Lotka’s Bibliometric Inverse-Power Law: Calibrating the Model and Exploring a Common Conceptual Error

David Rand Irvin

27 July 2024

Abstract: Two forms of Lotka’s Law appear in the literature. The first, which is the traditional form, holds that the number of authors who publish exactly *N* first-author papers in primary, technical journals is proportional to $1/N^2$. The second holds that the number of authors who publish at least *N* first-author papers is proportional to $1/N$, based on pseudo-integration of the first. We show that the two forms are not equivalent, despite assertions claiming that they are. We also show that the median contribution of authors who publish in primary journals is one paper, but that the theoretical average number of papers-per-author is undefined.

Introduction

According to Yale University’s historian of science, Derek J. De Solla Price in his classic work *Little Science, Big Science*,[1] Lotka’s Law

> *“. . . is an inverse-square law of productivity. The number of people producing N papers is proportional to $1/N^2$. For every 100 authors who produce but a single paper in a certain period, there are 25 with two, 11 with three, and so on. Putting it a little differently by permitting the results to accumulate, one achieves an integration that gives approximately an inverse-first-power law for the number of people who produce more than* [sic] *N papers; thus, about one in five authors produces five papers or more, and one in ten produces at least ten papers.”*

Price’s characterization has two shortcomings. The first may be simply a transcription error which eluded his editor at Columbia University Press -- the phrase “who produce more than *N* papers” should read “who produce *N*-or-more papers.” More importantly, however, although Price is careful to use the word “approximately,” his text leaves the impression that Lotka’s Law, which has the proportionality of $1/N^2$, can be equivalently restated as an integration that has the resulting proportionality $1/N$.

Although Price may well have understood the subtleties involved, the current Wikipedia entry[2] for Lotka’s Law makes a glaring error regarding equivalence. The entry states, quite incorrectly, that

> “*Equivalently, Lotka’s Law can be stated as* $Y' \propto X^{-(k-1)}$, *where* Y’ *is the number of authors with at least* X *publications. Their equivalence can be proved by taking the derivative.*”

Claiming equivalence between the two versions of Lotka's Law is incorrect, as demonstrated below. Also incorrect is the notion that equivalence of the two versions can be "proved" by taking the derivative of a function of this kind, wherein the domain of the function is the set of positive integers.

First, to help establish an intuitive foothold, the calibration of Lotka's Law is discussed below. This is followed by a demonstration that the $1/N^2$ form of Lotka's Law is not in agreement with its supposed equivalent having the $1/N$ form. Thus, the two forms are not equivalent; the Wikipedia "proof" involving differentiation is nonsense. Attention then turns to median and mean values, the difference between the two forms of Lotka's Law, and concluding remarks.

Calibrating Lotka's Law

Lotka[3] observed empirically that the number of authors who publish primary-journal papers in STEM fields approximates an inverse-square relationship: the number of authors contributing n first-author papers varies inversely as n-squared. A more general statement of Lotka's observation is:

$$y_n = c / n^a \qquad <1>$$

where y_n is the frequency of occurrence of authors with n contributions; c and a are constants that depend on the specific discipline. In many disciplines, $c = 1$; in the traditional form of Lotka's Law, $a = 2$.

To establish a sense of scale, let $c = 1$, and let K_s be the number of authors within a sample who contribute exactly one first-author paper each, wherein everyone in the sample contributes at least one first-author paper. Let T_s be the total number of authors in the sample that includes K_s. We then have:

$$T_s = K_s\,(y_1 + y_2 + y_3 + y_4 + \ldots)\,,$$

which becomes, with the substitution defined by <1> ,

$$T_s = K_s\,(1 + 1/2^a + 1/3^a + 1/4^a + \ldots)\,,$$

or, more conveniently,

$$T_s = K_s\,\Sigma\,(1/n^a),\ \text{sum from } n = 1 \text{ to } \infty\,. \qquad <2>$$

This infinite series factor of <2> is the Riemann Zeta function, $\zeta(a)$, which converges for all real $a > 1$. For the infinite series with $a = 2$, finding the sum is traditionally known as the "Basel problem." Leonhard Euler, in 1735, was the first to show that the infinite sum converges to $\pi^2/6$, i.e.,

$$\Sigma\,(1/n^2) = \pi^2/6\,.$$

Thus, for a = 2,

$$T_s = K_s \pi^2/6 . \qquad <3>$$

Media and Mean Values

For example, consider a sample wherein 100 authors each publish only one first-author paper: i.e., let K_{100} = 100 (with a = 2, as first proposed by Lotka). The total number of authors T_s in the sample is then given by:

$$T_{100} = 100\pi^2/6 ,$$

which is approximately 164.

