The Power of Lotka's Law Through the Eyes of R

PhD Alon FRIEDMAN
School of information, University of South Flordia, Tampa

ABSTRACT

The paper aims to outline the author impact based on Lotka's Law using R. Lotka's Law is well known as "the inverse square law of scientific productivity." It states that the number of authors publishing a certain number of articles is a fixed ratio to the number of authors publishing a single article. The study found authoring patterns do not match to Lotka's Law. The study summarizes the results using ggplot2. While the study fail to support eminent literary authors productivity, future tests should use full data set, and avoid using samples or grouping as alternative. This is the first attempt to utilize R to employ Lotka's law using R.

Keywords: *Bibliometrics, Authorship Distribution, Lotka's Law, R, OCLC WorldCat*

INTRODUCTION

The term Infometrics is often defined as the "science of processing data for storage and retrieval," according Egghe (2005). The term was developed as the modern substitute to term bibliometrics, where its roots found in the field of library science. The llibrarians use bibliometrics analysis to measaure the number of citations found in the library catalog. Many research professional and accademic fields use bibliometric methods to explore the impact a researcher or the impact of a particular paper in their fields. Bibliometrics also was used in wide range of others applications including: linguistics, therasuri development and even the evulation of reader usage of the resources offer by the library. According to Pritchard (1969), "The definition and purpose of bibliometics is to shed light on the process of written communication and of the nature and course of a discipline (in so far as this is displayed through written communication) by means of counting and analyzing the various facets of written communication." (Cited by Nicholas, D. & Ritchie, M. (1978)).

One of the most popular methods under bibliometrics analysis is citation analysis, which is used for the purposes of collection development, acquisitions and tenure promotion process (Budd, 1999, Budd & Gross and Gross, 1927). Many of the citation analysis studies explored the impact of gender, geographic location, library type and authors and academic researchers productivity (Budd & Seavey, 1990, Hart, 2000). One of the most discuss model under bibliometrics is Lotka's law.

Lotka's Law states that number of authors making n contributions is about 1\n_{2} of those making single publication. The contributions of authors making single contribution are about 60% of the entire publication in a specific field.

In today's R environment, many packages were created to provide the researcher solutions to calculate the author productivity. We found three specific packages that allow the user to calculate Zipf Law and Bradford's Law and Gaussian-Poisson distribution. For example, in 2003, Evert, S. developed the ZipfR's package that allows the user to calculate Zipf law and Branford law. However, We did not find any package or guidelines on how to calculate Lotka's Law in R. This study raises the question: can we convert Lotka's Law using R? and can we improve the graphical representation of the result based on Lotka's Law. We examined the eminent literary authors productivity, where we collected a sample data from Online Computer Library Center also known as OCLC to address this hypothesis: Can eminent literary authors productivity fit Lotka's Law?

VISUALIZATION

Visualization as a form of human communication can be traced back to the development of language, signs, and sounds in ancient times. In the present day, it is most often a visual-mapping process in which computer representations are mapped to perceptual representations, using encoding techniques to maximize human understanding of the data. Visualization is often defined as a tool or method for interpreting data fed into a computer and for generating images from complex multidimensional data sets. According to Heer et al. (2002), the goal of visualization is to aid our understanding of data by leveraging the human visual system's highly tuned ability to see patterns, spot trends, and identify outliers but it also could be a visual proof to judge the reasonableness of the data.

Visualization plays an important part of R community development. Many packages were created to support the researcher to visualize his/her findings. One in particular package was ggplot2. The package was created by Hadley Wickham in 2005. The advances of ggplot2 package is the ability of the user to add, remove or alter components in a plot at high level of abstraction, according to Smith (2011).

Lotka's Law

In 1926, Alfred J. Lotka examined the distribution of frequencies of researchers publications in the area of chemical and physics scientific productivity. He examined the quantitative relation among the authors and their scientific production publications listed in the Chemical Abstracts since 1907 to 1916. His observation later become Lotka's Law that show an asymmetric distribution with a concentration of articles among a few authors, while the remaining articles would be distributed among a great amount of authors with low distribution. Since then, many researchers from different fields employ Lotka's Law to examine author productivity and publications. The validity of Lotka's law also been studied by a number of researchers. The most notable researchers were Pao (1985, and1986) and Nicholls (1986 and 1987), reported that Lotka's model fitted the majority of the data sets they set to study. Both established a standard testing procedure for testing Lotka's Law by employing:

a) Data collection procedure,

b) Estimation of the unknown parameters in the model
c) and testing conformity of the observed data to the theoretical distribution by means of a goodness of fit test.

