Journal of Informetrics 12 (2018) 1072–1088

Contents lists available at ScienceDirect

Journal of Informetrics

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

Regular article

# Impact Factors and the Central Limit Theorem: Why citation averages are scale dependent☆

Manolis Antonoyiannakis[a,b,*]

[a] Department of Applied Physics & Applied Mathematics, Columbia University, 500 W. 120th St., Mudd 200, New York, NY 10027, United States
[b] American Physical Society, Editorial Office, 1 Research Road, Ridge, NY 11961-2701, United States

ARTICLE INFO

*Article history:*
Received 4 April 2018
Received in revised form 25 August 2018
Accepted 25 August 2018

*Keywords:*
Science of Science
Scholarly Publishing
Impact Factors
Journal size
Central Limit Theorem

ABSTRACT

Citation averages, and Impact Factors (IFs) in particular, are sensitive to sample size. Here, we apply the *Central Limit Theorem* to IFs to understand their scale-dependent behavior. For a journal of $n$ randomly selected papers from a population of all papers, we expect from the Theorem that its IF fluctuates around the population average $\mu$, and spans a range of values proportional to $\sigma/\sqrt{n}$, where $\sigma^2$ is the variance of the population's citation distribution. The $1/\sqrt{n}$ dependence has profound implications for IF rankings: The larger a journal, the narrower the range around $\mu$ where its IF lies. IF rankings therefore allocate an unfair advantage to smaller journals in the high IF ranks, and to larger journals in the low IF ranks. As a result, we expect a scale-dependent stratification of journals in IF rankings, whereby small journals occupy the top, middle, *and* bottom ranks; mid-sized journals occupy the middle ranks; and very large journals have IFs that asymptotically approach $\mu$. We obtain qualitative and quantitative confirmation of these predictions by analyzing (i) the complete set of 166,498 IF & journal-size data pairs in the 1997–2016 Journal Citation Reports of Clarivate Analytics, (ii) the top-cited portion of 276,000 physics papers published in 2014–2015, and (iii) the citation distributions of an arbitrarily sampled list of physics journals. We conclude that the Central Limit Theorem is a good predictor of the IF range of actual journals, while sustained deviations from its predictions are a mark of true, non-random, citation impact. IF rankings are thus misleading unless one compares like-sized journals or adjusts for these effects. We propose the $\Phi$ index, a rescaled IF that accounts for size effects, and which can be readily generalized to account also for different citation practices across research fields. Our methodology applies to other citation averages that are used to compare research fields, university departments or countries in various types of rankings.


## 1. Introduction

What do crime rates, cancer rates, high-school mean test scores, and Impact Factors have in common? They are all manifestations of the Central Limit Theorem, which explains why small populations (cities, schools, or research journals) score more often than one would expect at the top *and* bottom places of rankings, while large populations end up in less

☆ *Disclaimer*: The author is an Associate Editor at the American Physical Society. The opinions expressed here are his own.
* Correspondence to: Department of Applied Physics & Applied Mathematics, Columbia University, 500 W. 120th St., Mudd 200, New York, NY 10027, United States.
*E-mail address:* ma2529@columbia.edu

https://doi.org/10.1016/j.joi.2018.08.011
1751-1577/

remarkable positions. But if size affects one's position in a ranking, then rankings of population averages must be misleading. The Impact Factor is an average measure of the citation impact of journals. Therefore, it may seem perfectly justifiable to use it when ranking journals of different sizes, in the same vein we use averages to rank, say, the class size of schools, the GPA's of students, the fuel efficiency of engines, the life expectancy in countries, or the GDP per capita for various countries. However, underlying such comparisons is the tacit admission (De Veaux et al., 2014) that the distributions being compared are (approximately) symmetric and do not contain outliers (i.e., extreme values)—or if they do, that the sample sizes are large enough to absorb extreme values. If the distributions are highly skewed, with outliers, and especially if the populations are small, then rankings by averages can be misleading, because averages are no longer representative of the distributions. Impact Factors qualify for these caveats. So far, several studies drew attention to the skewness of the citation distribution, or various other features of the Impact Factor, such as the 'free' citations to front-matter items of journals, the need to normalize for different citation practices among fields, the citation time windows, the lack of verifiability in the citation counts entering the Impact Factor calculations, the mixing of document types with disparate citabilities (articles versus reviews), etc. (Adler et al., 2008; Antonoyiannakis, 2015a, 2015b; Bornmann and Leydesdorff, 2017; Fersht, 2009; Glänzel and Moed, 2013; Radicchi et al., 2008; Redner, 1998; Rossner et al., 2007; San Francisco Declaration on Research Assessment, 2012; Seglen, 1992, 1997; Wall, 2009). However, little attention has been paid (Amin and Mabe, 2004; Antonoyiannakis and Mitra, 2009) to the effect of journal scale on Impact Factors, which, as we will show, is substantial.

