Information Processing & Management Vol. 30, No. 2, pp. 239–251, 1994
Printed in Great Britain.

0306-4573/94 $6.00 + .00

THE POISSON-LOGNORMAL MODEL FOR BIBLIOMETRIC/SCIENTOMETRIC DISTRIBUTIONS

JOHN A. STEWART
Department of Sociology, University of Hartford, West Hartford, CT 06117

(*Received* 22 *December* 1992; *accepted in final form* 20 *April* 1993)

Abstract—The Poisson-lognormal model assumes that the intensity parameter of a Poisson process has a lognormal distribution in a sample of observations. This model can yield highly skewed, discrete distributions, but must be estimated by numerical methods. When applied to many of the empirical data sets related to the 'laws' of Lotka, Bradford, and Zipf, this compound Poisson model produces good to excellent fits. Discussion includes possible 'causal' processes and some implications for future bibliometric and scientometric studies.

Bibliometricians and scientometricians have shown considerable interest in applying and relating the 'laws' of Lotka, Bradford, and Zipf (e.g., Haitun, 1982a, 1982b; Chen & Leimkuhler, 1986, 1987a, 1987b; Nicholls, 1986, 1987a, 1989; Qiu, 1990). Among others, Haitun (1982a) suggests that these different laws can be reduced to a common form:

$$f_x = k/x^b, \quad x = 1,2,3,\ldots \tag{1}$$

where f_x is the number of scientists producing x articles (Lotka's law) or the number of journals containing x articles in a comprehensive bibliography on a particular topic (Bradford's law) or the number of words used x times in a document (Zipf's law). This equation gives the 'frequency form' of Zipf's law, where k and b are constants to be estimated (often by taking the log of eqn (1) to produce a linear log-log form). Lotka's 'inverse-square' law specifies that b is approximately 2.0, but allowing b to vary gives a 'generalized' form of Lotka's law. Haitun (1982a) states that Bradford's law reduces to the above form under certain conditions.

Attempts to test these laws have encountered a number of problems. First, there is considerable diversity in their precise formulation (e.g., Qiu, 1990, reviews and tests different forms of Bradford's law). Such variation in these laws often arises from pressures created by empirical tests; for example, extreme values might be ignored or treated differently (Pao, 1985; Brookes, 1969), additional parameters to the basic equations might be introduced (Brookes, 1969; Griffith, 1988), and adjustments are needed for the discrete nature of the counts and maximum possible scores (Tague & Nicholls, 1987). Second, researchers have used a variety of significance tests and estimation techniques, which range from simple inspections of graphs to maximum likelihood methods (Nicholls, 1987b). Third, the data sets often include a diverse mix of, say, scientists of different ages and disciplines, counts including or excluding coauthorships, or measurement over very different time intervals. This diversity makes it unlikely that stable (and comparable) parameters will be estimated. Finally, almost all of these applications use 'truncated' data sets missing the count for scientists with zero articles, $f_{x=0}$ (or non-used words or journals with no articles on the topic). This count is ignored because of data-collection difficulties and/or model-testing problems (e.g., division by zero in eqn 1 or taking the log of zero). Yet in a *representative* sample of, say, scientists, zero productivity would be both a legitimate *and* a conceptually important value. Thus, some researchers avoid the mathematical problems by adding a third parameter, c, and using $(x + c)^b$ in eqn (1).

A few researchers have fitted the lognormal distribution to the productivity of scientists (Shockley, 1957), to articles in different journals (Karmeshu *et al.*, 1984), or to word frequencies (Carroll, 1967); but these efforts also have problems. First, the lognormal is

a continuous distribution, but articles or word use are discrete events. Second, zero counts again present problems because the log of zero is negative infinity, so an adjustment, such as adding 0.5 to all scores, is needed. Finally, these two problems are particularly acute for the typical, 'reverse-J' shape, bibliometric distribution with a mode of one or zero. However, several factors favor the lognormal distribution. It can model very skewed distributions, it has simple parameters that are understandable in the context of normal distributions, and the mechanisms that might generate lognormal distributions have possible bibliometric/scientometric interpretations.

The Poisson-lognormal (PL) distribution retains the above advantages, but avoids the problems associated with fitting continuous, algebraic functions to discrete distributions that include $x = 0$ values. The PL is a compound distribution, where the underlying 'propensities', δ, to, say, publish an article follow a lognormal distribution across scientists. Given a *specific* scientist's publication propensity, δ, his or her probability of publishing x articles, P_x, follows a simple Poisson model:

$$P_x = \frac{\delta^x e^{-\delta}}{x!}, \quad x = 0,1,2,3,\ldots \tag{2}$$

The distribution of observed scores for all scientists *having the same* δ *value* will have a distribution with a mean *and* variance of δ.

In a sample of scientists, whose *logged* δs are normally distributed with a mean of μ and a standard deviation of σ, the P_x in the *total sample* is given by:

$$P_x = \frac{1}{\sigma\sqrt{2\pi}} \frac{1}{x!} \int_0^\infty e^{-\delta}\delta^{x-1} \exp\left\{-\frac{(\ln\delta - \mu)^2}{2\sigma^2}\right\} d\delta, \quad x = 0,1,2,3,\ldots \tag{3}$$

This equation must be estimated by numerical methods, which may explain why it has not been used for bibliometric distributions, even though it has been used in ecological research to estimate the distribution of species (e.g., Bulmer, 1974; Pielou, 1969; Dennis & Patil, 1988; Shaban, 1988). Note that x can take on values of zero without creating any estimation problems. If the zero values are missing from the data set, then a truncated distribution can be fit (Bulmer, 1974), which can estimate the number of scientists who had a greater than zero *propensity* to produce articles, but did not actually do so during the measurement interval. (A FORTRAN program based upon Bulmer's (1974) article is available from the author.)

In this article, I will illustrate that the PL distribution provides acceptable fits to a variety of bibliometric and scientometric distributions. I then consider some of the possible 'theories' or 'processes' that might produce a PL distribution for bibliometric data. Finally, I discuss some of the limitations of these analyses and possible implications and elaborations.