LOTKA'S LAW OF SCIENTIFIC PRODUCTIVITY

Lotka's law states that number of authors making n contributions is about 1\n_{2} of those making single publication. The contributions of authors making single contribution are about 60% of the entire publication in a specific field. The basic Lotka's formula outline the number of authors y_x each credited with x number of papers is inversely proportional to x, which is the output of each individual author. The relation is expressed as

i. X and 1{(n)}Y - (x)=C

where y_x is the number of authors making x contributions to the subject and n and c are the two constants to be estimated for the specific set of data. Lotka's noted that eqn applied to a variety of phenomena. In his analysis, he examined two different subject areas: physics and chemistry abstracts published in different journals. He then formulated this general rule for scientific productivity.

The main elements involved in fitting a Lotka's law are: measurement of the variables and tabulation, form of the model, parameter estimation and criterion for goodness-of-fit. We will follow Pao (1985) recommendations to pursue Lotka's law:

1. Measurement and tabulation: the number of authors' y^{x} contributing x paper are organized into a size frequency table of n, x, y pairs.
2. Model: the generalized inverse-power model where ii. Y_{x}= KX^{-b} is adopted
3. Estimation of slop b: The ordinary linear least squares estimate of b in the transformed model:

iii. log(Y_{x})= log K- b log = 1, 2, x/max

4. Estimation of constant C:

Based on: iv: Y_{x}=C/X_{n}

Pao (1985) recommend dividing both sides of eqn by□, \sum Y_{x} the total number of authors

y_{x}/\Sigma Y^{x}= C/(\Sigma y_{x})/(1/X^{n})

Let f(Y_{x})=Y_{x}/\sum Y_{x} provides the fraction of authors making x contributions and C = c/ \sum Y_{x} is the new constant, expressed as a fraction of the total sample of authors. Thus, equ

v: y_{x}/\Sigma Y^{x}= C/(\Sigma: y_{x})/(1/X^{n}) can be written as

(y_{x})/C(1/ y_{x})

According to Pao(1985), this equation is another form of Lotka's general law that stands for, the percentage of authors f[(y] ↓x), where each with x is the number of publications. This is inversely proposal to x raised to the nth power.

5. Extrapolating from Lotka's calculation of the special case for n = 2, the general formulation equ for any value of n is as follows

$y_{1} = c(1/1^{n})$
$y_{2} = c(1/2^{n})$
$y_{3} = c(1/3^{n})$
$y = c(1/x^{n})$

Summing both sides of these equations will provide us the following formula where according to Pao; we need to divide both sides by the total number of authors

vi: $\Sigma yx=y1=c(1/1n)+y1=c(1/2n)+y1=c(1/3n)+y=c(1/xn)$

$\Sigma y_{x}/\Sigma y_{x}= (c/\Sigma y_{x}) (\Sigma 1)/ x^{n}$

Since the summation Σy_{x} gives unity, and $c/ \Sigma y_{x} = C$ and as a result

$1=c(\Sigma/1/xn)$

$C=1(\Sigma/1/xn)$

6. Test: There are several statstical tests avabille for goodness-of-fit.Among those tests, including Kolmofrov Smirnov (K-S) test.

a) Kolmogorov –Smirnov (K-S) aims to accomplished by findings the theoretical cumulative frequency distribution which would be expected the null hpotyhesis and comparing it with the observed cumulative frequency distribution. The point at which the two observed distributions show the maximum deviation can be determined. The null hypothesis is then rejected if the caculated value of D is greater than critical value.

