Impact Factors and the Central Limit Theorem

Manolis Antonoyiannakis[a,b]

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

Abstract

In rankings by average metrics, smaller samples are more volatile: They can fluctuate to higher or lower scores more easily than larger samples, which are more stable. The range of these fluctuations depends on two factors: The disparity (variance) of values in the wider population, and sample size. We have used the celebrated Central Limit Theorem (CLT) of statistics to understand the behavior of citation averages (Impact Factors). We find that Impact Factors are strongly dependent on journal size. We explain the observed stratification in Impact Factor rankings, whereby small journals occupy the top, middle, *and* bottom ranks; mid-sized journals occupy the middle ranks; and very large journals converge to a single Impact Factor value. Further, we applied the CLT to develop an 'uncertainty relation' for Impact Factors, which provides an upper ($f^{th}_{max}(n)$) and lower bound for a journal's Impact Factor given its size, n. We confirm the functional form of $f^{th}_{max}(n)$ by analyzing the complete set of 166,498 journals in the 1997–2016 Journal Citation Reports (JCR) of Clarivate Analytics, the top-cited portion of 345,177 papers published in 2014–2015 in physics, as well as the citation distributions of an arbitrarily sampled list of journals. We conclude that the Impact Factor 'uncertainty relation' is a very good predictor of the range of Impact Factors observed for actual journals. Because the size-dependent effects are strong, Impact Factor rankings can be misleading, unless one compares like-sized journals or adjusts for these effects.

Keywords: Science of Science, Scholarly Publishing, Impact Factors, Journal Size, Central Limit Theorem

1. Introduction

Can rankings of population averages be misleading? The Journal Impact Factor (JIF) 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 GPAs of students, the fuel efficiency of engines, the life expectancy of citizens in countries, or the GDP per capita for various countries. However, underlying such comparisons is the tacit admission [1] 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,

Email address: ma2529@columbia.edu ()

and especially if the populations are small, then rankings by averages can be misleading, because averages are no longer representative of the distributions. Journal Impact Factors qualify for these caveats. So far, several studies drew attention to the skewness of the 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. [2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14]. However, little attention has been paid [15, 16] to the effect of journal scale on Impact Factors, which, as we will show, is substantial.

The Journal Impact Factor is defined as

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

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 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 Impact Factor 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 depending on its size, a journal has different chances to make it in the top, middle, or bottom ranks. *Second*, we apply the CLT 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 papers published. *Third*, we confirm our theoretical results, by analyzing Impact-Factor and journal-size data from 166,498 journals, as well as citation-distribution data from 345,177 physics papers and also from an arbitrarily sampled list of 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.

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 [17, 18]. Our analysis shows that Impact Factor rankings—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 (CLT) for citation averages (i.e., Journal Impact Factors)

The Central Limit Theorem (CLT) 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 [1]. The sample mean $\bar{x}_n$ will then approach the population mean μ, *in distribution*. More formally,

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

whence

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

where $N(0,1)$ is the normal distribution and the symbol "d" in the equality means *in distribution*. σ_n is the standard deviation of a sampling distribution, σ is the standard deviation of the entire population we wish to study (and which is often not known), and n the sample size. So, sample means vary less than individual measurements. The square of a standard deviation if called the *variance*.

The sampling distribution is a notional (imaginary) distribution from a very large number of samples, each one of size n, which approaches a normal distribution in the limit of large n. In practice, the CLT holds for n as low as 30, unless there are exceptional circumstances—e.g., when the population distribution is highly skewed—in which case higher values are needed. So, σ_n measures how widely the sample means of size n vary around the the population mean μ (which is approached in the limit of large n), while σ measures how widely the population values vary around μ.

Let us now examine the two quantities (σ and n) on which the sample standard deviation (σ_n) depends in Eq. (3).

2.1.1. Variance effects (dependence on σ)

Equation (3) shows that the sample variance is proportional to the population variance. High variance (i.e., variability, disparity of values) in the population causes high variance in the sample. This makes sense. For example, imagine that the world's richest and tallest persons simultaneously move into a neighborhood of a population of 1000 people. Because income disparity (variance) among the population is far greater than height disparity, we would expect the income means (averages) of various random samples drawn from the population to vary more (have higher variance) than height means. Note that citation 'wealth' is very unevenly distributed, like monetary wealth.

For populations of scientific research papers, the citation distributions have high variances (disparities), because the individual papers can be cited from 0 to a few thousand times. Therefore citation means (impact factors) will have a much higher variance at a given sample size, compared to, say, the height means for adults.

It is the high variance (σ) of citations in populations of scientific papers that makes the CLT highly relevant in Impact Factor rankings. Had σ been 100 times smaller for citation distributions, none of the effects described in this paper would be seen—they would be there, of course, but they would be too small to be of relevance and would not interfere with rankings of average metrics. Thus, the multiplier σ in the numerator of Eq. (3) acts as a 'switch' that turns on the size effects of the denominator.

