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