Physica A 482 (2017) 173–180

Contents lists available at ScienceDirect

Physica A

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

Impact factor distribution revisited

Ding-wei Huang

Department of Physics, Chung Yuan Christian University, Chung-li, Taiwan

HIGHLIGHTS

- Consistent frequency distribution is obtained.
- The goodness-of-fit is evaluated by chi-square value.
- A bell-shaped distribution is restored by a log transformation.
- The tail of distribution can be well described.

ARTICLE INFO

Article history:
Received 9 August 2016
Received in revised form 15 February 2017
Available online 22 April 2017

Keywords:
Journal impact factor
Citation dynamics
Lavalette distribution
Rank distribution
Frequency distribution
Log transformation

ABSTRACT

We explore the consistency of a new type of frequency distribution, where the corresponding rank distribution is Lavalette distribution. Empirical data of journal impact factors can be well described. This distribution is distinct from Poisson distribution and negative binomial distribution, which were suggested by previous study. By a log transformation, we obtain a bell-shaped distribution, which is then compared to Gaussian and catenary curves. Possible mechanisms behind the shape of impact factor distribution are suggested.

1. Introduction

For complex systems, distributions of relevant variables often reflect the underlying basic dynamics. Traditional physic has brought us to understand various kinds of distributions in nature phenomena. In the realm of sociophysics, the physics perspective is extended to social phenomena [1]. Different aspects of empirical data are often demonstrated by different distributions. When the system is complicated and the data are not accurate enough, different distributions might be used without being aware of their inconsistency. This work aims to study the consistency of typical distributions in the citation dynamics.

In the system of academic publishing, a scholarly journal is often ranked by its impact factor. The impact factor is a simple measure of the average number of citations to recent articles published in the journal. With the emphasis on peer opinions, citations reflect the impact, i.e., how often an article being cited by other scholars. The impact factor becomes a convenient measure of the journal's importance in the research field. The rank distribution of impact factors is widely used to compare journals, even with some weaknesses [2]. In many complex systems involving human activities, the rank distribution follows a simple power law with a negative exponent, i.e., the Zipf's law. The rank distribution of impact factors does not conform to the Zipf's law. The power law behavior can only be observed in a limited range in the lower ranks, i.e., the highly cited journals. A cut-off appears in the higher ranks. A recent modification known as the Lavalette distribution gave a satisfactory

E-mail address: dwhuang@cycu.edu.tw.

http://dx.doi.org/10.1016/j.physa.2017.04.038
0378-4371/

description of the empirical data [3,4]. Later, the two-exponent distribution had been proposed as a further modification [5,6]. Data from both science (SCI) and social science (SSCI) journals can be fairly described by the same formulation.

Ever since the impact factor was proposed in 1970s, the distribution has been a focus of study in bibliometrics and scientometrics. Besides the rank distribution, the frequency distribution is also widely used. The frequency distribution of journal impact factors has been utilized for the practical statistical research in library and information science [7]. Related to the Zipf's law, a power law in the frequency distribution is known as the Lotka's law. For the impact factor distribution, significant deviations from the Lotka's law are observed. The distribution is not monotonic but has a maximum. The distribution is positively skewed, SCI journals much more so than SSCI journals. The skewed distribution causes some inconvenience in applying the standard statistical methods, which are often based on the assumption of a normal distribution. Previous study indicated that SCI and SSCI journals follow different formulations: SCI journals follow the negative binomial distribution and SSCI journals follow the Poisson distribution [8].

Rank distribution and frequency distribution are correlated. The relation had been used to investigate the detailed shape of the distribution. An S-shaped rank distribution was related to a bell-shaped frequency distribution [9]. However, the bell-shaped curve cannot be justified in the empirical data of the frequency distribution [10]. Thus, the existence of the S-shaped curve caused a controversy in recent investigations of the rank distribution [11]. Later, a more general relationship between rank distribution and frequency distribution has been investigated [12]. We notice that neither the Poisson distribution nor the negative binomial distribution in the frequency distribution are consistent with the Lavalette distribution or the two-exponent distribution in the rank distribution. In this work, we propose a consistent frequency-distribution. We introduce a variable transformation to restore a non-skewed frequency-distribution. We also evaluate the goodness-of-fit of various formulations. Possible mechanisms behind the shape of impact factor distribution are suggested.

2. Lavalette distribution

The Lavalette distribution is given as

$$g(y) = k\left(\frac{N+1-y}{y}\right)^a, \tag{1}$$

where $k > 0$ and $a > 0$ are two parameters to control the shape of the rank distribution. Besides the application in scientometrics, this distribution has also been applied to other rank distributions [13,14]. Various data can be well fitted by the formula, yet the underlying mechanisms are seldom provided. The corresponding frequency distribution becomes

$$f(x) = \frac{k^{\frac{1}{a}}\, x^{\frac{1}{a}-1}}{a\left(x^{\frac{1}{a}} + k^{\frac{1}{a}}\right)^2}. \tag{2}$$

The relation between a rank distribution and a frequency distribution is summarized in the Appendix. When $a > 1$, Eq. (2) becomes a monotonically decreasing distribution. When $0 < a < 1$, the frequency distribution has a single peak. The median value locates at $x = k$. The most probable value locates at $x = k\left(\frac{1-a}{1+a}\right)^a$, which is less than the median value. The average value locates at $x = k\left(\frac{\pi a}{\sin \pi a}\right)$, which is larger than the median value. The distribution is skewed and distinctly different from the Poisson distribution and the negative binomial distribution. The power law behavior can be observed both in the small x and in the large x. When $x \ll 1$, the distribution increases as $x^{+(1-a)/a}$; when $x \gg 1$, the distribution decreases as $x^{-(1+a)/a}$. The impact factor distribution can be fairly described. Typical results are shown in Fig. 1, where data are taken from 2011 Journal Citation Report (JCR) published by Thomson Reuters. The database includes more than 10^4 scholarly journals, which are divided into 232 subject categories. Fig. 1 shows the typical distributions in 16 subject categories. The first row shows the SCI journals with high impact-factor; the second row shows the SCI journals with low impact-factor; the third row shows the SSCI journals with high impact-factor; the fourth row shows the SSCI journals with low impact-factor.

When the numerator and the denominator of Eq. (1) assume different exponents, the Lavalette distribution is extended to the two-exponent distribution as following,

$$g(y) = k\,\frac{(N+1-y)^b}{y^a}. \tag{3}$$

The two exponents b and a dictate the power law behavior of the frequency distribution in the small x and in the large x, respectively. When $x \ll 1$, the distribution increases as $x^{+(1-b)/b}$; when $x \gg 1$, the distribution decreases as $x^{-(1+a)/a}$. Compared to the Lavalette distribution, the two-exponent distribution has one more free parameter. The fitting of empirical data improves slightly, as shown by the dotted lines in Fig. 1.

To compare the goodness of fit among different distributions, we use the standard χ^2 value defined as following

$$\chi^2 = \frac{1}{n}\sum_{i=1}^{n}\frac{[f(x_i) - f_i]^2}{f_i}, \tag{4}$$