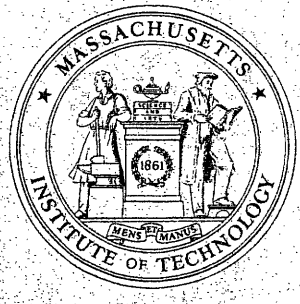
OPERATIONS RESEARCH CENTER

working paper



MASSACHUSETTS INSTITUTE OF TECHNOLOGY

THE UNDERLYING CHARACTERISTICS OF THE BRADFORD DISTRIBUTION

by

Philip M. Morse

OR 104-80 December 1980

THE UNDERLYING CHARACTERISTICS OF THE BRADFORD DISTRIBUTION.

by Philip M. Morse
MIT Oper'ns Res.Center

1. Definitions

The Bradford distribution differs from the classical probability distributions in several respects: in the range of variables of the basic variable, which is finite; and in that the engendering relationship for the distribution is between two functions of the basic variable, rather than with the basic variable itself.

To demonstrate these matters and then to display the characteristics of the distribution, we first need a few definitions. We are dealing with a finite number A of items, each having productivity n. For example, the items can be those technical journals which publish articles in a given specialty; their productivity would be the number of articles per year each would publish in the specialty. Or the items could again be journals, but the productivity would be the number of citations to a specific journal that appear in other journals. Or the items could be individual articles (or books) with the productivity being the number of citations amassed.

To define the Bradford distribution[1,2,3], we first rank-order all items in order of decreasing productivity n. There will be a scatter of items of very high n, which can be conveniently lumped together in what is called the "core", of mean productivity q_N. Below these items is a fairly continuous range, with few breaks in the n-scale, clear down to some minimal productivity M (which often equals 1). The first variable in the Bradford distribution is simply the rank-order S_n of the item, 1 plus the number of items above it in the rank-order, none of which have productivity less than it does (items with the same n can be ordered arbitrarily, but, once fixed, the order is not changed). The reason that S_n is not simply a linear function of n is that the number A_n of items with productivity n varies with n.

If A is large, S_n can be considered to be a continuous variable, whereas A_n, the number of items with productivity n, is a quite discontinuous function of n, particularly for small values of n. Perhaps this can be emphasized by using the symbol S to denote the rank-order of an item, where S goes by unit jumps from 1, for the

highest productive unit, up to A, for the lowest. Then S_n can be the rank-order of the last unit having productivity n, the next unit in order having productivity n-1. In that case

$$A_n = S_n - S_{n+1} \quad ; \quad \text{or} \quad S_n = \sum_{m=n}^{N} A_m \qquad (N \geq n \geq M) \qquad (1)$$

with A_N being the number of items in the core. The value of S_M, the sum of all items clear down to the minimal value of n, is of course equal to A, the total number of items in the collection. When M is large enough, even n may be considered to be a continuous variable, and

$$S = \int_n^N A_n dn + A_N \quad ; \quad A_n = -dS/dn \qquad (1 \ll M) \qquad (2)$$

The second variable is the cumulative production Q, the total production of the item numbered S plus that of all items above it in the rank-order. Related to it is the discontinuous function nA_n, the total production of all items with productivity n, and the cumulative function

$$\begin{aligned} Q_n &= \sum_{n=m}^{N-1} mA_m + Q_N \quad ; \quad Q_n - Q_{n+1} = nA_n \quad , \quad \text{or} \\ Q &= \int_n^N mA_m dm + Q_N \quad ; \quad nA_n = -dQ/dn \qquad (1 \ll M) \end{aligned} \qquad (3)$$

Thus Q_n is the total production of all items having production n or greater and Q_N is the total production of the core.

If S is considered to be the continuous rank-order number, then Q can also be considered to be continuous, as function of S changing slope from $dQ/dS = n$ to $dQ/dS = n-1$ as S passes from the last item of productivity n to the next unit, of productivity n-1.

It is usually convenient to normalize the range of the rank-order variable to 1, rather than A, letting $F = S/A$. F thus ranges continuously from 0 to 1 in steps of 1/A. The related production variable is $G = Q/A$. The mean productivity of all items with rank-order equal to or less than AF is then G/F and the mean productivity of the items between AF and A(F+dF) is dG/dF. We see that although F and G can be considered to be continuous variables, n, as a function of F (or G) is a discontinuous function, particularly for small values of n. Viewed as a function of F, G is a continuous variable with discontinuities in slope. For the discontinuous quantities of Eqs.(1) and (3) we have the corresponding formulas.

$$f_n = A_n/A \quad ; \quad F_n \equiv S_n/A = \sum_{m=n}^{N-1} f_m + F_N \quad ; \quad F_M = 1$$
$$F = S/A = \int_n^N f\,dn + F_N \quad ; \quad f = -\,dF/dn \quad (M \le n \le N) \tag{4}$$

$$G_n = Q_n/A = \sum_{m=n}^{N-1} m f_m + G_N \quad ; \quad G_n - G_{n+1} = n f_n$$
$$G \equiv Q/A = \int_n^N n f\,dn + G_N \quad ; \quad dG/dn = -n f \tag{5}$$

As mentioned earlier, $q_n = G_n/F_n$ is the mean productivity of all items with productivity equal to n or greater. In general it is greater than n. Also $q_M = G_M$ (M often equals 1) since $F_M \equiv A/A = 1$. so that G_M is the mean productivity of all items. F_n, G_n are the values of the continuous functions F,G, at the points where the slope of G changes discontinuously, i.e.,where productivity n changes to n-1.

2. The Bradford Requirement.

Bradford[1] noticed that S increased geometrically with Q for all three of the collections of informational items mentioned in the second paragraph of this report. He noted that if one divided up the rank-ordered items into sequential "zones", each with the same production as every other zone, then the number of items in each successive zone follows a geometric progression. If the number of units in the first zone is A(1) then the number in the k'th zone is $\alpha^{k-1}A(1)$. Put another way, if S is plotted against Q on a semi-log plot, the results closely approximate a straight line.

The closeness of typical fit is shown, for two of the three types of collections, in Figs. 1 and 2. Data are circles and the straight line and cross bars are theory. Details of the calculations, and possible reasons for the n = 1 discrepancy are discussed at the end of this paper. What is to be emphasized at this point is the remarkable degree with which the circles cluster to the straight line, over a wide range of S and of n. The geometric relationship between S and Q (or between F and G) is more accurately followed than is the detailed dependence on n (i.e., the circles are on the line, though each circle may not fall on its corresponding cross bar.

