Lotka's Law Revisited

WILLIAM GRAY POTTER

Introduction

THE ORIGINAL STATEMENT of what has come to be known as Lotka's law was made in Lotka's 1926 journal article, "The Frequency Distribution of Scientific Productivity": "...the number (of authors) making *n* contributions is about $1/n^2$ of those making one; and the proportion of all contributors, that make a single contribution, is about 60 per cent."[1] To derive his "inverse square law," Lotka used comprehensive bibliographies in chemistry and physics and plotted the percentage of authors making 1, 2, 3,...*n* contributions against the number of contributions with both variables on a logarithmic scale. He then used the least-squares method to calculate the slope of the line that best fit the plotted data, and he found that the slope was approximately -2.

Since the publication of Lotka's original article in 1926, much research has been done on author productivity in various subject fields. The publications arising from this research have come to be associated with Lotka's work and are often cited as proving or supporting his findings. However, a review of this literature reveals that Lotka's article was not cited until 1941, that his distribution was not termed "Lotka's law" until 1949, and that no attempts were made to test the applicability of Lotka's law to other disciplines until 1973. The present article will discuss the literature that has become associated with Lotka's law and will attempt to identify the important factors of Lotka's original methodology which should be considered when attempting to test the applicability of Lotka's law.

William Gray Potter is Acquisitions Librarian, University of Illinois Library at Urbana-Champaign.

Applying Lotka's Law

Russell C. Coile in 1977 admonished investigators who, "studying the applicability of 'Lotka's law' to the humanities and to map librarianship, may have misinterpreted Lotka's law and concluded erroneously that the law applies to these fields."[2] In a cogent exposition, Coile detailed the derivation of Lotka's law in Lotka's original article. He then proceeded to test the applicability of Lotka's law to data from Murphy's 1973 study of the humanities[3] and Schorr's 1975 study of map librarianship[4] using the Kolmogorov-Smirnov statistic. In both cases, it was found that, contrary to the authors' claim, Lotka's law did not apply to the observed data. Coile attributes Lotka's erroneous conclusion to a misinterpretation of Lotka's formulation, to the inclusion of coauthors (whereas Lotka counted only the senior author), and to the failure to use an appropriate statistical test of significance. Schorr also counted coauthors and then used the chi-square test to determine if Lotka's law held. Coile contends that the chi-square test is not an appropriate test in this case because the table entries for authors with five to nine contributions show fewer than five observations.

The reason these data do not fit Lotka's law may be simply that Lotka's law does not apply in the fields studied. However, the scope of the studies by Murphy and Schorr does not apppear to be comparable to that of Lotka's work. Lotka drew 6891 names from the 1907-16 *Decennial Index to Chemical Abstracts*[5] and 1325 names from Auerbach's *Geschichtstafeln der Physik*, which included outstanding contributions in physics throughout history up to 1900.[6] Murphy took 170 authors drawn from the first decade of *Technology and Culture*. Schorr used 326 authors publishing between 1921 and 1973 on map librarianship based on a bibliography he had compiled earlier. The bibliographic sources used by Murphy and Schorr do not approach the coverage, in terms of either subjects or time, of the sources used by Lotka. The same objections can also be applied to Schorr's 1974 study of library science[7] and Voos's 1974 study of information science.[8]

In order to test the applicability of Lotka's law to a set of data, a statistical test is needed. Coile recommends the Kolmogorov-Smirnov (K-S) statistic. The K-S test detemined the maximum deviation, D:

$$D = \text{Max} \mid F_o\ (X - S_n\ (X)\ ) \mid$$

where $F_o(X)$ is the theoretical cumulative frequency function and $S_n(X)$ is the observed cumulative frequency function of a sample of n observations. At a 0.01 level of significance, the K-S statistic is equal to $1.63/n^2$.

If D is greater than the K-S statistic, then the sample distribution does not fit the theoretical distribution.

The K-S statistic was used here to test the fit of Lotka's data to the law that now bears his name. Using Lotka's law as the theoretical distribution and the data from Lotka's study of *Chemical Abstracts* and Auerbach's *Geschichtstafeln der Physik* as the observed data, it was found that a portion of Lotka's data does not fit his law. As shown in table 1, D from the *Chemical Abstracts* data is 0.0287, and the K-S statistic is $1.63/\sqrt{6981}$ or 0.0195. The value of D is greater, and therefore Lotka's law does not apply to Lotka's sample from *Chemical Abstracts*. With the Auerbach figures, D is 0.0253 and the K-S statistic is $1.63/\sqrt{1325}$ or 0.0448 (see table 2). The value of D is less, and therefore Lotka's law does apply to Lotka's figures from Auerbach's *Geschichtstafeln der Physik*. Lotka's law, then, applies to only a portion of his data.

