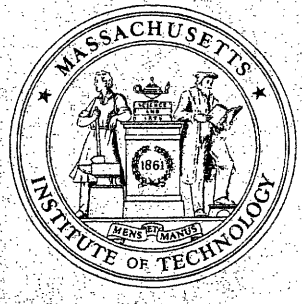
OPERATIONS RESEARCH CENTER

working paper



MASSACHUSETTS INSTITUTE OF TECHNOLOGY

USE OF DISTRIBUTION FUNCTIONS TO DESCRIBE THE FLOW OF SCIENTIFIC INFORMATION

by

Philip M. Morse

OR 087-79 March 1979

USE OF DISTRIBUTION FUNCTIONS TO DESCRIBE THE FLOW OF SCIENTIFIC INFORMATION.

March, 1979

by Philip M. Morse
MIT OR Center

Scientific information flows from source to ultimate user in a number of ways. This note concentrates on the flow via the published scientific literature and its use via university, project or industrial library. At each step semi-random matters of choice govern the rates of flow and thus determine the probability distributions describing the flow. For example, the individual scientist, or member of a research team, chooses the particular journal to which his paper is submitted; whether it appears in that journal depends on the choice of the editor and his referees. The library has a choice of which journals to subscribe to. And, finally, a scientist, in carrying out new research, has a choice of which journals, in the library and accessible to him, he peruses in the hope of finding ideas or data pertinent to his present research. In each step of this process the related distribution functions combine, resulting in a new distribution, descriptive of the next stage of flow.

Thus the effects of combining distribution functions are important in the study of this cyclic flow of scientific information. It is the purpose of this note to discuss the combining properties of the distribution functions that have been found useful in describing various aspects of the flow. In all such distributions we are describing how a number A of <u>items</u> (journals, users, articles, etc.) are distributed with respect to what we shall call <u>productivity</u> (number of articles in a given field in

a volume of a given journal, journal usage ("used" n times a year, or month) by specialists in a given field, probability that a scientist contributes n articles per year to a given journal, number of citations amassed by a given article during x years after its publication, etc.) The fraction of the A items that have productivity n will be called f_n (or, at times, P_n), satisfying the usual requirements

$$\sum f_n = 1 \quad ; \quad \sum n f_n = q \equiv \text{mean productivity of all items} \qquad (1a)$$

where the summations are over all allowed values of n. For the cumulative distribution function

$$F_n = \sum_{m=n} f_m \quad ; \quad F_{min} = 1 \quad ; \quad \sum_{min} F_m = q \qquad (1b)$$

where min is the lower limit of n and where the summations are over all allowed values of n not less than the lower limit shown.

The two distributions particularly (but not exclusively) studied in this note are the well-known <u>geometric</u> distribution,

$$f_n = (1-\gamma)\gamma^n \quad ; \quad F_n = \gamma^n \quad (0 \leq n < \infty) \quad ; \quad q = \gamma/(1-\gamma) \qquad (2)$$

and the <u>Bradford</u> distribution. The basic properties of this distribution have been described in the MIT-OR Center Working Paper Mo. OR 068-77. The main properties are given in Table I of this note, and a more extensive and more accurate set of values of the more important quantities are given in Table II.

The important properties of the Bradford distribution are:

1. The distribution index n ranges from 1 to some maximum productivity N; all items of productivity of N or greater are lumped together into one upper limit F_N, with mean productivity $G_N/F_N \equiv q_N \geq N$ (see Table I for definitions).

2. The number $AF_n = A\left[1 - (1/\beta)U_n\right] = A\left[1 - (1/\beta)Y_1 \exp(-V_n)\right]$,

of items with productivity greater or equal to n depends exponentially on AG_n ($V_n = \beta G - \beta G_n$), the total production of those items. This is the Bradford condition.

