Informetric Distributions, Part I: Unified Overview

A. Bookstein
University of Chicago, Center for Information and Language Studies, Chicago, IL 60637

This article is the first of a two-part series on the informetric distributions, a family of regularities found to describe a wide range of phenomena both within and outside of the information sciences. This article introduces the basic forms these regularities take. A model is proposed that makes plausible the possibility that, in spite of marked differences in their appearance, these distributions are variants of a single distribution; heuristic arguments are then given that this is indeed the case. That a single distribution should describe such a wide range of phenomena, often in areas where the existence of any simple description is surprising, suggests that one should look for explanations not in terms of causal models, but in terms of the properties of the single informetric distribution. Some of the consequences of this conclusion are broached in this article, and explored more carefully in Part II.

Introduction

For a period of almost a century, in a vast variety of fields, researchers have discovered surprising regularities when they count events or tabulate the sizes of things. These discoveries, often called laws and named after the most prominent persons associated with them, have tended to be treated as curiosities. Yet new versions continue to be discovered and more recent empirical investigations have confirmed many of the earlier regularities. The intellectual challenge to explain or understand these regularities has engaged the attention of some of the most prominent researchers in mathematics and the sciences; for example, the field of stochastic processes has developed out of a classical article by Yule (1924) that was written to explain such a regularity, discovered by Willis (1922), in the field of evolutionary biology. Still, after some 50 years of investigation, these regularities remain a puzzle. Recently there has been a resurgence of interest in Informetrics and related subjects, as evidenced by the formation of the journal *Scientometrics* and the inauguration by Leo Egghe and Ronald Rousseau (1988) of an international conference series on the subject.

Increasing the appeal and importance of research in these regularities is the impression that many of them are identical in form, and that the different formulations are approximations of one another. Thus, though the content of these regularities vary substantially, we may very well be considering differing manifestations of a single regularity, which, to have a single rubric, I shall call the informetric law.[1]

In general, when confronted with a regularity such as that which we are now considering, we can take one of two approaches. For one, we can try to understand the regularity as a manifestation of some underlying, substantively interesting phenomenon, much as Keppler's laws are consequences of the more basic laws of Newton; this is the more tempting and more common approach (Bookstein, 1979). But it is also possible that these regularities are the result of very general causes that cut across a wide range of different phenomena; the ubiquitous normal and lognormal regularities of statistics come to mind as prototypes here. If such is the case, then the presence of these regularities would have little to say about the underlying phenomena. Indeed, a wide range of underlying phenomena would, in this case, result in regularities having the same form. In brief, it is my belief that the form these informetric regularities take is very robust and will tend to appear as a consequence of a wide variety of underlying models. The implication of the existence of such robust regularities for the philosophy of science is very interesting, for it directly concerns how we learn from evidence. This is further explored in the concluding sections of this paper.

My discussion of the informetric regularities is divided into two articles. In this article (Part I) I shall provide an overview of the regularities, describing several of the most prominent and indicating why it is possible that these may be variants of a single, more general regularity even though they differ greatly in appearance. Specifically, a model of data generation will be introduced, and these regularities

Received August 21, 1989; revised November 20, 1989; accepted November 27, 1989.

[1] Terminology varies here. There is some confusion regarding the use of the terms *Informetrics*, *Bibliometrics*, and *Scientometrics*. In earlier papers I referred to the regularities discussed here as the *Bibliometric* regularities. My impression is that this term is being replaced in the literature by *Informetrics*, a term suggesting a wider range of applicability, and my usage in this paper is intended to be consistent with this evolution.

will be shown to represent different modes of description of the resulting body of numbers. This raises the possibility that differences in how the distribution of values are *described* may be adequate to account for the differences in *appearance* of the regularities. I shall then continue the process of unification by showing, by means of heuristic approximations, that these regularities are in some sense equivalent to one another. In the closing sections of Part I, I shall consider some of the broader implications of the robustness properties noted above. The companion article (Part II) is devoted entirely to the notion of robustness against ambiguity: it makes explicit what is meant by robustness in this context, it illustrates similar properties for other regularities found in the social sciences, and finally explores these properties with regard to the informetric regularities themselves.

Informetric Regularities

To illustrate the forms of these regularities, or laws, take I shall concentrate on examples from the information sciences, and then indicate in passing some other cases to give a sense of their breadth of application. In this section, a number of regularities describing phenomena in the social and biological sciences will be defined. Each of these regularities act as prototypes for many similar regularities, formally the same but describing different domains of phenomena. I shall, in the next section, argue that the regularities described in this section, though different in form and applied to different subject contents, are actually, to a good approximation, variants of one another. To prepare for this argument, I shall first define the regularities in a manner that emphasizes their similarity.