TABLE 1
LOTKA, *Chemical Abstracts* DATA
PROPORTION OF AUTHORS

No. Contributions	*Observed*	$S_n(X)$	*Expected*	$F_o(X)$	$\lvert F_o(X) - S_n(X)\rvert$
1	0.5792	0.5792	0.6079	0.6079	0.0287
2	0.1537	0.7329	0.1520	0.7599	0.0270
3	0.0715	0.8044	0.0675	0.8274	0.0230
4	0.0416	0.8460	0.0380	0.8654	0.0194
5	0.0267	0.8727	0.0243	0.8897	0.0170
6	0.0190	0.8917	0.0169	0.9066	0.0149
7	0.0164	0.9081	0.0124	0.9190	0.0109
8	0.0123	0.9204	0.0095	0.9285	0.0081
9	0.0093	0.9297	0.0075	0.9360	0.0063
10	0.0094	0.9391	0.0061	0.9421	0.0030

$D = \text{Max}\,\lvert F_o(X) - S_n(X)\rvert = 0.0287$
At 0.01 level of significance, K-S statistic $= 1.63/\sqrt{6891} = 0.0195$
$D > 0.0195$
Therefore, data from *Chemical Abstracts* do not fit Lotka's law.

It should be stressed that Lotka's inverse square law is a general, theoretical estimate of productivity. The appeal of a hard and fast distribution cannot be denied. However, Lotka's law is not a precise statistical distribution. Rather, it is a generalization based upon two samples.

TABLE 2
LOTKA, AUERBACH DATA
PROPORTION OF AUTHORS

No. Contributions	*Observed*	$S_n(X)$	*Expected*	$F_o(X)$	$\|F_o(X) - (S)_n(X)\|$
1	0.5917	0.5917	0.6079	0.6079	0.0162
2	0.1540	0.7457	0.1520	0.7599	0.0142
3	0.0958	0.8415	0.0675	0.8274	0.0141
4	0.0377	0.8792	0.0380	0.8654	0.0138
5	0.0249	0.9041	0.0243	0.8897	0.0144
6	0.0211	0.9252	0.0169	0.9066	0.0186
7	0.0143	0.9395	0.0124	0.9190	0.0205
8	0.0143	0.9538	0.0095	0.9285	0.0253
9	0.0045	0.9583	0.0075	0.9360	0.0223
10	0.0053	0.9636	0.0061	0.9421	0.0215

$D = \text{Max}\,|F_o(X) - SX)| = 0.0253$
At 0.01 level of significance, K-S statistic $= 1.63/\sqrt{1325} = 0.0448$
$D < 0.0448$
Therefore, the Auerback data fit Lotka's law.

Given Coile's analysis of the work of Murphy and Schorr, and given that even Lotka's data do not exactly fit his inverse square law, it would be useful to examine the literature on and associated with Lotka's law. Coile emphasizes that for statistical comparisons to be made to Lotka's work, Lotka's methodology should be followed. This leads to the problem of identifying which of the factors of Lotka's methodology are most significant. In the following review of the literature, an attempt is made to identify these factors.

Literature of Lotka's Law

Many discussions of Lotka's law begin with a statement to the effect that the distribution has previously been shown to hold in various subject fields. Turkeli, Krisciunas, Hubert, and Allison and Stewart are examples.[9] To quote from some of these authors:

> It (Lotka's law) has been shown to hold for the productivity patterns of chemists, physicists, mathematicians, and econometricians.[10]

> The productivity of scientists has been a subject of inquiry ever since the pioneering investigation of Lotka, and others have since carried out Lotka's type of investigation.[11]

> Lotka's "inverse square law" of scientific productivity has since been shown to fit data drawn from several widely varying time periods and disciplines.[12]

While some of these studies do not cite sources, those that do often cite Derek de Solla Price's *Little Science, Big Science*.[13] Those that go beyond Price cite Dresden, Dufrenoy, Davis, Williams, Zipf, Leavens, and Simon.[14] Several authors, following Price's lead, have assumed Lotka's law to have been proved and have proceeded to discuss why the distribution occurs, i.e., why some authors produce more or less than others. These include later works by Price, Bookstein, Allison et al., and Shockley.[15] These efforts to explain and refine Lotka's formulation are interesting and valuable. In looking at the work of these authors, however, it appears that some misunderstanding has developed, for, in fact, most of the studies cited as demonstrating Lotka's law do not mention Lotka and do not offer comparable data.

Dresden is the earliest author cited in relation to Lotka's law.[16] Although Hubert refers to Dresden's article as "subsequent" to Lotka's work,[17] it did, in fact, appear in 1922. Dresden lists authors who presented papers at the regular meetings of the Chicago section of the American Mathematical Society (AMS). While Dresden does mention that 59 percent of the papers were later published, he is not concerned with the publishing behavior of the authors involved. Hubert claims that Dresden studied the output of "American mathematicians." Actually, the authors studied were members of a regional section of AMS. Dresden's purpose is to provide a record of the work of the Chicago section of the AMS, not to make a generalization about the productivity of mathematicians. To do so from Dresden's figures would be misleading, because the Chicago section of the AMS may not be representative of all mathematicians, and because the figures apply to presented papers, not publications. Dresden's work is interesting, but its relation to Lotka's law is questionable.

Dufrenoy attempted to study the publishing behavior of biologists by anlayzing the index to the *Review of Applied Mycology* for 1932, 1934 and 1935, and papers published in volumes 115, 118 and 120 of *Comptes Rendus de la Société de Biologie* (1932, 1934, 1935).[18] He is interested in the publishing behavior of biologists on an annual basis, not in the rate of productivity over time as Lotka is. Dufrenoy does not even cite Lotka, let alone attempt to apply Lotka's inverse square to his data.

