J. Adv. Sci. Edu. Res. 2023: 4: 105-114, *ISSN*: 2583-0155 (ONLINE)
https://doi.org/10.56253/JASER.3.1.2023.105-114
Published: 22.11.2024*(http://jaser.rkmvccrahara.org/)*

Authorship Pattern in Lie Group of Mathematics Discipline: A Test with Simpson's 3/8 Rule

[1]Anindya Basu & [2]Bidyarthi Dutta

[1]Librarian, Maharani Kasiswari College

[2]Assistant Professor, Department of Library and Information Science, Vidyasagar University

*For Correspondence: *anindyabasuuit@gmail.com*

Abstract:

This paper aims to investigate the authorship distribution in field of Lie Group in Mathematics. Lie group is such a subject which combines two core fields of Mathematics i.e. Algebra and Geometry. The authors' publication frequency data has been collected from more than 100 journals and 10261 individual authors' name have been collected from the same. After preparing the authors' publication frequency data in ascending order, two methods have been applied to investigate whether Lotka's Law is applied on this distribution or not. The first method is classical Lee Pao Method and the second method is with Simpson's 3/8 Rule. This paper compared the values of the parameters of the Lotka's Law and derive the inferences regarding fitness of the authorship data.

Keyword: Scientometrics Law, Lotka's Law, Authors' Productivity, Lie Group

1. INTRODUCTION:

Sophus Lieis considered the pioneer of Lie group and Lie algebra during 1870s and his collaborator was Felix Klein. The name 'Lie Algebra' was first given by Hermann Weyl in 1930s and during the course of time, the subject has become a incubating bed of research among the mathematicians and physicists. The main contributors to this subject are Wilheim Killing, ElieCartan, Harish Chandra, VR Varadrajan, Lajos Pukanszky and Bertram Kostant. Lie Group and Lie Algebra have connected all the major areas of Mathematics [Analysis, Algebraic Topology, Algebraic Geometry, Combinatorics, Differential Geometry, Number Theory, Low-dimensional Topology, Riemann Geometry, Finite Group Theory etc.] and Physics [Particle Physics, Quantum Mechanics etc.](ICIAM, 2023).

The scientific productivity of authors is measured by Lotka's Law which states that, the number of authors making x contributionsis inversely proportional to square value ($1/n^2$) of those making one article (Fitzgerald, 2017). The distribution of authors' publications is highly skewed and handful number of scientists are actually real contributors in a subject. According to observation from

Lotka, only 6% authors in a subject/domain/discipline contribute more than ten articles in their life-time. Lotka's Law can be expressed as –

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

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(Qiu,2017; Patra,2006).

Here are some characteristics of the Lotka distribution function:

1. Power-law relationship: The Lotka distribution follows a power-law relationship, where the probability of an event occurring is proportional to its rank to the power of a constant exponent alpha.
2. Heavy-tailed distribution: The Lotka distribution has a heavy-tailed distribution, which means that the probability of observing extreme events is much higher than in other types of distributions, such as the normal distribution.
3. Scale invariance: The Lotka distribution is scale invariant, which means that the distribution remains the same regardless of the scale at which it is measured.
4. Pareto principle: The Lotka distribution is closely related to the Pareto principle, which states that a small number of events (the "vital few") are responsible for the majority of the results (the "trivial many").

Lotka's law involves the determination of two factors – a. calculation of exponent value and b. calculation of the value of the constant (Pulgarin, 2004).

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 –

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

Here,N = Number of Data Points as pair of x and y values, X= Logarithmic Transformation of x and Y= Logarithmic Transformation of y(Basu,2023)

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(Fernandez,2014). 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 value at the exponent and there is no specific method to estimate such things. As stated by Pao, dividing the both sides of equation-1 with total number of authors, i.e.

