Social Studies of Science, 6 (1976), 269-76

Notes and Letters

Lotka's Law: A Problem in Its Interpretation and Application

Paul D. Allison, Derek de Solla Price, Belver C. Griffith, Michael J. Moravcsik and John A. Stewart

[*Editors' Note:* The present paper grew out of a lengthy and frequently heated correspondence among these authors and the Editors of the Journal. In *Little Science, Big Science* (1963) Derek de Solla Price wrote that 'the total number of scientists goes up as the square, more or less, of the number of good ones', his text indicating that this quite provocative idea was consistent with a number of findings, including the empirical law named 'Lotka's law', after its discoverer.[1] In August 1974, John Stewart and Paul Allison submitted to us a manuscript questioning the consistency of Price's 'square root' law with Lotka's law. After this manuscript was referred to Price for comment, it emerged that Michael Moravcsik and Belver Griffith had also been concerned with this problem. Price's reply to Stewart and Allison clearly indicated the plausibility of his earlier reasoning. Moravcsik and Griffith, working separately and in concert, explicated the differences between the Stewart and Allison and Price papers, and the assumptions underlying those differences. The present paper is intended to identify and explicate the problem for our readers, and to indicate areas of actual or potential agreement.]

In this Note we discuss the predictions of the Lotka law of scientific authorship for the relative contribution of the most prolific authors.[2] This problem has been repeatedly discussed in the literature and we feel a need for a rigorous analysis of it on account of its widespread use in, and implications for, science

Authors' addresses (respectively): Department of Sociology, State University of New York, Stony Brook, New York 11794, USA; Department of History of Science and Medicine, Yale University, New Haven, Connecticut 06520, USA; Graduate School of Library Science, Drexel University, Philadelphia, Pennsylvania 19104, USA; Institute of Theoretical Science, University of Oregon, Eugene, Oregon 97403, USA; Department of Sociology, Social Science Building, University of Wisconsin, Madison, Wisconsin 53706, USA.

policy. In particular, we want to clarify the relationship between the Lotka law and a statement by one of us (D. deS. P.) that half of the scientific papers are contributed by the square root of the total number of scientific authors (henceforth referred to as 'the Price law').[3] Because of its broad ramifications, this statement has attracted much mention in the literature.[4]

We want to emphasize at the outset that we will *not* concern ourselves with whether either the Lotka law or the Price law is, in fact, in agreement with empirical data on authorships. Our exclusive focus will be on the mathematical consequences of the Lotka law and on the mathematical connection between the Lotka and Price laws, although when facing the necessity of making additional assumptions in order to be able to proceed, we will comment on these assumptions in terms of their plausibility.

We begin with the notation. The number a of those authors who contribute n papers will be denoted as $a(n)$. The number of authors A who contribute between (and including) n and n' papers will be denoted as $A(n,n')$. Clearly

$$A(n, n') = \sum_{i=n}^{n'} a(i). \tag{1}$$

The number of papers p contributed by authors who contribute n papers each will be denoted by $p(n)$. The number of papers P contributed by authors each of whom contributes between (and including) n and n' papers will be denoted by $P(n,n')$. Clearly one has

$$p(n) = n\, a(n) \tag{2}$$

and

$$P(n, n') = \sum_{i=n}^{n'} p(i). \tag{3}$$

The number of papers in all these quantities must be an integer larger or equal to 1. For a discussion of a possible upper limit on the number of papers, see below.

In terms of our notation, Lotka's law states that

$$a(n) = C/n^2, \tag{4}$$

with C a constant, while the Price law expresses the following relationship

$$\tfrac{1}{2} P\left(1, n_{\max}\right) = P\left(m, n_{\max}\right) = P(1, m), \tag{5}$$

where m satisfies the requirement

$$\left\{ A\left(1, n_{\max}\right) \right\}^{1/2} = A\left(m, n_{\max}\right) \tag{6}$$

in which $n_{\max}$ is the largest number of papers contributed by any single author,

a number that, at this point in our discussion, might be finite or infinite, and might or might not depend on C.

Substituting equation (4) into (2), we get as a consequence of Lotka's law

$$p(n) = C/n \tag{7}$$

and hence, from equations (3) and (A2) (see Appendix, below)

$$P(1, n) = \sum_{i=1}^{n} C/i = C\,(\ln n + 0.577 \ldots + \epsilon_n) \tag{8}$$

as another consequence of Lotka's law alone.

