Journal of Scientometric Research, 2024; 13(2):466-474.
https://www.jscires.org

Research Article

An Analytical Study of Alternative Method for Solving Lotka's Law with Simpson's 1/3 Rule

Anindya Basu[1], Bidyarthi Dutta[2,*]

[1]Department of Library, Maharani Kasiswari College, Kolkata, West Bengal, INDIA.
[2]Department of Library of Information Science, Vidyasagar University, Midnapore, West Bengal, INDIA.

ABSTRACT

This paper deals with a new stochastic method to solve Lotka's Law with higher degree of Newton-Cotes Quadrature Rule as an alternate method to existing solution of the value determination of the constant part of the power law; so far, M.L. Pao gave a solution with an equation to determine the area under the curve with numerical integration rule with degree=1 which is also known as Trapezoidal Rule. Here, next higher degree 2, popularly known as Simpson's 1/3 rule at closed interval [x_1,] has been used to establish a deterministic equation form to solve authors' productivity realized through Lotka's Law. Re-estimating the value of C with higher degree quadrature rule is very crucial as the probability of inclusion of more area and exclusion of unnecessary area under the curve is more precise. Another area of investigation is the determination of *p* value (Pao determined *p*=20), i.e. whether *p*=20 can be altered? Or equation derived through Simpson's 1/3 rule, whether it can give a minimal residual error beyond *p*=20. This paper is dedicated to build up a mathematical equation to solve the constant value(C) of the Lotka's law equation as well as enlighten all these investigating points.

Keywords: Lotka's Law, Trapezoidal Rule, Simpson's 1/3 Rule, Simpson's 3/8 Rule, Pao Method

Correspondence:
Bidyarthi Dutta
Department of Library of Information Science, Vidyasagar University, Midnapore, West Bengal, INDIA.
Email: bidyarthi.bhaswati@gmail.com
ORCID: 0000-0001-6049-5445

Received: 18-11-2023;
Revised: 29-03-2024;
Accepted: 14-04-2024.

INTRODUCTION

Fundamentals of Lotka's Law

In last century, researches in the field of Scientometrics was dominated by three empirical power laws: Bradford's Law, Lotka's Law and Zipf's Law and ultimately, they become classical laws in this subject and created a whirlpool of researches in last 7 decades. All of them are essentially behaving like Power Law and their forms can also be derived to interchange when they are put to representative forms under generalized framework. This article discusses the essential existing solutions of Lotka's Law[1] when applied to datasets and highlights other possible solutions.

Lotka's law, also known as law of authors' productivity, states that-the number of authors making contribution with x papers at certain time interval is inversely proportional to the number of authors making single contribution in their lifetime.

$$\varphi(x) = \frac{c}{x^n} \text{- } \mathbf{1}$$

Here, x denotes number of papers, $\varphi(x)$ denotes number of authors, C is a constant value and n is the exponent value of the variable x. According to Lotka, n is perceived to be around 2, if n=2, the equation becomes a perfect inverse square law. In other words, function $\varphi(x)$ is a power law function representing concentration of sources (here author) with items (papers) and x ϵ N {x≥1}. The form of Riemann-Zeta function is

$$\boldsymbol{\zeta(n)} = \sum_{x=1}^{\infty} \frac{1}{x^n} = \frac{1}{\tau(n)} \int_1^{\infty} \frac{x^{n-1}}{e^x - 1} \boldsymbol{dx} \text{ [}\boldsymbol{\tau(n)}\text{ Indicates Gamma Function] -2}$$

Expanding the summated form to integer numbers, it becomes,

$$\boldsymbol{\zeta(n)} = \sum_{x=1}^{\infty} \frac{1}{x^n} = \frac{1}{1^n} + \frac{1}{2^n} + \frac{1}{3^n} + \frac{1}{4^n} + \cdots \infty \text{ -3}$$

Thus, when n=2, it becomes a special case which is called as inverse-square law which invites Lotka's Law to fit in. Lotka's Law is a special case of Riemann-Zeta Function. Eqn. 2 is only solvable when n=2 and 4 only. The solution is

$$\zeta(2) = \sum_{x=1}^{\infty} \frac{1}{x^n} = \frac{1}{1^2} + \frac{1}{2^2} + \frac{1}{3^2} + \frac{1}{4^2} + \cdots \infty = \frac{\pi^2}{6} \text{ and}$$

$$\zeta(4) = \sum_{x=1}^{\infty} \frac{1}{x^n} = \frac{1}{1^4} + \frac{1}{2^4} + \frac{1}{3^4} + \frac{1}{4^4} + \cdots \infty = \frac{\pi^4}{90} \text{ - 4}$$

Lee Pao's Method

The main procedure of solving Lotka's function is deriving the value of the constant (C) as well as the exponent value (n). Exponent value is calculated using MLE (Maximum Likelihood Estimator) method. There is no room for doubting the efficacy of MLE while calculating the value of the exponent. Calculation of

DOI: 10.5530/jscires.13.2.37

Publishing Partner : EManuscript Tech. [www.emanuscript.in]

the constant is always very problematic. Essentially the constant C can be derived by the following equation form

$$C = \frac{1}{1/(\sum_{x=1}^{\infty} x^n)} \quad \text{-5}$$

C is the inverse of the summation of the series. The value of C is depending on the particular domain/subject, age of the subject, collaborating behaviour among researchers, research aptitude of individual scientist and many other factors.

Calculation of the exponent 'n'

The value of the exponent actually reflects many factors, on the reverse, we can hypothesize-there are several factors which actually governs the value of the exponent. The exponent value is calculated using simple linear least square estimator. The value of n can be found with –

$$\mathrm{n} = \frac{\mathrm{N}\sum \mathrm{XY} - \sum \mathrm{X} \sum \mathrm{Y}}{\mathrm{N}\sum \mathrm{X}^2 - (\sum \mathrm{X})^2} \quad \text{- 6}$$

Here, N=Number of Data Points as pair of x and y values, X= Logarithmic Transformation of x and Y=Logarithmic Transformation of y.

