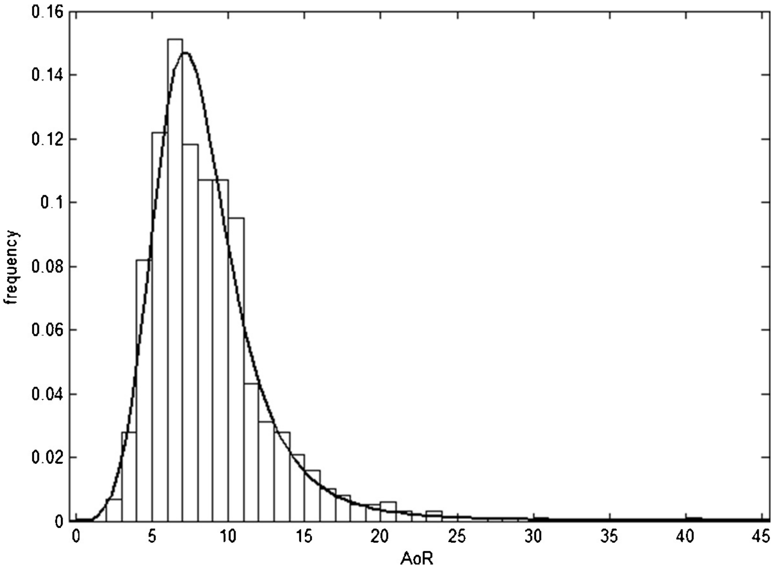

Fig. 1. A histogram of 1000 AoR's and fitted log–logistic density calculated, $n=50$ researchers and $\gamma=3.5$.

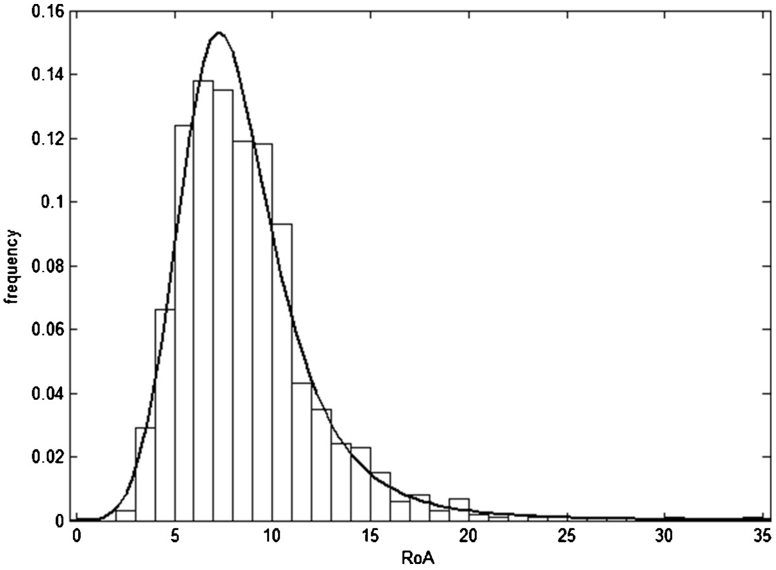

Fig. 2. A histogram of 1000 RoA's and fitted log–logistic density calculated, $n = 50$ researchers and $\gamma = 3.5$.

3. Simulation study

For say n researchers, a number of publications are simulated for each researcher according to the probabilities of the zeta distribution (1). The distribution was truncated at 5000 publications. This was done because of the slow convergence of the probabilities to zero, and it was checked that for example the mean of the truncated distribution is almost exactly equal to the theoretical distribution without truncation. The normalizing constant is $C = 1/\sum_{j=1}^{5000}[k^{-\gamma}/\zeta(\gamma)]$ and the probability for k publications is

$$p(k) = \frac{Ck^{-\gamma}}{\zeta(\gamma)}. \tag{6}$$

It can be noted that the theoretical mean of the zeta distribution is $\zeta(\gamma - 1)/\zeta(\gamma)$, $\gamma > 2$, and the mean using the truncated distribution with $\gamma = 3$ is 1.3683 compared to the theoretical mean of 1.3684, indicating a close approximation.