To begin, notice a similarity in the *type* of phenomena described by these regularities. These regularities usually start with a population of discrete entities, for example, businessmen, scientists, words, or journals. Each of these entities is *producing* something over a time-like variable—dollars earned, articles published, occurrences of articles in a given discipline, to continue with the above examples. Below, I shall refer to the items as generating events, or as having a yield.[2] Basically, these regularities chop off a segment of a time-like dimension and describe the distribution in the number of events generated in that segment by the members of the population, one yield per member. Often the segment is defined rather naturally; for example, the yield described by Pareto's law is the amount of dollars earned in a one-year period; conceptually, however, any time period could have been used. For other regularities, the time slice is determined by considerations of convenience or external constraint; for example, the five-year span that Lotka originally used was imposed on him by *Chemical Abstracts*' cumulation policy. Even Wyllis' law can be conceived in this way, the yields being the net number of species produced to date by a (variable) population of genera. Although I am not claiming that all regularities of the form I am discussing describe this type of phenomena, I do believe the association is a strong one, and that such phenomena are good candidates to test for a informetric law. Once one recognizes this underlying similarity among the regularities, it is natural to explore various ways that such phenomena can be described. I shall argue that the basic similarity among these regularities are obscured by (1) the different subject content and (2) the different ways that we can describe the distribution of yields over a population of entities. To stress the unity of these regularities, I shall refer to yields in general and use the names of these regularities only to distinguish the different modes of description and the specific mathematical forms the regularities take. I will also try to use a consistent notation below, especially for the critical quantities. Generally, I will use letters from the beginning of the alphabet to denote arbitrary constants; in different parts of the discussion, the same letters may refer to different values. Constants that are parameters intrinsic to the model may be denoted by other letters—for example, the Bradford multiplier is denoted by k in keeping with tradition. I shall use the letter r to refer to rank, y to refer to a yield per item, and Y to refer to a cumulative yield. The letter N shall refer to the cumulative number of entities associated with a given range of yields, for example, the number of journals in the core of Bradford's law. The letter f will denote the number of entities having a specific value for a yield (or a density for a continuous yield variable)—the notation suggesting a distribution function. Occasionally, when a quantity takes integer values and it is important to emphasize this fact, the letter n will be used, even though another letter, for example Y, may be more appropriate in that context. With this background, it is now possible to define the specific regularities.

Bradford's Law of Scattering

This is probably the most prominent of the informetric regularities within the information sciences, perhaps because of its promise for application to the control of literature. Bradford (1934) discovered this regularity when studying the extent to which literature in a single discipline is scattered over a range of journals. He found that he could form a core of journals of central interest to the discipline, and that if he formed rings of successively less productive journals, so that each ring contained the same number of relevant articles as the core, the number of journals in a ring divided by the number of journals in the preceding ring was approximately a constant, k. That is, if the core and each ring contained the same number of articles, then, if N_n is the number of journals in the nth ring,

$$N_n = k^n N_0, \qquad (1)$$

for N_0 the number of journals in the core. Although empirical investigation tends to be limited to a core and a

[2] A yield is a quantity, like income, that is possible to cumulate. This is very different from the types of variables, such as height, often treated in statistics. Awareness of this difference is critical for understanding the richness of the forms taken by the informetric regularities.

small number of rings, mathematical models of a Bradford population assume an infinite journal population.

Leimkuhler Variant

Leimkuhler (1967), and earlier Vickery (1948), found an "equivalent" version of this law: if we rank the journals in decreasing order of productivity for the concerned discipline, then N, the number of journals required to yield Y articles, is related to Y by:

$$Y = A \log(1 + BN), \tag{2a}$$

where A and B are constants. It is useful to also have this equation in its inverted form:

$$N = A'(\exp(B'Y) - 1). \tag{2b}$$

Strictly speaking, Leimkuhler's form of this law is stronger than Bradford's. Bradford only claimed that a core can be found that enjoyed the regularity he formulated. Leimkuhler proved his relation to be valid only at the boundaries of the rings formed by Bradford. However, if we take the Leimkuhler version as primary, and assume we can give meaning to the notion of fractional journals and articles, then I shall show that for *any* core, the Bradford regularity will hold. (See also the discussion in Egghe (1985).) Thus, in the Leimkuhler version, a distinguished *core* is not really required.

In form, Leimkuhler's distribution (especially in the form of equation (2b)) looks very much like the CDF of traditional statistics, but this is misleading. We note that Y denotes the *sum* of yields corresponding to the top N journals, not the actual yield of the Nth ranking journal. The sum of yields, while natural as a variable in this context, makes no sense in most statistical discussions (consider a discussion, for example, of *heights* of people rather than of their publications). A similar mode of description is used, however, in other contexts in which the distribution of yields is discussed; for example, the Lorenz curve is sometimes used by economists to illustrate disparities in the distribution of resources (Samuelson, 1970).