2.1.2. Size effects (dependence on n)

The inverse square root dependence of Eq. (3) with sample size n means that for small journals (small n), the impact factors can fluctuate widely around the population mean μ. Thus,

Impact Factor	Journal Size
High	Small
Moderate	Small — Medium
Average	Small — Medium — Large
Below average	Small — Medium
Low	Small

Table 1: Stratification of journals in Impact Factor rankings are sensitive to journal size.

for small journals we expect to see a wide range of impact factors, from very low to very large values. Small journals will thus dominate the high ranks of impact factor values, but also the low ranks! Actually, small journals will cover the entire range of Impact Factor values. With increasing n to medium-sized journals, the fluctuation σ_n of impact factors around the population mean μ decreases. Therefore, mid-sized journals will not be able to achieve as high impact factors as small journals but they will be spared from really low values too. So, mid-sized journals will do better than small journals in the low ranks but worse in the top ranks. Finally, for large journal sizes n, the fluctuation of sample means around the population mean μ is small, so all impact factors of large journals will asymptotically approach μ. Therefore, very large journals have no chance at all to populate even the middle ranks of Impact Factors; however, they will be ranked higher than many small (and a few mid-sized) journals.

We can codify the above discussion in a simple conceptual diagram. For simplicity, let us use three size classifications as follows. We classify journals with biennial publication count $n \leq 2000$ as 'small'; journals with $2000 < n \leq 10000$ as 'mid-sized'; and journals with $n > 10000$ as 'large'. We would then expect the scale-dependent stratification of journals, in terms of Impact Factor, that is shown in Table 1. By the way, such stratification effects have been reported for other average metrics—e.g., crime statistics, school performances, cancer rates, etc.—and explained in terms of the Central Limit Theorem [19, 20, 21].

2.1.3. Why is the Central Limit Theorem relevant for Impact Factors?

Journal sizes range typically from $100 - 100,000$ (biennial count, n), so the quantity $1/\sqrt{n}$ ranges from $10^{-1} - 10^{-5}$. Had the population standard deviation σ been roughly equal or less than 1, say, then impact factor fluctuations (strictly, their standard deviation σ_n) would be less than 0.1 and thus irrelevant for impact factor rankings (except for very-low Impact Factors). But if σ is of the order of 100, then σ_n lies in the range $0.3 - 10$, and impact factors are affected significantly, because random fluctuations (σ_n) are no longer insignificant compared to impact factors themselves. This is why the CLT is relevant here.

2.2. An 'uncertainty principle' for Impact Factors

Consider the population of citations in a certain year to all papers published in the previous two years. Imagine that we draw random samples ("journals") of size n from this population, and calculate their citation average, f_n, which for practical purposes is equal to the Impact factor of the n papers (apart from for the 'free' citations in the numerator). Because the sampling distribution of the sample means is normal (i.e., Gaussian), we can expect roughly 99% of f_n values to lie within $3\sigma_n$ of μ. So, for practical purposes, we can write

$$\mu - 3\sigma_n \leq f_n \leq \mu + 3\sigma_n. \tag{4}$$

We invoke the Central Limit Theorem, Eq. (3), to rewrite the above inequality as

$$\mu - 3\frac{\sigma}{\sqrt{n}} \leq f_n \leq \mu + 3\frac{\sigma}{\sqrt{n}}, \tag{5}$$

or even as

$$\Delta f_n \cdot \sqrt{n} \leq 3\sigma, \text{ where } \Delta f_n \equiv |f_n - \mu|. \tag{6}$$

The expression (5) or (6) can be regarded as an *uncertainty principle* for Impact Factors f_n as a function of the journal (biennial) size n, the population mean μ, and the population standard deviation σ. It says that the 'uncertainty' Δf_n (i.e., range of values of f_n) multiplied by the square root of size cannot exceed 3σ, statistically speaking. Therefore, for small journals, Δf_n is large, while for large journals, Δf_n has to be small.

While the expression (6) holds in a statistical sense—roughly in 99% cases for a distance of $3\sigma_n$ from μ—it can of course be made more precise by increasing the distance from μ to $4\sigma_n$, $5\sigma_n$, etc.

The uncertainty principle for Impact Factors has important practical implications, as we discuss below.

1. Expression (5) says that Impact Factor uncertainties—i.e., Δf_n mean μ—have a maximum value $3\sigma/\sqrt{n}$ that is inversely proportional to the square root of journal size. For small n sizes $3\sigma/\sqrt{n}$ is large and μ can be dropped, so the Impact Factor itself is inversely proportional to $\sqrt{n}$. This scale dependence is rather punitive for large journals: A 100-fold increase in journal size yields a 10-fold decrease in how high the impact factor *can* be as measured from μ.

2. The impact factor maximum increases with the population standard deviation σ, which is a measure of the disparity (variability) of citations among all papers in the population. But what is the population? We have assumed so far that it consists of all papers in all research fields, and this statement is true in a general sense. However, for research fields that do not cite each other (or do so with low intensity) one can claim that they are distinct populations, each with its own σ. In this case, expression (5) says that journals from the population with larger σ can reach higher impact factors. Indeed, normalizing citation averages to account for the different citedness of research fields is a standard practice in bibliometrics.

3. For large enough n there is an impact factor minimum, equal to $\mu - 3\frac{\sigma}{\sqrt{n}}$ (f_n has to be nonnegative, of course). That is, f_n is bounded *from below*.

Let us take a more detailed look at the Impact Factor uncertainty principle (5) for two limits of interest.