For a specific n, value of γ, $m = 1000$ publications and citation results were generated for each of the n researchers and the average of ratios and the ratio of averages were calculated. The number of publications was simulated using the zeta density and then for each of the individual publications, a number of citations was simulated using the empirical distribution of the ISI data. In other words for example, for a given γ, say $n = 25$, there were 25 numbers of publications simulated and for every individual publication a number of citations, AoR and RoA calculated and this was repeated $m = 1000$ times, leading to a sample of 1000 AoR's and RoA's.

A 95% bootstrap confidence interval (CI) for the difference between mean RoA and AoR was calculated using 500 simulated estimated ratios of averages and averages of ratios. The difference between the two averages was tested and not normally distributed. It can be mentioned that the difference of the logs give a more symmetric distribution, but also not normally distributed. Bootstrap methods where no distributional assumptions are made were used in the simulation study to calculate confidence intervals for the difference between mean AoR and RoA.

Using the data of 250 researchers with $\gamma = 3$ in the zeta density, the averages over 1000 simulations of the AoR and RoA both differ by less than 0.1 from 8.5733 which is the average number of citations per publication in the ISI data. It should be kept in mind that with $m = 1000$ and especially for the smaller values of n, there would still be much variation when estimating a means.

The results for $n = 25$, 50, 250, 1000 and $\gamma = 2.0$, 3.0, 3.5 is given in Tables 1–3.

It should be noted that theoretically for $\gamma = 2$ the mean is not finite although finite for the truncated distribution. More realistic values of γ larger than 2.5 was found for example by Goldstein, Morris, and Yen (2004) and Redner (1998).

In the following table with $\gamma = 3.5$, the correlations were calculated to investigate the result of Egghe (2012) concerning the regression between the ratios and denominators. Thus for a specific n, 1000 correlations between citations/publications and publications were calculated using n pairs each time. It was found that the correlation is on average very close to zero and decreasing as the sample becomes larger. The distribution of the observed correlations is skewed to the right in small samples. These results indicate that for large sample sizes AoR and RoA are approximately equal (Table 5).

Table 1
Summary statistics of RoA and AoR using $m = 1000$ simulated samples of size n each for each of the sample sizes n. Bootstrap CI for the difference between AoR and RoA included.

$\gamma = 2.0$	Mean AoR	Mean RoA	100 (AoR-RoA)/AoR	Lower bound, 95% Bootstrap CI	Upper bound, 95% Bootstrap CI
$n = 25$	8.5706	8.6511	0.9392	−0.3330	0.2198
$n = 50$	8.5025	8.5418	0.4622	−0.2315	0.1596
$n = 250$	8.6170	8.6363	0.2239	−0.1038	0.0692
$n = 1000$	8.5882	8.5844	−0.0442	−0.0394	0.0546

Table 2
Summary statistics of RoA and AoR using $m = 1000$ simulated samples of size n each for each of the sample sizes n. Bootstrap CI for the difference between AoR and RoA included.

$\gamma = 3.0$	Mean AoR	Mean RoA	100 (AoR-RoA)/AoR	Lower bound, 95% Bootstrap CI	Upper bound, 95% Bootstrap CI
$n = 25$	8.3142	8.3460	0.3824	−0.1400	0.0804
$n = 50$	8.5059	8.5497	0.5149	−0.1572	0.0487
$n = 250$	8.5140	8.5221	0.0951	−0.0568	0.0412
$n = 1000$	8.5590	8.5502	−0.1028	−0.0159	0.0335

Table 3
Summary statistics of RoA and AoR using $m = 1000$ simulated samples of size n each for each of the sample sizes n. Bootstrap CI for the difference between AoR and RoA included.