Lotka's Law

Lotka (1926) found that if he associated with each member of a group of chemists his article productivity, then the number of chemists, f, producing y articles was approximately given by

$$f = \frac{A}{y^{\alpha}}, \tag{3}$$

where A is an arbitrary constant and α is a constant approximately equal to 2. Thus plotting $\log(f)$ vs. $\log(y)$ produces a graph approximating a straight line, and this plot has formed the basis of many of the regularities referred to above. For example, substituting biological genera for scientists and species for articles would yield the important law of Willis mentioned above. In Lotka's law, the mode of description of this yield distribution is that of the traditional statistical probability distribution function; indeed, of the informetric regularities, this one is most consistent with traditional statistical description.

Zipf's Law

Many of the informetric regularities depend on ranked data. The most prominent belongs to Zipf (1935), who analyzed the frequency of word occurrence in natural text. He found that if he multiplied a word's frequency by its rank in number of occurrences in text, then the product was approximately constant:

$$r \times y = A, \tag{4a}$$

where r denotes a word's rank, y its frequency (or yield), and A an arbitrary constant. In fact, this regularity holds best after we pass over the highest ranking documents. Thus, plotting log-rank vs log-frequency produces a straight line over a wide range, but droops at the lowest values. Mandelbrot has pointed out that a law of the form

$$y = \frac{A}{(1 + Br)^{\alpha}} \tag{4b}$$

is more appropriate. As it is my goal to describe the relations among these laws rather than to be accurate to the historical formulations, I shall mean this last form when I refer to *Zipf's* law.

Zipf's law, like Lotka's law, can be related to the forms of description traditional in statistics: if we solve equation (4a) or (4b) for r in terms of y, and convert r to *percent* of population, the result is of the form $1 - CDF$, where CDF is the cumulative distribution function well known in statistics.

Pareto's Law

The final form I shall note is taken from economics. Pareto (1897) claimed that if we look at the incomes of the wealthiest members of a community, then the number of people, r, that earn more than y dollars a year is given approximately by:

$$r = \frac{A}{y^{\alpha}}, \tag{5}$$

for some constants A and α. Since r is what we otherwise refer to as the *rank* of the item, this law is a variant of Zipf's law. This identity is obscured by Pareto's law usually being formulated in terms of probabilities rather than ranks, and y being conceived of as a continuous variable.

Although the above-mentioned regularities are the most prominent, many other regularities of this type have been found. For example, Kendall (1960) found that the operations research literature also follows Lotka's law as described for chemists. Bradford's law is often used to refer to citations received by journals rather than the actual articles appearing in them. It is claimed that the distribution of lake sizes obeys Pareto's law and that the distribution of city sizes follows Zipf's law. (See Fairthorne (1969), Simon (1955), and Mandelbrot (1963) for discussions of the range of applicability of this law.)

Relations among the Regularities

Above I showed that the various informetric regularities can be seen as arising from different modes of description of the distribution of yields generated by a population of items. The question remains whether the difference in form of these regularities is due only to the difference in mode of description, or whether these regularities in fact describe differently distributed sets of values. I shall now present heuristic arguments that when one of these regularities is translated into the mode of description of another of these regularities, then the form of these distributions will be approximately the same. For other treatments, see also, for example, Egghe (1985), Yablonsky (1980), Haitun (1982), and Rousseau (in press).

Bradford–Leimkuhler

Leimkuhler showed the equivalence of Bradford's law to a log form. A more direct derivation, similar to that of Vickery (1948), follows.

Bradford, after ordering the entities in decreasing magnitudes of yield, established a core of the N_0 most productive items (so N_0 is the *rank* of the last item put into the core), collectively responsible for a yield of Y_0, and defined successive rings of items, each ring also collectively having a yield of Y_0. Leimkuhler related the rank of an item to the cumulative yield of all the lower ranking items. If we combine the first n rings, starting with the core as the 0th ring, we obtain a cumulative yield of $Y = nY_0$; the number of items required is given by

$$N = N_0 + kN_0 + k^2N_0 + \cdots + k^{n-1}N_0 = \frac{k^n - 1}{k - 1}N_0$$

items. But, since $n = Y/Y_0$, we find

$$\frac{N}{N_0} = \frac{k^{Y/Y_0} - 1}{k - 1}, \tag{6a}$$

or, solving for Y,

$$\frac{Y}{Y_0} = \frac{\ln\left(1 + \dfrac{(k-1)N}{N_0}\right)}{\ln(k)}. \tag{6b}$$

If we set $k - 1$ to b, we get the form given by Leimkuhler, which can, in turn, be directly related to equation (2a).