Calculation of constant 'C'

The crux of the problem of solving Lotka's law is the accuracy of fitting the observed data into the equation with the calculated numerical values. As stated above, if n is 2 or 4, solving equation-1 is pretty straightforward. But in real world, distribution of data rarely shows such sharp exponent value and there is no specific method to estimate such things. As stated by Pao,[2] dividing both sides of equation-1 with total number of authors,

$$\frac{\varphi(\mathrm{x})}{\sum \varphi} = \frac{\mathrm{c}/\sum \varphi}{\mathrm{x}^{\mathrm{n}}} \quad \text{-7}$$

Now, let F(x)= and C= is the new constant. So, equation-7 can be rewritten as –

$$\mathrm{F(x)=C.}\frac{1}{x^n} \quad \text{- 8}$$

Eqn.8 is another identical form of the original Lotka's equation as written in Eqn-1. Putting the numerical values in equation 1-

$$\varphi_1 = \frac{c}{1^n}, \varphi_2 = \frac{c}{2^n}, \varphi_3 = c/3^n, \varphi_4 = \frac{c}{4^n}\ldots\ldots\ldots\varphi_x = \frac{c}{x^n}$$

Summing each term –

$$\sum_1^x \varphi_x = c\left(\frac{1}{1^n} + \frac{1}{2} + \frac{1}{3^n} + \frac{1}{4^n} + \cdots + \frac{1}{x^n}\right) \quad \text{-9}$$

Dividing both sides with total number of authors

$$\sum_1^x \varphi_x / \sum_1^x \varphi_x = \frac{c}{\sum_1^x \varphi_x}\left(\frac{1}{1^n} + \frac{1}{2} + \frac{1}{3^n} + \frac{1}{4^n} + \cdots + \frac{1}{x^n}\right) \quad \text{-10}$$

$$1 = C * \left(\sum_1^x \frac{1}{x^n}\right) \quad \text{- 11}$$

The new constant $C = \frac{c}{\sum_1^x \varphi_x}$

$$C = \frac{1}{\sum_1^x \frac{1}{x^n}} \quad \text{- 12}$$

When $n=2$, $\sum_1^x \frac{1}{x^n} = \frac{\pi^2}{6}$ So, $C = \frac{6}{\pi^2} = 0.6074380$ - 13

LITERATURE REVIEW

The researches done so far on Lotka's Law can be segregated into two broad sub-divisions-a. Development of Methods to solve it, building mathematical relation among Bradford, Lotka and Zipf's law; b. Application of the law in different subjects to understand authors' productivity. As Lotka's law shows a power law form, there have been continuous efforts to fit different standard distributions on authors' productivity in different disciplines. Murphy[3] applied Lotka's law in Humanities discipline using inverse square law and the law did fit on the dataset. Hersh[4] analyzed authors' productivity on researches on Drosophila assuming a simple power law function: $y=bx^k$ and he stated that, prediction on the behaviour of exponential curve is risky due to uncertainty in reaching at the point of inflection at any moment. Sen[5] used a simple method to calculate constant value C by putting the first value of number of paper published (here x=1); he also calculated constant value C with Lee Pao Method.[2] His method showed C= 1.485 and Pao method showed the constant C to be 1.682 and he also showed his method by and large follows Sen's Method of deriving C (constant value). Empirical analysis made on the available standard datasets using inverse power model of Lotka's Law ($Y_x=kx^{-b}$) and estimated constant K using Pao's method and exponent with Maximum Likelihood Estimator (MLE) and subsequently K-S Test was done in order to test the goodness of fit.[6] Bookstein[7] made an important analysis on the intrinsic behaviour of classic bibliometric distributions as shown in Bradford's Law, Lotka's Law and Zipf's law. Bookstein concluded that, "Lotka's law is not sensitive to how we count articles, so that two people testing the law for a single population, but different count methods, will very likely to come up with the same law".[7] Krisciunas[8] made a short communication to the editor explaining his study on randomly chosen sample and he observed that, Lotka's law is biased towards the distribution of relatively more prolific authors and he also concluded that, the dataset must be prepared taking authors' publication over a long time period; then only Lotka's law can be best fitted on the dataset. Krisciunas fitted exponential distribution form over his collected data, but it did not decrease as fast as Lotka's distribution.[8] In another study,[9] Brookes proved that, all the empirical laws get reduced to simple form of hyperbolic law and its probability density function revolves around k/x^2 and operates upon certain intervals of X-axis. Bradford's law shows identical behaviour with other similar empirical laws (Lotka, Price and Zipf).[9] In another paper,[2] Pao used general inverse power law $x^n y=c$ and estimated the exponent value using linear least-square method. Constant C was approximated by numerical integration method using trapezoidal rule. As compared with the actual value of $\pi^2/6$, Pao's method produced error less than 1/110,000; for $\pi^4/90$, the error is less than 1/25,000,000. In another research paper,[6] Nicholls validated Lotka's law on 15 classical datasets after ground-breaking work of Pao[2] and he proposed two modification on Pao method. First proposal of modification was while calculating the exponent value, the data need to be truncated and Maximum Likelihood Estimator should be solved by numerical iterative method provided by Johnson and Kotz.[10] Second suggestion for modification was to also include multi-authorship fractional credit counts.[6] Nicholls[11] investigated the validity of Price's Law against the back-drop of the Lotka's law and he tried to prove consistency with respect to its theoretical and empirical behaviour between the two. But, the value of the constant k is always dependent on the exponent value b of the distribution and empirically on X_{max} (No. of Papers produced by most prolific authors) is not infinite and also doesn't follow a limiting value; in some datasets-number of authors with single publication vary considerably across different subjects and this actually disobeys Price's conjecture due to problem in indicating cut-off point between prolific and non-prolific authors. His observation was "the validity of the Price law need not depend on that of Lotka's law, the Price law is seen to be inconsistent with the generalized Lotka model. The Price law does not agree with empirical data very well; empirical results do not support the Price hypothesis. Since the empirical validity of Lotka's law has recently been more firmly established, it is not surprising that the empirical and theoretical findings are consistent".[11] In another study by Nicholls,[12] Pao method was applied on 70 datasets and his observation was, 90% of the cases followed generalized Lotka's method. In a ground-breaking study by Bailón-Moreno *et al.*,[13] they deduced a Unified Scientometric model by unifying all three classical scientometric laws and their variant forms through concept of Fractal theory and accumulated advantage models. Through the use of Index of fractality, they also showed with the difference of Fractality Index, how different forms of distributions (Zipf-Mandelbrot, Lotka, Leimkuhler Distribution form of Bradford's Law, Booth-Federowicz-Zipf Distribution, Condon-Zipf Distribution, Brooke's Law for Aging of Science, Price's Law of Exponential Growth, Generalized Model of Aging-Viability etc.) are created and some of them changed their equation forms as well.[13] Another important development in this field is the incorporation of the idea of Information Production Pro cess(IPP).[14-21] Egghe propounded two dimensional Information Production Process(IPP) with size-frequency function and size-frequency functions and this novel approach created a new domain of subject known as "Lotkaian Informetrics".[17,19] A general mathematical framework was developed for a continuous description of classical bibliometric laws.[22] Egghe[15] applied Pratt's measure for bibliometric distribution and proved that, 80/20 rule and Price's Square-root law basically can be derived from Lotka's law. In another study,[23] Egghe explained different consequences of Lotka's Law, for instance, the negative correlation between average number of papers and Lotka's parameters and change of Lotka's parameters for high concentration of authors. He also explained Herdan's law in linguistics and Heap's law in information retrieval on the basis of Lotkaian informetrics.[20] Egghe proved the fitness of Naranan's theorem in the framework of generalized IPP framework of power law and further interpreted Lotka's and Zipf's laws.[21] In the same line of research, Egghe derived mathematical relation between fraction of sources and produced items[15] and also studied inequality aspects of Zipfian and Lotkaian functions.[18]