So the data indicate that rank-order location among items is geometrically (i.e.,exponentially) related to cumulative production, that the data follow a law which can be expressed in terms of the normalized quantities of Eqs.(4) and (5) as

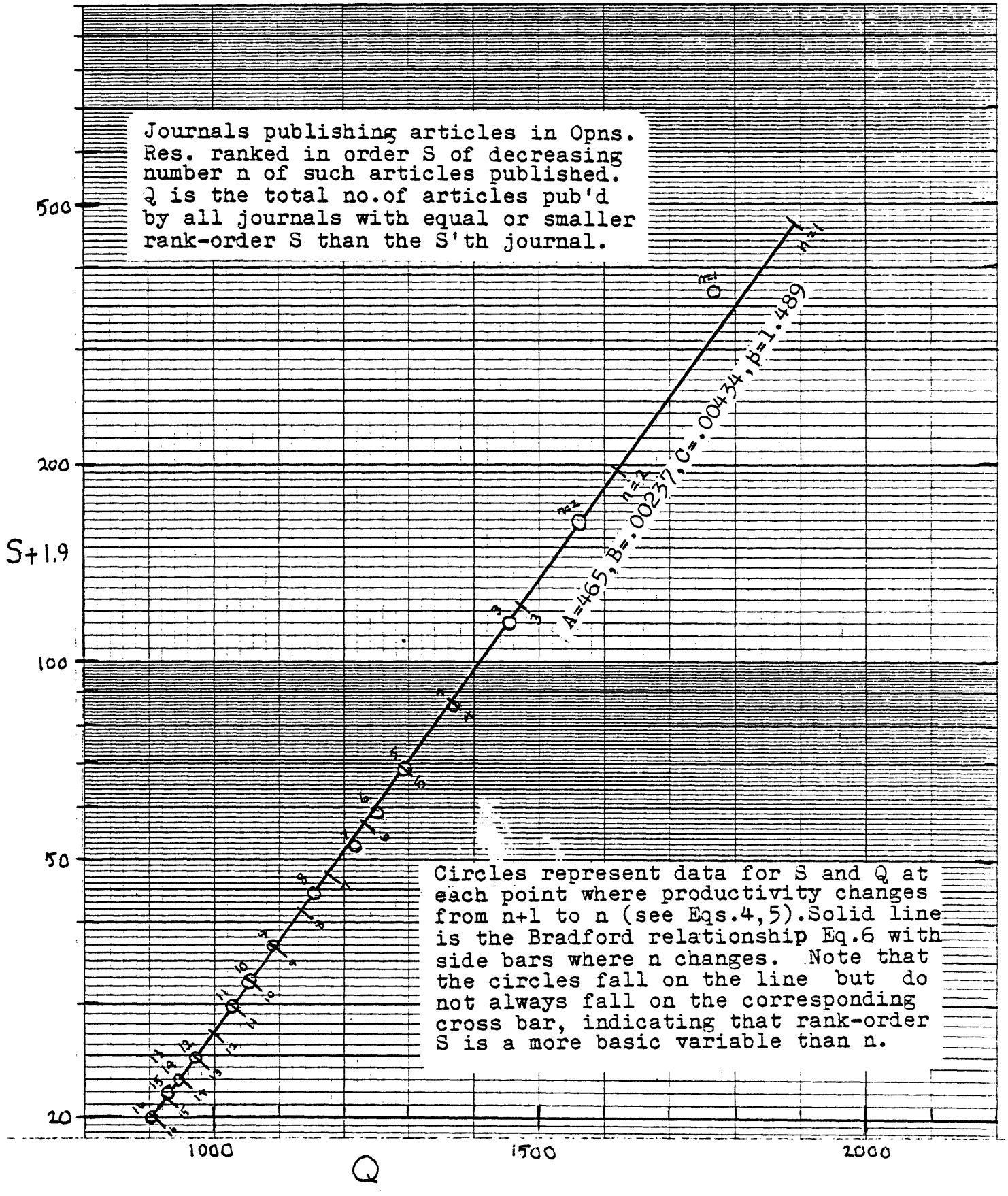


Figure 1.

Scatter of O/R articles among journals.