Note again that this form is strictly equivalent to Bradford's form only for Y/Y_0 an integer. However, in this form, it is natural to interpret the law as holding for arbitrary Y or N.

The Leimkuhler form, as presented above, seems to depend on having begun with a core of N_0 items and a yield of Y_0, since these parameters appear explicitly in the equations. That this is not the case is suggested by our being able to express the law in terms of transformed parameters: suppose, for example, that we wish Y_0 to become Y'_0. We can formally rewrite equation (6a) as

$$\frac{k-1}{k^{Y'_0/Y_0} - 1} \cdot \frac{N}{N_0} = \frac{k^{(Y'_0/Y_0)\cdot(Y/Y'_0)} - 1}{k^{Y'_0/Y_0} - 1},$$

or,

$$\frac{N}{N'_0} = \frac{k'^{(Y/Y'_0)} - 1}{k' - 1},$$

where

$$N'_0 = \frac{N_0(k^{Y'_0/Y_0} - 1)}{k - 1} \tag{7a}$$

and

$$k' = k^{Y'_0/Y_0}. \tag{7b}$$

Thus, the Leimkuhler form looks like it could as well have come from a core of N'_0 items having a yield of Y'_0, where N'_0, Y'_0, N_0 and Y_0 are related as above. I shall now show this to be true, that is, given Leimkuhler's form, with an apparent core of N_0, Bradford's law does hold, starting with an arbitrary core, N'_0. Indeed, if the Leimkuhler form does exactly describe a population, we could consider the consequences of beginning with a core of N'_0 instead of the designated value N_0. According to equation (6b), these N'_0 items will produce a yield, Y'_0, of

$$Y'_0 = \frac{Y_0 \ln\left(1 + (k-1)\dfrac{N'_0}{N_0}\right)}{\ln(k)}.$$

Direct algebraic manipulation shows that Y'_0 satisfies

$$N'_0 = N_0\frac{k^{Y'_0/Y_0} - 1}{k - 1},$$

as suggested by the formal manipulations given above (equation (7a))—that is, the current relation between N'_0 and Y'_0 is identical to the earlier relation, in fact, a variant of equation (2b).

If the core is ring zero, the nth ring with yield Y'_0 will contain $N_{n+1} - N_n$ items, where

$$N_n = \frac{N_0(k^{nY'_0/Y_0} - 1)}{k - 1}.$$

That is,

$$\begin{aligned} N_{n+1} - N_n &= \frac{N_0(k^{(n+1)Y'_0/Y_0} - 1)}{k - 1} - \frac{N_0(k^{nY'_0/Y_0} - 1)}{k - 1} \\ &= N_0(k^{Y'_0/Y_0} - 1)\frac{k^{nY'_0/Y_0}}{k - 1} = N'_0k'^n. \end{aligned}$$

Thus, if the Leimkuhler form is a valid description of our population, then we can define an arbitrary core of items and Bradford's law will hold for that core as well. That is, the appearance in the formula of parameters reflecting the initial choice of core does not support the existence of a fundamental core, since corresponding parameters can be associated with a different core and the identical formula be the result. Henceforth, when I mention Bradford's law, I shall assume any core is acceptable. If so, Bradford's law and Leimkuhler's variant are equivalents. I shall comment on the significance of this below.

An area of interest for commentators on Bradford's law is to characterize the difference between disciplines by means of Bradford's parameters. The constant, k, as a parameter measuring the intrinsic scatter in a discipline, is heavily used here. We now see that k is ambiguous, since it depends on the core, which is itself arbitrary. Even within the original Bradford formulation, in which we may not freely choose a core, there remains some ambiguity. Suppose, for example, assuming an infinite population of items, we combine successive sets of s rings, so that the new core and each new ring will contain s times as much yield as the original core. If the original core had N_0 items, the new core will have

$$N_0' = N_0 + kN_0 + \cdots + k^{s-1}N_0 = \frac{k^s - 1}{k - 1}N_0 \qquad (8)$$

items; the new first ring (that is, the old sth through $2s$–1th rings) will have

$$k^sN_0 + k^{s+1}N_0 + \cdots + k^{2s-1}N_0 = k^s\frac{k^s - 1}{k - 1}N_0$$

items, and the new nth ring,

$$(k^s)^n \cdot \frac{k^s - 1}{k - 1}N_0$$

items. Here, once again, the Bradford regularity is observed, but with k^s replacing k as the Bradford multiplier. Thus, if Bradford's law, even in the restricted formulation sometimes given, holds for any case, it holds for a variety of cases, with the Bradford multiplier changing as we redefine the case: whether we accept Bradford's narrow definition of his law or Leimkuhler's extension of it, the core is arbitrary; the Bradford scattering coefficient k depends on which core is chosen, and thus cannot be used to characterize a discipline.