7. Visualization of the result.

METHODS

In 2002, Newby et.al (2002) reports that many graduate students and library professionals find it hard to caculate Lotka's Law. In order to provide alternative choice, they developed an open source application using Linux Software Map (LSM) and Sourceforge. In this study, we followed Pao (1985) discussion on the procedure to measure Lotka's Law. The data collection of author impact and eminence is based on Burt (2009) study on the top of authors. A total of 198 author were randomly selected. With 29 of the authors coming from England, Ireland, or the United States, 19 authors alive in the 20^{th} century and only 6 were women. We used OCLC WorldCat as the measure platform. The criteria to evulaute the authors ranking was based on 4 principles: The first two was ranking were based on Bloom (2002) and Gottlieb et al. (1998) classification. The both published their own ranking and recived rdcongition from the library community for their effecrd. The next classification was based on By, About and By and About. By using Boolean search technique to refinoce the result of specific authors enables us to redefine the terms related to concepts of impact and eminence operationally in terms of the numbers of records by and about authors (as well as related logical combinations, such as about but not by).

The measures we have devised for a given individual *x* are:

1. The *eminence* of x*,* defined as the number of records attributed to *x* (by *x*).
2. The *impact* of *x*, defined as the number of records of items about x.
3. The *fame* of x is defined as the number of records by or about *x*.
4. The *auto-perpetuation* of *x* is the number of records both by and about *x*.

Using these terms we caculate: fame = eminence + impact – auto-perpetuation.

The OCLC WorldCat is a union catalog that itermize the collections of 72,000 libraries in 170 countries. The list was not constructed with an agenda to provide proportional representation to different constituencies; nor, because of the combined criteria of our different sources, does it include the most recent or ancient authors.

Data for this study were collected both in 2007 and then in 2014. With over 325 million bibliographic records at the time of writing, WorldCat is updated so rapidly that it is necessary for readings of a group of names to taken at approximately the same time for measurements to be comparable. Readings were taken from July 19 to 23, 2007. In updating the research for the present study, data on the same authors studied in 2007 were taken again from August 6 to 9, 2014. Single author might have multipli listings due to number of publications. From the two datasets, we took a sample to capture the data. In both cases, we used random sample. Table 1 represent our data 20 row of data.

Sample of our data set taken from WorldCat

Table 1

Bloom's rank	Gottlieb Bowers	author	By	About	By and about
19	15	Freud, Sigmund, 1856-1939	100	7671	440
0	19	Rousseau, Jean-Jacques	203	6521	734
9	30	Dante Alighieri, 1265-1321	237	17312	1395
5	34	Tolstoy, Leo	282	5932	652
0	36	Voltaire, 1694-1778	297	4946	686
2	44	Cervantes, Miguel	364	6360	517
4	53	Milton, John, 1608-1674	431	7834	730
10	62	Chaucer, Geoffrey, d. 1400	470	6677	540
96	70	Dickens, Charles, 1812-1870	506	9378	1259
35	73	Murasaki Shikibu, b. 978?	558	2037	377
97	77	Dostoyevsky, Fyodor, 1821-1881	585	5557	468
0	97	Erasmus, Desiderius, d. 1536	646	2289	280
0	112	Petrarca, Francesco, 1304-1374	767	4045	614
18	131	Goethe, Johann Wolfgang von, 1749-1832	786	21017	3529
56	163	Hugo, Victor, 1802-1885	802	3983	803
34	167	Austen, Jane, 1775-1817	810	3268	452
27	167	Ibsen, Henrik, 1828-1906	838	2785	372
63	171	Joyce, James, 1882-1941	849	6859	821
26	175	Molière, 1622-1673	895	4122	491

RESULTS

To calculate the value of N, we generate the following procedure in R:

```
> LotkasN2 <- function(Sums, FullTable, n)
   {
>N <- n
>xy <- Sums[5]
>loX <- Sums[3]
>loy <- Sums[4]
>x2 <- Sums[6]
>loX2 <- loX^2
>top <- (N*xy) - (loX*loy)
>bottom <- (N*x2) - (loX2)
>Nfinal <- top/bottom
return(Nfinal)

}
```

and then

```
>LotkasC2 <- function(p, N)

   {

    P <- p
    increm <- c(1:(P-1))
    sum <- sum(1/increm^N)
    part1 <- sum
    part2 <- 1/((N-1)*(P^(N-1)))
    part3 <-1/(2*(P^N))
    part4 <- N/(24*(P-1)^(N+1))
    result <- 1/(part1+part2+part3+part4)
    return(result)
    }
```

That allow us to caculaute the frequency table of the distribution of the eminent literary authors productivity. Table 2 represents the frequency distribution of our data.

The frequency distribution of the eminent literary authors productivity.