$\frac{\varphi(x)}{\sum \varphi} = \frac{c/\sum \varphi}{x^n}$-3

Now, let **F(x)** = $\frac{\varphi(x)}{\sum \varphi_x}$ and **C** = $\frac{C}{\sum \varphi_x}$ is the new constant. So, equation-1 can be rewritten as –

F(x) = **C** .$\frac{1}{x^n}$- 4

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

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

Summing each terms –

$\sum_1^x \varphi_x = c\left(\frac{1}{1^n} + \frac{1}{2} + \frac{1}{3^n} + \frac{1}{4^n} + \cdots + \frac{1}{x^n}\right)$ -5

Dividing both sides with total number of authors

$\sum_1^x \varphi_x / \sum_1^x \varphi_x = \frac{c}{\sum_1^x \varphi_x}\left(\frac{1}{1^n} + \frac{1}{2} + \frac{1}{3^n} + \frac{1}{4^n} + \cdots + \frac{1}{x^n}\right)$ -6

1 = **C** * ($\sum_1^x \frac{1}{x^n}$) - 7

Thus, the new constant **C** = $\frac{c}{\sum_1^x \varphi_x}$

$C = \frac{1}{\sum_1^x \frac{1}{x^n}}$ - 8

When n=2, $\sum_1^x \frac{1}{x^n} = \frac{\pi^2}{6}$

So, $C = \frac{6}{\pi^2} =$ 0.6074380 - 9

As the value of C can be determined by Riemann-Zeta function, Pao could derive an equation to get the constant value. From eq. 7, for upto∞,

$\sum_1^\infty \frac{1}{x^n} = \left[\sum_{x=1}^{P-1} \frac{1}{x^n} + \frac{1}{(n-1)(P^n - 1)} + \frac{1}{2P^n} + \frac{n}{24(P-1)^{(n+1)}}\right]$ -10

So, $C = \frac{1}{\sum_1^x \frac{1}{x^n}} = \frac{1}{\left[\sum_{x=1}^{P-1} \frac{1}{x^n} + \frac{1}{(n-1)(P^n - 1)} + \frac{1}{2P^n} + \frac{n}{24(P-1)^{(n+1)}}\right]}$ -11

M.L. Pao has developed the eq. 11 to use C as parameter to fit the Lotka'slaw(Pao,1985).

In a latest development, Simpson's 3/8 rule can be used to redesign the equation to derive the constant value C(Basu,2023).

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

$$= 1/\sum_{1}^{P-1}\frac{1}{x^n} + \frac{1}{2P^n} + \frac{4}{(n-1)P^{(n-1)}} - \frac{3}{2(n-1)}\left[\frac{1}{\left(p-\frac{2}{3}\right)^{(n-1)}} + \frac{1}{\left(p-\frac{1}{3}\right)^{(n-1)}}\right] + \frac{3n(n+1)(n+2)(n+3)(n+4)}{20(n+5)(P-1)^{(n+5)}} \text{-} 12$$

In this analytical study, both methods have been used in order to compare the performance of the area under the curve of Lotka distribution.

Data:-The authorship data has been downloaded from Web of Science in the subject of Lie Group and Lie Algebra spanning from 1972 to 2022.

Number of Papers	**Number of Authors**	**% of Publications**	**Cumulative Number of Authors**
1	5974	58.2204	58.22
2	1726	16.8210	75.04
3	816	7.9524	82.99
4	498	4.8533	87.85
5	321	3.1284	90.98
6	211	2.0563	93.03
7	158	1.5398	94.57
8	99	0.9648	95.54
9	85	0.8284	96.36
10	66	0.6432	97.01
11	50	0.4873	97.50
12	54	0.5263	98.02
13	30	0.2924	98.31
14	34	0.3314	98.65
15	19	0.1852	98.83
16	15	0.1462	98.98
17	24	0.2339	99.21
18	10	0.0975	99.31
19	6	0.0585	99.37
20	6	0.0585	99.43

21	6	0.0585	99.48
22	8	0.0780	99.56
23	7	0.0682	99.63
24	4	0.0390	99.67
25	6	0.0585	99.73
26	3	0.0292	99.76
27	2	0.0195	99.78
28	3	0.0292	99.81
29	2	0.0195	99.82
30	3	0.0292	99.85
33	1	0.0097	99.86
36	1	0.0097	99.87
37	3	0.0292	99.90
40	1	0.0097	99.91
41	1	0.0097	99.92
44	1	0.0097	99.93
47	2	0.0195	99.95
54	1	0.0097	99.96
55	1	0.0097	99.97
59	1	0.0097	99.98
74	1	0.0097	99.99
75	1	0.0097	100

Table-1:- Authors' Productivity Data from Lie Group Subject

1. DATA ANALYSIS:-

Linear least squares is a common method for estimating the power exponent α in the Lotka distribution, which describes the frequency distribution of a particular type of data, such as the number of publications by authors, the size distribution of cities, or the number of citations received by scientific papers(Björck,1990; Alapati,2000).