We can now calculate the value of m for which equation (5) is satisfied; that is, we can calculate the number of papers such that those authors (if distributed according to Lotka's law), each of whom contribute more than that number of papers, contribute collectively half of all the papers. We have, by inserting equation (8) into (5)

$$\tfrac{1}{2}\, C \left(\ln n_{\max} + 0.577 \ldots + \epsilon_{n_{\max}}\right) = C\,(\ln m + 0.577 \ldots + \epsilon_m) \tag{9}$$

or

$$\ln (n_{\max}^{1/2}/m) = 0.289 \ldots + \epsilon_m - \tfrac{1}{2}\epsilon_{n_{\max}} \tag{10}$$

Assuming now that $\epsilon_m \ll 0.289$ (we have $\tfrac{1}{2}\epsilon_{n_{\max}} \ll \epsilon_m$ anyway) we get

$$m = 0.749 \left(n_{\max}^{1/2}\right) \tag{11}$$

Note that this result is a consequence of the Lotka law alone (except for our assumption that ϵ_m can be neglected), and that the square root that appears in equation (11) has no direct relationship to the square root in (6) (which we have not discussed or derived yet). We will in fact see that equation (6) cannot be derived without further assumptions, while (11) holds just on account of Lotka's law, and holds regardless of whether $n_{\max}$ is finite or infinite, or regardless of how $n_{\max}$ is chosen.

For the sake of brevity, we will call 'the elite group' the group of the most prolific authors who supply half of all papers. So far we have seen that, as a result of Lotka's law alone, the least prolific member of the elite group produces 0.749 times the square root of the number of papers the most prolific member of the elite group produces. We will try to determine the size of the elite population compared to the whole authorship population.

The total authorship population, from equations (4), (1) and (A3), is

$$A\left(1, n_{\max}\right) = C\left\{\frac{\pi^2}{6} - \frac{1}{n_{\max}} + 0\left(\frac{1}{n_{\max}^2}\right)\right\} \tag{12}$$

while the elite population, from equations (4), (1) and (A5), is

$$A\left(m, n_{\max}\right) = C\left\{\frac{1}{m} - \frac{1}{n_{\max}} + 0\left(\frac{1}{m^2}\right) + 0\left(\frac{1}{n^2_{\max}}\right)\right\} \qquad (13)$$

which, by (11), gives

$$A\left(m, n_{\max}\right) = C\left\{\frac{1}{0.749\, n^{1/2}_{\max}} - \frac{1}{n_{\max}} + 0\left(\frac{1}{n^2_{\max}}\right)\right\} \qquad (14)$$

We will now assume that $n_{\max} \gg (n_{\max})^{1/2}$, and that $n_{\max} \gg 6/\pi^2$

in which case we get

$$R \equiv \frac{A\,(m, n_{\max})}{A\,(1, n_{\max})} = \frac{6/\pi^2}{0.749\, n^{1/2}_{\max}} = \frac{0.812}{n^{1/2}_{\max}} \qquad (15)$$

R is, of course, the ratio of elite authors to all authors; we will multiply by 100 to generate percentages below.

This result still depends only on Lotka's law, but this is as far as we can go on Lotka's law alone. To go beyond this we must make some assumptions about $n_{\max}$.

It is both natural and regrettable that further results depend on $n_{\max}$. It is natural because in the absence of an $n_{\max}$ the total number of papers, on account of equation (A6), is infinite, and almost all of them would be contributed by the very most prolific authors. In fact, these authors would be so tremendously prolific that their population need not be large to turn out such an overwhelming fraction of the papers. The presence of $n_{\max}$ is, however, not only natural but also regrettable because the further results of this analysis will therefore depend entirely on how we choose our cut-off value of $n_{\max}$. Empirically, this is bad because $n_{\max}$ will depend on the very few exceptionally prolific authors, to which Lotka's law may not apply, and even if it does, one can expect large statistical fluctuations in their numbers which would prevent us from making any reliable predictions except after a detailed analysis of these fluctuations.[5]

We will now consider several plausible models for $n_{\max}$.

(A) Choose $n_{\max}$ to be an absolute constant, independent of C. In this case, as equation (15) shows, the elite population is a fixed percentage of the total population. Using a few values in an empirically-plausible range, we get R = 8 per cent for $n_{\max}$=100; R = 3.7 per cent for $n_{\max}$= 500; R = 2.5 per cent for $n_{\max}$= 100; and R = 1.5 per cent for $n_{\max}$= 3000.