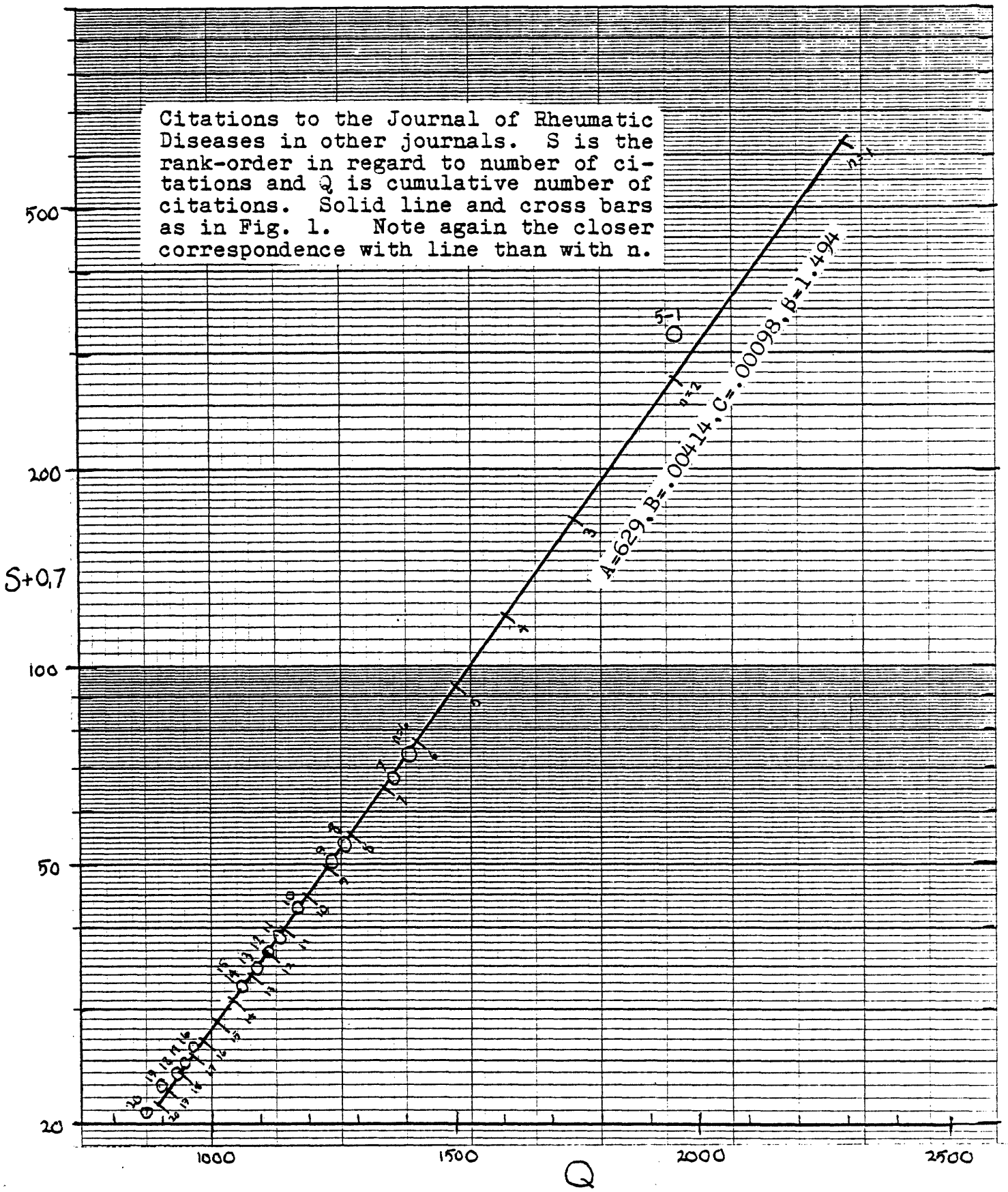


Figure 2.

Scatter of citations to a journal in other journals.

$$F = Be^{\beta G} - C$$
$$G = (1/\beta)\left[\ln(F + C) - \ln B\right] \qquad (6)$$

Remembering that F and G increase as n decreases in value, down from N, its upper limit, we see that as production (proportional to G) increases, the number of items required to provide this production, increases exponentially. This is not surprising. It of course takes more low-productivity items to provide the same production as do the high-productivity ones. What is characteristic of the Bradford distribution, and thus of a wide variety of informationally related systems, is that the increase is not linear, or proportional to some simple power, but is exponential.

It is difficult to resist speculating, at this point, as to what stochastic tendencies could so motivate the people responsible for these operational systems that the Bradford distribution would result. For example, the writers of papers on some specialty must somehow submit these papers among the appropriate journals so that the "scatter" of the articles among these journals conforms to this distribution. Similarly these writers must somehow find inspiration from other articles in journals so that their citations result in a Bradford distribution of citations among journals or articles.

The exact nature of these tendencies is not at all clear. All that can be said at present is to note what the shape of the Bradford distribution implies in this regard. From Eq.(6) we see that

$$(dF/dG) = \beta(F + C) \quad \text{or} \quad (dG/dF) = 1/\beta(F + C) \qquad (7)$$

But we have already pointed out that dG/dF is the mean productivity of the group of items with rank-order between AF and A(F + dF). We see that this mean productivity is inversely proportional to the rank-order of that group (with a small additive correction C to keep dG/dF for F = 1 from being too large). As F increases from F_N, for the core, to $F_M = 1$, for all items (i.e.,as n decreases from N to M), the mean productivity of the small group A dF of items is inversely proportional to F + C, decreasing steadily as F increases. The decrease is proportional to 1/(F + C), not to a - bF or to e^{-hF}.

In the case of the scatter of articles in a given specialty , one could imagine the tendency would operate as follows. The expected number of articles in the specialty published per year

per journal in the small group A dF is inversely proportional to the rank order established by the articles published in the previous year. If a journal has published many articles last year (its F is small) then its "popularity index", proportional to $1/(F + C)$, is large and it would expect to receive a large portion of the articles in the specialty submitted during the year. But other tendencies would prevent all articles from going to the most popular journal. Some of the other journals, with a smaller number of articles in the given specialty, may be popular with specialists in a related field, whom the writer of the paper may wish to influence.

The few conclusions we can draw from Eq.(7) regarding the tendency of authors to submit papers to journals are thus not very convincing, though they may be suggestive. Of course, even though the situation may be roughly the same from year to year, it does not mean that the journals will maintain their rank-order placement the same each year. Some will rise and some will fall. But stochastic steady state means that those journals, that comprise the group with rank order between AF and $A(F + dF)$ in a given year, will have mean production $1/\beta(F + C)$ during that year. And the consistent adherence of the data to Eq.(6) indicates that the decrease of "popularity" is inversely proportional to rank order, not proportional to n or to any other decreasing function of F. The mathematical similarity of Eqs.(6) and (7) to the Weber-Fechner law of psychophysics may be worth noting, though the parallel is probably fortuitous. Speculation regarding the other systems mentioned in the second paragraph of this paper is even more unsure.

The dependence on <u>rank-order</u> of the item in terms of decreasing productivity, rather than on the value of productivity n itself, results in a relationship between n and Q, or the normalized variable G, rather than a simple relationship between f_n and n. The distinction is analogous to the distinction between Lebesgue and Riemann integration, which is somewhat further strengthened by the fact that, if A is large, F is a continuous function of G, whereas n is a discontinuous function.

To carry the analogy further, we note that the usual probability distributions may be obtained[6] from a variational principle involving the "entropy function" $f \ln f$ of the probability $f(x)$. For example,

the exponential distribution is obtained by maximizing the entropy function integrated over x from 0 to ∞, subject to the constraints that the integral of f be unity and that the integral of xf be equal to L, the mean value of x. The integral that is to be maximized, by varying the shape of f, is

$\int_0^\infty [f \ln f - \mu f - \lambda x f]\,dx$, where μ and λ are the Lagrange multipliers, their value to be determined by the constraints. The solution of the corresponding Lagrange-Euler equation is

$$f = (1/L)e^{-x/L} ,$$ the familiar exponential distribution.

In the case of the Bardford distribution, instead of considering G to be a function of n, we consider n to be a function of G and set the "entropy function" $(1/n)\ln(1/n)$, with its integral to be maximized, subject to the constraints that the integral of f over n and that the mean value of G, averaged over n, both have constant values. The integral of the "entropy" over the full range of G is

$\int (1/n)\ln(1/n)\,dG$ and the constraints, when changed to integrals over the variable G, are

$\int dF = \int (1/n)\,dG$ and $\int f\,G\,dn = -\int (1/n)\,G\,dG$ using one of Eqs.(5). The variational integral is thus

$$\int [(1/n)\ln(1/n) - \gamma(1/n) - \beta(1/n)G]\,dG$$ with γ and β as Lagrange multipliers.