3. The distribution depends on one table of values of the quantities Y_n, y_n, U_n, and V_n, tabulated in Table II and (aside from the total number A of items and for N, the value of n for the core) depends on the single parameter β, which depends on N and on the lower and upper mean productivities q_1 and q_N and on the Tabulated quantities U_N and V_N

$$\beta = \frac{q_N U_N - V_N}{q_N - q_1} \quad ; \quad q_1 = q \quad ; \quad q_N = G_N/F_N \tag{4}$$

Moreover, the dependence on β is purely multiplicative;

$$\begin{aligned} f_n &= y_n/\beta \quad (1 \leq n < N) \quad ; \quad f_N = 1 - (U_N/\beta) \equiv F_N \\ F_n &= 1 - (U_n/\beta) \quad ; \quad G_n = G_1 - (V_n/\beta) \equiv q_1 - q_n F_n \quad (1 \leq n \leq N) \end{aligned} \tag{5}$$

where AG_n is the total production of all items equal to or greater than n and $q_n = G_n/F_n$ is the mean productivity of all these items. This last property gives the Bradford distribution useful properties when combining with other distributions.

Thus, if the A items are also distributed according to some other criterion j (number of times per year or month a particular journal is "used" in a library, or the probability that an author submits an article to a particular journal, for example) and this distribution depends on n, the productivity of the item, then the probability that an item have property j <u>if</u> it has productivity n, can be defined as p(j if n), where $\sum_j p(j \text{ if } n) = 1$. (In many cases p(j if 0) = 0, interest in zero productivity items is zero), Then the combined probability that an item of A has property j <u>as well as</u>

productivity n is

$$p(j \text{ and } n) = p(j \text{ if } n) f_n = p(n \text{ if } j) p_j \quad (1 \leq n \leq N) \tag{6}$$

where p_j is the unconditional probability that an item of A has property j and p(n if j) is the conditional probability that the item has productivity n if it has property j $\left[\sum_n p(n \text{ if } j) = 1 \text{ and } \sum_j p(j \text{ if } n) = 1 \right]$.

Because of the structure of the Bradford distribution f_n, the summation over n to find p_j is particularly simple;

$$\begin{aligned} p_j &= \sum_{n=1}^{N-1} p(j \text{ if } n)(y_n/\beta) + p(j \text{ if } N)\left[1 - (U_N/\beta)\right] \\ &= p(j \text{ if } N) + (1/\beta) R_j(N) \qquad \text{where} \end{aligned} \tag{7}$$

$$R_j(N) = \sum_{n=1}^{N-1} p(j \text{ if } n) y_n - p(j \text{ if } N) U_N$$

Since $\sum_j R_j(N) = 0$, we are assured that $\sum_j p_j = 1$. The quantities $R_j(N)$ are independent of β, they are the same for all Bradford distributions with the same N. The mean value of j (mean usage or number of articles submitted, etc.) for those items of productivity n, would be $J(n) = \sum_j j\, p(j \text{ if } n)$ and the mean value of j for all A items, for all values of n, would be

$$\bar{J} = \sum_j j\, p_j = \sum_{n=1}^{N} J(n) f_n = J(N)\left[1 - \frac{1}{\beta}(Y_1 - Y_N)\right] + \frac{1}{\beta} \mathcal{Q}(N) \tag{8}$$

$$\text{where} \quad \mathcal{Q}(N) = \sum_{n=1}^{N-1} J(n) y_n$$

Thus the mean value of j can be expressed in terms of the quantities in Table II, plus the values of J(N), determined from the known p(j if n), finally using the value of β, determined from the N, q_1 and q_N for the appropriate Bradford distribution.