$\gamma = 3.5$	Mean AoR	Mean RoA	100 (AoR-RoA)/AoR	Lower bound, 95% Bootstrap CI for difference in means	Upper bound, 95% Bootstrap CI for difference in means
$n = 25$	8.5485	8.4946	−0.6345	−0.0525	0.1543
$n = 50$	8.4118	8.4837	0.84750	−0.1496	0.0003
$n = 250$	8.5866	8.6160	0.34122	−0.1164	0.0104
$n = 1000$	8.5525	8.5597	0.08411	−0.0335	0.0127

A histogram of the 1000 observed AoR's is given in Fig. 1 and a histogram of the observed RoA's given in Fig. 2. The fitted log–logistic densities are shown also in these figures. The parameters for this simulated sample are $n = 50$, $\gamma = 3.5$ Approximately 415 of the AoR's were larger than the RoA's. The observed mean of AoR for the 1000 simulated values was 8.6488 and the RoA mean was estimated as 8.6179. The 95% bootstrap confidence interval for the difference between these two means is (−0.0386, 0.0920) which includes zero, and show there is no significant difference. In Figs. 1 and 2 histograms are shown of the AoR's and RoA's together with the fitted log–logistic distributions. The estimated parameters of the log–logistic distribution for the AoR's are $\hat{\alpha} = 4.4489$, $\hat{\beta} = 7.9282$ and for the RoA's, $\hat{\alpha} = 4.6696$, $\hat{\beta} = 7.9780$ (Fig. 3).

4. An application to country citation ranks

Scimago (2007) give research outputs of countries for the period 1996–2011. The results are given in terms of totals, thus total number of citable documents and total number of citations. If the totals are used to approximate the average number of citations per researcher, the result is given in Table 4, showing that some of the smaller countries with respect to population produce high quality researcher if measured by citations. These averages are calculated over all subject fields.

The sample sizes on which the results were calculated is very large and as can be seen, these ratios of averages give a reliable ranking, keeping in mind that these are random variables.

Table 4
Ranking of top 10 countries in terms of average citations per researcher together with ranking in terms of research output.

Rank with respect to research output	Country	Citations per document
17	Switzerland	22.46
24	Denmark	21.17
14	Netherlands	20.82
1	United States	20.51
18	Sweden	19.78
25	Finland	18.28
7	Canada	18.19
3	United Kingdom	18.03
21	Belgium	17.81

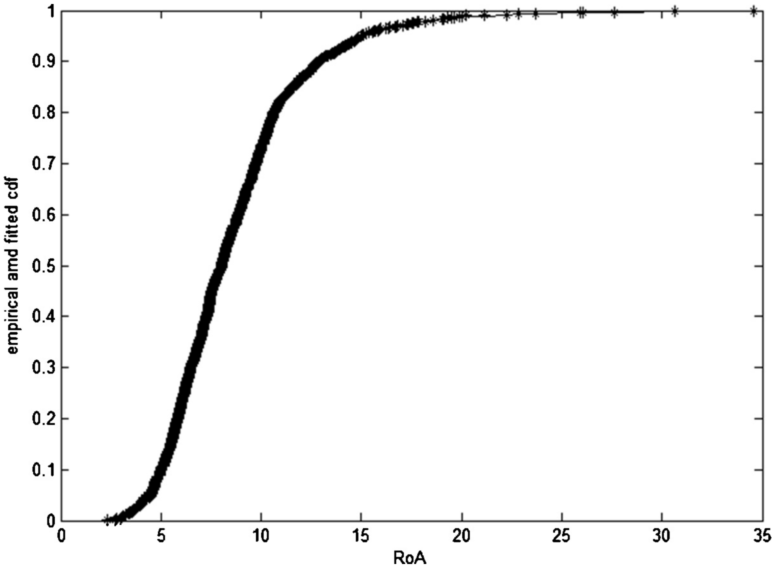

Fig. 3. Empirical and fitted log–logistic distribution function for 1000 RoA's, $n = 50$ researchers and $\gamma = 3.5$.

Table 5
Summary results of $m = 1000$ simulations for the correlation between citations/publications and publications.

$\gamma = 3.5$	Mean correlation	Variance	Minimum	Maximum
$n = 25$	0.0142	0.0263	−0.2550	0.7945
$n = 50$	0.0095	0.0127	−0.1825	0.7885
$n = 250$	0.0018	0.0015	−0.0780	0.2228
$n = 1000$	0.0013	0.0003	−0.0426	0.1096

5. Conclusions

A simulation study was conducted and by using resampling methods (i.e., bootstrap) confidence intervals were obtained showing that the difference between the average of ratios and the ratio of average is not statistically significant. It was found that there are no significant difference between the estimated AoR's and RoA's calculated in the simulation study, since the confidence intervals covered zero in all cases considered. The confidence intervals were constructed by using bootstrap methods and thus no distributional assumptions were made.