1. FITTING THE POISSON-LOGNORMAL DISTRIBUTION

Preliminaries

The criteria for selecting the empirical distributions included (a) preference for those with $x = 0$ frequencies and ungrouped score categories, (b) used or discussed in recent literature, (c) sampled from fairly specific populations and time periods, and (d) previously fit to the Generalized Inverse Gaussian-Poisson Distribution (GIGP) by Sichel (1975,1985, 1992a,1992b). This last criterion is important because the GIGP model provides the best empirical fits to a diverse set of relevant distributions.

Two 'significance' tests are used to assess the fit between the observed and predicted frequencies in the following tables. The chi-square test is commonly used, but it requires grouping of scores when the predicted scores fall much below 5.0. In particular, large samples can lead to high chi-square values, even though the fit is quite good on the basis of the second test, the Kolmogorov-Smirnov (K-S) test. The K-S test uses the maximum difference (D_{max}) between the observed and predicted cumulative probability distributions.

The critical values of the K-S test at the .20 level were calculated with the following formula:

$$D_{K\text{-}S} = 1.07/\sqrt{N + \sqrt{N/10}}, \quad (4)$$

where N is the sample size. This test does not require arbitrary groupings, is conservative for discrete distributions, is less sensitive to sample size, and provides a simple comparison of the fits for different theoretical models to the same distribution. However, both tests assume random sampling, which is seldom the case. For these reasons, I report the results of both tests, but regard them more as measures of fit than as tests of significance.

Lotka's law: Article productivity by scientists

The first distribution in Table 1 approaches the 'ideal test' because it is based upon random sampling from a well defined population (chemists receiving their Ph.D.s between 1955-61), coauthored articles are counted, zero productivity counts are included, and the measurement interval is similar for all chemists: the first six years after the Ph.D. degree. (See Reskin, 1977, for a description of the data set.) The PL distribution fits very well, even slightly better than the Generalized Inverse-Gaussian Poisson (GIGP) distribution (Sichel, 1985). Most empirical distributions lack $f_{x=0}$ information, but one can fit a truncated PL distribution, as done in Table 1 for the same distribution. This estimated the number of chemists with zero productivity to be 43.3 (with a standard error of 12.4, which includes the actual value of 37 chemists).

Table 1. Observed and estimated number of scientists producing x articles

Chemists[a]				Entomologists[b]		
x Observed number of articles	f_x Observed number of chemists	Estimated number of chemists using $f_{x=0}$	Estimated number of chemists W/O $f_{x=0}$	x Observed number of articles	f_x Observed number of scientists	Estimated number of scientists
0	37	38.8	–	1	320	311.9
1	50	46.3	48.9	2	92	111.5
2	37	39.1	39.8	3	63	54.7
3	31	29.6	29.4	4	32	31.8
4	24	21.5	21.1	5	24	20.5
5	13	15.5	15.1	6	10	14.2
6	10	11.2	10.9	7	11	10.4
7	7	8.2	7.9	8	7	7.8
8	7	6.0	5.9	9	7	6.1
9	3	4.5	4.4	10-11	10	8.8
10-11	5	6.0	5.9	12-13	7	6.0
12+	13	10.5	10.7	14-16	6	6.0
				17-20	8	5.0
				21-30	5	6.1
				31+	6	7.3
Total	237	237.1	200.0		608	608.0
Chi-square		3.01	2.46			9.58
df & p value		10 & .98	9 & .98			13 & .73
Observed D_{max}		.016	.014			.019
$D_{K\text{-}S}$ @ $p = .20$		.069	.075			.043
Mean log δ		.880	.826			−1.953
St. Dev. log δ		.866	.899			1.964
Mean RAW count		3.460	4.100			3.319
St. Dev. RAW count		3.789	3.794			5.734
Maximum x value		19	19			66

[a]The data set is described by Reskin (1977), but the distribution and score groupings are given in Allison (1980a) and Sichel (1985).
[b]The distribution is from Gupta (1987).

Table 1 also presents the article distribution from Gupta's (1987) study of publishers on Nigerian entomology during a 73-year period, and includes coauthored articles. The PL fit is still very good, and the D_{max} value is less than half of that obtained by Gupta, who used the 'generalized' Lotka law. The fit to this truncated distribution estimated that 1646 (with a standard error of 435) scientists had a greater than zero *propensity* to contribute to this literature, but failed to do so during this period.

The PL fit to the original Lotka (1926) data from *Chemical Abstracts* (not shown here) had an acceptable D_{max} ($p > .20$), but a chi-square probability less than .01, whereas the GIGP fit was excellent by both criteria (Sichel, 1985). However, Lotka's data pertain to a very diverse group, which is unlikely to be characterized by a shared mean and variance. Sichel (1985) had an excellent GIGP fit to one other productivity distribution (from Coile), but the PL fit (not shown) was equally good.

More extensive comparisons to the fits produced by Lotka's law are possible. Nicholls (1986) fit the generalized Lotka's law to 15 different productivity distributions using maximum likelihood estimation. The PL fit to these same distributions all had D_{max} scores with probabilities greater than .20, including the two distributions that Nicholls could not fit at the .01 level. All the obtained D_{max} scores were less than or equal to those obtained by Nicholls and averaged less than half.