However, substituting from equations (7a) and (7b) for k' and N_0', we see:

$$\frac{k' - 1}{N_0'} = (k^{Y_0'/Y_0} - 1)\frac{k - 1}{N_0(k^{Y_0'/Y_0} - 1)} = \frac{k - 1}{N_0};$$

similarly (substituting from equation (7b) for k'):

$$\frac{Y_0'}{\ln(k')} = \frac{Y_0'}{\ln(k^{Y_0'/Y_0})} = \frac{Y_0}{\ln(k)}.$$

Thus we see it is $(k - 1)/N_0$ and $\ln(k)/Y_0$ that are quantities invariant against changes in the core size. It is from these quantities, rather than from the Bradford multiplier, k, that we ought to construct statistics intended to describe a discipline and its scatter. More generally, since Bradford's regularity is often observed, these quantities might be useful as a means to create measures of concentration (Rousseau, 1989).

Relationship of Lotka and Zipf

Bradford's law represents one way of describing the output of a population of entities, each producing a stream of events. Lotka, looking at the same population, asked, how many items have a yield of y? He concluded that this quantity, f, is given by equation (3):

$$f = \frac{A}{y^\alpha},$$

where α is a constant approximately equal to two. Lotka considered y to be an integer; I shall also assume this, though we can generalize to continuous y if we interpret f as a density function. The maximum yield, y_0, is estimated as the output of only one item ($f = 1$), so $A = y_0^\alpha$; below I shall substitute y_0^α for A to simplify the form taken by the equations, though the argument does not depend on this substitution. Thus,

$$f = \left(\frac{y_0}{y}\right)^\alpha. \qquad (3a)$$

To translate this form into the Zipf form, we ask how many items have a yield of y or more; this number will be the rank, r, of the items with yield y. This quantity is given by $\Sigma_{x=y}^{y_0} (y_0/x)^\alpha$, which we approximate by (since $\alpha \neq 1$)

$$\int_{y-(1/2)}^{y_0+(1/2)} \left(\frac{y_0}{x}\right)^\alpha dx.$$

Thus

$$r = \frac{y_0}{\alpha - 1}\left[\left(\frac{y_0}{y - \frac{1}{2}}\right)^{\alpha-1} - \left(\frac{y_0}{y_0 + \frac{1}{2}}\right)^{\alpha-1}\right]. \qquad (9a)$$

Solving for y, we find that the yield of the rth ranking item could be expressed in the form:

$$y = \frac{A}{[B + Cr]^{\alpha'}}, \qquad (9b)$$

ignoring an additive constant, 1/2. As usual, A, B and C denote constants, as does α'. Here,

$$A = y_0,$$

$$B = \left(\frac{y_0}{y_0 + \frac{1}{2}}\right)^{\alpha-1},$$

$$C = \frac{\alpha - 1}{y_0},$$

and

$$\alpha' = \frac{1}{\alpha - 1}.$$

This is, indeed, a generalization of Zipf's law, for when $\alpha' = 1$ and $B = 0$, we find $r \cdot y$ is constant. More reasonably, though α' is approximately equal to 1, we do not expect $B = 0$; however, for larger r, where $Cr \gg B$, the Zipf condition does hold, and available data indicates that it is only for larger r that Zipf's law should be expected.

The converse is obtained from equation (9a). The number, f, of items with yield y is the difference in the ranks of the items having a yield of $y + (1/2)$ and a yield of $y - (1/2)$:

$$f = \frac{y_0}{\alpha - 1}\left[\left(\frac{y_0}{y}\right)^{\alpha-1} - \left(\frac{y_0}{y-1}\right)^{\alpha-1}\right] \qquad (10)$$

This can be approximated by the derivative:

$$f \approx \frac{y_0}{\alpha - 1} \cdot \frac{d}{dy}\left(\frac{y_0}{y}\right)^{\alpha-1} = \left(\frac{y_0}{y}\right)^{\alpha}.$$

That is, Lotka's law is returned.