Research gap and purpose of the study

After literature review, it is realized that, no research has yet been done to find out new alternate methods apart from trapezoidal rule and that is probably due to minimisation of error approximation. Most of the researches have been focussed on the applications of Lotka's law on different subjects in order to investigate whether the authors' productivity distribution follows this law or not. This study derived the formula for Lotka's constant (Equation 32) on the basis of Simpson's 1/3rd rule.

METHODOLOGY

The central part of the calculation is to divide the whole curve into two regions, the first part is to calculate the area under the curve upto a certain P and then from P to infinity. Under this method, the curve representing the function f(x) is approximated to compute the given integral form. Any approximating method always contains some error and the error must be evaluated along with its integration. The method, so far used is to evaluate the constant value, is the use of Newton-Cotes Quadrature formulas.

The integral form is,

$$I = \int_{x_1}^{x_n} f(x)dx = \int_{x_1}^{x_n}[P_n(x) + \epsilon_n(x)] = \int_{x_1}^{x_n} P_n(x)dx + \int_{x_1}^{x_n} \epsilon_n(x)dx = I_n + EI_n \quad -14$$

Here I_n is essentially a Lagrangian polynomial equation and EI_n is the error part. Under Newton-Coates formulas, Pao approximated the value of constant C with the principle of calculating the area under the curve using Trapezoidal rule of numerical integration method and here n=1, that means only 2 points (x_1 & x_2) are connected. Our approach is to refine the calculation of approximating the area with several other higher degree numerical integration methods. That's why, in the same line with Pao, we are applying Simpson's 1/3 rule to calculate the area under the curve and here n=2. The function can be estimated by fitting a parabola passing through $(x_0,f(x_0))$, $(x_1,f(x_1))$ and $(x_3,f(x_3))$.

The Integral form of the Simpson's 1/3 rule is -

$$I = \int_a^b f(x)dx = \frac{(b-a)}{3}\left[f(a) + 4f\left(\frac{(a+b)}{2}\right) + f(b)\right]\frac{1}{2}$$

$$= \frac{(b-a)}{6}\left[f(a) + 4f\left(\frac{(a+b)}{2}\right) + f(b)\right] + \frac{M}{90(b-a)^5} \qquad \text{-15}$$

Error estimate is - $\frac{-h^5}{90}f''''(x_1)$ -16

If $f(x) = \frac{1}{x^n}$ then, $f''''(x) = \frac{n(n+1)(n+2)(n+3)}{x^{n+5}}$

So, M=$\frac{n(n+1)(n+2)(n+3)}{x^{n+5}}$ -17

Error estimate in case of simpson's 1/3 rule is -

$$\int_a^b f(x)dx - \frac{(b-a)}{6}\left[f(a) + 4f\left(\frac{(a+b)}{2}\right) + f(b)\right] < \frac{M}{90(b-a)^5} \qquad \text{-18}$$

When we define the limit as [x,x+1] on [a,b] and putting the value of M in eqn. 18

$$\therefore \int_x^{x+1}\frac{1}{x^n}dx - \frac{1}{6}\left[\frac{1}{x^n} + \frac{4}{\left(x+\frac{1}{2}\right)^n} + \frac{1}{(x+1)^n}\right] < \frac{n(n+1)(n+2)(n+3)}{90x^{(n+5)}} \qquad \text{-19}$$

$$\Rightarrow 0 < \frac{n(n+1)(n+2)(n+3)}{90x^{(n+5)}} - \left[\int_x^{x+1}\frac{dx}{x^n} - \frac{1}{6}\left(\frac{1}{x^n} + \frac{4}{\left(x+\frac{1}{2}\right)^n} + \frac{1}{(x+1)^n}\right)\right] \qquad \text{-20}$$

$$\Rightarrow 0 < \frac{1}{6}\left[\frac{1}{x^n} + \frac{4}{\left(x+\frac{1}{2}\right)^n} + \frac{1}{(x+1)^n}\right] - \int_x^{x+1}\frac{dx}{x^n} < \frac{n(n+1)(n+2)(n+3)}{90x^{(n+5)}} \qquad \text{-21}$$

If x=P, P+1, P+2, P+3 ....∞, so summing the inequalities

$$\therefore 0 < \sum_P^\infty \frac{1}{6}\left[\frac{1}{x^n} + \frac{4}{\left(x+\frac{1}{2}\right)^n} + \frac{1}{(x+1)^n}\right] - \int_P^\infty \frac{dx}{x^n} < \sum_P^\infty \frac{n(n+1)(n+2)(n+3)}{90x^{(n+5)}} \qquad \text{-22}$$