Distribution of citations to articles and scientists

Despite some reservations, there is increasing interest in using citations to study the distribution of the 'recognition' or 'influence' of scientists (e.g., Smith, 1981; Seglen, 1992; Peritz, 1992). However, citations are generally to *specific articles*, which provide the most fundamental unit of analysis for citation studies. Thus Table 2 first gives the PL fits for

Table 2. Observed and estimated number of articles receiving x citations and number of chemists receiving x citations

Israeli physics articles[a]			1970 Geoscience articles[b]			Chemists[c]		
x Observed number of citations	f_x Observed number of articles	Estimated number of articles	x Observed number of citations	f_x Observed number of articles	Estimated number of articles	x Observed number of citations	f_x Observed number of chemists	Estimated number of chemists
0	60	54.5	0	7	8.1	0	102	100.5
1	34	43.2	1	11	9.4	1	36	38.0
2	27	30.7	2	8	8.5	2	16	20.3
3	28	22.1	3	5	7.3	3	15	12.9
4	16	16.4	4	10	6.2	4	9	9.1
5	12	12.6	5	6	5.3	5	6	6.8
6	7	9.9	6-7	8	8.4	6	6	5.3
7	13	7.9	8-9	4	6.3	7	6	4.3
8	9	6.4	10-12	7	6.8	8-9	11	6.5
9	6	5.3	13-16	6	6.1	10-12	4	6.7
10-11	8	8.2	17-22	6	5.7	13-17	7	6.9
12-13	9	6.0	23-32	6	5.1	18-24	6	5.6
14-16	4	6.4	33-65	5	5.4	25-33	5	4.1
17-20	2	5.6	66+	2	2.5	34+	10	12.3
21-29	6	6.8						
30-60	7	6.5						
61+	3	2.5						
Total	251	251.0		91	91.1		239	239.2
Chi-square		14.82			4.77			7.09
df & *p* value		15 & .464			12 & .97			12 & .85
Observed D_{max}		.0294			.0332			.0195
$D_{K\text{-}S}$ @ $p = .20$		.0669			.1104			.0685
Mean log δ		.890			1.733			−.026
St. Dev. log δ		1.374			1.276			2.164
Mean RAW count		6.092			12.648			Grouped data
St. Dev. RAW count		12.130			22.527			Grouped data
Maximum x value		99			153			Grouped data

[a]This distribution is from Arunachalam *et al.* (1984).
[b]The data set is described in Stewart (1987).
[c]The data are from Reskin (1977). The distribution is given in Allison (1980b) and Sichel (1985).

citations to articles in two disciplines (physics and geosciences). The third distribution is for citations to *individual chemists*. None of these distributions is zero-truncated.

Arunachalam *et al.*'s (1984) data on 251 articles published by Israeli physicists in 1977 includes citation counts for five years after publication. The PL fit has a chi-square probability of .46. This increases to .97 for the random sample of geoscience articles published in 1970 with citations counts from 1971–1974—see Stewart (1987). Sichel (1985) has not fit the GIGP distribution to citations *to articles*, only for the 1966 citations for Reskin's sample of chemists. He obtained an excellent fit, but as shown in the last column of Table 2, the PL model yields equally good results. However, these citations are only to first-author publications, and those chemists without such publications *must have zero citations*, and should be excluded from the distribution. Although this number is not available, Table 1 indicates that it should be approximately 37. The PL fit (not shown) to the chemists in Table 2 improves slightly if $f_{x=0}$ is changed to 65 (=102 − 37).

Bradford's law: Journal scattering

Bradford's law was developed to describe the distribution of articles on a specific subject among different journals. Table 3 gives the fit of the PL distribution to three such distributions, which have also been fit by the GIGP distribution (Sichel, 1985, 1992a). The PL distribution provides a good fit to all three distributions, and equals or beats the GIGP fit on all but the Bradford data, where the GIGP's chi-square probability is .83 (Sichel, 1985), compared to .56 for the PL distribution.

Table 3. Observed and estimated number of journals containing x articles in Bradford's geophysics, Kendall's operational research, and Rao's economics data sets

Bradford's data set[a]			Kendall's data set[b]			Rao's data set[c]		
x Observed number of articles	f_x Observed number of journals	Estimated number of journals	x Observed number of articles	f_x Observed number of journals	Estimated number of journals	x Observed number of articles	f_x Observed number of journals	Estimated number of journals
1	169	163.8	1	203	200.0	1	229	234.4
2	49	56.9	2	54	60.1	2	138	132.6
3	23	28.3	3	29	28.8	3	88	83.2
4	17	16.9	4	17	16.9	4	61	56.4
5	12	11.1	5	10	11.2	5	40	40.4
6	11	7.9	6	6	7.9	6	29	30.2
7	7	5.9	7	8	5.9	7	20	23.3
8	8	4.5	8	8	4.6	8	14	18.5
9–10	8	6.5	9	4	3.7	9	10	14.9
11–12	3	4.4	10–14	11	11.0	10	12	12.3
13–15	6	4.5	15–19	7	5.4	11	7	10.2
16–20	6	4.5	20–29	6	5.4	12	9	8.6
21+	7	10.9	30–49	2	4.1	13	11	7.4
			50+	5	5.0	14	6	6.4
						15–16	10	10.3
						17–18	8	8.0
						19–21	12	9.0
						22–24	8	6.6
						25–27	8	5.0
						28+	24	26.2
Total	326	326.1		370	370.0		744	743.9
Chi-square		9.66			6.15			9.07
df & p value		11 & .56			12 & .91			18 & .96
Observed D_{max}		.0246			.0159			.0127
$D_{K\text{-}S}$ @ $p = .20$		.0588			.0552			.0390
Mean log δ		−2.484			−4.212			.120
St. Dev. log δ		2.253			2.740			1.573
Mean RAW count		4.086			4.765			Grouped data
St. Dev. RAW count		9.049			16.296			Grouped data
Maximum x value		93			242			Grouped data

[a]The distribution is from Bradford (1948), as grouped by Sichel (1985).
[b]The Kendall (1960) distribution was taken from Chen and Leimkuhler (1986), as grouped by Sichel (1985).
[c]The distribution is taken from Sichel (1992a), who gives Rao (1989) as the source.

Although not shown, the PL distribution was also fit to Sichel's (1985) summary of the Goffman and Warren (1969) distribution on 'mast cells'. This gave an acceptable probability of .23, but the GIGP probability was three times larger (Sichel, 1985). Both models provided equally good fits ($p = .93$) to the Goffman and Warren (1969) distribution on 'schistosomiasis' articles (see Sichel, 1992a). Two additional journal scattering distributions were fit with the PL model—Bradford's (1948) 'lubrication' data and Lawani's (1973) agriculture data (from Basu, 1992). The PL fits had chi-square probabilities above .40.