Relationship of Leimkuhler to Lotka

Lotka's law relates a yield to the number of items having that yield, with the maximum yield, y_0, being experienced by a single item. Thus, the cumulative yield, Y, up to the item of rank r, which itself has a yield of y, is given by

$$Y = \sum_{n=y}^{y_0} n \cdot f_n,$$

for f_n the number of items having a yield of n. Given Lotka's law, this quantity is given by

$$Y = \sum n\left(\frac{y_0}{n}\right)^{\alpha} = y_0^{\alpha} \sum n^{1-\alpha},$$

which we again approximate by an integral:

$$Y \, \alpha \, y_0^{\alpha} \int_{y-(1/2)}^{y_0+(1/2)} x^{1-\alpha}\, dx =$$

$$\begin{cases} \dfrac{y_0^{\alpha}}{2-\alpha}\left[\dfrac{1}{\left(y_0 + \dfrac{1}{2}\right)^{\alpha-2}} - \dfrac{1}{\left(y - \dfrac{1}{2}\right)^{\alpha-2}}\right] & \text{if } \alpha \neq 2 \\ y_0^2 \ln\left(\dfrac{y_0 + \dfrac{1}{2}}{y - \dfrac{1}{2}}\right) & \text{if } \alpha = 2 \end{cases} \qquad (11a)$$

Since, by equation (9a),

$$\frac{1}{y - \dfrac{1}{2}} = \left\{\frac{(\alpha - 1)r}{y_0^{\alpha}} + \left[\frac{1}{y_0 + \dfrac{1}{2}}\right]^{\alpha-1}\right\}^{1/(\alpha-1)},$$

Y is of the form

$$Y \approx \begin{cases} A(\{B + Cr\}^{\alpha'} + 1) & \text{for } \alpha \neq 2 \\ A \ln(1 + Br) & \text{for } \alpha = 2 \end{cases} \qquad (11b)$$

where, for example, $A = y_0^2$ and $B = (y_0 + (1/2))/y_0^2$ for the second version. The second version is identical in form to Leimkuhler's representation of Bradford's law, and, since α tends to be close to 2, is the form we would usually expect. The above result, however, also shows the representation for $\alpha \neq 2$.

Relation between Zipf and Leimkuhler

Since Lotka's law is equivalent to Zipf's and, if $\alpha = 2$ implies Leimkuhler's, we expect that Zipf and Leimkuhler would be related to each other. It is instructive, and easy, to show this directly:

Given Leimkuhler's log function, the yield of the rth ranking item is given by

$$y_r = A \ln(1 + Br) - A \ln(1 + B(r - 1))$$
$$= A \ln\left(1 + \frac{B}{1 + B(r-1)}\right); \qquad (12a)$$

that is, to first order in $B/(1 + B(r - 1))$,

$$y_r \approx \frac{AB}{(1 - B + Br)} = \frac{AB}{1 - B}\,\frac{1}{1 + \left(\dfrac{B}{1-B}\right)r} \qquad (12b)$$

or

$$y_r \approx \frac{A'}{(1 + B'r)}, \qquad (12c)$$

with constants A' and B' defined by equation (12b). This is, indeed, the modified Zipf form of the law, equation (9b), with the exponent, α', equal to one. This value of α' reflects the restriction $\alpha = 2$ in equations (11a), (11b). Thus, for $\alpha = 2$, Leimkuhler implies Zipf; since Zipf was seen to imply Lotka, Leimkuhler implies (and is thus equivalent to) Lotka for $\alpha = 2$.

Lotka–Pareto

Pareto asked, how many items r have a yield greater than y? For a population observing Lotka's law, this quantity is given by

$$r = \sum_{x=y}^{\infty} N_x = \sum_{x} \frac{A}{x^{\alpha}} \approx \int_{y}^{\infty} Ax^{-\alpha}\, dx.$$

Thus

$$r \approx \frac{A'}{y^{\alpha'}}$$

for $A' = A/(1 - \alpha)$ and $\alpha' = \alpha - 1$. This form is, of course, that suggested by Pareto. Again note that Pareto's law is very similar to Zipf's law; however we are now considering the highest yield to be infinite in comparison to the yields in the range we are interested in.

Discussion

My conclusion that the various informetric distributions are very similar has depended upon heuristic arguments. In this section I shall describe an approach that can form the basis of a more precise, though numerical, comparison of the regularities. The reader can use the results of this section to get a better sense of how closely the heuristic results match a more exact evaluation. I shall examine, by way of example, how one might relate Zipf's law to Lotka's law.

Suppose, then, that we have a vocabulary of $r_{\max}$ word types, and that the probability of occurrence of the word ranking rth in probability of occurrence is given by A/r (Zipf's law). The probability, then, that the rth ranking word in our population will occur exactly y times in our sample is given by the binomial distribution and approximated by the following Poisson distribution:

$$Pr\{y|r\} = e^{-A'/r}\frac{\left(\frac{A'}{r}\right)^{y}}{y!}$$

where A' is the value A multiplied by the sample size. We would like to calculate the number of words expected to occur y times. Let $\tilde{d}_{y,r}$ be the indicator random variable defined by

$$\tilde{d}_{y,r} = \begin{cases} 1 & \text{if the } r\text{th ranking word occurs y times} \\ 0 & \text{otherwise} \end{cases}$$

Thus the number of words occurring y times is given by $\tilde{d}_y = \Sigma_r \tilde{d}_{y,r}$ and $d_y \equiv E(\tilde{d}_y)$, the expected value of $\tilde{d}_y$, equals $\Sigma_r E(\tilde{d}_{y,r}) = \Sigma_r Pr\{y|r\}$.