It should be kept in mind that the ratios are random variables and even if one mean is less than the other theoretically, if the means are close the smaller mean could yield a larger sample mean for a specific sample. With this restriction kept in mind, it can be seen that the use of RoA is a very good approximation of AoR.

Based on the simulation study, it can be seen that the use of summary data and RoA's to approximate averages of individual ratios, is quite reliable even in small samples, for the number of publications and citations generated according to the laws governing the distribution of publications and citations.

This is a specific model and approach based on reasonable assumptions and one can expect very similar results if publications and citations follow approximately the same type of distributions. Within reasonable bound one can assume that similar patterns would be found in other sets of data, although the parameters of the distribution of citations and number of publications might be different.

References

Arnold, B. C. (1983). *Pareto distributions*. International Co-operative Publishing House: MD.
Egghe, L. (2012). Averages of ratios compared to ratios of averages: Mathematical results. *Journal of Informetrics*, *6*, 307–317.
Gibbons, J. D., & Chakraborti, S. (2011). *Nonparametric statistical inference*. Boca Raton: Chapman & Hall/CRC.
Goldstein, M. L., Morris, S. A., & Yen, G. G. (2004). Problems with fitting to the power-law distribution. *European Physical Journal B*, *41*, 255–258.
Johnson, N. L., Kemp, A. W., & Kotz, S. (2005). *Univariate discrete distributions* (3rd ed.). Hoboken, NJ: Wiley.
Kleiber, C., & Kotz, S. (2003). *Statistical size distributions in economics and actuarial sciences*. New Jersey: John Wiley & Sons.
Larivière, & Gingras. (2011). Averages of ratios vs. ratios of averages: An empirical analysis of four levels of aggregation. *Journal of Informetrics*, *5*, 392–399.
Limbert, E., Stahel, W., & Abt, W. (2001). Log–normal distributions across the sciences: Keys and clues. *BioScience*, *51*(5), 341–352.
Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences*, *16*(12), 317–324.
Mardia, K. V. (1962). Multivariate Pareto distributions. *The Annals of Mathematical Statistics*, *33*(3), 1008–1015.
Markovich, N. M. (2009). Nonparametric estimation of copula with application to Web traffic data. In *Proceedings of the VIII International conference System identification and control problems SICPRO-09, Selected papers* Moscow, January 26–30, 2009, (pp. 35–43).
Newman, M. E. J. (2005). Power laws, Pareto distributions and Zipf's law. *Contemporary Physics*, *46*(5), 323–351.

Opthof, T., & Leydesdorff, L. (2011). Caveats for the journal and field normalization in the CWTS (Leiden) evaluations for research performance. *Journal of Informetrics*, *4*(3), 423–430.

Peterson, G. J., Pressè, S., & Dill, K. A. (2010). Nonuniversal power law scaling in the probability distribution of scientific citations. *PNAS*, *107*(37), 16023–16027.

Radicchi, F., Fortunato, S., & Castellano, C. (2008). Universality of citation distributions: Toward an objective measure of scientific impact. *PNAS*, *105*(45), 7268–17272.

Redner, S. (1998). How popular is your paper? An empirical study of the citation distribution. *European Physical Journal B*, *4*, 131–134.

Redner, S. (2005). Citation statistics from 110 years of physical review. *Physics Today*, *58*(6), 49–54.

SCImago. (2007). *SJR—SCImago Journal & Country Rank*. Retrieved from:. www.scimagojr.com

Van Zyl, J. M., & van der Merwe, S. (2012). *An empirical study to order citation statistics between subject fields*. arXiv:1210.2246v2

Waltman, L., van Eck, N. J., van Leeuwen, T. N., Visser, M. S., & van Raan, A. F. J. (2011). Towards a new crown indicator: Some theoretical considerations. *Journal of Informetrics*, *5*(1), 37–47.