If we expand the equation, we can get the following expression-

$$0 < \left[\frac{1}{6P^n} + \frac{4}{6\left(P+\frac{1}{2}\right)^n} + \frac{1}{6(P+1)^n}\right] + \left[\frac{1}{6(P+1)^n} + \frac{4}{6\left(P+\frac{3}{2}\right)^n} + \frac{1}{6(P+2)^n}\right] + \left[\frac{1}{6(P+2)^n} + \frac{4}{6\left(P+\frac{5}{2}\right)^n} + \frac{1}{6(P+3)^n}\right] + \cdots - \int_P^\infty \frac{dx}{x^n} < \sum_P^\infty \frac{n(n+1)(n+2)(n+3)}{90x^{(n+5)}} \qquad \text{-23}$$

$$\Rightarrow 0 < \frac{1}{6}\left[\frac{1}{P^n} + 4\sum_{\mathbf{p}}^\infty \frac{1}{(\mathrm{x}+\frac{1}{2})^{\mathrm{n}}} + 2\sum_{P+1}^\infty \frac{1}{x^n}\right] - \int_P^\infty \frac{dx}{x^n} < \sum_P^\infty \frac{n(n+1)(n+2)(n+3)}{90x^{(n+5)}} \qquad \text{-24}$$

Now we rearrange the whole equation form-

$$\int_P^\infty \frac{dx}{x^n} - \frac{1}{6\mathrm{p}^{\mathrm{n}}} < \frac{2}{3}\sum_{\mathbf{p}}^\infty \frac{1}{(\mathrm{x}+\frac{1}{2})^{\mathrm{n}}} + \frac{1}{3}\sum_{P+1}^\infty \frac{1}{x^n} < \int_P^\infty \frac{dx}{x^n} - \frac{1}{6\mathrm{p}^{\mathrm{n}}} + \sum_P^\infty \frac{n(n+1)(n+2)(n+3)}{90x^{(n+5)}} \qquad \text{- 25}$$

As we know, $\sum_P^\infty \frac{1}{x^n} < \int_{P-1}^\infty \frac{dx}{x^n}$

Estimation of $\sum_P^\infty \frac{n(n+1)(n+2)(n+3)}{90x^{(n+5)}}$, take A=$\sum_P^\infty \frac{n(n+1)(n+2)(n+3)}{90x^{(n+5)}}$

$$\sum_P^\infty \frac{n(n+1)(n+2)(n+3)}{90x^{(n+5)}} < \frac{n(n+1)(n+2)(n+3)}{90}\int_{P-1}^\infty \frac{dx}{x^{(n+5)}} \qquad \text{-26}$$

$$\Rightarrow \sum_P^\infty \frac{n(n+1)(n+2)(n+3)}{90x^{(n+5)}} < \frac{n(n+1)(n+2)(n+3)}{90}\left[\frac{x^{-(n+4)}}{(n+4)}\right]_{P-1}^\infty \qquad \text{-27}$$

$$\Rightarrow \sum_P^\infty \frac{n(n+1)(n+2)(n+3)}{90x^{(n+5)}} < \frac{n(n+1)(n+2)(n+3)}{90(n+4)(P-1)^{(n+4)}} \qquad \text{-28}$$

In the same line, we need to estimate the term $\sum_{\mathbf{p}}^\infty \frac{1}{(\mathbf{x}+\frac{1}{2})^{\mathbf{n}}}$

$$\sum_{\mathrm{p}}^{\infty} \frac{1}{(\mathrm{x}+\frac{1}{2})^{\mathrm{n}}} < \int_{\mathrm{p}-1}^{\infty} \frac{\mathrm{dx}}{(\mathrm{x}+\frac{1}{2})^{\mathrm{n}}}$$

$$\Rightarrow \sum_{\mathbf{p}}^\infty \frac{1}{(\mathbf{x}+\frac{1}{2})^{\mathbf{n}}} < \left[\frac{1}{-(\mathrm{n}-1)\left(x+\frac{1}{2}\right)^{(\mathrm{n}-1)}}\right]_{P-1}^\infty = \frac{1}{(n-1)(p-\frac{1}{2})^{(n-1)}} \qquad \text{-29}$$

It is implied that $\sum_P^\infty \frac{1}{x^n} < \int_{P-1}^\infty \frac{dx}{x^n}$

Now again re-arranging eqn. 25 with Value of 28 and 29

$$\int_P^\infty \frac{dx}{x^n} - \frac{1}{6\mathrm{p}^{\mathrm{n}}} - \frac{2}{3(n-1)(p-\frac{1}{2})^{(n-1)}} < \frac{1}{3}\sum_{P+1}^\infty \frac{1}{x^n} < \int_P^\infty \frac{dx}{x^n} - \frac{1}{6\mathrm{p}^{\mathrm{n}}} - \frac{2}{3(n-1)(p-\frac{1}{2})^{(n-1)}} + \frac{n(n+1)(n+2)(n+3)}{90(n+4)(P-1)^{(n+4)}} \qquad \text{-30}$$

$$\Rightarrow \frac{1}{(\mathrm{n}-1)\mathrm{p}^{(\mathrm{n}-1)}} - \frac{1}{6\mathrm{p}^{\mathrm{n}}} - \frac{2}{3(n-1)(p-\frac{1}{2})^{(n-1)}} < \frac{1}{3}\sum_{P+1}^{\infty}\frac{1}{x^n} < \frac{1}{(\mathrm{n}-1)\mathrm{p}^{(\mathrm{n}-1)}} - \frac{1}{6\mathrm{p}^{\mathrm{n}}} - \frac{2}{3(n-1)(p-\frac{1}{2})^{(n-1)}} + \frac{n(n+1)(n+2)(n+3)}{90(n+4)(P-1)^{(n+4)}}$$

$$\Rightarrow \frac{3}{(\mathrm{n}-1)\mathrm{p}^{(\mathrm{n}-1)}} - \frac{1}{2\mathrm{p}^{\mathrm{n}}} - \frac{2}{(n-1)(p-\frac{1}{2})^{(n-1)}} < \sum_{P+1}^\infty \frac{1}{x^n} < \frac{3}{(\mathrm{n}-1)\mathrm{p}^{(\mathrm{n}-1)}} - \frac{1}{2\mathrm{p}^{\mathrm{n}}} - \frac{2}{(n-1)(p-\frac{1}{2})^{(n-1)}} + \frac{n(n+1)(n+2)(n+3)}{30(n+4)(P-1)^{(n+4)}}$$