Note that in this case R is independent of the total number of authors. This fact can be used as an objection to this way of choosing the cut-off, because one would intuitively believe that in a larger sample of authors one would have a better chance of finding some exceptional authors of great prolificacy.

(B) Choose $n_{\max}$ so that $a(n_{\max}) = \alpha$, where α is a given small integer. The argument in favour of this type of cut-off is in the assertion that there is no point in talking about the statistical distribution in the realm of small

numbers. In this case we have from equation (4)

$$C/n_{\max}^2 = \alpha \quad \text{or} \quad n_{\max} = (C/\alpha)^{1/2} \tag{16}$$

and so in this case $n_{\max}$ increases with C, as intuition would dictate.
We than have from equation (15)

$$R = 0.812\,\alpha^{1/4}/C^{1/4} \tag{17}$$

and since $A(1, n_{\max}) = C\,\pi^2/6$, we get

$$A\left(m, n_{\max}\right) = 0.90\,\alpha^{1/4}\left[A\left(1, n_{\max}\right)\right]^{3/4}. \tag{18}$$

Using $\alpha = 1$, we get $n_{\max} = 1000$ for $C = 10^6$, which is not altogether unreasonable. For this group we then have an elite population of about 5×10^4, or 3 per cent.

(C) Choose $n_{\max}$ so that $A(n_{\max}, \infty) = \beta$, where β is a given small integer. The argument for this type of cut-off says that one should stop at an $n_{\max}$ *beyond* which the total number of all more prolific authors is a given small integer, because the statistical treatment in that range is unreliable. The justification for this type of a cut-off is very similar to that of case (B). The fact that we get a quite different result for R, however, demonstrates the vulnerability of our result to small changes in the treatment of the most prolific authors.

In any case, we obtain, under these assumptions, using equation (13)

$$C/n_{\max} = \beta \quad \text{or} \quad n_{\max} = C/\beta \tag{19}$$

so again $n_{\max}$ increases with C, in fact faster than in the case (B). We thus have

$$R = \frac{0.812\,\beta^{1/2}}{C^{1/2}} \tag{20}$$

and so

$$A\left(m, n_{\max}\right) = 1.04\,\beta^{1/2}\left[A\left(1, n_{\max}\right)\right]^{1/2} \tag{21}$$

which, for $\beta = 1$, gives almost exactly equation (6) and hence the Price law. Note that for other values of β, the functional dependence is still the same as for the Price law, and only the coefficient changes.

We see, therefore, that Price's law follows from Lotka's law if an additional, plausible, but non-unique assumption is made about $n_{\max}$. Other plausible and non-unique assumptions about $n_{\max}$ give different laws. The ambiguity is due to the overwhelming role the most prolific authors play.

As mentioned earlier, a more reliable assumption about $n_{\max}$ can be obtained by a detailed study of the statistics of the relatively few very prolific authors in different fields and over different periods of time. Another promising direction is the derivation of the law from individual patterns of productivity. Some of us are pursuing such problems.

It must also be emphasized that the practical significance of these quantitative results may be questionable in any case. The broad validity of Lotka's law has

not been established,[6] and in fact, there begins to be some evidence that, especially in the crucial range of prolific authors, the law itself will have to be modified.[7] As we have demonstrated in this paper, however, the validity of Price's law does not necessarily depend on the validity of Lotka's law, and hence can be judged on the basis of empirical evidence alone. Finally, we should keep in mind the limitations of our measure. Certainly, a far broader basis than mere productivity in publication will be required in any final analysis of scientific achievement.

APPENDIX

The following mathematical results are used in the text:

Let us define

$$S_j(n, n') \equiv \sum_{i=n}^{n'} \frac{1}{i^j} \qquad (n, n', i, j \text{ are integers}). \tag{A1}$$

Then

$$S_1(1, n) = \ln n + 0.577 \ldots + \epsilon_n \tag{A2}$$

Here 0.577 . . . is the Euler constant, and ϵ_n is a correction term which decreases as n increases. In particular, $S_1(1,10) = 2.9290$ with an ϵ_{10} of 0.051 (or 2 per cent), while $S_1(1,50) = 4.4992$, with an ϵ_{50} of 0.010 (or 0.2 per cent).

Also we have

$$S_2(1, n) = \frac{\pi^2}{6} - \frac{1}{n} + 0\left(\frac{1}{n^2}\right). \tag{A3}$$

Using the first two terms only, this formula is accurate to about 0.1 per cent even at $n = 50$, and correspondingly more accurate for higher n.