We are to adjust the shape of (1/n) as function of G to render the variation of the integral zero. The corresponding Lagrange-Euler equation becomes

$$\ln(1/n) + 1 - \gamma - \beta G = 0$$ with solution

$$(1/n) = \beta B e^{\beta G} \qquad (8)$$

or, if we wish to bring in the probability function f_n of Eq.(4),

$$f \equiv -dF/dn = \beta B n f e^{\beta G} = -\beta B\, e^{\beta G}(dG/dn) \quad , \quad \text{or}$$

$$F = B e^{\beta G} - C \qquad (9)$$

which is the same as Eq.(6). Thus the variational principle differs from that of the usual probability distributions by using the cumulative production as the independent variable and (1/n) as the dependent one, analogous to the change from Riemann to Lebesgue integration.

Discrete Variable n.

For small values of productivity n, n must be considered to be a discrete variable, even though, for large values of A, the total number of items, F and G can still be considered to be continuous variables. For example, if there are A_n items, all with productivity n, they can be arranged in some arbitrary order. While each of them has the same value of n, each item has its own rank-order number S, with S rising from S_{n+1} linearly to S_n, as one passes over the whole A_n items. The corresponding Q rises linearly from Q_{n+1} to Q_n, with a slope dQ/dS equal to n.

The relationships between F_n, G_n and n, for n small, are not quite as simple as they are for n large. The fundamental relationship between F_n and G_n (or between $S_n = AF_n$ and $Q_n = AG_n$), Eq.(6),

$$F_n = B \exp(\beta G_n) - C$$

will of course, still hold. Returning to Eq.(4), we see that this results in an equation for $f_n = A_n/A$;

$$f_n = F_n - F_{n+1} = Be^{\beta G_n}(1 - e^{-\beta n f_n}) \qquad (10)$$

Constants B, β and G_N depend on the particular system of items chosen for study. They can be removed, to arrive at more fundamental quantities. We set

$$y_n = \beta f_n \; ; \quad \beta G_n = \sum_{m=n}^{N-1} m y_m + \beta G_N = V_N + \beta G_N - V_n \qquad (11)$$

where $V_n = \sum_{m=1}^{n-1} m y_m$; and also set

$$\beta B \exp(V_N + \beta G_N) = Y_1 \quad \text{a constant}$$

so that Eq.(10) becomes

$$y_n = Y_1 e^{-V_n}(1 - e^{-n y_n}) = Y_1(e^{-V_n} - e^{-V_{n+1}}) = Y_n - Y_{n+1} \qquad (12)$$

if $Y_n = Y_1 e^{-V_n}$ But the last equation of (12) shows that

$$Y_n = \sum_{m=n}^{\infty} y_m = \beta F_n + Y_1 - \beta = \beta F_n + Y_N - \beta F_N \qquad (13)$$

since $\beta F_n = \sum_{m=n}^{N-1} y_m + F_N$ and $F_1 = 1$

Therefore Eq.(6) becomes

$$F_n = (Y_1/\beta)e^{-V_n} + 1 - (Y_1/\beta) \qquad (14)$$

so that $B = (Y_1/\beta)\exp(-V_N + \beta G_N)$ and $C = (Y_1/\beta) - 1$ are constants, values determined by the values of N, β and G_N chosen to fit a particular collection of items.

We begin the calculations for y_n and thus for Y_n and V_n for n large, and work down to n = 1. For n very large, the asymptotic solution of the first of Eqs.(12) is $Y_1\exp(-V_n) \longrightarrow (1/n)$ $(y_n \longrightarrow 0)$, so that $y_n = Y_n - Y_{n+1} \longrightarrow \frac{1}{n} - \frac{1}{n+1} \longrightarrow 1/n^2$

To improve the asymptotic series, we set

$$y_n \simeq (1/n^2) + (c/n^3) + (d/n^4) .$$

From Eq.(13) we see that

$$Y_n = T_2(n) + cT_3(n) + dT_4(n) \quad \text{where} \quad T_k(n) = \sum_{m=n}^{\infty} (1/m^k)$$

By direct calculation, using the known values of $T_k(1)$ and subtracting powers of (1/m), we can obtain $T_k(n)$ for $n \geqslant 20$ and thus work out their asymptotic series

$$T_2(n) \simeq (1/n) + (1/2n^2) + (1/6n^3) - (0/n^4)$$
$$T_3(n) \simeq (1/2n^2) + (1/2n^3) + (1/4n^4)$$
$$T_4(n) \simeq (1/3n^3) + (1/2n^4)$$

Inserting these formulas into the equation $y_n = Y_n(1 - e^{-ny_n})$ enables us to solve for c and d and thus to obtain the asymptotic formulas for all important quantities. For seven decimal accuracy for $n > 25$ we have

$$\begin{aligned} ny_n &\quad (1/n) + (0/n^2) - (1/4n^3) + (0/n^4) \\ Y_n &\simeq (1/n) + (1/2n^2) + (1/12n^3) - (1/8n^4) \end{aligned} \qquad (15)$$

To carry on to lower values of n we start from a rewrite of Eq.(12),

$$(nY_1e^{-V_n}/e^{z_n})\left[(e^{z_n} - 1)/z_n\right] = 1 \qquad (z_n = ny_n)$$

which allows us to define an adjoint function λ_n as

$$\lambda_n = (1/z_n)(e^{z_n} - 1) \quad \text{or} \quad z_n = \ln(1 + \lambda_n z_n) \qquad (16)$$

or else as

$$\lambda_n = (e^{z_n}/nY_n) = (1/nY_1e^{-V_{n+1}}) \quad \text{or}$$
$$\lambda_{n-1} = \left[1/(n-1)Y_1e^{-V_n}\right] = \left[n\lambda_n/(n-1)\right] e^{-z_n} \qquad (17)$$

The asymptotic formula for λ_n is then

$$\lambda_n \simeq 1 + (1/2n) + (1/6n^2) - (1/12n^3) - (3/40n^4) \qquad (18)$$

From the asymptotic formulas (15) for n = 30, say, we calculate Y_{30} and z_{30}. We can then use Eq.(17) to compute λ_{29} and then solve Eq.(16) by successive approximations to obtain a self-consistent value of z_{29}. Eq.(17) can then be used to obtain λ_{28} and so on, for decreasing values of n down to n = 1. One can then use the value of Y_1 and/or y_1 to compute the values of

$$U_n = Y_1 - Y_n = \sum_{m=1}^{n-1} y_m \quad \text{and} \quad V_n = \ln Y_1 - \ln Y_n = \sum_{m=1}^{n-1} m y_m \qquad (19)$$