-31

After putting the sum $\sum_{P+1}^\infty \frac{1}{x^n}$ in Eqn. 31

$$\Rightarrow \sum_1^\infty \frac{1}{x^n} = \sum_1^P \frac{1}{x^n} + \frac{3}{(\mathrm{n}-1)\mathrm{p}^{(\mathrm{n}-1)}} - \frac{1}{2\mathrm{p}^{\mathrm{n}}} - \frac{2}{(n-1)(p-\frac{1}{2})^{(n-1)}} + \frac{n(n+1)(n+2)(n+3)}{30(n+4)(P-1)^{(n+4)}}$$

$$\Rightarrow \sum_1^\infty \frac{1}{x^n} = \sum_1^{P-1} \frac{1}{x^n} + \frac{1}{2P^n} + \frac{3}{(\mathrm{n}-1)\mathrm{p}^{(\mathrm{n}-1)}} - \frac{2}{(n-1)(p-\frac{1}{2})^{(n-1)}} + \frac{n(n+1)(n+2)(n+3)}{30(n+4)(P-1)^{(n+4)}}$$

So, from Lotka's law, the constant C has been transformed into the following form -

$$C = \frac{1}{\sum_1^\infty \frac{1}{x^n}}$$

$$=1/\left[\sum_1^{P-1}\frac{1}{x^n} + \frac{1}{2P^n} + \frac{3}{(\mathrm{n}-1)\mathrm{p}^{(\mathrm{n}-1)}} - \frac{2}{(n-1)(p-\frac{1}{2})^{(n-1)}} + \frac{n(n+1)(n+2)(n+3)}{30(n+4)(P-1)^{(n+4)}}\right] \qquad \text{- 32}$$

RESULTS AND ANALYSIS

Data

As this paper is solely focussed to develop a new method to derive the value of the constant, no new dataset has been collected here. In the celebrated paper by Lotka, he coined 6891 unique contributor/author starting with letter A and B from 1907-1916 of Decennial Index from Chemical Abstract of volume 1-10. The Auerbach data is the collection of 1325 number of prominent physicist from **Geschichtstafeln der Physik** until 1900. Only those physicists' names were covered who made significant contributions in Physics.[24] For the sake of proof of concept, old datasets [Chemical Abstract and Auerbach Data originally compiled by Lotka himself in 1926] used by Pao,[2] have been used here in order to continue the legacy.

Data Analysis

After deriving the mathematical form, it is necessary to test whether the threshold of the point P to be fixed at 20 or something else. $\sum_1^{\inf} 1/\mathrm{x}^{\mathrm{n}} = \frac{\pi^2}{6} = =1.646258503$. After setting exponent value n=2, we are putting several values of P to check upto what the error minimises. The table given below shows that, using equation-10, at P=21 the error minimises. After p=21, difference goes negative which means, using Simpson's 1/3 rule, estimating the inverse square infinite sum does not fit. Using trapezoidal rule, at p=20 the error or difference is minimised and with our method, the threshold can be fixed at 21 and this is a significant

Table 1: Value of $\frac{\pi^2}{6}$ determined by Equation 10 and Difference from the Actual value of $\frac{\pi^2}{6}$.

Value of *p* (Column A)	Value of summation of $1/x^n$calculated with Simpson's 1/3 Rule (Column B)	Difference (B -)
17	1.655419	-0.009161428
18	1.652355	-0.006097007
19	1.649603	-0.003344507
20	1.647117	-0.0008587292
21	1.644861	0.001397218
22	1.642804	0.003453736
23	1.640922	0.005336103
24	1.639192	0.007065505

development in this estimation process. Inclusion of more data points and better error minimization are more acceptable.

As mentioned by Pao, Lotka decided to use the simplest possible solution as to use to calculate constant because of its mathematical elegance and subsequently concluded as –"the proportion of all contributors that make a single contribution is about 60%". Although, it is needless to say that, more approximation is needed even though inverse power relation exists between authors and their contributions. The testing with higher degree numerical integration resulting into inclusion of more data points as expressed though *p* value. Whereas, Coile's testing on Murphy's data and Vlachy's effort was based on inverse square function rather than re-calculation on values of *n* and *c*. The Table 1 shows the successive calculation of 1/ with contiguous *p* values. The values of the constant 'C' at different exponent values are presented in Table 2. The results of Kolomogorov-Smirnov Test (KS Test) of observed and expected distributions of Chemical Abstract data are presented in Table 3. The results of KS Test of the Observed and Expected Values of Authors' Productivity Distribution of Auerbach Data are presented in Table 4. The comparative study between the values of 'C' (The Constant) and 'N' (The Exponent) for both of the Chemical Abstract data and Auerbach data are presented in Table 5. The methods applied and the results of hypothesis testing are also indicated in Table 5.

The value on the exponent (n) depend on multiple factors-Size of the bibliography, authors' collaboration behaviour, nature of subject (Physical Sciences, Social Sciences, Biological Sciences, Humanities and Arts) etc. Small changes in decimal places can change a lot in C value.

Calculation of Constant C (For Chemical Abstract Data):-

$$C = \frac{1}{\sum_1^{\infty}\frac{1}{x^n}} = 1\Big/\left[\sum_1^{P-1}\frac{1}{x^n} + \frac{1}{2P^n} + \frac{3}{(\mathrm{n}-1)\mathrm{p}^{(\mathrm{n})1}} - \frac{2}{(n-1)(p-\frac{1}{2})^{(n-1)}} + \frac{n(n+1)(n+2)(n+3)}{30(n+4)(P-1)^{(n+4)}}\right]$$

$$= 1/\ 1.764142$$

$$= 0.5668478$$

Table 2: Value of Constant C at different exponent values.