Case I. Small journals. If there are values σ, μ such that $n \ll 9\sigma^2/\mu^2$, then μ can be left out and Eq. (5) simplifies to

$$0 \leq f_n \leq 3\frac{\sigma}{\sqrt{n}}, \quad \text{for} \quad \mu \ll \sigma/\sqrt{n}. \tag{7}$$

So, for small journals the impact factor can range from 0 to a maximum value that is inversely proportional to $\sqrt{n}$, which can become quite large for small enough size. In other words, small journals are highly volatile, and they will populate all positions in impact factor ranks, from the lowest to the highest.

Case II. Very large journals, i.e., $n \gg 9\sigma^2/\mu^2$.

Here, expression (5) reduces to

$$\mu - \delta \leq f_n \leq \mu + \delta, \quad \text{where } \delta = \frac{3\sigma}{\sqrt{n}} \ll 1, \tag{8}$$

and the Impact Factor asymptotically approaches the population mean, μ. This is both good and bad news for very large journals in impact factor rankings: They will neither populate the low ranks nor the high ranks. These journals sample the population, so to speak, so they are stable and insensitive to size effects. Their impact factors are bounded from above and below.

3. Materials and Methods

3.1. Approximating $N_{2Y} \approx 2N_Y$ for easier data retrieval

We collected data on Journal Impact Factors and citable items N_Y from Clarivate Analytics Journal Citation Reports (JCR), in the 20-year period 1997–2016. The citable items (N_Y) data refer to the JCR year: They are the sum of articles and reviews published by a journal in that year. From the original data, we removed those journals whose Impact Factors or citable items were listed as either non-available or zero, as well as duplicate entries. A total of 166,498 journals were thus obtained. The N_Y values range from 1 to 31,496, while the Impact Factor values range from 0.027 to 187.04.

For the purposes of this paper we need data on the JIF and its denominator, N_{2Y}, the biennial publication count in the two years prior to the JCR year. A practical difficulty arises here. While the JCR list JIF values and yearly publication counts (N_Y) in the JCR year, they do not list N_{2Y} values. To obtain N_{2Y} data we must check each journal individually in the Web of Science—a conceptually trivial but nevertheless cumbersome procedure for tens of thousands of journals. However, it is reasonable to assume that the publication count does not change appreciably over the 3-year window spanned by N_Y and N_{2Y}, and write

$$N_{2Y} \approx 2N_Y. \tag{9}$$

If the approximation (9) holds for all journals, we would be justified to use $2N_Y$ data as a substitute for N_{2Y}. For a qualitative analysis, it would suffice that N_{2Y} be no greater (or smaller) than a few times the product $2N_Y$. That is, it would suffice that (9) be valid for all journals up to a multiplicative constant, so that

$$N_{2Y} = n\ 2N_Y, \text{where } 0.1 \lessapprox n \lessapprox 10. \tag{10}$$

We test Eq. (9) for the 2016 JCR year and the 8710 journals in the Science Citation Index Expanded (SCIE) list. As we can see from Fig. 1, $2N_Y$ and N_{2Y} are strongly correlated (slope = 0.96, R^2=0.82, Pearson correlation coefficient = 0.90). Of the 8710 journals, all but 5 (or 99.94%) satisfy Eq. (10), while even for the 5 remaining journals, n remains small ($n < 18$). Therefore, we are justified to use the approximation $N_{2Y} \approx 2N_Y$, provided we are interested in the broad, overall relationship of Impact Factors with journal size. But when we analyze *individual* journals, especially with respect to each other (as in *ranking*), then we must use N_{2Y}. Certainly, the only reason we may prefer to use $2N_Y$ instead of N_{2Y} is the ease of data retrieval from JCR, but where and when necessary, the value N_{2Y} should be used.

4. Data Analysis

4.1. A boundary curve for Impact Factors from citation distribution data

We have looked at the 345,177 physics papers published in 2014–2015, and cited in 2016, in the Web of Science Core Collection. We calculated the total citations, $C_{max}(n)$, as a function of

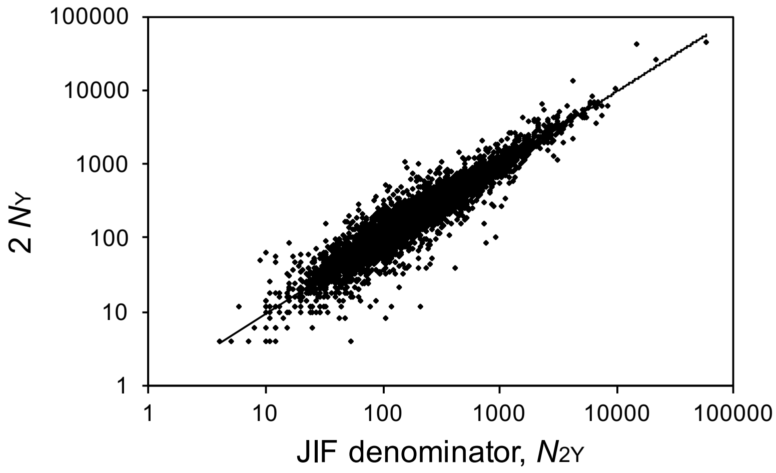

Figure 1: How good is the approximation of Eq. (9)? We test this from the 2016 JCR SCIE data. The biennial publication count in the two years prior to the JCR year (N_{2Y}), is plotted against twice the annual publication count in the JCR year ($2N_Y$).