If one wished to carry out this process numerically, one would, for a given $r_{\max}$ and sample size:

(1) Evaluate the constants, A and A'. A is determined by the requirement that probabilities sum to one;
(2) compute d_y for a range of values y, and
(3) compare the results with those predicted by Lotka's law.

By varying $r_{\max}$ and sample size, the reader can see the impact of population and sample size on the accuracy of the approximations. How satisfying the approximations are will vary from person to person. The results of such experiments as carried out by the author seemed to confirm the reasonableness of the heuristics: for low values of y, the simulations diverged from the Lotka form. This became pronounced when I assumed a small vocabulary (small $r_{\max}$). However, for realistic vocabulary sizes, and for y greater than 3 or 4, the two results become quite close.

Consequences for the Social Sciences

A striking feature of the informetric regularities is that many of them occur in the biological and social sciences, fields otherwise resistant to the discovery of mathematical regularity. One reason that we have been unsuccessful in discovering regularities in the social sciences is that, unlike the physical sciences, the social sciences have not been able adequately to conceptualize and precisely to define those variables which would enter into the regularities. The contrast with the physical sciences is striking. Momentum, for example, is mass times velocity, not some undefined function of these quantities; and both mass and velocity are well defined. The regularities of physics in which momentum appear depend on exactly this definition. In contrast to this, the social sciences are burdened by ambiguities and arbitrariness in the definitions of measurements. For example, the five year time span chosen by Lotka in his study of publication patterns within chemistry was determined by the cumulation period of Chemical Abstracts, a unit chosen for convenience and independent of Lotka's research concerns. Biological classification, on which Willis' law is based, is notorious for the arbitrariness with which genera and species are defined (Sneath & Sokal, 1973). Even publication counts have an element of ambiguity, most strongly evidenced by the problem of how to count papers with multiple authors, but also, papers of different size and significance.

Under these conditions, it is surprising that any regularity at all can be found, for even if regularities existed, we would not expect them to be discovered until the appropriate concepts were defined and appropriate measuring techniques developed. Recognizing these difficulties increases the surprise with which the informetric regularities are met. Yet it is precisely this observation that may suggest an explanation for the prevalence of these regularities. For, in searching for scientific regularities under the conditions described above, a minimum condition that we must demand is that the law be resistant to (or stable against) ambiguity and imperfect measurement. Thus, a reasonable question to ask is, can we in some way characterize regularities that enjoy this type of resilience, so that when we seek regularities, we should restrict our search to members taken from this class. This is the approach I shall take in Part II: I shall assume that a regularity can be expressed in terms of a function that keeps the same form when we change the conditions of measurement; I shall show that the informetric regularities do have this resilience; and I shall show that for a range of modification (ambiguity), those functions that are resilient to the changes form a very restricted class. This is consistent with our findings above: the regularities we have examined, though superficially very different, were found to be variant expressions of a single distribution. This suggests that there is something special about this single form that permits it to be found in domains where such regularities would be surprising. I shall argue that that special quality is its resiliency to ambiguity.

General Implications of Informetrics

I have defined a number of regularities reported in a wide range of different disciplines, and have argued that we can think of these as different manifestations of a single regularity. In the preceding section, I suggested that, given the nature of the social and biological sciences, we should expect that such a regularity be very resilient to ambiguity in definition. In this last section, I shall consider the implication of the existence of such regularities for scientific investigation in general.

The importance of the existence of robust laws for science in general can be best seen by examining the logic of scientific learning. The paradigm:

If H then C

H

therefore C

is well understood as the basis of deductive reasoning. Similarly,

If H then C
C
therefore H

is generally acknowledged as a logical fallacy. On the other hand, if one is predisposed toward believing H, and C is a consequence of H, then observing C does increase our confidence in H. The scientific method is the process by which hypotheses are tested by examining the validity of their consequences.