Exponent n	Constant C	Theoretical Percentage of Single Contribution
1.5	0.3828302	38.28
1.75	0.5096365	50.96
1.80	0.531319	53.13
1.85	0.5519344	55.19
1.90	0.5715422	57.14
1.95	0.5901981	59.01
2.0	0.6079542	60.79
2.05	0.6248595	62.48
2.10	0.64096	64.09
2.15	0.6562987	65.62
2.20	0.6709161	67.09
2.25	0.6848502	68.48
2.30	0.6981368	69.81
2.35	0.7108096	71.08
2.40	0.7229002	72.29
2.45	0.7344385	73.44
2.50	0.7454524	74.54
2.55	0.7559687	75.59
2.60	0.7660121	76.60
2.65	0.7756063	77.56
2.70	0.7847736	78.47
2.75	0.7935349	79.35
2.80	0.8019102	80.19
2.90	0.8175766	81.75
3.0	0.831911	83.19
3.25	0.8627016	86.27
3.5	0.8875221	88.75
3.75	0.9076138	90.76

So, the Lotka's equation becomes -

$$\varphi(x) = \frac{c}{x^n} = \varphi(x) = \frac{0.5668478}{x^{1.8878}} \quad - 33$$

Statistical Test

So, the maximum difference lies beyond the critical value at 0.01 significance level of Kolomogrov-Smirnoff test. Therefore, null hypothesis is rejected that means the given distribution is different from the theoretical distribution given by ϕ(x)=0.566847/x1.8878. Our method exactly complies with the exact phenomenon explained by Pao, w.r.t. Chemical Abstract data.[2]

Table 3: Kolomogorov-Smirnov Test of observed and expected distributions of chemical abstract data.

articles_x	authors_y	Proportions of Authors	Cumulative of Column C	Fitted Value with Eqn. 14	Cumulative of Column E	Difference Between Column D and F
1	3991	0.579161	0.579161	0.566848	0.566848	0.012313
2	1059	0.153679	0.73284	0.153173	0.720021	0.012819
3	493	0.071543	0.804383	0.071245	0.791266	0.013117
4	287	0.041649	0.846032	0.04139	0.832656	0.013376
5	184	0.026701	0.872733	0.027161	0.859817	0.012916
6	131	0.01901	0.891743	0.019252	0.879069	0.012674
7	113	0.016398	0.908141	0.014391	0.89346	0.014681
8	85	0.012335	0.920476	0.011184	0.904644	0.015832
9	64	0.009287	0.929763	0.008955	0.913599	0.016164
10	65	0.009433	0.939196	0.007339	0.920938	0.018258
11	41	0.00595	0.945146	0.006131	0.927069	0.018077
12	47	0.00682	0.951966	0.005202	0.932271	0.019695
13	32	0.004644	0.95661	0.004473	0.936744	0.019866
14	28	0.004063	0.960673	0.003889	0.940633	0.02004
15	21	0.003047	0.96372	0.003414	0.944047	0.019673
16	24	0.003483	0.967203	0.003022	0.947069	0.020134
17	18	0.002612	0.969815	0.002695	0.949764	0.020051
18	19	0.002757	0.972572	0.00242	0.952184	0.020388
19	17	0.002467	0.975039	0.002185	0.954369	0.02067
20	14	0.002032	0.977071	0.001983	0.956352	0.020719
21	9	0.001306	0.978377	0.001809	0.958161	0.020216
22	11	0.001596	0.979973	0.001657	0.959818	0.020155
23	8	0.001161	0.981134	0.001523	0.961341	0.019793
24	8	0.001161	0.982295	0.001406	0.962747	0.019548
25	9	0.001306	0.983601	0.001301	0.964048	0.019553
26	9	0.001306	0.984907	0.001209	0.965257	0.01965
27	8	0.001161	0.986068	0.001125	0.966382	0.019686
28	10	0.001451	0.987519	0.001051	0.967433	0.020086
29	8	0.001161	0.98868	0.000983	0.968416	0.020264
30	7	0.001016	0.989696	0.000922	0.969338	0.020358
31	3	0.000435	0.990131	0.000867	0.970205	0.019926
32	3	0.000435	0.990566	0.000817	0.971022	0.019544
33	6	0.000871	0.991437	0.000771	0.971793	0.019644
34	4	0.00058	0.992017	0.000728	0.972521	0.019496
36	1	0.000145	0.992162	0.000654	0.973175	0.018987
37	1	0.000145	0.992307	0.000621	0.973796	0.018511
38	4	0.00058	0.992887	0.00059	0.974386	0.018501
39	3	0.000435	0.993322	0.000562	0.974948	0.018374
40	2	0.00029	0.993612	0.000536	0.975484	0.018128
41	1	0.000145	0.993757	0.000512	0.975996	0.017761
42	2	0.00029	0.994047	0.000489	0.976485	0.017562

articles_x	authors_y	Proportions of Authors	Cumulative of Column C	Fitted Value with Eqn. 14	Cumulative of Column E	Difference Between Column D and F
44	3	0.000435	0.994482	0.000448	0.976933	0.017549
45	4	0.00058	0.995062	0.000429	0.977362	0.0177
46	2	0.00029	0.995352	0.000412	0.977774	0.017578
47	3	0.000435	0.995787	0.000395	0.978169	0.017618
49	1	0.000145	0.995932	0.000365	0.978534	0.017398
50	2	0.00029	0.996222	0.000352	0.978886	0.017336
51	1	0.000145	0.996367	0.000339	0.979225	0.017142
52	2	0.00029	0.996657	0.000327	0.979552	0.017105
53	2	0.00029	0.996947	0.000315	0.979867	0.01708
54	2	0.00029	0.997237	0.000304	0.980171	0.017066
55	3	0.000435	0.997672	0.000294	0.980465	0.017207
57	1	0.000145	0.997817	0.000275	0.98074	0.017077
58	1	0.000145	0.997962	0.000266	0.981006	0.016956
61	2	0.00029	0.998252	0.000242	0.981248	0.017004
66	1	0.000145	0.998397	0.000208	0.981456	0.016941
68	2	0.00029	0.998687	0.000197	0.981653	0.017034
73	1	0.000145	0.998832	0.000172	0.981825	0.017007
78	1	0.000145	0.998977	0.000152	0.981977	0.017
80	1	0.000145	0.999122	0.000145	0.982122	0.017
84	1	0.000145	0.999267	0.000132	0.982254	0.017013
95	1	0.000145	0.999412	0.000105	0.982359	0.017053
107	1	0.000145	0.999557	8.40E-05	0.982443	0.017114
109	1	0.000145	0.999702	8.10E-05	0.982524	0.017178
114	1	0.000145	0.999847	7.40E-05	0.982598	0.017249
346	1	0.000145	0.999992	9.00E-06	0.982607	0.017385