The asymptotic formula for V_n is then

$$V_n \simeq 0.40243649 + \ln n - (1/2n) + (1/24n^2) + (1/8n^3) + (1/48n^4) \qquad (20)$$

Values of λ_n, z_n, y_n, Y_n, V_n and U_n are given to 7 decimals in Table I for $1 \leq n \leq 30$. The formulas of greatest use in fitting the data for a particular system are

$$f_n = A_n/A = y_n/\beta \quad (1 \leq n < N)$$
$$F_n = \sum_{m=1}^{N-1} f_m + F_N = 1 - (U_n/\beta) \equiv S_n/A = (Y_1/\beta)(e^{-V_n} - 1) + 1$$
$$G_n \equiv \mathcal{Q}_n/A = G_1 - (V_n/\beta) = (1/\beta)(V_N - V_n) + G_N \qquad (21)$$
$$q_n \equiv \mathcal{Q}_n/S_n = G_n/F_n \;;\; F_1 = 1 \;;\; G_1 = q_1 \;;\; G_n = q_n F_n$$
$$B = (Y_1/\beta)\exp(-\beta q_1) \;;\; C = (Y_1/\beta) - 1$$

where q_n is the mean productivity of those items with productivity not less than n; thus $q_1 = G_1$ is the mean productivity of all items.

Fitting the Data.

Data usually comes as a series of values of A_n, the number of items, in the particular collection, that have productivity n. Often, but not always, the data run clear down to n = 1, but above n = some value N, they begin to skip more and more values of n. One should pick a value N of n below which most (or all) values of n have non-zero values of A_n; the choice of N is not particularly crucial . Lump all data for $n \geq N$ into the core and calculate

TABLE I.

The Bradford Functions.

n	λ_n	z_n	y_n	Y_n	U_n	V_n
1	1.5915533	0.8671469	0.8671469	1.4954639	0	0
2	1.2805101	.4756952	.2378476	.6283170	.8671469	.8671469
3	1.1817695	.3252195	.1084065	.3904694	1.1049945	1.3428421
4	1.1339713	.2463951	.0615988	.2820629	1.2134010	1.6680616
5	1.1059305	.1981048	.0396210	.2204641	1.2749998	1.9144567
6	1.0875396	.1655524	.0275921	.1808432	1.3146208	2.1125615
7	1.0745654	.1421491	.0203070	.1532511	1.3422128	2.2781139
8	1.0649278	.1245221	.0155653	.1329441	1.3625198	2.4202630
9	1.0574903	.1107745	.0123083	.1173788	1.3780851	2.5447851
10	1.0515773	.0997533	.0099753	.1050706	1.3903934	2.6555595
11	1.0467654	.0907238	.0082476	.0950953	1.4003687	2.7553128
12	1.0427727	.0831898	.0069325	.0868476	1.4086163	2.8460366
13	1.0394078	.0768106	.0059085	.0799152	1.4155488	2.9292264
14	1.0365325	.0713378	.0050956	.0740066	1.4214573	3.0060370
15	1.0340483	.0665934	.0044396	.0689111	1.4265529	3.0773748
16	1.0318796	.0624390	.0039024	.0644715	1.4309924	3.1439682
17	1.0299709	.0587732	.0034573	.0605691	1.4348949	3.2064072
18	1.0282772	.0555126	.0030840	.0571118	1.4383521	3.2651804
19	1.0267650	.0525956	.0027682	.0540278	1.4414361	3.3206930
20	1.0254058	.0499686	.0024984	.0512596	1.4442043	3.3732886
21	1.0241783	.0475924	.0022663	.0487612	1.4467028	3.4232572
22	1.0230636	.0454310	.0020650	.0464949	1.4489691	3.4708495
23	1.0220473	.0434580	.0018895	.0444298	1.4510341	3.5162805
24	1.0211165	.0416485	.0017354	.0425404	1.4529236	3.5597385
25	1.0202612	.0399840	.0015994	.0408050	1.4546589	3.6013869
26	1.0194724	.0384473	.0014787	.0392057	1.4562583	3.6413709
27	1.0187427	.0370243	.0013713	.0377269	1.4577370	3.6798182
28	1.0180658	.0357029	.0012751	.0363556	1.4591083	3.7168426
29	1.0174360	.0344725	.0011887	.0350805	1.4603834	3.7525455
30	1.0168487	.0333241	.0011108	.0338918	1.4615721	3.7870180

$$S_N = \sum_{m=N}^{\infty} A_m \qquad \text{and} \qquad Q_N = \sum_{m=N}^{\infty} m A_m \tag{22}$$

with all values of $m \geq N$, for which A_m differs from zero, included in the sum.

Then, for each value of n between N and 1 (or M, if the data does not extend down to 1) calculate

$$S_n = \sum_{m=n}^{N-1} A_m + S_N \quad ; \quad Q_n = \sum_{m=n}^{N-1} m A_m + Q_N \quad \text{and} \quad q_n = Q_n/S_n \tag{23}$$

By combining some of Eqs.(21) we obtain the basic equation relating the three parameters specifying a particular Bradford distribution, q_1, q_N and β (and, of course, N)

$$\begin{aligned} q_n F_n = q_n\left[1 - (U_n/\beta)\right] = G_n = q_1 - (V_n/\beta) \quad &\text{or} \\ q_1 + (1/\beta)(q_n U_n - V_n) = q_n \quad &\text{or} \\ \beta = \frac{q_n U_n - V_n}{q_n - q_1} \quad \text{or} \quad q_n = \frac{\beta q_1 - V_n}{\beta - U_n} \end{aligned} \tag{24}$$

Thus a choice of any pair of these three parameters determines the distribution. Which pair should be used as basic, to compute from the data, depends on the nature and accuracy of the data. Often $q_N = Q_N/S_N$, the mean productivity of the core items, and $q_1 = Q_1/A = Q_1/S_1$, the mean productivity of the whole collection, may be computed directly from the data. In this case a best value value for β may be computed by the use of the next to last of Eqs.(24). For each value of n between N and 2 we calculate

$$B_n = \frac{(Q_n/S_n)U_n - V_n}{(Q_n/S_n) - (Q_1/A)} \tag{25}$$

obtaining values of U_n and V_n from Table I. If the values of the B_n's cluster randomly about some value, then the data fit the Bradford distribution without any adjustment and the best value of β is the average value of the B_n's.