Davis in 1941[19] is the first author to cite Lotka in the fifteen years following Lotka's original article. He also used Dresden's data, thus linking the two authors. Davis was interested in presenting data to show that the distribution of individuals in one of a variety of endeavors would approximate a Pareto distribution when the measure of that endeavor is sufficiently large. The ability to publish is one such endeavor. Another example used by Davis plots the billiards scores of seventy-nine faculty members at Indiana University. Davis plots the data from Lotka and Dresden and finds that they resemble the Pareto distribution, although the slope of their data is closer to -2 than to the expected Pareto exponent of -1.5. No statistical tests for goodness of fit are applied. Davis offers no new data on author productivity and is not concerned that Dresden is describing papers presented at meetings, while Lotka is describing published articles. He does provide a valuable service by citing both Dresden and Lotka for the benefit of later researchers. (Incidentally, the slope for the plotted billiards scores is -1.867.)

Williams uses Dufrenoy's data from the *Review of Applied Mycology* for 1935 and compiles his own figures from volume 1 (1913) and volume 24 (1936) of the *Review of Applied Entomology*.[20] As with Dufrenoy, Williams analyzes publishing behavior of authors in individual years of individual journals and does not discuss the rates of author productivity over time. Williams also does not cite Lotka and does not appear to be familiar with Lotka's work.

In *Human Behavior and the Principle of Least Effort*, Zipf has a chapter titled "The Distribution of Economic Power and Prestige." Zipf discusses the authorship of scientific articles as an indication of prestige and cites Lotka, Dresden and Davis. Zipf is the first to call the inverse square rule "Lotka's law" and discusses it as an approximation, not a rigid distribution. Accepting Lotka's formulation and Davis's interpretation of Dresden, Zipf also speculates on why some authors publish more than others.[21] No new data are presented and no statistical tests are made of the available data of Dresden and Lotka.

Leavens in 1953 based his study of econometricians on the work of 721 authors who presented papers at meetings of the Econometric Society or had articles published in the first twenty volumes of *Econometrica* (1933-52). He does not cite or mention Lotka. While his data cover an extensive period of time, they represent only one journal in a relatively small field compared to Lotka's study of physics and chemistry. Leavens counts unpublished papers read at meetings and counts all authors where Lotka counted only the senior author. Still, using the K-S test, Leavens's data do fit Lotka's law (see tables 3 and 4).[22] The major factor that Lotka and Leavens have in common is that both of their studies cover a substantial period of time.

TABLE 3
LEAVENS, PAPERS PRESENTED AT MEETINGS OF THE ECONOMETRICS SOCIETY OR IN *Econometrica*, 1933-52

No. Contributions	*No. Contributors*	*% Contributors*	*Total No. Contributions*
1	436	60.47	436
2	107	14.84	214
3	61	8.46	183
4	40	5.55	160
5	14	1.94	70
6	23	3.19	138
7	6	0.83	42
8	11	1.53	88
9	1	0.14	9
11	4	0.55	44
12	2	0.28	24
13	3	0.42	39
14	2	0.28	28
16	1	0.14	16
17	2	0.28	34
18	1	0.14	18
23	1	0.14	23
24	1	0.14	24
28	2	0.28	56
30	1	0.14	30
37	1	0.14	37
46	1	0.14	46
TOTAL	721	100.00	1,759

Simon, in an article appearing in *Biometrika* in 1955 and reprinted in his *Models of Man* in 1957, cites Davis and Leavens.[23] In observing how these and other data culled from many sources and involving word frequencies, city sizes and income distribution fit the Yule distribution, Simon uses the figures compiled by Lotka and Dresden, but cites neither writer directly and does not mention Lotka. Rather, he provides a reference to Davis. Lotka is listed in the index to *Models of Man*, but for an article on a different topic. Establishing a theoretical distribution for the data from Lotka, Dresden and Leavens, Simon claims that "the fit is reasonably good" without applying any statistical tests. As with Davis and Zipf, Simon offers no new data and does not attempt to find statistical support for what has become known as Lotka's law.

In 1963, Price's *Little Science, Big Science* appeared. Price claims that Lotka and several others have shown that whenever data are drawn from an index extending: "over a number of years sufficient to enable

TABLE 4
LEAVENS
PROPORTION OF AUTHORS

No. Contributions	*Observed*	$S_n(X)$	*Expected*	$F_o(X)$	$\lvert F_o(X) - S_n(X) \rvert$
1	0.6047	0.6047	0.6079	0.6079	0.0032
2	0.1484	0.7531	0.1520	0.7599	0.0068
3	0.0846	0.8377	0.0675	0.8274	0.0103
4	0.0555	0.8932	0.0380	0.8654	0.0278
5	0.0194	0.9126	0.0243	0.8897	0.0229
6	0.0319	0.9445	0.0169	0.9066	0.0379
7	0.0083	0.9528	0.0124	0.9190	0.0338
8	0.0153	0.9681	0.0095	0.9285	0.0396
9	0.0014	0.9695	0.0075	0.9360	0.0335