decreasing citation rank, n, for the 2089 papers cited at least 30 times in 2016. The dependence of $C_{max}(n)$ on n is found to be (see Fig. 2, inset)

$$C_{max}(n) = \lambda\, n^{0.55},\ \ 1 \le n \le 2089, \tag{11}$$

where $\lambda = 1774$ is a scale factor in the order of the number of citations received by the most cited paper in this distribution, which in this case is $c_{max} = 2121$. Note that $C_{max}(n)$ grows much slower than linearly with n. The impact factor $f^{fit}_{ph,max}(n)$ is defined as the ratio $C_{max}(n)/n$. (We denote results obtained from data fitting with a 'fit' superscript, as opposed to results from theory 'th'. For actual Impact Factor data no superscript is used.) So we can write

$$f^{fit}_{ph,max}(n) = \lambda\, n^{-\alpha}, \tag{12}$$

with $\alpha = 0.45$ and $\lambda = 1774$ here. Clearly, $C_{max}(n)$ grows less than linearly with n, which results in a size-dependent upper bound for $f^{fit}_{ph,max}(n)$. See Fig. 2.

The fact that Eq. (12) has been deduced from 'only' the top 2089 papers should not distract us from recognizing the generality of the conclusion: Equation (12) agrees well with the impact factor uncertainly relation (7). It shows that Impact Factor rankings strongly favor small journals, since $f^{fit}_{ph,max}(n)$ drops abruptly with n. In fact, had we continued the analysis to lower-ranked papers, the exponent in Eq. (11) would have surely decreased, since the $C_{max}(n)$ curve would cave downward to account for lower-cited papers; consequently, the α value in Eq. (12) would approach 0.5.

To further explore the generality of Eq. (12), we looked at the citation distributions from an arbitrarily sampled list of 15 journals—namely, *CA-A Cancer Journal for Clinicians, Nature, Nat. Physics, Cell, Appl. Microb. & Biotech., New J. Phys., Eur. J. Phys., Phys. Rev. X, Phys. Rev. Lett., Phys. Rev. A, Phys. Rev. B, Phys. Rev. C, Phys. Rev. D, Phys. Rev. E*, and *Rev. Mod. Phys.* For these 15 journals, Eq. (12) holds with α ranging from 0.44 to 0.85 (the 0.85 value was

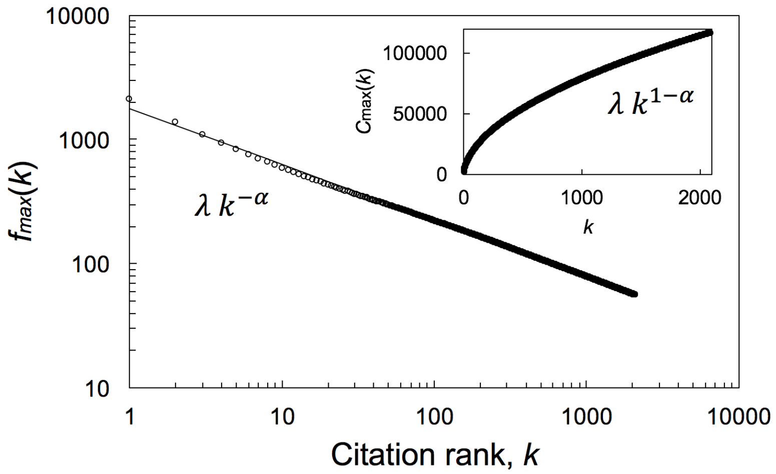

Figure 2: Impact Factor, $f^{fit}_{ph,max}(n)$, of the n top-cited papers among the 345,177 papers published in physics in 2014–2015, versus citation rank, n. Inset: Total citations of n top-cited papers versus n. Here, λ = 1774 and α = 0.45.

an extreme outlier, corresponding to the highest Impact Factor value ever recorded, namely, for *CA-A Cancer Journal for Clinicians*). Most α values were in the 0.45–0.55 range, with a median of 0.50, and a mean 0.53 ± 0.03. As for the coefficient, λ, it is typically in the order of the most cited paper in the distribution ($\lambda/c_{max} \approx$ 1–5).

4.2. A boundary curve for Impact Factors from Impact-Factor & journal-size data

We have analyzed all 166,498 journals with nonzero values for the Impact Factor and number of annual citable items (N_Y) in the Clarivate Analytics Journal Citation Reports in the 1997–2016 period. Before we proceed, let us present two important features of Impact Factors and journal sizes, which may not be widely known.

4.2.1. Small journals are extremely common

In Fig. 3 we plot the frequency distribution of journals vs. their annual size, i.e., the number of citable items (articles and reviews) published in the JCR year, N_Y. Small journals are extremely common. The most common journal size is 24 citable items per year. 50% of all journals publish 60 or fewer citable items per year, while 90% of all journals publish 250 or fewer citable items annually.

4.2.2. Most journals have small Impact Factors

In Fig. 4 we plot the frequency distribution of journals vs. their Impact Factors, f. As is evident from the figure, most Impact Factors are quite small: The most commonly occurring value is 0.5. In the range $0.5 < f < 7$, which covers 85% of all journals, the frequency distribution can be approximated by an exponentially decreasing function (see dotted line).

We are now ready to analyze Impact-Factor and journal-size data to obtain a boundary curve for Impact Factors.