If the B_n's vary widely in value as n goes from N to 2, then the data do not correspond to a Bradford distribution. But if there is a secular change of the B_n's with n, a small, regular change in value, it may be that the data for the smallest values of n are incomplete and our value of (Q_1/A) may represent incomplete

data. It may be difficult to count all the journals that have just one article per year in the given specialty, or have just one or two citations to a given journal. In this case we can consider both β and q_1 to be unknown and solve for their best values by least squares. We assume N and $q_N = Q_N/S_N$ as given by the data are accurate (results are not very sensitive to the choice of N). We then set down the series of equations

$$q_1 + (1/\beta)\left[(Q_n/S_n)U_n - V_n\right] = (Q_n/S_n)$$

for all values of n from N down to M, the minimal value of n for which one trusts the values of Q_n and S_n (or below which there is no data). By the usual methods of least squares the best values of β and q_1 to fit these equations are

$$\beta = \frac{K(N-M+1) - J^2}{H(N-M+1) - JL} \quad ; \quad q_1 = \frac{KL - HJ}{K(N-M+1) - J^2} \tag{26}$$

where

$$J = \sum_{n=M}^{N}\left[(Q_n/S_n)U_n - V_n\right] \quad ; \quad K = \sum_{n=M}^{N}\left[(Q_n/S_n)U_n - V_n\right]^2$$

$$L = \sum_{n=M}^{N}(Q_n/S_n) \quad ; \quad H = \sum_{n=M}^{N}(Q_n/S_n)\left[(Q_n/S_n)U_n - V_n\right]$$

Having obtained the best values of β and of q_1 for these data, we still have to determine the best value of A, to estimate how many low productivity items were missed, to see what the distribution predicts should be the values of S_n and Q_n for n less than M. We do this by using the sequence of equations

$$(S_n/A) \equiv F_n = 1 - (U_n/\beta) \quad \text{or} \quad A\left[1 - (U_n/\beta)\right] = S_n$$

for n from N down to M, using the value of β obtained from Eq.(26). The best value of A is then the mean value of $S_n/\left[1 - (U_n/\beta)\right]$,

$$A = \frac{1}{N-M+1}\sum_{n=M}^{N}\frac{S_n}{1 - (U_n/\beta)} \tag{27}$$

To show how closely data on the scatter of specialty articles among journals fits the Bradford distribution, Table II gives the counted values of S_n and Q_n for O/R articles published in various journals[4], for all values of n from 16 to 1 for which A_n differs from zero. The value of Q_1/S_1 is 4.765. If this is taken to be the value of q_1, then the values of B_n of Eq.(25) are given in the 4th column. We see that the values change secularly from 1.51 to 1.61 and then leap to 3.00 for n = 2. Next we try to see

TABLE II.

Scatter of O/R Articles among Journals.

n	S_n	Q_n	B_n $q_1=4.8$	B_n $q_1=4.1$	S_n Theor	Q_n Theor
16	18	902	1.51	1.491	18	906
15	20	932	1.52	1.492	20	927
14	21	946	1.52	1.491	21	949
13	21	946	1.51	1.487	21	973
12	23	970	1.51	1.486	25	999
11	28	1025	1.48	1.492	28	1027
10	31	1055	1.53	1.495	31	1058
9	35	1091	1.53	1.491	35	1093
8	43	1155	1.55	1.500	40	1132
7	51	1211	1.57	1.510	46	1176
6	57	1247	1.56	1.494	55	1228
5	67	1297	1.56	1.490	67	1290
4	84	1365	1.57	1.475	86	1366
3	113	1452	1.61	1.483	120	1468
2	167	1560	3.00		194	1616
1	370	1763			465	1887

$(N-M+1) = 14$; $L = 451.7$; $J = 588.1$; $H = 21921$; $K = 29093$

$\beta = 1.489$; $q_1 = 4.060$; $A = 465$

if the fit would be better if we assumed that the data for n = 1 and 2 were incomplete. Using Eqs.(26) and (27) for N = 16 and M = 3 we compute values of $\beta = 1.489$, $q_1 = 4.06$ and A = 465 (instead of 370 as counted). Column 5 gives values of B_n, using this value of q_1; we see that the values cluster very closely to the best value of 1.489 for β. Using these best values, columns 6 and 7 give values of $A[1 - (U_n/\beta)]$ and $A[q_1 - (V_n/\beta)]$, which should equal the S_n and Q_n of columns 2 and 3. We see that the check is quite good, except for n = 1 and 2, of course. It is not impossible that 27 journals out of 194, having but two O/R articles per year, were missed and about 70 out of 270 journals,with only one such article, were not counted.

A graphical comparison shows an even closer fit. In Fig.1 we plot $S_n + A[(Y_1/\beta) - 1] = S_n + 1.9$ against Q_n (see Eqs.14 and 21) and compare tham with the straight line between the points $AF_n + A[(Y_1/\beta) - 1]$ for A = 465, $\beta = 1.489$, on which cross bars have been marked for each value of n. We notice that, except for the circle for n = 1 (which we have already called into question by using Eq.26) the circles fall more closely on the line than they do on cross bar for the corresponding n. In other words the logarithmic relationship between rank-order S = AF and Q = AG, as given in Eq.(6), is adhered to more closely than is the apportionment of A_n's, the exact number of journals with a particular value of n. This is another illustration that, somehow, the rank-order S of a journal is more important, in deciding author's prediliction for that journal, than is the exact number of articles in the specialty its editor publishes each year.

Table III gives the same analysis of data [5] for the scatter of citations to the Journal of Rheumatic Diseases in other journals, a case picked at random from the 1977 Citation Index. Exact data stopped at n = 6, but there was an estimate that 252 other journals had had less than 6 citations each, with 533 total citations from these low-yield journals. Here, of course, we have to use Eqs.(26) and (27), with N = 20 and M = 6. The calculations indicate that the best values are $\beta = 1.494$, $q_1 = 3.673$ and A = 629. The fourth column, listing B_n for $q_1 = 1940/325 = 5.97$, displays a secular change of B_n from 1.59 to 1.75 for n = 6, which indicates that that the counts for $n < 6$ are likely incomplete, and that the value of 3.67 for q_1,

TABLE III.