Table 2

Author name	Paper	Authors	Log X	Log Y	Xy	X^2
Cervantes, Miguel	1000	30	3.000000	1.4771213	4.43124	9.0000
Petrarca, Francesco	2000	46	3.301030	1.6627578	5.48813	10.89680
Eliot, George	3000	27	3.602060	1.3979400	4.86007	12.09037
Thoreau, Henry David	4000	27	3.602060	1.413638	5.15585	12.97484
Byron, George	5000	20	3.698970	1.301030	4.81242	13.68238
La Fontaine, Jean de	6000	10	3.778151	1.000000	3.77815	14.27443
Irving, Washington	7000	11	3.845090	1.041397	4.00425	14.27443
Baudelaire, Charles	8000	6	3.903090	0.778151	3.03719	15.23411
	9000	5	3.954243	0.698970	2.76389	15.63603
Rousseau, Jean-Jacques	10000	6	4.000000	0.778151	3.11260	16.00000
Freud, Sigmund	11000	2	4.041393	0.301030	1.21658	16.33285
Harold Cartland	12000	3	4.079181	0.477213	1.94626	16.63972
Austen, Jane	13000	2	4.146128	0.300100	1.23842	16.92452
Dr. Seuss	14000	1	4.146128	0.000000	0.00000	17.19038
Pushkin, A.	16000	1	4.204120	0.000000	0.00000	17.67462
Williams, Tennessee	20000	1	4.301030	0.000000	0.00000	18.4986
Christie, Agatha	25000	1	4.397940	0.000000	0.00000	19.34188
Shakespeare, W.	64000	1	4.806180	0.000000	0.00000	23.09937

We then follow Pao (1985) recommendation for K-S test, a goodness-of-fit statistical test to assert that the observed author productivity distribution is not significantly different from a theoretical distribution

The maximum deviation was equaled 0.0811 which exceeds the critical value of 0.0076 at the 0.01 level of significance. Therefore, the null hypothesis must be rejected and concluded that the data of eminent literary authors productivity do not fit Lotka's Law.

\max F_{0}(x) - S_{n}(x) = 0.0811

The critical value at the 0.01 level of significance: 1.63/\sqrt{\sum Y } = 1.64/213.8059

D>0.0076

In R, we caculate K-S tesrt as follow:

```
>ks.test(x, y)
>qqplot(x, y)
>abline(0, 1)
>ks.test(x, "pnorm", mean = mu.hat, sd = sigma.hat)
```

The result D> 0.0076

The visualization summary

We also employ ggplot2 to create our visualization. In this a boxplot graph where the frequency distribution categories are plotted. X is the number of papers and Y stands for the categories of authors. In compare to any common visual used to report on Lotka's Law, this visualization summarizes the data based on groups and add the element of colors to differentiate between the groups. In this case, we used ggplot2 to generate this bar plot.

Figure 1

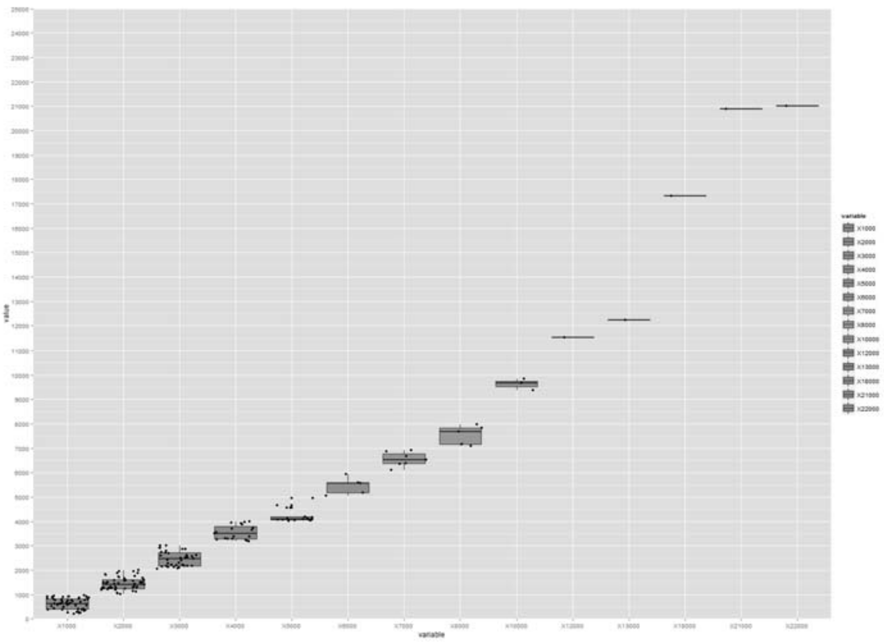

Figure 1 represent boxport graph captures the frequency distribution between the number of papers and the categories of authors.