Zipf's law: Word frequencies

Zipf's law was developed to describe the frequency of word use in documents. Table 4 provides the PL fit to three classic sets of word frequency data: Eldridge's distribution of word use in four American newspaper articles, Brugmann's study of four plays in Plautine Latin, and noun frequency in Macaulay's essay on Bacon (Yule, 1944). Again the PL provides acceptable to excellent fits to these data sets. Although not shown, the PL adequately fit the distribution of Chinese words (Zipf, 1935) with a probability of .20. Sichel (1975) reports the results of fitting the GIGP to these four frequency distributions, among others. His chi-square probabilities (vs. those for the PL model) were .023 (vs. .88) for newspaper English, .191 (vs. .64) for Latin words, .832 (vs. .24) for nouns in Macaulay's essay, and .912 (vs. .20) for Chinese words. Thus, the PL model seems to offer equally acceptable, perhaps complementary, results.

Fits to other distributions

Some other distributions are simple extensions of the more traditional topics covered above. For example, Tables 1 and 2 indicated that total productivity (N) and total citations (C) *among scientists* are both lognormal, so the *average citations per article* (C/N) across scientists might be as well. The simple lognormal distribution provided a good fit (not shown) to the C/N data for those chemists with one or more publications, even though the discrete nature of both articles and citations made the C/N distribution very uneven.

Using a single C/N value for each scientist ignores variation in the quality of the articles produced by an *individual* scientist, but few scientists have sufficient productivity for statistical modeling. Seglen (1992), however, presents a distribution of citations to over a hundred publications by a single biomedical researcher, and the PL model provided a good fit (not shown) to this skewed distribution. Thus, the lognormal distribution may characterize variation in both the *average* quality *between scientists* and the quality among a *single scientist's* articles.

Sichel (1985) fits the GIGP distribution to several other types of data distributions. He obtained a very good fit ($p = .95$) to the number of Lending Library of England journals receiving x requests from other organizations. The PL fit was not as good, but certainly adequate ($p = .63$). He gives (in his Table 4) GIGP fits to three distributions on 'in-house' use of chemistry or physics journals in different libraries. All PL fits were better: .92 vs. .31, .62 vs. .35, and .99 vs. .95.

Finally, Sichel (1992b) examines the distribution of the number of references in a very diverse sample of articles: 10,000+ articles with at least one Hungarian (co)author. The GIGP model provided an acceptable fit with a probability of .51, whereas the PL fit was very poor with a probability less than .001. However, the selected articles are multidisciplinary and unlikely to be characterized by a single mean and variance. Very acceptable PL fits to similar data were obtained for more consistently defined, but much smaller, samples (e.g., the 1970 geoscience publications analyzed in Table 2).

In summary, the PL model seems to work quite well for a great variety of bibliometric/scientometric distributions, especially for more narrowly defined samples. Overall, it seems to produce equally good fits as the GIGP model. However, empirical adequacy is only one of several criteria that might be used to judge bibliometric models. Other criteria might include the plausibility of the causal mechanisms implied by the model, the implications of the model for future studies in bibliometrics, and possible elaborations of the model. These are considered below for the PL model.

Table 4. Observed and estimated number of words used x times in American newspaper articles, Plautine Latin plays, and nouns in Macaulay's essay on Bacon

American newspaper English[a]			Plautine Latin plays[a]			Nouns in Macaulay[b]		
x Observed frequency of use	f_x Observed number of words	Estimated number of words	x Observed frequency of use	f_x Observed number of words	Estimated number of words	x Observed frequency of use	f_x Observed number of words	Estimated number of words
1	2976	3003.1	1	5429	5429.4	1	990	975.8
2	1079	1032.0	2	1198	1177.6	2	367	374.2
3	516	514.8	3	492	509.0	3	173	191.8
4	294	307.9	4	299	285.1	4	112	115.1
5	212	204.6	5	161	182.9	5	72	76.1
6	151	145.7	6	126	127.5	6	47	53.7
7	105	108.9	7	87	94.1	7	41	39.8
8	84	84.4	8	69	72.3	8	31	30.5
9	86	67.2	9	54	57.3	9	34	24.0
10	45	54.8	10	43	46.6	10	17	19.4
11	40	45.4	11	44	38.6	11	24	15.9
12	37	38.3	12	36	32.5	12	19	13.2
13	25	32.6	13	33	27.8	13	10	11.2
14	28	28.1	14	31	24.0	14	10	9.6
15	26	24.5	15	13	20.9	15	13	8.2
16	17	21.5	16	25	18.4	16-20	31	28.4
17	18	19.0	17	21	16.3	21-30	31	26.5
18	10	16.9	18	21	14.6	31+	26	34.7
19	15	15.1	19	11	13.1			
20	16	13.6	20	15	11.8			
21	13	12.3	21	10	10.7			
22	11	11.2	22	8	9.8			
23	6	10.2	23	8	8.9			
24	8	9.3	24	9	8.2			
25	6	8.6	25	11	7.6			
26	10	7.9	26-30	27	30.4			
27	9	7.3	31-35	18	21.8			
28	6	6.7	36-40	21	16.4			
29	5	6.3	41-45	9	12.7			
30	4	5.8	46-50	8	10.2			
31	6	5.4	51-55	8	8.3			
32	4	5.1	56-61	5	8.2			
33-34	8	9.2	62+	71	67.7			
35-36	8	8.0						
37-38	2	7.2						
39-40	6	6.5						
41-44	13	11.0						
45-50	15	13.0						
51+	81	71.6						
Total	6001	6001.0		8421	8420.9		2048	2048.1
Chi-square		28.16			27.63			20.65
df & p value		37 & .85			31 & .64			16 & .19
Observed D_{max}		.0057			.0028			.0126
$D_{K\text{-}S}$ @ $p = .20$		.0138			.0116			.0236
Mean log δ		−2.702			−9.055			−1.676
St. Dev. log δ		2.355			3.562			1.983
Mean RAW count		(Censored at 61+)			(Censored at 62+)			3.928
St. Dev. RAW count		(Censored at 61+)			(Censored at 62+)			8.627

[a]The distribution is from Zipf (1935).
[b]The distribution is from Chen and Leimkuhler (1987b). Score groupings follow Sichel (1975).