From these relations we get immediately

$$S_1(n, n') = S_1(1, n') - S_1(1, n) = \ln(n'/n) \div \eta, \tag{A4}$$

where η is smaller than $\epsilon_{n'}$, and

$$S_2(n, n') = S_2(1,n) - S_2(1, n') = \frac{1}{n} - \frac{1}{n'} + 0\left(\frac{1}{n^2}\right) + 0\left(\frac{1}{n'^2}\right) \tag{A5}$$

The last relationship must be used with care since $0(1/n^2)$ can be of the same order as $1/n'$.

We note that

$$S_1(1, \infty) = \infty \tag{A6}$$

and

$$S_2(1, \infty) = \pi^2/6 = 1.645 \ldots . \tag{A7}$$

We also note that already for $n = 100$

$$S_2(1,100) \doteq S_2(1,\infty) = \pi^2/6, \qquad \text{to better than 1 per cent,} \qquad \text{(A8)}$$

and for $n > 100$ the equality is even more accurate.

NOTES

The work of one of us (B.C.G.) is supported by the US Public Health Service, and of another of us (M.J.M.) by the US National Science Foundation.

1. Derek de Solla Price, *Little Science, Big Science* (New York: Columbia University Press, 1963), 53; Alfred J. Lotka, 'The Frequency Distribution of Scientific Productivity', *Journal of the Washington Academy of Sciences,* Vol. 16 (19 June 1926), 317.
2. Lotka, op.cit. note 1.
3. Price, op.cit. note 1.
4. Derek de Solla Price, 'The Scientific Foundations of Science Policy', *Nature,* Vol. 206 (17 April 1965), 233-38, and 'The Structures of Publication in Science and Technology', in W. H. Gruber and D. R. Marquis (eds), *Factors in the Transfer of Technology* (Cambridge, Mass.: MIT Press, 1969), 91-104; Jonathan R. Cole, 'Patterns of Intellectual Influence in Scientific Research', *Sociology of Education,* Vol. 43 (Fall 1970), 377-403; Joel Yellin, 'A Model for Research Problem Allocation among Members of a Scientific Community', *Journal of Mathematical Sociology,* Vol. 2 (1972), 1-36.
5. At this point, one of us (D. de S.P.) wishes to make the following comment:

> I do not consider we have the liberty to accept any model for n_{max} other than that which would seem to follow from its definition. Since this is supposed to be the score of the highest-scoring author, we must have $A(n_{max},\infty) = 1$, which, by equations (1) and (A5), gives $n_{max}=C$. I am therefore forced to reject options (A) and (B), and accept only the special case of $\beta = 1$ of option (C), which (as noted) agrees with the Price law. It should be added that a more precise and somewhat more probabilistic formulation is that the actual score of the highest-scoring author would be indeterminate.
>
> Moreover, by setting $A(n_r, \infty) = r$, we may show that the rth highest-scoring author has a score given by $n_r = C/r$. I consider this inverse first power law governing score as a function of rank for the elite to be a direct consequence and equivalent of the Lotka law. Again, more exactly, the expected score is in the range from C/r to $C/(r - 1)$. Thus the most prolific authors form a rather regular and predictable series whose maximum scores are (∞, C, $C/2$, $C/3$, $C/4$, etc.) and whose minimum scores are (C, $C/2$, $C/3$, $C/4$, $C/5$,

etc.). The infinity at the first maximum shows that it is only the *most* prolific author (not several of them) who gives the inconvenience of an infinity that leads to a divergent series – and also to the heated composition of this Note. Taking only the minimum scores of the most prolific authors, one shows easily that this elite group (as defined by the Price law) contributes *at least* half the total production of papers.

6. It should be noted, however, that several workers have suggested statistical theory for a suitable infrastructure: for example, Herbert Simon, 'On a Class of Skew Distribution Functions', in his *Models of Man* (New York: John Wiley and Sons, 1957), Chapter 9; and William Shockley, 'On the Statistics of Individual Variations of Productivity in Research Laboratories', *Proceedings of the Institute of Radio Engineers,* Vol. 45 (March 1957), 279-90.

7. Such a modification continuous with the Lotka law but giving an inverse second power law for the scores of the elite, in agreement with empirical findings, has been suggested in Price, op. cit. note 1, 48, footnote 8. The modification agrees excellently with all data for lifetime scores, but seems to break down for short time invervals.