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

PhD Alon FRIEDMAN
School of information, University of South Florida, 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 librarians use bibliometrics analysis to measure the number of citations found in the library catalog. Many research professional and academic 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, thesauri development and even the evaluation of reader usage of the resources offer by the library. According to Pritchard (1969), “The definition and purpose of bibliometrics 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 $\frac{1}{n^2}$ 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 \( \ln(n) \) 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 \( \ln(n) \) \( Y - (x) = C \)

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. \( \log(Y_\{x\}) = \log K - b \log x = 1, 2, x/\text{max} \)
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
   \( y_\{x\}/\Sigma \Sigma Y^\{x\} = C/(\Sigma y_\{x\})/(1/X^\{n\}) \)
   Let \( RY_\{x\} = Y_\{x\}/\sum Y_\{x\} \) 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: \( y_\{x\}/\Sigma \Sigma Y^\{x\} = C/(\Sigma y_\{x\})/(1/X^\{n\}) \) 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 \( \Sigma (C^{\{x\}} X^{\{n\}}) \), 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: \( \Sigma y_x = y_1 = c(1/1^n) + y_2 = c(1/2^n) + y_3 = c(1/3^n) + y = c(1/x^n) \)

\( \Sigma y_x/\Sigma y_x = (c/\Sigma y_x) (\Sigma 1)/x^n \)

Since the summation \( \Sigma y_x \) 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 statistical tests available for goodness-of-fit. Among those tests, including Kolmogorov Smirnov (K-S) test.

a) Kolmogorov –Smirnov (K-S) aims to accomplish by finding the theoretical cumulative frequency distribution which would be expected the null hypothesis 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 calculated 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 calculate 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 20th century and only 6 were women. We used OCLC WorldCat as the measure platform. The criteria to evaluate 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 recognition from the library community for their effort. The next classification was based on By, About and By and About. By using Boolean search technique to refine 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 calculate: fame = eminence + impact – auto-
perpetuation.

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

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

```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

```r
> LotkasC2 <- function(p, N)
   {
     P <- p
     increm <- c(1:(P-1))
     sum <- sum(1/increm^N)
     part1 <- sum
     part2 <- 1/(P*(N-1))
     part3 <- 1/(2*(P^N))
     part4 <- 1/(N*(P-1)*(N+1))
     result <- 1/(part1+part2+part3+part4)
     return(result)
   }

That allow us to calculate 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

<table>
<thead>
<tr>
<th>Author name</th>
<th>Paper</th>
<th>Authors</th>
<th>Log X</th>
<th>Log Y</th>
<th>Xy</th>
<th>X²</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cervantes, Miguel</td>
<td>1000</td>
<td>30</td>
<td>3.000000</td>
<td>1.4771213</td>
<td>4.43124</td>
<td>9.0000</td>
</tr>
<tr>
<td>Petrarcha, Francesco</td>
<td>2000</td>
<td>46</td>
<td>3.301030</td>
<td>1.6627578</td>
<td>5.48813</td>
<td>10.89680</td>
</tr>
<tr>
<td>Eliot, George</td>
<td>3000</td>
<td>27</td>
<td>3.602060</td>
<td>1.3979400</td>
<td>4.86007</td>
<td>12.09037</td>
</tr>
<tr>
<td>Thoreau, Henry David</td>
<td>4000</td>
<td>27</td>
<td>3.602060</td>
<td>1.413638</td>
<td>5.15585</td>
<td>12.97484</td>
</tr>
<tr>
<td>Byron, George</td>
<td>5000</td>
<td>20</td>
<td>3.698970</td>
<td>1.301030</td>
<td>4.81242</td>
<td>13.68238</td>
</tr>
<tr>
<td>La Fontaine, Jean de</td>
<td>6000</td>
<td>10</td>
<td>3.778151</td>
<td>1.000000</td>
<td>3.77815</td>
<td>14.27443</td>
</tr>
<tr>
<td>Baudelaire, Charles</td>
<td>8000</td>
<td>6</td>
<td>3.903090</td>
<td>0.778151</td>
<td>3.03719</td>
<td>15.23411</td>
</tr>
<tr>
<td>Rousseau, Jean-Jacques</td>
<td>9000</td>
<td>5</td>
<td>3.954243</td>
<td>0.698970</td>
<td>2.76389</td>
<td>15.63603</td>
</tr>
<tr>
<td>Freud, Sigmund</td>
<td>11000</td>
<td>2</td>
<td>4.041393</td>
<td>0.301030</td>
<td>1.21658</td>
<td>16.33285</td>
</tr>
<tr>
<td>Harold Cartland</td>
<td>12000</td>
<td>3</td>
<td>4.079181</td>
<td>0.477213</td>
<td>1.94626</td>
<td>16.63972</td>
</tr>
<tr>
<td>Austen, Jane</td>
<td>13000</td>
<td>2</td>
<td>4.146128</td>
<td>0.300100</td>
<td>1.23842</td>
<td>16.92452</td>
</tr>
<tr>
<td>Dr. Seuss</td>
<td>14000</td>
<td>1</td>
<td>4.146128</td>
<td>0.000000</td>
<td>0.00000</td>
<td>17.19038</td>
</tr>
<tr>
<td>Pushkin, A.</td>
<td>16000</td>
<td>1</td>
<td>4.204120</td>
<td>0.000000</td>
<td>0.00000</td>
<td>17.67462</td>
</tr>
<tr>
<td>Williams, Tennessee</td>
<td>20000</td>
<td>1</td>
<td>4.301030</td>
<td>0.000000</td>
<td>0.00000</td>
<td>18.4986</td>
</tr>
<tr>
<td>Christie, Agatha</td>
<td>25000</td>
<td>1</td>
<td>4.397940</td>
<td>0.000000</td>
<td>0.00000</td>
<td>19.34188</td>
</tr>
<tr>
<td>Shakespeare, W.</td>
<td>64000</td>
<td>1</td>
<td>4.806180</td>
<td>0.000000</td>
<td>0.00000</td>
<td>23.09937</td>
</tr>
</tbody>
</table>

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 equal to 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 calculate K-S test as follow:

```r
>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