As an example we could assume that the distribution of usage j of a journal in the library was geometric, with the mean usage rate J(n) for a journal with productivity n (n articles on a given subject per volume) such that

$$p(j \text{ if } n) = (1 - \gamma_n)\gamma_n^j \quad ; \quad \gamma_n \equiv \frac{J(n)}{1 + J(n)} = \alpha(1 - e^{-\mu n})$$

$$J(n) = \frac{\gamma_n}{1 - \gamma_n} = J(\infty)\frac{1 - e^{-\mu n}}{1 + J(\infty)e^{-\mu n}} \quad ; \quad \alpha = \frac{J(\infty)}{1 + J(\infty)} \tag{3a}$$

A few values of the quantities $R_j(10)$, $J(N)$ and $\mathcal{Q}(N)$ are given in Table III. All of these can be computed independently of β, and thus of q_1 and q_N. We see that $R_j(10)$, β times the difference between p_j and $p(j \text{ if } 10)$, adds to $p(j \text{ if } n)$ for $j = 0$ and subtracts from it for $j > 0$. Thus the chance p_o of a journal, chosen at random from the library, not being used at all by specialists in the designated field, is greater than it is for those journals that have N articles in the field per volume -- a not surprising result. We see also that the new distribution p_j is neither geometric nor Bradford. Its form, however, is one that can be combined further with little more trouble than would be encountered in combining $p(j \text{ if } n)$ with another distribution.

A still greater simplification can be obtained if the combination of distributions is the other way around. Suppose, for example, that the fraction of scientific journal volumes in a given library, which have k articles in a given scientific field, is P_k and suppose that the fraction of these journals that are consulted (used) by specialists in the field is distributed according to the Bradford distribution, with a value of β_A that depends on k (q_1 and/or q_N depend on k). Then the fraction, of all A journal volumes in the library, that have k articles in the given field and are consulted n times a year (or month) by specialists in the field is

$$p(n \text{ and } k) = p(n \text{ if } k)P_k = p(k \text{ if } n)f_n \qquad (9)$$

where we have assumed that

$$p(n \text{ if } k) = y_n/\beta_k \quad (1 \leq n < N) \quad ; \quad p(N \text{ if } k) = 1 - (1/\beta_k)U_N$$

In addition we probably should assume that the specialist knows what journals have articles in his specialty, so he looks only at journals with at least one article in his specialty, neglecting the fraction P_o of journal volumes in A that contain no article in the field, i.e., we should set $p(n \text{ if } 0) = 0$. Thus we should consider, not all A journal volumes in the library, but only those that have at least one article in the field, and consider the distribution $P'_k = P_k/(1 - P_o)$ $(k \geq 1)$ of the $A' = (1-P_o)A$ volumes with one or more articles in the field. Then $\sum_{k=1} P'_k = 1$.

If this is the case, then the unconditional probability that one of the A' journal volumes is used n times a year (month) by specialists in the field is, from Eq. (9),

$$f_n = \sum_{k=1} p(n \text{ and } k) = y_n/\beta \quad (1 \leq n < N) \quad ; \quad f_N = 1 - (1/\beta)U_N \qquad (10)$$

$$\text{where} \quad (1/\beta) = \sum_{k=1} (1/\beta_k)P'_k$$

Therefore the distribution in use, by specialists in the field, of the journals in the field in the library, is also a Bradford distribution, with a β equal to the reciprocal mean, of all the β_k's, over the distribution P'_k of articles per journal (of productive journals).

An interesting special case is where P_k also is a Bradford distribution. For example, suppose the distribution in papers published, among the A journals that publish at all in some field, is Bradford and suppose further that the fraction of those journals that published n articles last year (or month), which had k

articles in the field submitted to and accepted for publication in its next year (or month), p(k if n), also is a Bradford distribution, with parameter β_n. That is

$$f_n = y_n/\beta \quad (1 \leq n < N) \quad ; \quad f_N = 1 - (U_N/\beta)$$
$$p(k \text{ if } n) = y_k/\beta_n \quad (1 \leq n \leq N) \quad ; \quad p(N \text{ if } n) = 1 - (U_N/\beta_N) \tag{11}$$

Then the unconditional probability that a journal in the field publishes k articles in the field next year (or month) is

$$f'_k = \sum_{n=1} p(k \text{ if } n) f_n = y_k/\beta' \quad (1 \leq k < N) \quad ; \quad 1 - (U_N/\beta'_N) = f'_N$$
$$\text{where } (1/\beta') = (1/\beta)\left[\sum_{n=1}^{N-1}(y_n/\beta_n) - (U_N/\beta_N)\right] + (1/\beta_N) \tag{12}$$

which is another Bradford distribution (as it should be) with another value β' of the parameter. If $\beta' = \beta$ then the distribution of articles published is steady-state, which sets some conditions for the values of the parameters β_n for the conditional probabilities p(k if n). One could carry this further by considering the p(k if n) to be Markov transition probabilities.