The basic idea of the linear least squares method is to take the logarithm of both sides of the Lotka distribution equation:

$$f(x) = \frac{C}{x^n} \quad \text{-13}$$

$$\ln\big(f(x)\big) = \ln(C) - n * \ln(x) \quad \text{-14}$$

This transforms the power-law relationship into a linear relationship that can be fit using linear regression. The exponent α can be estimated by fitting a straight line to the logarithm of the frequency and rank data, using a method such as the ordinary least squares method (Golub,1965;Alapati,2000).

The steps to estimate n using the linear least squares method are as follows:

1. Collect data on the frequency of occurrences (f) and their rank (x).
2. Take the logarithm of both f and x.
3. Fit a straight line to the logarithmic data using linear regression, such that ln(f) = b - n * ln(x), where b is the intercept of the line.
4. Estimate the exponent n as the slope of the line.

The linear least squares method can be useful when the data does not follow a strict power-law distribution and when there are a limited number of data points. However, it is important to note that this method can be sensitive to outliers and may not be as accurate as other methods when the data has a large number of points and follows a strict power-law distribution(Björck,1990). Therefore, it is recommended to use multiple methods to estimate n and to carefully interpret the results of any statistical analysis.

$$\mathbf{n} = \frac{\mathbf{N}\sum\mathbf{XY} - \sum\mathbf{X}\sum\mathbf{Y}}{\mathbf{N}\sum\mathbf{X^2} - (\sum\mathbf{X})^2} = \frac{\mathbf{-3780.972}}{\mathbf{1600.854}} = \mathbf{-2.36184}$$

The power value of the equation is 2.36184

Statistical Test- The Kolmogorov-Smirnov (KS) test is a statistical test that can be used to assess the goodness-of-fit of a distribution to a set of data(Massey,1951). In the context of fitting Lotka's Law to a dataset, the KS test can be used to determine whether the empirical distribution of the data follows a power-law distribution as predicted by Lotka's Law.

To perform a KS test, we first fit the Lotka distribution to the data using a maximum likelihood estimation method to estimate the value of the exponent α(Savanur,2014). We then compare the cumulative distribution function (CDF) of the Lotka distribution to the empirical CDF of the data using the KS test statistic, which measures the maximum distance between the two curves(Dahiru,2008).

If the p-value of the KS test is less than a predetermined significance level (e.g., 0.05), we can reject the null hypothesis that the data follow a Lotka distribution and conclude that the fit is not good (Brezinski,2015). However, if the p-value is greater than the significance level, null hypothesis can be accepted and conclude that the Lotka distribution is a good fit for the data (Smolinsky,2017).

It's worth noting that the KS test is just one of several methods that can be used to assess the goodness-of-fit of a distribution to a set of data, and it has some limitations, such as being sensitive

to sample size and not being able to distinguish between different types of departures from the theoretical distribution (Tsiatis,1980). Therefore, it is important to use multiple methods and to carefully interpret the results of any statistical test(Adigwe,2016).