$n = 721$
$D = \text{Max} \lvert F_o(X) - S_n(X) \rvert = 0.0396$
At the 0.01 level of significance, K-S statistic $= 1.63/\sqrt{721} = 0.0607$
$D < 0.0607$
Therefore, Lotka's law holds for Leaven's data.

those who can produce more than a couple of papers to do so,...the result...is an inverse square law of productivity."[24] He discussed Lotka's data from *Chemical Abstracts* and refers the reader to Simon for "a fuller analysis and justification." Price plots data from an analysis of the abridged *Philosophical Transactions of the Royal Society of London* for the seventeenth and early eighteenth centuries. He suggests that these new data fit Lotka's law, but he does not provide the actual figures or perform a statistical test for goodness of fit. Price's principal interest is in discussing how to modify Lotka's law in order to account accurately for authors of high productivity, i.e., those who produce fifteen or more papers. This refinement is necessary, Price says, "since otherwise the maximum scores of published papers in a lifetime would be thousands and even tens of thousands rather than the several hundreds that seem to represent even the most prolific scientific lives."[25] The modification of Lotka's law is, as mentioned earlier, the subject of several articles, notably those by Bookstein and by Allison et al.[26]

In a 1969 review article, Fairthorne is the first to link the distributions of Bradford, Zipf, Mandelbrot, and Lotka. While he does not cite Price, Fairthorne does mention that Lotka's "relation underestimates the number of more prolific authors but applies fairly well for the less prolific."[27] Naranan and Bookstein also observe that many bibliometric distributions are essentially the same.[28]

With the exception of Leavens, no new data fitting Lotka's law are found in the above articles, and the figures from Leavens could be suspect. Yet presumably these studies are the ones invoked as proof of the applicability of Lotka's law by later authors, e.g., "It has been shown to hold for the productivity patterns of chemists, physicists, mathematicians, and econometricians."[29] In point of fact, no published article attempts to apply or test Lotka's law until Murphy in 1973. A critique of Murphy's article is provided by Coile and is described above; Hubert also faults Murphy.[30]

After Murphy, the next published application of Lotka's law is Voos in a 1974 study of information science. Taking his data from all articles indexed in *Information Science Abstracts* for 1966-70, Voos proposes that the inverse square law does not hold for information science and that -3.5 is a better constant for this particular discipline.[31] The error Voos makes is pointed out by Coile in a subsequent letter to the editor.[32] Voos lists the five years under study separately and then simply adds the tabulations for the individual years to arrive at a total for the five years: i.e., the number of authors publishing one paper in 1966, 1967, 1968, 1969, and 1970 were added together to arrive at a figure for all authors publishing one paper. Thus, an author publishing one paper per year would be credited with only one paper for the five years and not five, as he should be. As Coile points out, Voos is studying single years of data whereas Lotka studied a number of years. Like Dufrenoy, Voos defines an important area for research in analyzing author productivity on an annual basis.

Schorr has published three articles dealing with Lotka's law in library science, history of legal medicine,[33] and map librarianship. The faults of the last article are documented by Coile as described earlier. The first article is similarly flawed because, as Tudor points out in a subsequent letter to the editor,[34] Schorr uses only two journals, *College & Research Libraries* and *Library Quarterly*, for 1963-72. Schorr concludes that the data on the history of legal medicine do not fit Lotka's law. Tudor terms Schorr's article a "frivolous bagatelle," but it did reawaken interest in Lotka. However, the choice of such a restricted subject field constrasts sharply with Lotka's use of the topics of physics and chemistry.

Rogge attempts to apply Lotka's law to the literature of anthropology. He cites Lotka and claims that "Lotka's law has been tested positively many times."[35] Using the 40-year cumulative index of the *American Anthropologist* (1888-1928) and the 30-year cumulative index of *American Antiquity* (1935-65), Rogge concludes that "it was clear that at least this portion of the anthropological literature was produced in accordance with Lotka's law."[36] However, Rogge does not provide the data or even a summary of his statistical findings. Even with data, the study would cover only two periodicals and not the whole body of literature in anthropology.

The most recent attempt to apply Lotka's law was made in 1979 by Radhakrishnan and Kernizan in the field of computer science.[37] These authors studied papers published during 1968-72 in *Communications of the Association for Computing Machinery (CACM)* and in the *Journal of the ACM (JACM)*. The same objection applied to Schorr's and Rogge's articles applies here—data are drawn from two journals only. The authors admit that this is a problem but contend that their finding it noteworthy that, for a single journal, the fitted line will have a slope of approximately -3. This is, of course, interesting, and might be linked to Dufrenoy's and Williams's studies of a single journal. In a second experiment, the authors selected two random samples of three hundred authors, one sample each from *CACM* and *JACM*, and checked these authors in the cumulaive index to *Computer and Control Abstracts* covering 1969-72 to determine the number of publications per author. They found that Lotka's law did not apply, but wisely caution against drawing a "negative conclusion about the satisfaction of Lotka's law from this single experiment."[38] They go on to point out the need for a large-scale test of Lotka's law using a large, comprehensive machine-readable file, such as *Engineering Index*. To date, no such test has been reported.