It would be interesting to know how restrictive are the requirements on the form of the distribution for it to satisfy the combining properties expressed in Eqs.(11) and (12). In any case, since the distribution over journals of articles in a given specialty has been shown to be Bradford in many cases, it appears that the combined choice of authors plus referees results in a conditional probability p(k if n), of k articles per year submitted and accepted by the journal with n articles last year, that also is Bradford, as was assumed in Eq.(11). This suggests speculation as to whether there is some socio-stochastic reason for the submission of papers to journals following a Bradford distribution. A tendency to submit papers that increases exponentially with the productivity of the journal would produce such a distribution.

TABLE I. THE BRADFORD FUNCTION.

A = Total number of productive items. Af_n = No. with prod. n.

$F_n = \sum_{m=n}^{N-1} f_m + F_N = 1 - \frac{1}{\beta}(Y_1 - Y_n) = 1 - (U_n/\beta)$ $(1 \leq n \leq N)$ is the fraction of items with productivity n or greater. All items with productivity N or greater are lumped together in the core.

$Y_n = \sum_{m=n}^{\infty} y_n$; $U_n = \sum_{m=1}^{n-1} y_m$; $f_n = y_n/\beta$ $(1 \leq n < N)$; $f_N = F_N$; $\sum_{m=1}^{N} f_m = 1$

f_n is the fraction of items with productivity n $(1 \leq n \leq N)$.

f_N is the fraction of items in the core.

$G_n = \sum_{m=1}^{n-1} m f_m + G_N = q_1 - (V_n/\beta)$ $(1 \leq n \leq N)$; $V_n = \sum_{m=1}^{n-1} m y_m = nU_n - \sum_{m=1}^{n} U_m$

AG_n is the total production of all items with productivity equal to n or greater. $G_1 = q_1$

$q_n = G_n/F_n$ is the mean productivity of items with productivity $\geq n$.

q_1 = mean productivity of all A items; q_N = mean productivity of core.

$Y_n = Y_1 \exp(-V_n)$ or $F_n = 1 - (Y_1/\beta)\left\{1 - \exp\left[\beta(G_n - G_1)\right]\right\}$ (Bradford rule).

For $n \geq 20$,

$y_n \cong (1/n^2) - (1/4n^4)$; $z_n \equiv n y_n \cong (1/n) - (1/4n^3)$

$Y_n \cong (1/n) + (1/2n^2) + (1/12n^3) - (1/8n^4)$; $U_n \cong 1.4954639 - Y_n$

$V_n \cong 0.4024434 + \ln(n-1) + (1/2n) + (1/2n^2) + (1.76/n^3) - (11/n^4)$

β is given in terms of N (items with productivity N or greater belong to the core), q_N mean productivity of core and $q_1 = G_1$ mean productivity of all A items.

$$\beta = \frac{q_N U_N - V_N}{q_N - q_1} \cong \frac{1}{q_N - q_1}\Big[1.4954639 q_N - 0.4024484 - \ln(N-1) - (1/2n)(2q_N+1) - (1/2n^2)(q_N+1) - (1/12n^3)(q_N+21.12) + (1/8n^4)(q_N+38)\Big] \quad \text{for } n \geq 20.$$

There are several limitations on the ranges of N, q_1 and q_N. First, q_N must be larger than q_1. Also q_N must be larger than V_N/U_N; in fact it is only reasonable to require that q_N, the mean productivity of the core, be larger than N-1, the productivity of the most productive items outside the core.