Table-2:- Kolomogrov-Smirnov Test of Observed and Expected Distributions of authors in Lie Group Subject

Articles	Authors	Percentage of Authors	Cumulative of Percentage of Authors	Expected Percentage of Authors	Cumulative of the Expected Percentage of Authors	Difference between Column 4 and Column 6
1	5974	0.582204	0.582204	0.714303	0.714303	-0.132099
2	1726	0.16821	0.750414	0.138962	0.853265	-0.102851
3	816	0.079524	0.829938	0.053333	0.906598	-0.07666
4	498	0.048533	0.878471	0.027034	0.933632	-0.055161
5	321	0.031284	0.909755	0.01596	0.949592	-0.039837
6	211	0.020563	0.930318	0.010376	0.959968	-0.02965
7	158	0.015398	0.945716	0.007209	0.967177	-0.021461
8	99	0.009648	0.955364	0.005259	0.972436	-0.017072
9	85	0.008284	0.963648	0.003982	0.976418	-0.01277
10	66	0.006432	0.97008	0.003105	0.979523	-0.009443
11	50	0.004873	0.974953	0.002479	0.982002	-0.007049
12	54	0.005263	0.980216	0.002018	0.98402	-0.003804
13	30	0.002924	0.98314	0.001671	0.985691	-0.002551
14	34	0.003314	0.986454	0.001403	0.987094	-0.00064
15	19	0.001852	0.988306	0.001192	0.988286	2.00E-05
16	15	0.001462	0.989768	0.001023	0.989309	0.000459
17	24	0.002339	0.992107	0.000887	0.990196	0.001911
18	10	0.000975	0.993082	0.000775	0.990971	0.002111
19	6	0.000585	0.993667	0.000682	0.991653	0.002014
20	6	0.000585	0.994252	0.000604	0.992257	0.001995
21	6	0.000585	0.994837	0.000538	0.992795	0.002042
22	8	0.00078	0.995617	0.000482	0.993277	0.00234
23	7	0.000682	0.996299	0.000434	0.993711	0.002588

24	4	0.00039	0.996689	0.000393	0.994104	0.002585
25	6	0.000585	0.997274	0.000357	0.994461	0.002813
26	3	0.000292	0.997566	0.000325	0.994786	0.00278
27	2	0.000195	0.997761	0.000297	0.995083	0.002678
28	3	0.000292	0.998053	0.000273	0.995356	0.002697
29	2	0.000195	0.998248	0.000251	0.995607	0.002641
30	3	0.000292	0.99854	0.000232	0.995839	0.002701
33	1	9.70E-05	0.998637	0.000185	0.996024	0.002613
36	1	9.70E-05	0.998734	0.000151	0.996175	0.002559
37	3	0.000292	0.999026	0.000141	0.996316	0.00271
40	1	9.70E-05	0.999123	0.000118	0.996434	0.002689
41	1	9.70E-05	0.99922	0.000111	0.996545	0.002675
44	1	9.70E-05	0.999317	9.40E-05	0.996639	0.002678
47	2	0.000195	0.999512	8.00E-05	0.996719	0.002793
54	1	9.70E-05	0.999609	5.80E-05	0.996777	0.002832
55	1	9.70E-05	0.999706	5.50E-05	0.996832	0.002874
59	1	9.70E-05	0.999803	4.70E-05	0.996879	0.002924
74	1	9.70E-05	0.9999	2.70E-05	0.996906	0.002994
75	1	9.70E-05	0.999997	2.70E-05	0.996933	**0.003064**

Lotka'sexponent – 2.361846

$$\varphi(x) = \frac{C}{x^{2.361846}}$$

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

$$= 1/\sum_1^{P-1} \frac{1}{x^n} + \frac{1}{2P^n} + \frac{4}{(n-1)P^{(n-1)}} - \frac{3}{2(n-1)}\left[\frac{1}{\left(p-\frac{2}{3}\right)^{(n-1)}} + \frac{1}{\left(p-\frac{1}{3}\right)^{(n-1)}}\right] + \frac{3n(n+1)(n+2)(n+3)(n+4)}{20(n+5)(P-1)^{(n+5)}}$$

Putting the n=2.361846 and P=21, the Lotka's equation becomes with Simpson's 3/8 Rule–

$$f_{expected} = \frac{0.7143032}{x^{2.361846}}$$

At the 0.01 level of significance,

The critical value $= \frac{1.63}{\sqrt{\sum Y_x}} = \frac{1.63}{\sqrt{1637}} = \frac{1.63}{40.4598} = 0.04028685$

Maximum Difference - 0.003064

With Lee Pao Method, The Constant value C = 0.7137128, the maximum difference is 0.025859 and the critical value is 0.04028685.With Lee Pao method, the equation becomes -

$$f_{expected} = \frac{0.7137128}{x^{2.361846}}$$

In both the cases, as the maximum difference is less than the critical value, it can be inferred that, the fitted equation is well obeying the Lotka's Law(Basu,2023). But, the maximum difference with Simpson's 3/8 rule is 0.003064 and that of with Lee Pao method is 0.025859. So, it can be easily inferred that, Simpson's 3/8 gives a better fit.