Critical Value is $= \frac{1.63}{\sqrt{\sum y_x}} = = \frac{1.63}{\sqrt{6891}} = =0.0196357$ (At 0.01 Significance Level).

And Maximum Difference(D_{max}) is-0.020719.

For Auerbach Data

$$C=1/\left[\sum_1^{P-1}\frac{1}{x^n}+\frac{1}{2p^n}+\frac{3}{(n-1)p^{(n)1a}}-\frac{2}{(n-1)(p-\frac{1}{2})^{(n-1)}}+\frac{n(n+1)(n+2)(n+3)}{30(n+4)(P-1)^{(n+4)}}\right]=1/1.625606=0.6151553$$

So, the Lotka's equation becomes -

$$\varphi(x)=\frac{c}{x^n}=\varphi(x)=\frac{0.6151553}{x^{2.021}}\quad -34$$

Here, we can see that, the maximum difference is less than the critical value at 0.01 significant level and thus null hypothesis is accepted. That means,$\varphi(x) = 0.6151553/x^{2.021}$ is fair enough to fit the observed values in Auerbach data.

DISCUSSION AND CONCLUSION

The null hypothesis is rejected for Chemical Abstracts data in both Trapezoidal Rule and Simpson's 1/3 Rule, whereas the null hypothesis is accepted for Aurbach's data in both Trapezoidal Rule and Simpson's 1/3 Rule. This paper has presented the possible scope of remodelling and restructuring the Pao method in order to fit Lotka's law upon any authors' productivity dataset in any subject domain. The Pao method is very vigorous and precise that has been used in order to validate Lotka's Law in different domains since more than two decades. The accuracy of every method can't be taken as 100% perfect and every method has some error term up to some finite level.

The main challenge is how to minimise the error and how much accuracy can be achieved for the area under the curve calculation. Pao method is based on Trapezoidal rule where two points are connected through a straight line fitted between two successive points. In Simpson's 1/3 rule, quadrature is fitted within interval of [a,b] where *n*=2. As a quadrature is fitted in a curve at an interval of [a,b]; the chance of exclusion of area under a curve gets minimised. Inversely, in its effect the constant value C gets

Table 4: KS Test of the Observed and Expected Values of Authors' Productivity Distribution of Auerbach Data.

articles_x	authors_y	Proportions of Authors	Cumulative of Column C	Fitted Value with Eqn. 14	Cumulative of Column E	Difference Between Column D and F
1	784	0.591698	0.591698	0.615155	0.615155	0.023457
2	204	0.153962	0.74566	0.151566	0.766721	0.021061
3	127	0.095849	0.841509	0.066792	0.833513	0.007996
4	50	0.037736	0.879245	0.037344	0.870857	0.008388
5	33	0.024906	0.904151	0.023788	0.894645	0.009506
6	28	0.021132	0.925283	0.016457	0.911102	0.014181
7	19	0.01434	0.939623	0.012052	0.923154	0.016469
8	19	0.01434	0.953963	0.009201	0.932355	0.021608
9	6	0.004528	0.958491	0.007252	0.939607	0.018884
10	7	0.005283	0.963774	0.005861	0.945468	0.018306
11	6	0.004528	0.968302	0.004834	0.950302	0.018
12	7	0.005283	0.973585	0.004055	0.954357	0.019228
13	4	0.003019	0.976604	0.003449	0.957806	0.018798
14	4	0.003019	0.979623	0.002969	0.960775	0.018848
15	5	0.003774	0.983397	0.002583	0.963358	0.020039
16	3	0.002264	0.985661	0.002267	0.965625	0.020036
17	3	0.002264	0.987925	0.002006	0.967631	0.020294
18	1	0.000755	0.98868	0.001787	0.969418	0.019262
21	1	0.000755	0.989435	0.001309	0.970727	0.018708
22	3	0.002264	0.991699	0.001191	0.971918	0.019781
24	3	0.002264	0.993963	0.000999	0.972917	0.021046
25	2	0.001509	0.995472	0.00092	0.973837	0.021635
27	1	0.000755	0.996227	0.000787	0.974624	0.021603
30	1	0.000755	0.996982	0.000636	0.97526	0.021722
34	1	0.000755	0.997737	0.000494	0.975754	0.021983
37	1	0.000755	0.998492	0.000417	0.976171	0.022321
48	2	0.001509	1.000001	0.000246	0.976417	0.023584

Critical Value is $= \frac{1.63}{\sqrt{\sum y_x}} = = \frac{1.63}{\sqrt{1325}} = =0.04477954$ (At 0.01 Significance Level)

And Maximum Difference(D_{max}) is - 0.023584

Table 5: Comparison between conclusions obtained after applying two different methods.