TABLE II

THE BRADFORD FUNCTION.

n	z_n	y_n	Y_n	V_n	U_n
1	0.8671469	0.8671469	1.4954639	0	0
2	.4756952	.2378476	.6283170	0.8671469	0.8671469
3	.3252195	.1084065	.3904694	1.3428421	1.1049945
4	.2463951	.0615988	.2820629	1.6680616	1.2134010
5	.1981048	.0396210	.2204641	1.9144567	1.2749998
6	.1655524	.0275921	.1808432	2.1125615	1.3146208
7	.1421491	.0203070	.1532511	2.2781139	1.3422128
8	.1245221	.0155653	.1329441	2.4202630	1.3625198
9	.1107745	.0123083	.1173789	2.5447851	1.3780851
10	.0997533	.0099753	.1050706	2.6555595	1.3903934
11	.0907238	.0082476	.0950953	2.7553128	1.4003687
12	.0831898	.0069325	.0868476	2.8460366	1.4086163
13	.0768106	.0059085	.0799152	2.9292264	1.4155488
14	.0713378	.0050956	.0740066	3.0060370	1.4214573
15	.0665934	.0044396	.0689111	3.0773748	1.4265529
16	.0624390	.0039024	.0644715	3.1439682	1.4309924
17	.0587732	.0034573	.0605691	3.2064072	1.4348949
18	.0555126	.0030840	.0571113	3.2651803	1.4383521
19	.0525956	.0027682	.0540278	3.3206929	1.4414361
20	.0499686	.0024984	.0512596	3.3732885	1.4442043
21	.0475923	.0022663	.0487612	3.4232571	1.4467028
22	.0454310	.0020650	.0464949	3.4708495	1.4489691
23	.0434579	.0018895	.0444298	3.5162804	1.4510341
24	.0416484	.0017354	.0425404	3.5597384	1.4529236
25	.0399842	.0015994	.0408050	3.6013868	1.4546589
26	.0384472	.0014787	.0392057	3.6413710	1.4562583
27	.0370245	.0013713	.0377269	3.6798182	1.4577370
28	.0357029	.0012751	.0363556	3.7168427	1.4591083
29	.0344724	.0011887	.0350805	3.7525456	1.4603334
30	.0333241	.0011108	.0338918	3.7870180	1.4615721

TABLE II

continued

n	z_n	y_n	Y_n	V_n	U_n
32	.0312424	.0009763	.0317407	3.3525921	1.4637232
34	.0294054	.0008649	.0298463	3.9141309	1.4656176
35	.0285656	.0008162	.0289815	3.9435365	1.4664824
36	.0277724	.0007715	.0281653	3.9721023	1.4672986
38	.0263112	.0006924	.0266635	4.0268973	1.4688004
40	.0249961	.0006249	.0253138	4.0788457	1.4701502
45	.0222195	.0004938	.0224700	4.1980127	1.4729939
50	.0199980	.0004000	.0202006	4.3044810	1.4752633
60	.0166655	.0002778	.0168059	4.4884659	1.4736580
70	.0142850	.0002041	.0143880	4.6438045	1.4810759
80	.0124995	.0001562	.0125783	4.7732275	1.4823856
100	.0099998	.0001000	.0100501	5.0026199	1.4854138
120	.0083332	.0000694	.0083681	5.1857742	1.4870958
140	.0071428	.0000510	.0071684	5.3405199	1.4882955
150	.0066666	.0000444	.0066889	5.4097508	1.4887750
160	.0062499	.0000391	.0062696	5.4744975	1.4891943
180	.0055555	.0000309	.0055710	5.5926277	1.4898929
200	.0050000	.0000250	.0050125	5.6982659	1.4904514
250	.0040000	.0000160	.0040080	5.9219094	1.4914559
300	.0033333	.0000111	.0033389	6.1045643	1.4921250
350	.0028571	.0000082	.0028612	6.2589530	1.4926027
400	.0025000	.0000062	.0025031	6.3926630	1.4929608
500	.0020000	.0000040	.0020020	6.6160565	1.4934619
600	.0016667	.0000028	.0016681	6.7985447	1.4937958
800	.0012500	.0000016	.0012508	7.0364351	1.4942131
1000	.0010000	.0000010	.0010005	7.3097037	1.4944634

TABLE III.