With this scaling, 64 of the 164 authors contribute more than one first-author paper. Of these 64, twenty-five should have two papers each to their credit ($100/2^2$), about eleven should have three papers ($100/3^2$), one should have 10 papers ($100/10^2$), and so on.

Another way to look at this is to note that 100/164, or about 60.8%, of the entire sample of authors contribute only one first-author paper each. Thus, the median contribution is one first-author paper, per author. This holds true for the more-general case as well, since, by inverting *<3>,*

$$K_s/T_s = 6/\pi^2,$$

or approximately 0.608.

Although the median number of papers per author is easily shown to be one, as just demonstrated, the average number of papers per author cannot be found. Because the number of authors producing exactly n papers is given by K_s/n^2, the number of papers produced by such authors is $(n)(K_s/n^2) = K_s/n$. Thus, the total number of papers produced by all of the authors of the sample is given by:

$$K_s \Sigma (1/n), \text{ sum from } n = 1 \text{ to } \infty,$$

which does not converge.

The Two Statements of Lotka's Law are not Equivalent

As mentioned above, a common misinterpretation of Lotka's Law asserts that an accumulation or integration of the $1/n^2$ distribution leads to an equivalent first-power inverse relationship: the number of authors contributing n-or-more papers varies as one-over-n. The assertion of equivalence can be shown to be incorrect by deriving the number of authors who produce **exactly** n papers from the assumption that $1/n$ produce **at least** n papers, as follows:

Let:

Q_n = the fraction of authors contributing exactly *n* first-author papers under the assumption that Q*n* = 1/*n*^{2}. This is Lotka's observation.

R_n = the fraction of authors contributing *n*-or-more papers under the assumption that $R_n = 1/n$. This is the supposed equivalent.

If the fraction of authors contributing at least *n* papers is given by

$$R_n = 1/n,$$

then the fraction of authors contributing at least *n* + 1 papers is given by

$$R_{n+1} = 1/ (n + 1).$$

From this, the fraction of authors contributing exactly *n* papers, which has been called here Q_n, is given by

$$Q_n = R_n – R_{n+1} ,$$

which, with a bit of algebra, reduces to

$$R_n – R_{n+1} = 1/n – [1/ (n + 1)] = . . . = 1 / (n^2 + n) .$$

Under Lotka's original formulation, however, the fraction of authors contributing exactly *n* papers is given by

$$Q_n = 1/n^2 .$$

Clearly,

$$1/n^2 \neq 1 / (n^2 + n).$$

Hence, the formulation based on the number of authors producing at least *n* papers is not consistent with the accepted form of Lotka's Law. Therefore, the two formulations are not equivalent.

Quantifying the Discrepancy

To evaluate the discrepancy between the two formulations, let Δ be their difference:

$$\Delta = 1/n^2 - [1 / (n^2 + n)] ,$$

which can be reduced to:

$$\Delta = 1 / (n^3 + n^2) ,$$

again wherein n is a positive integer.

Thus, the difference Δ is largest for small values of n, as might be the case in realistic applications. Clearly, $\Delta > 0$, and the limit of Δ is 0 as n approaches ∞. Thus, the estimate of the number of authors who produce exactly n papers, as derived from premise R (i.e., the supposed equivalent found by pseudo-integration), is bounded monotonically from above by the estimate of the same, given premise Q (i.e., Lotka's original observation). The two converge for large values of n.

Concluding Remarks

Lotka's Law itself is only an approximation devised by fitting a curve to noisy empirical data. The supposedly equivalent form of Lotka's Law, which is characterized by 1/*n* proportionality, is a not-so-good approximation of Lotka's approximation. All told, Lotka's Law, and to a greater extent its supposed equivalent, might be thought of as more metaphor than mathematics.

[1] Derek J. De Solla Price presents Lotka's Law gives a number of refinements in *Little Science, Big Science*; Columbia University Press, New York, 1963; Chapter 1.

[2] "Lotka's Law," *Wikipedia, The Free Encyclopedia*; https://en.wikipedia.org/wiki/Lotka%27s_law (22 July 2024)

[3] Lotka, A. J., "The frequency distribution of scientific productivity," *Journal of the Washington Academy of Sciences*, vol. 16 (1926), p. 317