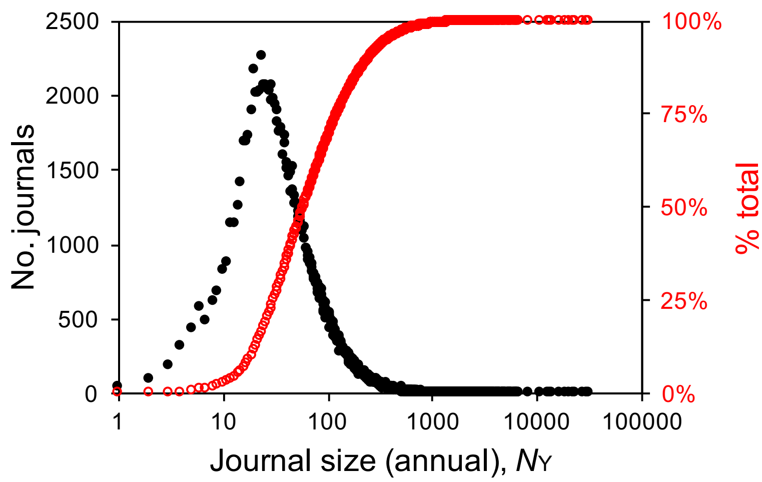

Figure 3: Frequency (filled dots) and percentage (hollow dots) of journals with annual publication count N_Y. Data for 166,498 journals in the 1997–2016 JCR.

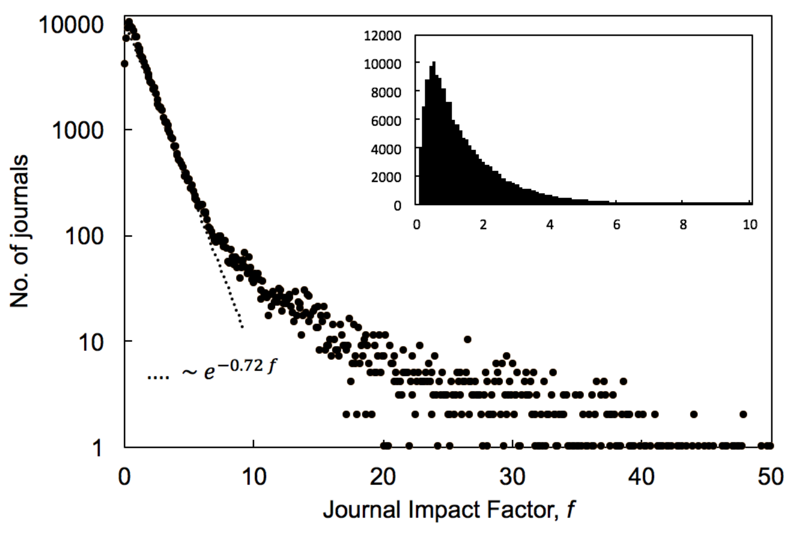

Figure 4: Frequency of journals with Impact Factor f. The dotted line is an exponential fit, which is valid ($R^2 = 0.99$) in the range $0.5 < f < 7$. Inset: Same data but in the range $0 < f < 10$, plotted as a histogram in linear scale. Data for 166,498 journals in the 1997–2016 JCR. The bin width of f values used in the distribution is 0.1.

4.2.3. Impact Factor vs. journal size

In Fig. 5, we plot the Impact Factor versus (annual) journal size for all 166,498 journals in our set. A glance at the figure confirms the penalizing effect of journal size that we described earlier. We observe a global (i.e., large-scale) trend whereby large journals cannot have high impact factors: Higher JIF values tend to occur for smaller ($N_Y < 1000$) than larger journals. In broad terms, we observe that of the 166,498 journals, (a) no journal with $N_Y > 2000$ has a JIF > 20; (b) no journal with $N_Y > 1000$ has a JIF > 40; (c) no journal with $N_Y > 500$ has a JIF > 80; etc. As we zoom in at smaller scales, we notice local irregularities, most notably three local peaks (groups of high-JIF data points) centered at around $N_Y = 25, 350,$ and 900 (see Fig. 5, left inset). The first peak ($N_Y = 25$) results from very small and selective journals that publish a few mega-cited papers, most notably *CA-A Cancer Journal for Clinicians*. The second peak ($N_Y = 350$) results from highly selective monodisciplinary journals such as the *New England Journal of Medicine*, *Lancet*, *Chemical Reviews*, *Journal of the American Medical Association*, *Cell*, *Nature Reviews Molecular Cell Biology*, *Nature Materials*, *Nature Nanotechnology*, etc. And the third peak ($N_Y = 900$) is due to highly selective multidisciplinary journals, such as *Nature* and *Science*.

We predicted earlier (§2.1.2) a scale-dependent stratification of journals, in terms of Impact Factor, and this is what we observe. For example, if we rank the 11765 journals in the 2016 JCR, we find that all the top 50 ranks are occupied by small journals. Out of the top 100 ranks, 99 are occupied by small journals and 1 by a mid-sized journal. Among the top 500 ranks, 94.8% are occupied by small journals, and 5.2% by mid-sized journals. Among the bottom 500 ranks, all 100% are small journals. Among the middle 5000 ranks, we find 99.3% occupied by small and 0.7% by mid-sized journals. Finally, three megajournals ($N_Y > 10,000$) occupy the ranks 1058, 2130, and 2592.