2. PROCESSES THAT COULD GENERATE LOGNORMAL DISTRIBUTIONS

Sichel (1975) provides a justification for why word frequencies should follow a compound Poisson process, and Allison (1980b) does the same for scientists' productivity. If we accept this aspect of the PL model, the remaining issue is what 'causal processes' might produce a lognormal mixing distribution.

A possible model for the lognormal productivity of scientists is given by the 'law of proportionate effects' (Aitchison & Brown, 1957; Shimizu & Crow, 1988; Karmeshu *et al.*, 1984), where the underlying propensity to publish is a *multiplicative* function of many inde-

pendently distributed factors, such as intelligence, training, motivation, and available resources. That is, such factors *do not add* together, but are *multiplied* together, so a weakness in any one factor reduces the effects of all the other factors. Empirical support is given by Reskin's (1977) analyses of her chemists' productivity levels, where she found many significant interactions among her predictors. If the multiplicative model is correct, she should have used the log of productivity in her *linear* regressions.

The same model could apply to citations to articles (e.g., for an article to be cited highly it must not only have good methods, but also good theory *and* be published in a good journal *and* on an important topic, etc.). Empirical support comes from Stewart's (1983) use of the Box-Cox test (Maddala, 1977), which indicated that logged citations was the best functional form for predicting citations to articles. The multiplicative model also might be appropriate for word frequencies, where the probability of using a particular word is a multiplicative function of, say, the topic, the specific context, author preferences, recent usage in the text, and its length (to follow Zipf's 'least effort' explanation).

Bulmer's (1974) version of the 'broken stick' model might apply to the bibliometric scattering of articles among journals. Applying this model to bibliometric scattering would suggest that the collection of articles on a particular topic is broken into groups, and then each group is broken into subgroups in such a way that the number of created subgroups is independent of the original group size, and the subgroups have the same relative sizes as existed in the first 'breakage'. This process is repeated for each subgroup, and so on. This model might apply to how articles on a particular topic get distributed among journals, as the number of articles and journals expand to meet the needs of specialized researchers. Karmeshu *et al.* (1984) use this model for a lognormal distribution and relate it to Bradford's law, but they do not mix it with a Poisson process. Basu (1992) employs a similar model and relates it to Bradford's law, but uses neither the lognormal nor a Poisson process. Koch (1966) provides some models that would produce lognormal distributions of biological variables 'directly' from underlying normal distributions, so other analogies for bibliometrics may be possible. See also Shimizu and Crow (1988).

Thus there are plausible causal models that could produce a lognormal distribution for the intensity parameter in a compound Poisson process. Sichel (1975) argues that plausible causal processes are less important than empirical adequacy in the comparison of models in bibliometrics. However, when two models are equal empirically, as are the PL and GIGP models, then this secondary criterion becomes more important. This aspect appears to be lacking for the GIGP model.

3. LIMITATIONS, IMPLICATIONS, AND ELABORATIONS

Even if the PL distribution provides very good fits to diverse bibliometric/scientometric distributions, this does not imply that the above causal processes are at work. Many possible processes might produce a lognormal distribution of propensities, and these processes might have very different implications (e.g., intrinsic heterogeneity vs. social reinforcement processes). Furthermore, other distributions, such as Sichel's GIGP, might provide equally good fits with different implications about causal processes. At best, one can say that the empirical results are generally 'consistent' with the PL model and the various causal processes.

What is to be made of the results clearly inconsistent with the PL model (e.g., Lotka's *Chemical Abstracts* data and the number of references in articles with Hungarian (co)authors)? Although these are particularly diverse samples, some researchers (e.g., Potter, 1981; Haitun, 1982c) argue that Lotka's and Zipf's laws apply only to large, diverse samples with long-term, even life-time, measurement intervals. An alternative view is also plausible: The PL model is valid, but these particular samples were mixtures of diverse populations, so it is unreasonable to expect a common mean and variance or a good PL fit.

The case for the PL model may be advanced by considering some of its useful implications. The first two characterize all compound Poisson distributions, which include both the PL and GIGP distributions. First, there is a simple way to measure how reliably the observed scores reflect the underlying propensities to publish articles, be cited, use partic-

ular words, or contain articles on a particular topic. For all compound Poisson processes the *observed* variability in, say, publications has two sources: (a) a 'systematic' component based on the variance of the *underlying* propensities, and (b) a 'random' component due to the Poisson process. Our data are reliable when most of the observed variability is due to variability in the underlying propensities. Allison (1978b) has shown that for compound Poisson processes this reliability (R) can be estimated from the mean (M) and variance (V) of the *observed* scores with the following formula:

$$R = 1 - M/V. \tag{5}$$

This formula requires the complete distribution of scores, including $f_{x=0}$. (With zero-truncated data the above formula always underestimates the reliability.) Applying this formula to the productivity data of chemists and entomologists in Table 1, we find reliabilities of 0.76 and (at least) 0.90, respectively. The lower reliability of the chemists' productivity probably stems from the shorter measurement interval (Allison, 1978b): six years vs. 70+ years for the entomologists.

This reliability calculation might provide useful information in bibliometrics (e.g., Allison, 1978b, illustrates how to estimate the length of the measurement interval for collecting productivity data of a desired reliability). Perhaps it provides a simple measure of the 'completeness' of the compilation of a bibliography. For example, the estimated reliability of the Bradford and Kendall data sets in Table 3 are both *at least* 0.95 or higher, whereas Bradford's (1948) 'lubrication' data (not shown here) has a minimum possible reliability of 0.75, which suggests this bibliography is less complete or a longer observation period is needed to allow journals to show their propensity to publish articles on lubrication.