Scatter among Journals of Citations to Jour. Rheum. Dis.

n	S_n	Q_n	B_n q_1=6.0	B_n q_1=3.7	S_n Theor	Q_n Theor
20	20	872	1.59	1.494	21	889
19	22	910	1.59	1.496	22	912
18	23	928	1.59	1.496	23	935
17	24	945	1.60	1.493	25	960
16	25	961	1.60	1.493	26	986
15	32	1066	1.63	1.500	28	1014
14	32	1066	1.63	1.500	30	1044
13	34	1092	1.63	1.496	33	1076
12	36	1116	1.63	1.495	36	1111
11	38	1138	1.64	1.492	39	1150
10	42	1178	1.65	1.494	43	1192
9	49	1241	1.67	1.495	49	1238
8	53	1273	1.68	1.493	55	1291
7	67	1371	1.74	1.500	64	1350
6	73	1407	1.75	1.494	75	1420
5					92	1503
4					118	1607
3					164	1744
2					264	1944
1	325	1940			629	2309

$(N-M+1) = 15$; $L = 479.8$; $J = 634.4$; $H = 21509$; $K = 28649$

$\beta = 1.494$; $q_1 = 3.673$; $A = 629$

obtained by solving Eq.(26), is probably better. Column 5 shows that the values of B_n, using this value of q_1, cluster relarkably closely around the "best value" 1.494 for β. Columns 6 and 7 check quite well with columns 2 and 3, the data, down to n = 6. The values given in columns 6 and 7 for n less than 6 indicate the expected values of S_n and $\mathcal{Q}_n$ if the Bradford distribution were really to hold clear down to n = 1.

Fig.2 again shows the closeness with which the data fit the straight line representing the curve for AF + 0.7 versus AG, with cross bars indicating the values of $(AF_n + 0.7, AG_n)$ at the points where productivity changes from n+1 to n. The circles are the corresponding data points for different n's, down to n = 6. The circle n = 1-5 shows that the reported estimate that 252 journals produced 533 citations does not fall on the curve. Either the Bradford distribution does not hold for n less than 6, or the data for n less than 6 are incomplete.

Again we note that the circles fall on the straight line more closely than they fall on the cross bars, once more indicating that n is less important in guiding the author's interest in a journal's contents than is the rank-order of the journal, among those in his specialty.

3. A Markov Process.

In part of the previous discussion we postulated that last year's productivity of an item must somehow inspire the people responsible for this year's productivity (contributers of articles, makers of citations, etc.) so that this year's productivity scatter also is Bradford. The dynamics of such a process may be represented as a Markov process.[7] The probability that an item, which had productivity m last year, has productivity j this year can be written as a Markov transition matrix p(j if m). In this case the probability $f_j(t+1)$ that an item has productivity j this year is related to the probability $f_m(t)$ that it had productivity m last year is the sum

$$f_j(t+1) = \sum_m p(j \text{ if } m) f_m(t) \qquad (28)$$

When a stochastic steady state exists, $f_j(t+1) = f_j(t)$. This does not mean that each individual item maintains its productivity from year to year; all it means is that the same number of items

each year enter the n'th productivity group as leave it, so that the fraction f_n that have productivity n remain the same from year to year. Of course the Markov process can deal with systems that change from year to year, but our discussion will generally be concerned with the steady state case.

First it should be noted that p(0 if m) and p(j if 0) are not necessarily zero; a journal, for example, that did not publish an article in the specialty last year may publish m next year and, vice versa, one that published j articles last year may not publish any next year. However the Bradford distribution does not include items with zero productivity, so the marginal transition probabilities p(o if m) and p(j if 0) must somehow be incorporated into the summations for f_j, which range from j = N (items in the core) to M (usually 1). This can easily be done when roughly the same <u>number</u> of journals (not necessarily the <u>same</u> journals) publish articles in the specialty each year, for then the number of items that enter the distribution each year must equal the number that drop out. Thus we have the equation

$$\sum_{m=1}^{N} p(0 \text{ if } m) f_m = \sum_{j=1}^{N} p(j \text{ if } 0) f_j \equiv T \qquad (29)$$

where T represents the temporary members of the collection of items, which move in and out of the collection each year, changing membership as they do so, but as many moving in as out, if A is to remain roughly the same from year to year.

Therefore the items that move into or out of the distribution f_n need only be counted while they are being productive. They need not be counted separately from the f_n's, for they are are only present when they are productive and are thus among those measured by $f_m(t)$ or $f_j(t+1)$. The collection of active items remains about the same each year, but the components change somewhat.

However the transition probabilities p(j if n) need to be modified if Eqs.(28) are to range over j and m from N to 1. For example, we have, for steady state,

$$\begin{aligned} f_j(t+1) = f_j(t) &= p(j \text{ if } 0)T + \sum_{m=1}^{N} p(j \text{ if } m) f_m(t) \\ &= \sum_{m=1}^{N} \left[p(j \text{ if } m) + p(j \text{ if } 0) p(0 \text{ if } m) \right] f_m(t) \end{aligned} \qquad (30)$$

from Eq.(29). Thus we can confine our distribution and our transition probabilities to the range of indices from N to 1, inclusive, those included in the Bradford distribution, by redefining the

transition probabilities (and incidentally relabelling them more in line with Markov process literature)

$$P_{mj} = p(j \text{ if } m) + p(j \text{ if } 0)p(0 \text{ if } m) \qquad (31)$$

for all values of m and of j from 1 to N inclusive. The matrix P_{mj} is thus an N by N matrix, each row, designated by m, giving the probability that a unit that had productivity m last year will have productivity j next year. For steady state we thus have

$$f_j = y_j/\beta = \sum_{m=1}^{N} f_m P_{mj} \qquad (y_N = 1 - U_N/\beta) \qquad (32)$$

where we need not distinguish between t and t+1 for steady state. We are thus in the unusual position of knowing the steady-state distribution f_n and not knowing what Markov transition probability produces it. As mentioned before, P_{mj} is the probability that an item, having productivity m last year, will have productivity j next year, including those that dropped out of m and came into j from the universe of inactive items. Since probability distributions do not specify which item belongs where, but simply assign the number of items that have productivity m, we can consider the items that move into the collection to be the same as those that move out, and allocate them to the m's and j's as though no items moved in or out.