Perhaps the most ambitious work to date in the study of Lotka's law has been done by Jan Vlachý. In an article appearing in 1972, Vlachý observes the role of several variables which might influence how appropriate Lotka's law is to a given set of data.[39] He examined bibliographies in many subject areas and listed the number of years covered by each source, the number of papers and authors represented, and the slope of the fitted line. While the data presented are interesting, Vlachý does not attempt to test the applicability of Lotka's law, nor does he provide sufficient data for others to perform statistical tests on his data. In this and a later article,[40] Vlachý discusses how the slope of the fitted line varies both according to the number of years covered and according to Vlachý's "division of the communities [of authors]...into universal, national, [international,] and those in journals."[41] Vlachý is mainly concerned with how these two variables affect the slope of the fitted line, i.e., the exponent in Lotka's formulation, and not with the appropriateness of Lotka's law. He also evaluated earlier studies as follows: "By analyzing the results of the previous studies, however, it was found that their scope and applicability is limited, since, first, their sampling background does not go much beyond the original data brought by Lotka and his early followers and, second, some basic concepts involved in these studies are anticipated without ever being thoroughly investigated."[42] Vlachý also compiled "A Bibliography of Lotka's Law and Related Phenomena."[43] This comprehensive bibliography lists works of interest not only on Lotka but also on the related laws of Bradford and Zipf, as well as bibliometrics and frequency distributions in general.

In a 1975 letter to the editor of the *Journal of Documentation*, Coile criticizes Kochen's discussion of authorship in the latter's *Principles of Information Retrieval*.[44] In this letter, Coile offers some useful insights into how the work of Leavens, Simon, Davis, and Dresden came to be associated with Lotka.[45]

Lotka's Law and Monograph Productivity

From this review of the literature, it can be argued that there have been no studies that replicate Lotka's methodology closely enough to be compared to Lotka's original work. Few of the authors of these studies should be faulted for this, because until Murphy's paper in 1973, no one attempted to compile new data to compare to Lotka's findings. Rather, earlier work by Dresden, Dufrenoy, Davis, Williams, and Leavens became associated with Lotka's work by subsequent authors and cited by some as providing proof of Lotka's law. Murphy, Schorr, Voos, and others in the 1970s sought to test Lotka's law in various disciplines, but failed to match the conditions under which Lotka conducted his study, usually because a suitable bibliographic source was not available.

Vlachý identified two variables which influence the distribution of author productivity: (1) the time period under study, and (2) the community of authors involved. None of the studies discussed above match Lotka's study in both these variables. Lotka's study covered ten years for the *Chemical Abstracts* figures, and all of history up to 1900 for Auerbach. Those that do match or surpass Lotka in time period, notably Rogge, do not match him in the selection of a community of authors. In Lotka's study of *Chemical Abstracts*, the community consists of all senior authors whose work was included in the 1907-16 decennial index. In his study of Auerbach, the community of authors consists of authors of the most notable works in the field of physics up to 1900. In most studies of author productivity, it is usually the subject field that defines a community of authors, because that is how journals and bibliographies are organized and because researchers are often interested in studying a particular field. Most subsequent studies single out one or two journals or study only a few years. These works are often significant in and of themselves, and contribute greatly to our understanding of author productivity and behavior. However, they should not be compared to Lotka's work without much caution.

There have been two recent studies which might be comparable to Lotka's work in terms of the time period and the community of authors. However, both deal with monographic literature, not journal articles. One is a study done by the Library of Congress (LC) of all author headings on its MARC tapes.[46] The other is a study of personal authors in the University of Illinois Library card catalog.[47] Both studies differ from Lotka's in that all authors, not just the senior authors, are counted. Lotka never discloses why he counted only senior authors. A look at the first decennial index to *Chemical Abstracts* reveals a possible explanation. If an article has four or fewer authors, all authors are indexed. However, the second, third and fourth authors will have only a "see" reference to the first author, not to the number of articles written by the authors together. Thus, to compile all authors, Lotka would have had to refer to the first author. A quick sample shows that over 20 percent of the author entries have "see" references. Considering that Lotka tabulated all authors whose surnames began with *A* or *B*, and that from 272 pages this resulted in 6891 authors, it is not surprising that he might have balked at this added chore.

The data from the University of Illinois Library catalog are shown in table 5. The Illinois catalog contains records for about 2.5 million titles. A random sample of 2345 personal authors was drawn. Plotting the first 29 observations on a log scale, the slope for the data is -2.0903, very close to Lotka's theoretical slope. The K-S test in table 6 shows that the Illinois data do indeed fit. It should be pointed out that the five most prolific authors in the Illinois study are Shakespeare, Milton, Goethe, Balzac, and Dickens. None of these authors write currently, but their works continue to be published, a feature Lotka did not face.

The LC study of its MARC tapes covers 1,336,182 machine-readable catalog records established between 1969 and 1979, with 695,074 unique personal name headings. The results are shown in table 7. Plotting the first 10 points, the slope of the data is -2.3450. Intuitively, this will not fit Lotka's theoretical distribution. Applying the K-S test to the first observation, D is 0.6565 - 0.6079 = 0.0486; the K-S statistic is $1.63/\sqrt{695{,}074} = 0.0020$. The value of D is greater than the K-S statistic; therefore, the data do not fit Lotka's law.