Data Source	C [Constant]	N [Exponent]	Method	Conclusion
Chemical Abstract	0.5669	1.8878	Trapezoidal Rule	Null Hypothesis rejected.
	0.566847	1.8878	Simpson's 1/3 Rule	Null Hypothesis rejected.
Auerbach Data	0.6151	2.021	Trapezoidal Rule	Null Hypothesis Accepted.
	0.6151553	2.021	Simpson's 1/3 Rule	Null Hypothesis Accepted.

more optimized and that's give better approximation where there is a very minute difference between the critical value and the maximum difference. The pattern of the result with Simpson's 1/3 rule and that with Pao method are same. Using Simpson's 1/3 method, it was found that the Lotka's law is not obeyed in case of Chemical Abstract data, but, Lotka's Law gets fit in case of Auerbach Data. The basic difference between Pao method and our method lies in the use of higher degree Newton-Coates formula. Fundamentally, trapezoidal rule deploys linking two intermediate points and then include/exclude areas as the curve moves upto its terminal distribution point. In our method, Simpson's 1/3 rule is used where 3 points are connected, 2 intermediate points and the mid-point between the said points and by this rule more area under the curve can be covered resulting better precision. Whatever Pao Method[2] explained about Chemical Abstract data and Auerbach's data, same phenomena can be explained by this method. This research may trigger to experiment with higher degrees of numerical integration methods. Similar efforts has been recently seen by Basu and Dutta[25] on implementing Simpson's 3/8 rule on similar datasets as used by Pao.

According to Chen and Leimkuhler,[26] common functional relationship can be established between Bradford, Zipf and Lotka's Law. Basically, Lotka's Law is about Frequency-Size approach; Bradford's Law is about cumulative-frequency log-rank approach and Zipf's Law is all about Frequency-Rank approach. Though mathematically, functional relations can be established but, the very nature of data and its distribution are different. An attempt was made by the authors of this paper to implement such Simpson's rule in order to develop new methods for calculation of Bradford's and Zipf's law, which was failed. There are other numerical integration methods to calculate area under the curve e.g. Lagrange interpolation functions, Bisection Method, Cubic Spline Methods etc., which are untouched areas till date. The future research may explore these domains for Lotka's Law and other scientometric laws as well.

ACKNOWLEDGEMENT

The necessary academic and technical assistance from The Department of Library and Information Science, Vidyasagar University, Midnapore, West Bengal and from The Department of Library, Maharani Kasiswari College, Kolkata, West Bengal has been hereby thankfully acknowledged.

CONFLICT OF INTEREST

The authors declare that there is no conflict of interest.

ABBREVIATIONS

Eqn.: Equation; **KS Test:** Kolomogorov-Smirnov Test; **w.r.t:** with respect to; **Inf:** Infinity.

REFERENCES

1. Lotka AJ. The frequency distribution of scientific productivity. Journal of the Washington Academy of Sciences. 1926;16(12):317-23.
2. Pao ML. An empirical examination of Lotka's law. Journal of the American Society for Information Science. 1986; 37(1): 26-33.
3. Murphy LJ. Lotka's law in the humanities? Journal of the American Society for Information Science. 1973;24(6):461-2.
4. Hersh AH. Drosophila and the course of research. Ohio Journal of Science.1942;42(5): 198-200.
5. Sen, BK. Lotka's Law: A View Point. Annals of Library and Information studies. 2010;57(2):166-7.
6. Nicholls PT. Empirical validation of Lotka's law. Information Processing and Management. 1986;22(5):417-9.
7. Bookstein A. The bibliometric distributions. The Library Quarterly. 1976;46(4):416-23.
8. Krisciunas K. Lotkas Law Year by Year. Journal of the American Society for Information Science (pre-1986). 1977;28(1):65.
9. Brookes BC. Ranking techniques and the empirical log law. Information Processing and Management. 1984;20(1-2):37-46.
10. Johnson NL, Kotz S. Discrete distributions: Distributions in statistics. Houghton Mifflin; 1969.
11. Nicholls PT. Price's square root law: empirical validity and relation to Lotka's Law. Information processing and management. 1988;24(4):469-77.
12. Nicholls PT. Bibliometric modelling processes and the empirical validity of Lotka's law. Journal of the American Society for Information Science. 1989;40(6):379-85.
13. Bailón-Moreno R, Jurado-Alameda E, Ruiz-Baños R, Courtial JP. The unified scientometric model. Fractality and transfractality. Scientometrics. 2005;63(2):231-57.
14. Egghe L. Consequences of Lotka's law for the law of Bradford. Journal of Documentation. 1985;41(3):173-89.
15. Egghe L. Pratt's measure for some bibliometric distributions and its relation with the 80/20 rule. Journal of the American Society for Information Science. 1987;38(4):288-97.
16. Egghe L. An exact calculation of Price's law for the law of Lotka. Scientometrics. 1987;11(1-2):81-97.
17. Egghe L, editor. Power laws in the information production process: Lotkaian Informetrics. Emerald Group Publishing Limited; 2005.
18. Egghe L. The power of power laws and an interpretation of Lotkaian informetric systems as self-similar fractals. Journal of the American Society for Information Science and Technology. 2005;56(7):669-75.
19. Egghe L. Zipfian and Lotkaian continuous concentration theory. Journal of the American Society for Information Science and Technology. 2005;56(9):935-45.
20. Egghe L. Untangling Herdan's law and Heaps' law: Mathematical and informetric arguments. Journal of the American Society for Information Science and Technology. 2007;58(5):702-9.
21. Egghe L. A new short proof of Naranan's theorem, explaining Lotka's law and Zipf's law. Journal of the American Society for Information Science and Technology. 2010;61(12):2581-3.
22. Rousseau R. Relations between continuous versions of bibliometric laws. Journal of the American Society for Information Science. 1990;41(3):197-203.
23. Egghe L. Theoretical evidence for empirical findings of A. Pulgarin on Lotka's law. Malaysian Journal of Library and Information Science. 2012;17(3):1-5.
24. Coile RC. Lotka's frequency distribution of scientific productivity. Journal of the American Society for Information Science. 1977;28(6):366-70.
25. Basu A, Dutta B. Redesigning of Lotka's Law with Simpson's 3/8 Rule. Journal of Scientometric Research. 2023;12(1):197-203.
26. Chen YS, Leimkuhler FF. A relationship between Lotka's law, Bradford's law and Zipf's law. Journal of the American Society for Information Science. 1986;37(5):307-14.

Cite this article: Basu A, Dutta B. An Analytical Study of Alternative Method for Solving Lotka's Law with Simpson's 1/3 Rule. J Scientometric Res. 2024;13(2):466-74.