A second result for compound Poisson distributions is the simple formula for $E(\delta|x)$, the expected value of δ given an observed x score:

$$E(\delta|x) = (x + 1)P_{x+1}/P_x, \tag{6}$$

where the P_x terms are those in eqn (3). In effect, this estimates the true score, or actual propensity by correcting for the unreliability of the observed count due to the randomness of the Poisson component. (This estimate has some error, because eqn (6) applies to compound Poisson distributions with known mixing distribution and parameters, whereas these are assumed and estimated in practice.) In particular, when the observed score is zero, $E(\delta|x = 0) = P_1/P_0$ is greater than zero and can be logged, so one can avoid adding an arbitrary constant to the observed counts before taking logs. (This adjustment could be used for all the observed x scores, especially if the estimated PL distribution had a mode at or above 1.0, which would give an $E(\delta|x = 0)$ above 1.0.)

The specific lognormal aspect of the PL model has additional implications when coupled with the above distinction between *observed* x scores and *underlying* propensities (δ scores). First, when comparing different distributions, it is better to work with the estimated parameters of the *logged* propensities. For example, in Table 2 the variance of the observed citations to the geoscience articles is over three times larger than the variance for the physics articles, which suggests that the geoscience articles had much more variation than the physics articles in their 'basic tendency' to be cited. However, since the variance of a simple Poisson process is equal to its mean, some of the observed x score variability for the geoscience articles arises from their higher mean propensity to be cited. In fact, the variances of the *logged* propensities indicate slightly more variability (or inequality) among the physics articles (1.89 vs. 1.628). Statistical tests for differences between fields that rest upon assumptions of normal distributions (e.g., F tests for differences in variances) are more appropriate for the logged parameters. Similarly, attempts to standardize scores between disciplines before studying the determinants of, say, productivity (e.g., Cole, 1978) should use estimated parameters for the propensities. For example, an article with four citations would have negative z scores in both the *observed* physics and geoscience citation distributions in Table 2, but a positive z score in the physics distribution of logged propensities.

The PL model provides information related to indices of concentration or inequality in bibliometrics (e.g., Egghe & Rousseau, 1990; Hustopecky & Vlachy, 1978). Since the

variance of a logged variable provides a 'scale-free' measure of the *inequality* in the variable's distribution (Allison, 1978a), the estimated variance of the logged propensities (σ^2) measures the *intrinsic* inequality (or concentration) existing in the sampled population. 'Intrinsic' is used because the estimated σ^2 avoids contamination from random variability due to the Poisson component, which would be included if one used the variance of the logs of the *observed* scores. This would also apply to studies using inequality in citations to measure the level of 'consensus' in fields (e.g., Cole *et al.*, 1978). (Allison, 1980b, uses the negative binomial model to remove Poisson variability from the coefficient of variation—another good measure of inequality, but the PL model generally provided better fits than the negative binomial model. Also Bulmer's discussion of PL-based measures of species 'diversity' may be relevant to concentration measures.)

The PL model also has some implications for the traditional scatterplots that stimulated the 'discovery' of the various bibliometric 'laws' (e.g., plots of $\text{Log}\, x$ vs. $\text{Log}\, f_x$). There is considerable disagreement on what variables should be plotted to produce a linear plot, how to define the transitions from 'core' journals or the start of the Groos' droop, and the meaning of different shapes in such plots. For simple lognormal distributions, a linear plot is produced with ($\text{Log}\, x$) vs. $INVCDF(CP_x)$ plots, where CP_x is the cumulative proportion of the sample with scores of x or less, and *INVCDF* is the inverse cumulative distribution function for a standard normal curve (Aitchison & Brown, 1957). The regression line intercepts at μ and has a slope equal to σ. Unfortunately, the Poisson component of the PL model invalidates this linear relationship for typical bibliometric distributions, especially at the lowest x scores. Despite this problem, the correlation between these two variables is generally above .98, and higher than the correlation between other variables that are commonly plotted. Further research is needed to see how much the differences in the shapes and other features of Bradford plots are simply the effects of discrete counts, zero truncation, different reliabilities, plotting different variables, and differences in the basic PL parameters.

A final implication of the PL model is that it keeps the study of bibliometric/scientometric topics in a Gaussian context (i.e., working with normal distributions, whose features are summarized by the well understood mean and standard deviation). The GIGP model has three parameters—one for the shape of each tail and one influencing the overall shape. However, the meaning of these parameters is less clear (e.g., see Sichel, 1992a, pp. 7–8), and the GIGP model is mathematically complex. Haitun (1982b) argues that Zipf's law applies to the distribution of many social characteristics, and that this law implies that social phenomena are inherently non-Gaussian. The good fits by the PL model in this paper suggest that this may be a premature conclusion, but only further tests can determine whether these initial successes of the PL model will generalize to other variables and samples. Thus it is worthwhile to consider some of the possible elaborations available for the PL model.

The three-parameter lognormal distribution introduces an adjusted 'origin' by subtracting a constant (the third parameter) from all the scores before taking the logarithm of the scores (Cohen, 1988). This might be compounded with a Poisson process to model distributions of word or sentence lengths having minimum possible scores above zero. The delta lognormal distribution (Aitchison & Brown, 1957; Dennis & Patil, 1988) might be used when 'excessive' zero counts are present. Here some proportion of the $f_{x=0}$ scores are assigned to a $\delta = 0$ status, with *no chance* of an event occurring, and the remainder are assumed to have small, but greater than zero, δ values. The fraction assigned to the $\delta = 0$ status would be a third parameter. An easy way to estimate this would be to ignore the zero counts and fit a truncated distribution, which would yield an estimate of $f_{x=0}$, so those in a $\delta = 0$ status could be found by subtraction. Sichel (1992b) used this method for his GIGP fit to the number of references in articles with Hungarian authors.