TABLE 5
UNIVERSITY OF ILLINOIS LIBRARY AT URBANA-CHAMPAIGN
STUDY OF PERSONAL AUTHORS IN THE CARD CATALOG

No. Works	*No. Authors*	*% Total Sample*	*Total No. Entries*
1	1,489	63.50	1,489
2	343	14.63	686
3	160	6.82	480
4	92	3.92	368
5	44	1.88	220
6	35	1.49	210
7	27	1.15	189
8	18	0.77	144
9	12	0.51	108
10	11	0.47	110
11	10	0.43	110
12	9	0.38	108
13	2	0.09	26
14	6	0.26	84
15	9	0.38	135
16	8	0.34	128
17	3	0.13	51
18	2	0.09	36
19	2	0.09	38
20	5	0.21	100
21	5	0.21	105
22	1	0.04	22
23	1	0.04	23
24	2	0.09	48
26	1	0.04	26
27	1	0.04	27
28	4	0.17	112
30	2	0.09	60
31	1	0.04	31
32	3	0.13	96
33	1	0.04	33
34	1	0.04	34
35	1	0.04	35
36	3	0.13	108
38	2	0.09	76
39	1	0.04	39
40	2	0.09	80
42	2	0.09	84
44	2	0.09	88
47	1	0.04	47
48	1	0.04	48
49	1	0.04	49
51	1	0.04	51
58	1	0.04	58
63	1	0.04	63
66	1	0.04	66

TABLE 5—*Continued*

No. Works	*No. Authors*	*% Total Sample*	*Total No. Entries*
70	1	0.04	70
90	1	0.04	90
111	1	0.04	111
115	1	0.04	115
149	1	0.04	149
167	1	0.04	167
231	1	0.04	231
266	1	0.04	266
298	1	0.04	298
379	1	0.04	379
592	1	0.04	592
652	1	0.04	652
835	1	0.04	835
1,374	1	0.04	1,374
1,490	1	0.04	1,490
TOTALS	2,345	100.00	13,148

TABLE 6
UNIVERSITY OF ILLINOIS LIBRARY AT URBANA-CHAMPAIGN
PROPORTION OF AUTHORS

Titles/ Author	*Theoretical (Lotka)*	$F_o(X)$	*Observed (Illinois)*	$S_n(X)$	$\|F_o(X) - S_n(X)\|$
1	0.6079	0.6079	0.6350	0.6350	0.0271
2	0.1520	0.7599	0.1463	0.7813	0.0214
3	0.0675	0.8274	0.0682	0.8495	0.0221
4	0.0380	0.8654	0.0392	0.8887	0.0233
5	0.0243	0.8897	0.0188	0.9075	0.0178
6	0.0169	0.9066	0.0149	0.9224	0.0158
7	0.0124	0.9190	0.0115	0.9339	0.0149
8	0.0095	0.9285	0.0077	0.9416	0.0131
9	0.0075	0.9360	0.0051	0.9467	0.0107

$D = \text{Max}\,|F_o(X) - S_n(X)| = 0.0271$
At the 0.01 level of significance, K-S statistic $= 1.63/\sqrt{2345} = 0.0337$
$D < 0.0337$
Therefore, UI Library data fit Lotka's law.

Why the LC figures do not fit, while the Illinois figures do, is open to conjecture. One reason might be that the LC data include persons occurring as subjects as well as authors. Another possible cause is that

TABLE 7
LIBRARY OF CONGRESS ANALYSIS OF PERSONAL NAME HEADINGS ON MARC TAPES

No. Occurrences	*No. Distinct Headings*	*% Distinct Headings*
1	456,328	65.65
2	119,681	17.22
3	46,247	6.65
4	23,951	3.45
5	13,820	1.99
6	8,790	1.26
7	5,827	0.84
8	4,056	0.58
9	2,998	0.43
10	2,153	0.31
11-13	4,116	0.59
14-20	3,748	0.54
21-50	2,678	0.39
51-100	448	0.06
101-200	149	0.02
201-300	47	0.01
301-400	19	0.00
401-500	11	0.00
501-1000	5	0.00
1001+	2	0.00
Total	695,074	99.99

the Illinois figures cover authors from the beginning of history to the present, while LC figures cover catalog records established over ten years. This could also be the reason Lotka's Auerbach figures fit, but not the *Chemical Abstracts* data. In any event, the fact that an exact fit is lacking in the Library of Congress figures is not as important as the emergence of a general rule which implies that a sufficiently large sample of a broad community of authors and a large time span will approximate Lotka's law.

It is of further interest to note that both the LC and Illinois figures were compiled for a practical management problem—planning for the implementation of the second edition of the Anglo-American Cataloging Rules. It is not uncommon for other bibliometric formulations to be used for practical planning, notably Bradford's distribution for planning periodical collections. This, however, is the first known case where Lotka's law has been useful in planning.