The problem here is that if H implies C, so may other hypotheses H', H'', etc. If we wish to understand the role of evidence in supporting one of a number of competing hypotheses, a Bayesian framework is most suitable. From a Bayesian point of view, we begin with a sense of likelihood regarding the truth of a hypothesis, expressed as a probability, $P_0(H)$. We then try to increase this probability by examining evidence, E. The effect of evidence, E, is given by

$$\frac{P_1(H)}{P_1(\overline{H})} = \frac{P_0(\mathrm{H})}{P_0(\overline{H})} \frac{P(E\,|\,H)}{P(E\,|\,\overline{H})}$$

where $P_1(H)/P_1(\overline{H})$ is the odds in favor of the hypothesis after examining the evidence, and $P_0(H)/P_0(\overline{H})$ is the odds before seeing the evidence. The effect of evidence, in the Bayesian model, is given by the factor $P(E\,|\,H)/P(E\,|\,\overline{H})$, where $P(E\,|\,H)$ and $P(E\,|\,\overline{H})$ are the probabilities of E conditional on the hypothesis being true (H) and not true ($\overline{H}$), respectively. The value of this model is that (1) it makes explicit that evidence favoring a hypothesis is not a proof that the hypothesis is true, a well-understood, but often overlooked, consequence of deductive logic; and (2) it compels us to consider the possibility that the evidence may be the consequence of other models. In particular, even though H implies E, if $\overline{H}$ implies E almost as strongly, the existence of E provides very little support for H. The deeply ingrained, though unstated, assumption behind much scientific investigation is that $P(E\,|\,\overline{H})$ will, in general, be small, and that the more precisely E is formulated, the less likely it will be a consequence of hypotheses other than that from which it was derived (Nagel, 1961).

It is necessary to understand fully this mode of argument in order to appreciate properly the implications of the approach to the informetric distributions taken in this paper and its companion. For the position I am taking is that in certain areas of research, $P(E\,|\,\overline{H})$ approaches $P(E\,|\,H)$ in size, and that this is, in particular, true of the phenomena studied in informetrics. In such situations, although we may have a causal model, however reasonable, that predicts that a member of the family of informetric distributions will describe the data we shall obtain, verifying this expectation is in fact very weak support for this model. In other words, searching for informetric distributions may be a very poor way of trying to support a theory that predicts the existence and form of these distributions.

Acknowledgment

This article reports research supported by NSF Grant IST-8026564 and is based on the report submitted in 1982. The author wishes to acknowledge the NSF support.

References

Bookstein, A. (1979). Explanations of the bibliometric laws. *Collection Management, 3,* 151–161.

Bradford, S. C. (1934). Sources of information on specific subjects. *Engineering, 137,* 85–86.

Egghe, L. (1985). Consequences of Lotka's law for the law of Bradford, *Journal of Documentation, 41,* 173–189.

Egghe, L., & Rousseau, R. (Eds.). (1988). *Informetrics 87/88.* Amsterdam: Elsevier Science Publishers, 1988.

Fairthorne, R. A. (1969). Empirical hyperbolic distributions (Bradford–Zipf–Mandelbrot) for bibliometric description and prediction, *Journal of Documentation, 25.*

Haitun, S. D. (1982). Stationary scientometric distributions. *Scientometrics, 4,* 5–25.

Kendall, M. G. (1960). The bibliography of operations research. *Operations Research Quarterly, 11,* 31–36.

Leimkuhler, F. F. (1967). The Bradford distribution. *Journal of Documentation, 23,* 197–207.

Lotka, A. J. (1926). The frequency distribution of scientific productivity. *Journal of the Washington Academy of Sciences, 16,* 317–23.

Mandelbrot, B. (1963). New methods in statistical economics. *Journal of Political Economy, 71,* 421–440.

Nagel, E. (1961). *Structure of science: Problem in the logic of scientific explanation.* New York: Harcourt, Brace and World.

Pareto, V. (1897). *Cours d'Economie Politique.* Vol. 2. Lausanne: 1 Universite de Lausanne.

Rousseau, R. (1989). Elements of concentration theory, presented at the *Second International Conference of Bibliometrics, Scientometrics and Informetrics,* July 5–7, 1989, London, Ontario.

Rousseau, R. (in press). Relation between continuous versions of bibliometric laws. *Journal of the American Society for Information Science.*

Samuelson, P. (1970). *Economics* (2nd Ed.). New York: McGraw-Hill.

Simon, H. A. (1955). On a class of skew distribution functions. *Biometrika, 42,* 425–440.

Sneath, P. H. A., & Sokal, R. R. (1973). *Numerical taxonomy: Principles and practice of numerical classification.* San Francisco: W. H. Freeman.

Vickery, B. C. (1948). Bradford's law of scattering. *Journal of Documentation, 41,* 198–203.

Willis, J. C. (1922). *Age and ares: A study of geographical distribution and origin of species.* Cambridge University Press.

Yablonsky, A. I. (1980). On fundamental regularities of the distribution of scientific productivity. *Scientometrics, 2,* 3–34.

Yule, G. U. (1924). A mathematical theory of evolution based on the conclusions of Dr. J. C. Willis. *Philosophical Transactions of the Royal Society of London.* Series B, *213,* 21–87.

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