Combination of Geometric and Bradford Distributions. See 7, 8, 8a.

Values of γ_{10}, J(10) and $R_j(10)$ vs. values of μ and J(∞).

μ	0.05	0.05	0.05	0.1	0.1	0.1	0.15	0.15	0.15
J(∞)	2	4	6	2	4	6	2	4	6
γ_{10}	0.2623	0.3148	0.3373	0.4214	0.5057	0.5418	0.5179	0.6215	0.6659
J_{10}	.3556	.4594	.5089	.7284	1.0230	1.1825	1.0743	1.6420	1.9931
R_0	0.2829	0.3395	0.3638	0.4338	0.5206	0.5578	0,5068	0.6081	0.6516
R_1	-.1949	-.2128	-.2183	-.2116	-.2006	-.1904	-.1797	-.1372	-.1109
R_2	-.0640	-.0852	-.0944	-.1238	-.1499	-.1581	-.1472	-.1601	-.1583
R_3	-.0176	-.0283	-.0336	-.0561	-.0824	-.0936	-.0845	-.1130	-.1216
R_4	-.0047	-.0091	-.0115	-.0243	-.0429	-.0523	-.0454	-.0734	-.0850
R_5	-.0012	-.0029	-.0039	-.0104	-.0220	-.0288	-.0239	-.0465	-.0579
R_6	-.0003	-.0009	-.0013	-.0044	-.0112	-.0157	-.0125	-.0293	-.0390
R_7	-.0001	-.0003	-.0005	-.0019	-.0057	-.0086	-.0065	-.0183	-.0262
R_8	-	-.0001	-.0002	-.0008	-.0029	-.0047	-.0034	-.0114	-.0175
R_9	-	-	-.0001	-.0003	-.0015	-.0025	-.0018	-.0071	-.0117
R_{10}	-	-	-	-.0001	-.0007	-.0014	-.0009	-.0044	-.0078

For $j > 10$, $R_j(10) \triangleq R_{10}(10)(\gamma_{10})^{j-10}$

Values of J(N) and Q(N) vs. N, μ and J(∞).

μ	J(∞)		N = 2	3	4	5	6	7	8	9	10
.1	4	J	0.170	0.202	0.358	0.459	0.565	0.674	0.787	0.904	
		Q	.071	.112	.140	.162	.180	.196	.210	.222	0.233
.1	6	J	.184	.286	.394	.509	.631	.759	.894	1.036	
		Q	.077	.121	.152	.176	.196	.214	.229	.243	.256
.1	8	J	.192	.299	.415	.538	.670	.810	.959	1.116	
		Q	.080	.126	.158	.184	.205	.224	.240	.255	.269
.2	4	J	.358	.565	.787	1.023	1.268	1.517	1.766	2.010	
		Q	.147	.232	.293	.342	.383	.418	.448	.476	.501
.2	6	J	.394	.631	.894	1.183	1.494	1.823	2.165	2.514	
		Q	.160	.253	.322	.377	.423	.465	.502	.535	.566
.2	8	J	.415	.670	.959	1.283	1.640	2.027	2.441	2.875	
		Q	.167	.265	.338	.397	.448	.493	.534	.572	.607
.3	4	J	.565	.904	1.268	1.642	2.010	2.356	2.669	2.941	
		Q	.227	.361	.459	.537	.602	.658	.706	.747	.783
.3	6	J	.631	1.035	1.494	1.993	2.514	3.035	3.533	3.988	
		Q	.248	.398	.510	.602	.681	.750	.812	.867	.916
.3	8	J	.670	1.116	1.640	2.232	2.875	3.546	4.215	4.853	
		Q	.260	.419	.540	.641	.729	.809	.881	.946	1.006