Conclusion

It has been seen that Lotka's law fits only a portion of the data from his 1926 study and that his most-cited figures, those for *Chemical Abstracts* from 1907 to 1916, do not fit his distribution. Later studies assume that Lotka's law had been proven to apply in a variety of subject areas, when in fact it had not. No data were compiled for the express purpose of verifying the law until the 1970s, and these recent studies, while valuable and useful, are not comparable to Lotka's study in terms of the time period covered and the community of authors involved.

Recent studies of monograph productivity suggest that Lotka's law might reflect an underlying pattern in the behavior of those people who produce publications, whether those publications are books or journal articles. It would appear that when the time period covered is ten years or more and the community of authors is defined broadly, author productivity approximates the frequenty distribution that Lotka observed and that has become known as Lotka's law. If this is correct, then there is a universal community of all authors who have ever published whose pattern of productivity might approximate Lotka's law. Within this universal community, there are many subcommunities defined, as Vlachý points out, by discipline, nation, institution, journal, etc. Even time could be used as a dimension to define a subcommunity. All studies of author productivity are concerned with a subset of the universal community of authors. The smaller the subset, the less likely it will be that the measurements of productivity reflect the measurements for the universal community, although these measurements may be useful and valuable in studying that particular subset. However, the larger and more representative the subset, the more closely it will resemble the universal community. The subsets studied by Lotka and those represented in the Library of Congress study of its MARC tapes and in the study of the University of Illinois Library card catalog are the largest yet considered, and the similarity of their patterns of author productivity and behavior suggest that broader patterns do indeed exist.

The above review of literature associated with Lotka's law suggests several areas for future research. First, the work of Dufrenoy and others on the annual productivity of authors points to an interesting measure of author behavior. Second, Radhakrishnan and Kernizan make a convincing argument for the use of large-scale machine-readable data bases in the study of author productivity. They suggest that the machine version of *Engineering Index* could be used, and this would be especially interesting in that *Engineering Index* is a multidisciplinary data base with records that are well indexed. Thus, subsets could be defined by a number of factors—subject, date, country, etc.—and the productivity of authors within these subsets could be determined and compared relatively easily. Studies of large bibliographic data bases could also lead to some standardization of methodology. Third, the concept derived from Vlachý of a universal community of authors needs to be explored further. Given that such a universal community exists, and that all studies of author productivity are based upon subsets, or subcommunities, of this universal community, then some work could be done on which factors used to define the subsets are most important—i.e., time, subject, language, format of publication, etc. Finally, the use of a univariate model like Lotka's law, where the response of one variable to another is measured, may oversimplify the complex subject of author productivity. The factors mentioned above that serve to define communities of authors, as well as other factors, might be included as variables in a more sophisticated model for measuring and predicting author productivity. More complex models will be more difficult to understand, but the inclusion of relevant variables in a multivariate model may result in a model that better simulates reality and thus is more useful.

References

1. Lotka, Alfred J. "The Frequency Distribution of Scientific Productivity." *Journal of the Washington Academy of Sciences* 16(19 June 1926):323.

2. Coile, Russell C. "Lotka's Frequency Distribution of Scientific Productivity." *Journal of the ASIS* 28(Nov. 1977):366.

3. Murphy, Larry J. "Lotka's Law in the Humanities?" *Journal of the ASIS* 24(Nov.-Dec. 1973):461-62.

4. Schorr, Alan E. "Lotka's Law and Map Librarianship." *Journal of the ASIS* 26(May-June 1975):189-90.

5. *Decennial Index to Chemical Abstracts*, vols. 1-10, 1907-1916. Easton, Pa.: American Chemical Society, n.d.

6. Auerbach, Felix. *Geschichtstafeln der Physik*. Leipzig: Barth, 1910.

7. Schorr, Alan E. "Lotka's Law and Library Science." *RQ* 14(Fall 1974):32-33.

8. Voos, Henry. "Lotka and Information Science." *Journal of the ASIS* 25(July-Aug. 1974):270-72.

9. Turkeli, Arif. "The Doctoral Training Environment and Post-Doctorate Productivity Among Turkish Physicists." *Science Studies* 3(1973):311-18; Krisciunas, Kevin. Letter to the editor in *Journal of the ASIS* 28(Jan. 1977):65-66; Hubert, John J. Letter to the editor in *Journal of the ASIS* 28(Jan. 1977):66; and Allison, Paul D., and Stewart, John A. "Productivity Differences among Scientists: Evidence for Accumulative Advantage." *American Sociological Review* 39(Aug. 1974):596-606.