4.2.4. The Maximum Impact Factor scales as the inverse square root of journal size

Let us now focus on the maximum Impact Factor value at any given journal size for the data in Fig. 5. First, we plot (dotted line) the theoretical impact factor maximum from (5),

$$f^{th}_{max}(n) = \mu + 3\frac{\sigma}{\sqrt{n}}, \quad \text{where} \;\; n = 2N_Y,\ \mu = 2.5,\ \sigma = 100. \tag{13}$$

Why did we choose $\mu = 2.5$ and $\sigma = 100$? First, the population mean μ can be estimated rather accurately from the Journal Citation Reports (JCR), by summing up all Impact Factor numerators and dividing by all Impact Factor denominators. Doing so for the 1997–2016 JCR we obtain $\mu = 2.5$. Of course, we expect a slight annual increase of μ with inflation, and indeed for the years 2014, 2015, and 2016, we obtain $\mu = 3.00, 3.07,$ and 3.18, respectively. Estimating the population standard deviation σ is a little trickier. We performed a random selection of 5 samples (500 papers each) from all papers published in 2015–2016, and cited in 2017, and obtained $\sigma \simeq 100$. So we will use this value here. It is certainly not in discord with our experience from analyzing journal citation datasets. We note that the $\sigma \simeq 100$ estimate is consistent with the observed large-journal behavior of impact factors (Fig. 5, inset), which seem to become stable for $N_Y > 10,000$. Had σ been 10 times larger (say, $\sigma \simeq 1000$), then $f^{th}_{max}(n) \sim 14.7$ for $n = 60,000$ and there would be considerable variation in JIF values for the megajournals in the inset.

As can be seen from the figure, the theoretical curve $f^{th}_{max}(n)$ captures the general trend of the data of impact factor maxima, f_{max}. Next, we extract the dependence of $f_{max}(n)$ on N_Y from data

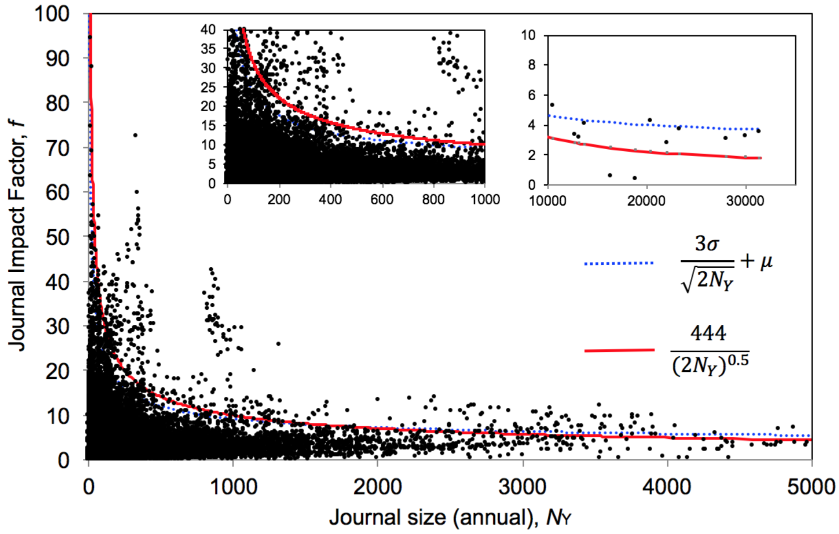

Figure 5: JIF values versus annual journal size N_Y. 166,498 points shown, corresponding to all journals with a nonzero JIF and N_Y from 1997–2016. Data from Journal Citation Reports, Clarivate Analytics. Left inset: Detail of main plot, for $N_Y \leq 1000$. Right inset: Extended N_Y values, to 35000. The dotted line is the theoretical maximum by the Central Limit Theorem $f^{th}_{max}(n)$, see Eq. (13). The solid line is the fitted maximum from the binned data, $f^{fit}_{max}(n)$ of Eq. (14).

(recall $N_Y = n/2$). So we bin the journal size data of Fig. 5 in groups (bins) of 10 citable items each (N_Y =1–10, 11–20, etc.). For the journals in each bin, we calculate f_{max} and plot it against N_Y in Fig. 6 (filled dots). The global downward trend is clearly visible: The journal size, N_Y, has an adverse effect on f_{max}. Also shown in Fig. 6 and in Fig. 5 is a best-fit curve calculated from the binned data

$$f^{fit}_{max}(n) = 444 \cdot n^{-0.5}, \; n = 2N_Y \tag{14}$$

The details of binning have some effect on the exponent of Eq. (14). Clearly, the bin size should not be too large, as it can affect the effect we are trying to measure, which is size-dependent.

Since the expression (14) has an exponent of 0.5, we plot the product ($f_{max} \cdot \sqrt{2N_Y}$) in Fig. 6. Evidently, this product is independent of N_Y (horizontal dotted line), which means that its variance is size-independent (as opposed to the variance of impact factors).