Another elaboration that stays within the 'traditional' scientometric approach would use a *bivariate* PL distribution to model bivariate distributions of discrete variables (Aitchison & Ho, 1989). For example, this might work for scientists classified by number of articles published and total number of citations or journals classified by article counts on two subject areas.

A reviewer suggested another elaboration: Assume the basic propensities represent a mixture from two distributions. In this case, one would have to estimate five parameters: the means and variances for two distributions and the fraction of the sample in each. A natural extension of this approach would try to identify specific articles with each distribution, which would permit relating differences in population parameters to common features of the articles in each population. In other words, one would have a simple causal model.

O'Connor and Voos (1981) carry this idea even further. They argue that rather than trying to infer the 'causal' processes by fitting univariate distributions, "[t]he analyses of bibliographic information should culminate in a [multivariate] causal model that accounts for variabilities in such phenomena as author productivity. . . ." (p. 15). This advice applies *regardless* of which statistical model best fits the univariate distributions. In fact, if one accepts the causal model assumption that productivity (or word usage, etc.) depends on *many* features of the scientists (or words) and their environments, then *it is unreasonable to expect univariate models to fit bibliometric distributions* because it is unlikely that all of the determinants have compatible univariate distributions.

The PL model does suggest that the dependent variable in these multivariate, causal models should be the *log* of the observed counts of articles, citations, or words, or perhaps the Log of $E(\delta|x)$. Logged scores are more likely to be a linear function of the predictors and to have normally distributed residuals, which are assumed in most significance tests. However, even in the sociology of science, where causal modeling of scientific productivity is common, the use of logged counts is rare; Stewart (1983, 1990) and Lovaglia (1991) are exceptions.

4. CONCLUSIONS

This article illustrates that the Poisson-lognormal model provides good fits to a diverse set of distributions commonly studied in bibliometrics and scientometrics. Only the GIGP model seems to provide equally good fits for such a diverse set of distributions, but it lacks development of possible 'causal' processes, is more complex mathematically, and its parameters are less well understood and do not directly relate to such key topics as inequality levels or how to standardize across fields.

If further studies generalize the applicability of the PL model, then it would replace the 'laws' of Lotka, Bradford, and Zipf. The substantial efforts to test and relate these laws would be available for the development of multivariate causal analyses of bibliometric processes. The more likely outcome, of course, is that the simple PL model will fail to fit some distributions that it 'should' fit (i.e., from reasonably homogeneous samples). For a causal modeler, this would only imply that one or more determinants has an inconsistent distribution, so causal analyses could proceed. Those working in the more traditional univariate mode might increase the available univariate models by developing some of the PL elaborations mentioned earlier.

Acknowledgements—NSF Grant SES-8706348 and a Vincent Coffin Grant from the University of Hartford supported collection of some of the data used in this article. Paul Nicholls at the University of Western Ontario's School of Library and Information Science generously provided comments on an earlier draft and copies of numerous distributions of scientists' productivity. Paul Bugl at the University of Hartford provided helpful comments.

REFERENCES

Aitchison, J., & Brown, J.A.C. (1957). *The lognormal distribution.* Cambridge: Cambridge University Press.

Aitchison, J., & Ho, C.H. (1989). The multivariate Poisson-lognormal distribution. *Biometrika, 76,* 643–653.

Allison, P.D. (1978a). Measures of inequality. *American Sociological Review, 43,* 865–880.

Allison, P.D. (1978b). The reliability of variables measured as the number of events in an interval of time. In K. Schussler (Ed.), *Sociological Methodology 1978* (pp. 238–253). San Francisco: Jossey-Bass.

Allison, P.D. (1980a). Estimation and testing for a Markov model of reinforcement. *Sociological Methods and Research, 8,* 434–453.

Allison, P.D. (1980b). Inequality and scientific productivity. *Social Studies of Science, 10,* 163–179.

Arunachalam, S., Rao, M.K.D., & Shrivasta, P.K. (1984). Physics research in Israel—A preliminary bibliometric analysis. *Journal of Information Science, 8,* 185–195.