Even if steady state does not prevail, the fact that both $f_m(t)$ and $f_j(t+1)$ must be Bradford distributions (i.e., must equal y_m/β and y_j/β' with the y's of Table I) puts a very stringent restriction on the form of the matrix P_{mj}. One possibility is that each row of the matrix be itself a Bradford distribution

$$P_{mj} = y_j/\beta_m \quad (1 \leq j \leq N-1) \quad = 1 - (U_N/\beta_m) \quad (j = N) \qquad (33)$$

where the values of the N constants β_m can vary with m and differ from the value of the β for the whole distribution f_j.

If the system is in steady state, so that Eq.(32) holds, there is a single constraint narrowing the choice of the β_m's;

$$\begin{aligned} f_j &= \sum_{m=1}^{N-1} (y_j/\beta_m)(y_m/\beta) + (y_j/\beta_N)\left[1 - (U_N/\beta)\right] \quad (1 \leq j < N) \\ f_N &= \sum_{m=1}^{N-1} \left[1 - (U_N/\beta_m)\right](y_m/\beta) \\ &\qquad + \left[1 - (U_N/\beta_N)\right]\left[1 - (U_N/\beta)\right] \quad (j = N) \end{aligned} \qquad (34)$$

where $\beta_m \rightarrow \beta_N$ is the parameter for those items that had productivity m last year (the m-group). The equations are satisfied for all values of j, from 1 to N inclusive, if

$$\beta = \beta_N + U_N - \beta_N \sum_{m=1}^{N-1} (y_m/\beta_m) \qquad (35)$$

which relates the β_m's for the m-groups to the steady state β for the whole collection.

It is far from clear that the transition probability P_{jm} should (or, if it does, why it should) follow the Bradford distribution; that journals, for example, that had m articles in a specialty last year should have a scatter of such articles in accord with the Bradford distribution this year. If study of year to year changes in data shows that this is indeed the case, it must mean, for example, that authors somehow rank-order all the journals that had m articles in the specialty last year and then scattered their submissions among these journals so that the relationship between that rank order F(m) and the cumulative production variable G(m) is that given by Eq.(16) with $\beta = \beta_m$. Since authors scatter their articles among all the active journals in this manner (as the data shows) perhaps they also treat the individual m-groups the same way.

One thing is certain; if the rows of the transition matrix P_{mj} (the scatter of the individual m-groups) are all Bradford distributions then the next year's distribution $f_j(t+1)$ must be Bradford, no matter what distribution $f_m(t)$ is. Thus the fact that $f_m(t)$ is Bradford, as well as $f_j(t+1)$, in a way strongly suggests that the rows of P_{mj} are Bradford.

There are, of course, other forms of P_{jm} that transform a Bradford distribution into itself. For example

$$P_{mj} = \begin{cases} \alpha_{mj} & (j < m) \\ 1 - \sum_{n=1}^{m-1} \alpha_{mn} - \sum_{n=m+1}^{N} \alpha_{nm}(y_n/y_m) & (j = m) \\ \alpha_{jm}(y_j/y_m) & (N \geq j > m) \end{cases} \qquad (36)$$

where y_m is given in Table I (and we let $y_N = 1 - U_N/\beta$, for convenience in writing the equation), β has the appropriate value for the steady state distribution f_j, and the (N/2)(N-1) different α's have any values thst allow P_{jj} to be non-negative. Inserting this form into Eq.(32) for steady state Bradford distribution, we have

$$f_j = (y_j/\beta) = \sum_{m=1}^{N} f_m P_{mj} = (1/\beta)\sum_{m=1}^{N} y_m P_{mj} \qquad (37)$$

$$= (1/\beta)\left\{\sum_{m=1}^{j-1}\alpha_{jm}y_j + \left[y_j - \sum_{n=1}^{j-1} y_j\alpha_{jn} - \sum_{n=j+1}^{N}\alpha_{nj}y_n\right] + \sum_{m=j+1}^{N} y_m\alpha_{mj}\right\}$$

$$= (y_j/\beta) = f_j$$

since the sums cancel out in pairs. This verifies the fact that this form of P_{mj} does indeed transform a Bradford distribution into itself.

However this is a rather specially constructed transition matrix. Unless the relationship between P_{mj} $(j < m)$ and its reflection across the main diagonal, $P_{jm} = P_{mj}(y_j/y_m)$ $(j > m)$ are maintained for each m and j the matrix will not convert a Bradford distribution into another Bradford distribution, let alone into itself. It would appear that such a matrix is unlikely to represent the dynamics of the scatter process we are investigating. The matrix given by Eq. (33) is more likely to correspond to actuality, for even if there is no steady state, and Eq.(35) does not hold, it does ensure that $f_m(t+1)$ is a Bradford distribution, in accord with the data.

But, of course, speculation is unproductive, what is needed is data on the yearly change of the distributions we have been analyzing.

Finally, for steady state, in addition to Eq.(35), we must arrange to have the mean productivity q_1 to remain the same from year to year. Therefore the mean productivities $q_1(\text{if } m)$ for each m-group must be related to the mean productivity q_1 of the whole collection by the equation

$$q_1 = \sum_{m=1}^{N} q_1(\text{if } m) f_m$$

$$\sum_{m=1}^{N-1} q_1(\text{if } m)(y_m/\beta) + q_1(\text{if } N)\left[1 - (U_N/\beta)\right] \qquad (38)$$

If the two parameters β_m and $q_1(\text{if } m)$ are determined for each m-group, then the transition probabilities P_{mj} are determined and the parameters β and q_1 for the whole collection are also determined.

When data, on yearly change of the informational collections we have been discussing, are collected and analyzed, we can begin to understand the dynamics of the flow of scientific information.

November 1980.

References.

1. S.C.Bradford, Documentation, Crosby Lockwood, London, 1948

2. B.C.Brookes,"The Derivation and Application od the Bradford-Zipf Distribution", J.Documentation,24,247-265 (1968).

3. P.M.Morse and F.F.Leimkuhler,"Exact Solution for the Bradford Distribution and its Use in Modelling Informational Data", Ops.Res. 27, 187-198 (1979)

4 M.G.Kendall,"The Bibliography of Operational Research", Opnl.Res.Quart. 11, 31-36 (1960)

5 1977 Citation Index.

6 P.M.Morse, Thermal Physics, 2nd Ed. Chapter 17, W.A.Benjamin, New York, 1969

P.M.Morse, "Information Theory and Probability Distributions" Working Paper 07-74, MIT O/R Center (1974).

7 A.T.Bharucha-Reid, Elements of the Theory of Markov Processes and their Applications, McGraw-Hill Book Co., NY, 1960