10. Krisciunas, letter to the editor, p. 65.

11. Turkeli, "Doctoral Training Environment," p. 311.

12. Allison and Stewart, "Productivity Differences among Scientists," p. 596.

13. Price, Derek de Solla. *Little Science, Big Science*. New York: Columbia University Press, 1963.

14. Dresden, Arnold. "A Report on the Scientific Work of the Chicago Section, 1897-1922." *American Mathematical Society Bulletin* 28(July 1922):303-07; Dufrenoy, Jean. "The Publishing Behavior of Biologists." *Quarterly Review of Biology* 13(June 1938):207-10; Davis, Harold T. *The Analysis of Economic Time Series*. Bloomington, Ind.: Principia Press, 1941. *See also* ________. *Theories of Econometrics*. Bloomington, Ind.: Principia Press, 1941, pp. 45-50; Williams, C.B. "The Numbers of Publications Written by Biologists." *Annals of Eugenics* 12(1944):143-46; Zipf, George K. *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. Cambridge, Mass.: Addison-Wesley, 1949; Leavens, Dickson H. Letter to the editor in *Econometrica* 21(Oct. 1953):630-32; and Simon, Herbert A. "On a Class of Skew Distribution Functions." *Biometrika* 42(Dec. 1955):425-40; reprinted in ________. *Models of Man*. New York: Wiley, 1957.

15. Price, Derek de Solla, and Gursey, S. "Studies in Scientometrics. Part I. Transcience and Continuance in Scientific Authorship." *International Forum on Information Documentation* 1(1976):17-24; Bookstein, Abraham. "The Bibliometric Distributions." *Library Quarterly* 46(Oct. 1976):416-23; Bookstein, Abraham. "Patterns of Scientific Productivity and Social Change." *Journal of the ASIS* 28(July 1977):206-10; Allison, Paul D. et al. "Lotka's Law: A Problem in Its Interpretation and Application." *Social Studies of Science* 6(1976):269-76; and Shockley, William. "On the Statistics of Individual Variations of Productivity in Research Laboratories." *Proceedings of the Institute of Radio Engineers* 45(March 1957):279-90.

16. Dresden, "Report on the Scientific Work."

17. Hubert, letter to the editor, p. 66.

18. Dufrenoy, "Publishing Behavior of Biologists."

19. Davis, *Analysis of Economic Time Series*.

20. Williams, "Publications Written by Biologists."

21. Zipf, *Human Behavior*.

22. Leavens, letter to the editor.

23. Simon, "On a Class of Skew Distribution Functions."

24. Price, *Little Science*, p. 43.

25. Ibid., pp. 48-49.

26. Bookstein, "Patterns of Scientific Productivity"; and Allison, et al., "Lotka's Law."

27. Fairthorne, Robert A. "Progress in Documentation." *Journal of Documentation* 25(Dec. 1969):325.

28. Naranan, S. "Power Law Relations in Science Bibliography—A Self-Consistent Interpretation." *Journal of Documentation* 27(June 1971):83-97; and Bookstein, "Bibliometric Distributions."

29. Krisciunas, letter to the editor, pp. 65-66.

30. Coile, "Lotka's Frequency Distribution"; and Hubert, letter to the editor.

31. Voos, "Lotka and Information Science."

32. Coile, Russell C. Letter to the editor in *Journal of the ASIS* 26(March-April 1975):133.

33. Schorr, Alan E. "Lotka's Law and the History of Legal Medicine." *Research in Librarianship* 30(Sept. 1975):205-09.

34. Tudor, Dean. Letter to the editor in *RQ* 14(Winter 1974):187.

35. Rogge, A.E. "A Look at Academic Anthropology." *American Anthropologist* 78(Dec. 1976):835.

36. Ibid.

37. Radhakrishnan, T., and Kerinzan, R. "Lotka's Law and Computer Science Literature." *Journal of the ASIS* 30(Jan. 1979):51-54.

38. Ibid., p. 54.

39. Vlachý, Jan. "Variable Factors in Scientific Communities (Observations on Lotka's Law)/" *Teorie a Metoda* 4(1972):91-120.

40. ________. "Time Factor in Lotka's Law." *Probleme de Informare si Documentare* 10(1976):44-87.

41. Ibid., p. 48.

42. Ibid., p. 46.

43. ________ , comp. "Frequency Distribution of Scientific Performance: A Bibliography of Lotka's Law and Related Phenomena." *Scientometrics, Bibliography Section* 1(1978):109-30.

44. Coile, Russell C. Letter to the editor in *Journal of Documentation* 31(Dec. 1975):298-301; *see also* Kochen, Manfred. *Principles of Information Retrieval.* Los Angeles: Melville, 1974.

45. Other works which cite and discuss Lotka to some extent include: Aiyepeku, Wilson O. "The Productivity of Geographical Authors: A Case Study from Nigeria." *Journal of Documentation* 32(June 1976):105-17; Cole, Jonathan R., and Cole, Stephen. "The Ortega Hypothesis." *Science* 178(Oct. 1972):368-75; Mantell, Leroy H. "On Laws of Special Abilities and the Production of Scientific Literature." *American Documentation* 17(Jan. 1966):8-16; and Narin, Francis, et al. *Evaluative Bibliometrics: The Use of Publication and Citation Analysis in the Evaluation of Scientific Activity.* Cherry Hill, N.J.: Computer Horizons, Inc., 1976. (CH Project No. 704R)

46. McCallum, Sally H., and Godwin, James L. "Statistics of Headings in the MARC File." Network Development Office, Library of Congress, unpublished paper, 5 Jan. 1981.

47. Potter, William G. "When Names Collide: Conflict in the Catalog and AACR 2." *Library Resources & Technical Services* 24(Winter 1980):3-16.

This Page Intentionally Left Blank