Equation (14) confirms the uncertainly relation (5) from impact factor data, just like Eq. (12) did from citation data from physics papers. Furthermore, it is clear from Fig. 5 that expression (5) reduces to (7) for journal (annual) sizes of $N_Y \leq 5000$, since $f^{th}_{max}(n)$ and $f^{fit}_{max}(n)$ practically coincide in this range. However, for larger journals ($N_Y \geq 10000$) the data agrees better with $f^{th}_{max}(n)$ (see inset of Fig. 5), while $f^{fit}_{max}(n)$ is clearly an underestimate of the observed values. Indeed, for these journal sizes we are in the realm of practically stable impact factors (Case II of §2.2 where the sample means approach the population mean μ). Note, incidentally, the two very low values of JIF, equal to 0.5 and 0.4 for $N_Y \simeq 16,000$ and $N_Y \simeq 19,000$ respectively. These data points belong to the 2004 and 2005 JIF values for *Lecture Notes in Computer Science*. Because this is a specialized venue that publishes conference proceedings, it corresponds to a distinct population with different citation features (different σ) than other large journals— see point 2 in §2.2. Indeed, this journal is now classified in the Web of Science as Book Series

Journal	N_{2Y}	f_{max}^{th}	$(f_{max}^{th} - \mu) \cdot \sqrt{N_{2Y}}$
J1	100	32.5	300
J2	1000	12	300
J3	10,000	5.5	300

Table 2: Maximum impact factors expected f_{max}^{th} for three journals of different (biennial) size N_{2Y}.

and does not receive an impact factor value. So it is a special case and does not invalidate the uncertainly relation (5) or, more specifically, $f_{max}^{th}(n)$.

Once again, we observe that the frequency distribution of actual citation averages (impact factor data, $f_{max}(n)$) agrees with the (notional) sampling distribution from the Central Limit Theorem, $f_{max}^{th}(n)$.

While Eqs. (12), (13) and (14) have been 'derived' by different means, they all point to a similar power-law dependence of $f_{max}^{th}(n)$ with n. We have thus identified, for the 166,498 journals in our set, a global boundary curve for Impact Factors as a function of journal size, in the form of Eq. (13), with the understanding that we could make it more statistically accurate by increasing the factor 3 to, say, 5. The only question is the value of the parameter σ in Eq. (13). We will use the value $\sigma \sim 100$ which is consistent with our experience. Of course, σ does not need to stay constant in time, in fact it is reasonable to expect that it increases every year, as the citation disparity among papers grows.

The implications of the inverse-square-root-of-size constraint of Eq. (13) on impact factors are remarkable. First, the very notion of a constraint implies bias: If two athletes are subjected to different constraints for how high they can score, then surely we cannot speak of a level playing field. Second, the rapidly decaying form of $1/\sqrt{n}$ means that the bias is strongest for small journals, which can thus reach very high impact factors. This effect is much stronger than has been previously reported [15, 6]. It was anticipated though not fully analyzed in [16, 22]. But another (weaker) bias works also in favor of very large journals, since their impact factors are bounded *from below*. See Table 2 for some examples that elucidate these behaviors.

5. Conclusions and outlook

In this paper, we have used the Central Limit Theorem (CLT) of statistics to understand the behavior of citation averages (Journal Impact Factors, JIF). We find that citation averages are strongly dependent on sample (journal) size. We explain the observed stratification of journals in Impact Factor rankings, whereby small journals occupy the top, middle, *and* bottom ranks; mid-sized journals occupy the middle ranks; and very large ("mega") journals converge to a single Impact Factor value. Further, we applied the CLT to develop an 'uncertainty principle' for Impact Factors, which provides an upper ($f_{max}^{th}(n)$) and lower bound for a journal's JIF given its size, n. We confirm the functional form of $f_{max}^{th}(n)$ by analyzing the complete set of 166,498 journals in the 1997–2016 Journal Citation Reports (JCR) of Clarivate Analytics, the top-cited portion of 345,177 papers published in 2014–2015 in physics, as well as the citation distributions of an arbitrarily sampled list of journals.

The direct implication of our work is that, because the scale-dependent effects are strong, Impact Factor rankings are misleading. It is thus strongly suggested to compare like-sized journals. If one *must* compare different-sized journals, it would be better to use the ratio JIF/$f_{max}^{th}(n)$.

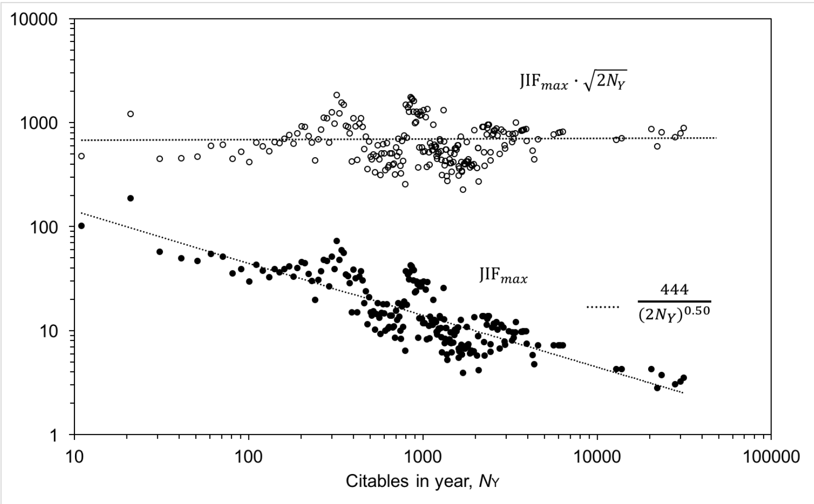

Figure 6: Maximum Impact Factor data, f_{max} (filled dots) for each journal size, N_Y, in the Journal Citation Reports from 1997–2016. The data is drawn from Fig. 5, with binning as described in the text. The dotted line (best fit, $R^2 = 0.67$) is Eq. (14). Also shown (hollow dots) is the product (Impact Factor) $\cdot \sqrt{2N_Y}$, which, clearly, does not depend on journal size.