SUMMARY

Lotka's Law of author productivity is regarded as one of the classical laws of bibliometrics. In this study, we followed Pao's modified methodology, where the value of the exponent *n* for eminent literary authors productivity is calculated 1.272X and the constant *c* was equaled 0.234501 Using the K-S test it's found that at the 0.01 level of significance the maximum deviation is 0.0073 which is lower than the critical value of 0.0076. Therefore, it can be concluded that the result of this study does not fit to Lotka's Law. However, the aim of the study was not to measure the validity of Lotka's Law but to examine if it can be applied to R. We so found R is easy platform to study again the subject of validity of Lotka's Law. We recommend future to examine the full data set, and avoid using samples or grouping as alternative.

REFERENCE

1. Allison, P.D., D., de Solla Price, B.B. Griffith, M.J. Moravsik, J.A., Strwart. (1967). "Lotka's Law: A problem in its interpretation and application." In:_ Social Studies of Science Vol. 6(2). pp. 269-276.
2. Budd, J.M. (1999). "Scholary productiviety of U.S LIS faculty: An update." In:_Library Quarterly Vol.(70) pp 230-245
3. Budd, John M., and Charles A. Seavey, (1990). Characteristics of Journal Authorship by Academic Librarians. In:_College and Research Libraries 51:5 (September 1990): 463-470.
4. Gross, P. L. K. & E. M. Gross (1927). Collaboration and author productivity: A study with a new variable in Lotka's law. In:_ Scientometrics, 44 129-134.
5. Hadley, Wickham (2010) ggplot2: Elegant Graphic for Data Analysis. In:_ Journal of statistical software. 35(1)
6. Hart, R. L. (2000). Co authorship in the academic library literature: A survey of attitudes and behaviour. In:_Journal of Academic Librarianship, 26(5), 339-345
7. Egghe, L. (2005) Expansion of the field of Infometrics: Origins and consequence. In:_ Information Processing and Management 41(6) pp. 603-610.
8. Ken, B. K. (2010) Lotka's Law: a viewpoint. In:_Annals of library and information Studies Vol. 57(3). pp. 166-167
9. Kyvik, S. Prodctuvuty differences fields of learning, and Lotka's Law. In:_Scientometrics 15(3-4) pp. 205-214.
10. Lotka, A. J. (1926) The frequency distribution of scientific productivity. In:_Journal of the Washington Academy of Sciences. 16(12), pp. 317-323.
11. Narendra, K (2010). Applicability to Lotka's Law to research productivity of Council of [11] Scientific and Industrial Research (CSIR), India. In:_Annals of Library and Information Studies 57(1).
12. Nicholas, D. & Ritchie, M. (1978). Literature and bibliometics. London: Clive Bingley.
13. Nicholls, P.T. (1986) Empirical validation of Lotka's Law. In:_Information Processing and Management. 22, 417-419.
14. Nicholls, P.T. (1986) Emprical validation of Lotka's law. In:_Information Processing and Manahement 22(1). pp. 417-419
15. Pitchard, A. (1968) Computers, statistical bibliography and abstracting services. (unpublished)
16. Pao, M.L. (1985) Loktka's Law: a testing procedure. In:_Information Processing & Management, 21(4), 305-320.
17. Pao, M. L. (1986) An empirical examination of Lotka's Law. In:_JASIS. 37(1), 1986, 26-33
18. Potter, W. G. Lotka's Law revisited. In:_Library Trends. 30(1), 1981, 21-39.
19. Smith, D. (2011) Create beautiful statistical graphics with ggplot2. Revolutions Analytics website. Retrieve: Feb 11, 2015. http://www.revolutionanalytics.com/
20. Zainab A. N. anyi, K.W. U. Anuar, N. B. (2009) A single journal study: malaysian journal of computer science. In:_ Malaysian Journal of Computer Science. 22(1).