Basu, A. (1992). Hierarchical distributions and Bradford's law. *JASIS, 43*, 494–500.
Bradford, S.C. (1948). *Documentation*. London: Crosby Lockwood.
Brookes, B.C. (1969). Bradford's law and the bibliography of science. *Nature, 224*, 953–956.
Bulmer, M.G. (1974). On fitting the Poisson lognormal distribution to species-abundance data. *Biometrics, 30*, 101–110.
Carroll, J.B. (1967). On sampling from a lognormal model of word-frequency distributions. In H. Kucera & W.N. Francis (Eds.), *Computational analysis of present-day American English* (pp. 406–424). Providence, RI: Brown University Press.
Chen, Y., & Leimkuhler, F.F. (1986). A relationship between Lotka's law, Bradford's law, and Zipf's law. *JASIS, 37*, 307–314.
Chen, Y., & Leimkuhler, F.F. (1987a). Bradford's law: An index approach. *Scientometrics, 11*, 183–198.
Chen, Y., & Leimkuhler, F.F. (1987b). Analysis of Zipf's law: An index approach. *Information Processing & Management, 23*, 171–182.
Cohen, C. (1988). Three-parameter estimation. In E.L. Crow & K. Shimizu (Eds.), *Lognormal distributions: Theory and applications* (pp. 113–137). New York: Marcel Dekker.
Cole, S. (1978). Scientific reward systems: A comparative analysis. *Research in the Sociology of Knowledge, Science, and Art, 1*, 167–190.
Cole, S., Cole, J., & Dietrich, L. (1978). Measuring the cognitive state of scientific disciplines. In Y. Elkana *et al.* (Eds.), *Toward a metric of science* (pp. 209–251). New York: John Wiley & Sons.
Dennis, B., & Patil, G.P. (1988). Applications in ecology. In E.L. Crow & K. Shimizu (Eds.), *Lognormal distributions: Theory and applications* (pp. 303–330). New York: Marcel Dekker.
Egghe, L., & Rousseau, R. (1990). Elements of concentration theory. In L. Egghe & R. Rousseau (Eds.), *Informetrics 89/90* (pp. 97–137). Amsterdam: Elsevier.
Goffman, W.M., & Warren, K.S. (1969). Dispersion of papers among journals based on a mathematical analysis of two diverse medical literatures. *Nature, 221*, 1205–1207.
Griffith, B.C. (1988). Exact fits to large ranked, bibliometric distributions. *JASIS, 39*, 423–427.
Gupta, D.K. (1987). Lotka's law and productivity patterns of entomological research in Nigeria for the period, 1900–1973. *Scientometrics, 12*, 33–46.
Haitun, S.D. (1982a). Stationary scientometric distributions. Part I. Different approximations. *Scientometrics, 4*, 5–25.
Haitun, S.D. (1982b). Stationary scientometric distributions. Part II. Non-Gaussian nature of scientific activities. *Scientometrics, 4*, 89–104.
Haitun, S.D. (1982c). Stationary scientometric distributions. Part III. The role of the Zipf distribution. *Scientometrics, 4*, 181–194.
Hustopecky, J., & Vlachy, J. (1978). Identifying a set of inequality measures for science studies. *Scientometrics, 1*, 85–98.
Karmeshu, N., Lind, C., & Cano, V. (1984). Relationales for Bradford's law. *Scientometrics, 6*, 233–241.
Kendall, M.G. (1960). The bibliography of operational research. *Operational Research Quarterly, 11*, 31–36.
Koch, A.L. (1966). The logarithm in biology, 1. Mechanisms generating the log-normal distribution exactly. *Journal of Theoretical Biology, 12*, 276–290.
Lawani, S.M. (1973). Bradford's law and the literature of agriculture. *International Library Review, 5*, 341–350.
Lotka, A.J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences, 16*, 317–323.
Lovaglia, M.J. (1991). Predicting citations to journal articles: The ideal number of references. *American Sociologist, 21*, 49–64.
Maddala, G.S. (1977). *Econometrics*. New York: McGraw-Hill.
Nicholls, P.T. (1986). Empirical validation of Lotka's law. *Information Processing & Management, 22*, 417–419.
Nicholls, P.T. (1987a). *The Lotka hypothesis and bibliometric methodology*. Ph.D. Thesis, School of Library and Information Science, The University of Western Ontario, London, Canada.
Nicholls, P.T. (1987b). Estimation of Zipf parameters. *JASIS, 38*, 443–445.
Nicholls, P.T. (1989). Bibliometric modeling processes and the empirical validity of Lotka's law. *JASIS, 40*, 379–385.
O'Connor, D.O., & Voos, H. (1981). Empirical laws, theory construction and bibliometrics. *Library Trends, 30*, 9–20.
Pao, M.L. (1985). Lotka's law: A testing procedure. *Information Processing & Management, 21*, 305–320.
Peritz, B.C. (1992). On the objectives of citation analysis: Problems of theory and method. *JASIS, 43*, 448–451.
Pielou, E.C. (1969). *An introduction to mathematical ecology*. New York: John Wiley & Sons.
Potter, W.G. (1981). Lotka's law revisited. *Library Trends, 30*, 21–40.
Qiu, L. (1990). An empirical examination of the existing models for Bradford's law. *Information Processing & Management, 26*, 655–672.
Rao, I.K. (1989). *Journal productivity in economics*. Paper presented at the Second International Conference on Bibliometrics, Scientometrics and Informetrics. London, Ontario, Canada.
Reskin, B. (1977). Scientific productivity and the reward system of science. *American Sociological Review, 42*, 491–504.
Seglen, P.O. (1992). The skewness of science. *JASIS, 43*, 628–638.
Shaban, S.A. (1988). Poisson-lognormal distributions. In E.L. Crow & K. Shimizu (Eds.), *Lognormal distributions: Theory and applications* (pp. 195–210). New York: Marcel Dekker.
Shimizu, K., & Crow, E.L. (1988). History, genesis, and properties. In E.L. Crow & K. Shimizu (Eds.), *Lognormal distributions: Theory and applications* (pp. 1–25). New York: Marcel Dekker.
Shockley, W. (1957). On the statistics of individual variations of productivity in research laboratories. *Proceedings of the Institute of Radio Engineers, 45*, 279–290.
Sichel, H.S. (1975). On a distribution law for word frequency. *Journal of the American Statistical Association, 70*, 542–547.
Sichel, H.S. (1985). A bibliometric distribution which really works. *JASIS, 36*, 314–321.

Sichel, H.S. (1992a). Anatomy of the generalized inverse Gaussian-Poisson distribution with special applications to bibliometric studies. *Information Processing & Management, 28*, 5–12.

Sichel, H.S. (1992b). Note on a strongly unimodal bibliometric size frequency distribution. *JASIS, 43*, 299–303.

Smith, L.C. (1981). Citation analysis. *Library Trends, 30*, 83–106.

Stewart, J.A. (1983). Achievement and ascriptive processes in the recognition of scientific articles. *Social Forces, 62*, 166–189.

Stewart, J.A. (1987). *Discipline integration, citations patterns, and the plate tectonics revolution.* Paper presented at the 1987 annual meeting of the Society for the Social Studies of Science, Wochester, MA.

Stewart, J.A. (1990). *Drifting continents and colliding paradigms.* Bloomington, IN: Indiana University Press.

Tague, J., & Nicholls, P. (1987). The maximal value of a Zipf size variable: Sampling properties and relationship to other parameters. *Information Processing & Management, 23*, 155–170.

Yule, G.U. (1944). *A statistical study of literary vocabulary.* Cambridge: Cambridge University Press.

Zipf, G.K. (1935). *The psycho-biology of language.* Boston: Houghton Mifflin.