The Impact Factor is defined as

$$\mathrm{IF} = \frac{C}{N_{2Y}} = \frac{\sum_{i=1}^{i=N_{2Y}} c_i}{N_{2Y}}, \tag{1}$$

where $C$ are the citations received in year $y$ to journal content published in years $y-1, y-2$, and $N_{2Y}$ is the biennial publication count, i.e., the number of citable items (articles and reviews) published in years $y-1, y-2$. As can be verified from the Journal Citation Reports (JCR) of Clarivate Analytics, the annual publication count of journals ranges from a few papers to a few tens of thousands of papers. At the same time, individual papers can collect from zero to a few thousand citations in the JCR year. With a span of 4 orders of magnitude in the numerator and 5 orders of magnitude in the denominator, the IF is a quantity with considerable room for wiggle.

In this paper, *first*, we apply the Central Limit Theorem (the celebrated theorem of statistics) to understand and predict the behavior of Impact Factors. We find that Impact Factor rankings produce a scale-dependent stratification of journals, as follows. (a) Small journals occupy all ranks (top, middle *and* bottom); (b) mid-sized journals occupy the middle ranks; and (c) very large journals ("megajournals") converge to a single Impact Factor value—the population mean—almost irrespective of their size. Impact Factors are thus sensitive to journal size, and Impact Factor rankings do not provide a 'level playing field,' because size affects a journal's chances to make it in the top, middle, or bottom ranks. *Second*, we apply the Central Limit Theorem to arrive at an *Impact Factor uncertainty relation:* an expression that limits the expected range of Impact Factor values for a journal as a function of journal size and the citation variance of the population of all published papers. *Third*, we confirm our theoretical results, by analyzing 166,498 IF & journal-size data pairs, the citation-distribution data from 276,000 physics papers, and an arbitrarily sampled list of physics journals. We observe the predicted scale-dependent stratification of journals. We find that the Impact Factor uncertainty relation is a very good predictor of the range of Impact Factors observed in actual journals. And *fourth*, we argue that sustained deviation from the expected IF range is a mark of non-random citation impact. We thus propose to normalize IFs with regard to the theoretically expected maximum at a given size (using appropriate offsets), as a scale-independent index of citation impact.

Why does all this matter? Because statistically problematic comparisons can lead to misguided decisions, and Impact Factor rankings remain in wide use (and abuse) today (Gaind, 2018; Stephan et al., 2017).

Our analysis shows that Impact Factor comparisons—even for similar fields and document types—for different-sized journals can be misleading. We argue that it is imperative to a seek metrics that are immune from or correct for this effect.

## 2. Theoretical background

### 2.1. The Central Limit Theorem for citation averages (i.e., Impact Factors)

The Central Limit Theorem is the fundamental theorem of statistics. In a nutshell, it says that for independent and identically distributed data whose variance is finite, the sampling distribution of any mean becomes more nearly normal (i.e., Gaussian) as the sample size grows (De Veaux et al., 2014). The sample mean $\bar{x}_n$ will then approach the population mean $\mu$, *in distribution*. More formally,

$$\lim_{n\to\infty}\left(\sqrt{n}\left(\frac{\bar{x}_n-\mu}{\sigma}\right)\right) \stackrel{d}{=} N(0,1) \tag{2}$$

whence

$$\sigma_n = \frac{\sigma}{\sqrt{n}}, \tag{3}$$