Thus, contrary to what one might expect and what is largely accepted in bibliometrics[23], the process of averaging does not guarantee size independence. But then, what is the point in using averages? Should we avoid them altogether? Of course not. When there is low disparity (variance) within the population, then averages are fine. But in citation wealth, as in actual wealth, the vast disparity within the population makes averages misleading.

One could conceive analyzing the uncertainty relation Eq. (5) by research fields, which would result in field-specific (upper and lower) bounds for Impact Factors. It would be also interesting to study the effect of citation inflation, which we have ignored here, on μ and σ for the citations population of research papers, even though it is reasonable to expect that this would be rather small [24].

Acknowledgments: I am grateful to Jerry I. Dadap (Columbia University), Albyn Jones (Reed College), and Amine Triki (Clarivate Analytics) for stimulating discussions, and to Richard Osgood Jr. (Columbia University) for encouragement and hospitality.

Disclaimer: The author is an Associate Editor at the American Physical Society. The opinions expressed here are his own.

References

References

[1] R. D. De Veaux, P. D. Velleman, D. E. Bock, Stats: Data and Models (3rd Edition), Pearson Education Limited, 2014.

[2] P. O. Seglen, The skewness of science, J. Am. Soc. Inf. Sci. 43 (1992) 628638.

[3] P. O. Seglen, Why the Impact Factor of journals should not be used for evaluating research, British Medical Journal 314 (1997) 498502.
[4] S. Redner, How popular is your paper? An empirical study of the citation distribution, European Physical Journal B 4 (1998) 131134.
[5] M. Rossner, H. Van Epps, E. Hill, Show me the data, Journal of Cell Biology 179 (2007) 10911092.
[6] R. Adler, J. Ewing, P. Taylor, Citation Statistics, Joint Committee on Quantitative Assessment of Research, Available at http://www.mathunion.org/fileadmin/IMU/Report/CitationStatistics.pdf, 2008.
[7] F. Radicchi, S. Fortunato, C. Castellano, Universality of citation distributions: Toward an objective measure of scientific impact, Proc. Natl. Acad, Sci. USA 105 (2008) 1726817272.
[8] H. J. Wall, Don't get skewed over by journal rankings, The B.E. Journal of Economic Analysis & Policy 9 (2009) article 34.
[9] A. Fersht, The most influential journals: Impact Factor and Eigenfactor, Proc. Natl. Acad. Sci. USA 106 (2009) 68836884.
[10] W. Glänzel, H. F. Moed, Opinion paper: Thoughts and facts on bibliometric indicators, Scientometrics 96 (2013) 381394.
[11] M. Antonoyiannakis, Impact Facts & Impact Factors: A view from the Physical Review. PQE - The 45th Winter Colloquium on the Physics of Quantum Electronics, Available at http://www.pqeconference.com/pqe2015/Abstracts/345p.pdf, 2015.
[12] M. Antonoyiannakis, Median Citation Index vs Journal Impact Factor. APS March Meeting, Available at http://meetings.aps.org/link/BAPS.2015.MAR.Y11.14, 2015.
[13] San Francisco Declaration on Research Assessment,Available at http://www.ascb.org/dora/, 2012.
[14] L. Bornmann, L. Leydesdorff, Skewness of citation impact data and covariates of citation distributions: A large-scale empirical analysis based on Web of Science data, Journal of Informetrics 11 (2017) 164175.
[15] M. Amin, M. Mabe, Impact factors: Use and abuse, International Journal of Environmental Science and Technology 1 (2004) 16.
[16] M. Antonoyiannakis, S. Mitra, Editorial: Is PRL too large to have an 'impact?, Physical Review Letters 102 (2009) 060001.
[17] N. Gaind, Few UK universities have adopted rules against impact-factor abuse, Nature News, Available at https://www.nature.com/articles/d41586-018-01874-w, 2018.
[18] P. Stephan, R. Veugelers, J. Wang, Reviewers are blinkered by bibliometrics, Nature 544 (2017) 411412.
[19] H. Wainer, H. Zwerling, Legal and empirical evidence that smaller schools do not improve student achievement, The Phi Delta Kappan 87 (2006) 300303.
[20] H. Wainer, The most dangerous equation, American Scientist May-June (2007) 249256.
[21] A. Gelman, D. Nolan, Teaching Statistics: A Bag of Tricks, Oxford University Press, 2002.
[22] M. Antonoyiannakis, Citation analysis: Beyond the Journal Impact Factor. APS March Meeting, Available at http://meetings.aps.org/link/BAPS.2008.MAR.U39.8, 2008.
[23] L. Waltman, A review of the literature on citation impact indicators, Journal of Informetrics 10 (2016) 365391.
[24] B. M. Althouse, J. D. West, C. T. Bergstrom, T. Bergstrom, Differences in Impact Factor across fields and over time, J. Amer. Soc. Inf. Sci. & Tech. 60 (2009) 2734.