2. CONCLUSION:

From 1986, the research who worked with authors' productivity data in different subjects used to apply Lee Pao method. Till then, no other method has been discovered. Simpson's 3/8 rule is a numerical integration technique used to approximate the definite integral of a function over a given interval. It is a higher-order method than the simpler trapezoidal rule and provides a more accurate approximation of the integral. The basic idea of Simpson's 3/8 rule is to divide the interval of integration into subintervals and approximate the integral over each subinterval using a quadratic polynomial. The integral over the entire interval is then approximated by summing the integrals over each subinterval. The dataset obeys Lotka's distribution. Both Lee Pao method and Simpson's method both gave similar approximation in constant C. But the new method gives the better fit.

3. REFERENCES:

1. Golub, G. (1965). Numerical methods for solving linear least squares problems. *NumerischeMathematik*, *7*, 206-216.
2. Basu, A., & Dutta, B. (2023). Redesigning of Lotka's Law with Simpson's 3/8 Rule. *Journal of Scientometric Research*, *12*(1), 197-203.
3. Björck, Å. (1990). Least squares methods. *Handbook of numerical analysis*, *1*, 465-652.
4. Alapati, S., & Kabala, Z. J. (2000). Recovering the release history of a groundwater contaminant using a non-linear least-squares method. *Hydrological processes*, *14*(6), 1003-1016.
5. Pao, M. L. (1985). Lotka's law: a testing procedure. *Information processing & management*, *21*(4), 305-320.

6. https://iciam2023.org/accepted_ms
7. Fitzgerald, S. R. (2017). *Information seeking of scholars in the field of higher education*. Retrieved May 1, 2023, from https://doi.org/doi:10.25335/M5P479
8. Qiu, J., Zhao, R., Yang, S., & Dong, K. (2017). *Informetrics: theory, methods and applications*. Springer.
9. Patra, S. K., & Mishra, S. (2006). Bibliometric study of bioinformatics literature. *Scientometrics*, *67*, 477-489.
10. Pulgarín, A., & Gil-Leiva, I. (2004). Bibliometric analysis of the automatic indexing literature: 1956–2000. *Information processing & management*, *40*(2), 365-377.
11. Carrera-Fernandez, M. J., Guardia-Olmos, J., & Peró-Cebollero, M. (2014). Qualitative methods of data analysis in psychology: An analysis of the literature. *Qualitative Research*, *14*(1), 20-36.
12. Massey Jr, F. J. (1951). The Kolmogorov-Smirnov test for goodness of fit. *Journal of the American statistical Association*, *46*(253), 68-78.
13. Savanur, K., Devi, S. G., & Konnur, P. V. (2014). Lotka's Law and Authorship Distribution in the Journal of 'Columbia Law Review'. *COLLNET Journal of Scientometrics and Information Management*, *8*(1), 193-208.
14. Dahiru, T. (2008). P-value, a true test of statistical significance? A cautionary note. *Annals of Ibadan postgraduate medicine*, *6*(1), 21-26.
15. Brzezinski, M. (2015). Power laws in citation distributions: evidence from Scopus. *Scientometrics*, *103*, 213-228.
16. Tsiatis, A. A. (1980). A note on a goodness-of-fit test for the logistic regression model. *Biometrika*, *67*(1), 250-251.
17. Smolinsky, L. (2017). Discrete power law with exponential cutoff and Lotka's law. *Journal of the Association for Information Science and Technology*, *68*(7), 1792-1795.
18. Adigwe, I. (2016). Lotka's Law and productivity patterns of authors in biomedical science in Nigeria on HIV/AIDS: A bibliometric approach. *The Electronic Library*, *34*(5), 789-807.