A General Formulation of Bradford's Distribution: The Graph-Oriented Approach

Isao Asai
Department of Industrial Engineering, University of Osaka Prefecture, Mozu, Sakai, Osaka, 591, Japan

From the detailed analysis of eight previously published mathematical models, a general formulation of Bradford's distribution can be deduced as follows: $y = a \log(x + c) + b$, where y is the ratio of the cumulative frequency of articles to the total number of articles and x is the ratio of the rank of journal to the total number of journals. The parameters a, b, and c are the slope, the intercept, and the shift in a straight line to log rank, respectively. Each of the eight models is a special case of the general formulation and is one of five types of formulation. In order to estimate three unknown parameters, a statistical method using root-weighted square error is proposed. A comparative experiment using 11 databases suggests that the fifth type of formulation with three unknown parameters is the best fit to the observed data. A further experiment shows that the deletion of the droop data leads to a more accurate value of parameters and less error.

Introduction

The mathematical model describing Bradford's distribution [1] takes either of two approaches: graph-oriented or inference-oriented. The former, based on the graph from the observed data, throws light on the static structure of scatter [e.g., 2-9]. The inference-oriented approach based on the inference from a general principle, mostly derived from Lotka's distribution, throws light on the dynamic process of scatter [e.g., 10-13].

The curve of Bradford's distribution consists of the following: the nucleus section containing the most productive journals, the linear section, and the droop section containing the least productive journals [14-16]. It is difficult to formulate a mathematical model which expresses all these sections, but for the purpose of this article, the nucleus and linear sections of the frequency distribution, as being the most significant data for practical application, will be focused on.

The purpose of this article is to present a general formulation of Bradford's distribution with the graph-oriented approach, to propose a statistical method for estimating three parameters using root-weighted square error, to demonstrate a comparative experiment for examining the characteristics of five types of formulation using 11 databases, and to discuss the effect of deleting the "droop" data for better estimation.

A General Formulation

From the detailed analysis of eight previously published mathematical models [2-9], a general formulation of Bradford's distribution can be deduced as follows:

$$y = a \log(x + c) + b, \tag{1}$$

or

$$R(n)/R = a \log(n/N + c) + b,$$

where $R(n)$ is the cumulative frequency of articles, R is the total number of articles, n is the rank of journal, and N is the total number of journals. The parameters a, b, and c are the slope, the intercept, and the shift in a straight line to log rank, respectively. y is a continuous variable represented by the ratio of $R(n)$ to R. x is a continuous variable represented by the ratio of n to N.

The general formulation is classified into five types of formulation defined by the known-unknown combinations of the three parameters. The conditions of the parameters in each type of formulation are shown in Table 1. Each of the eight mathematical models, however, is one of the five types of formulation. Table 2 shows the corresponding models of each type of formulation, including the relationships between parameters in the model and in the general formulation (see Appendix A), and the source reference. The original notation of parameters in Table 2 is altered in order to avoid confusion. The characteristics of each type of formulation are discussed below from the graphical point of view.

The first type of formulation is a single equation expressed by Cole [2]. The parameter d, called the reference-scattering coefficient, is a measure of literature usage. The straight line always passes the last point (N,R), and the shift parameter is zero. This means that the deviation between the observed and estimated data becomes greater in proportion to the increase of the droop section, and the formulation of the nucleus section is negligible.

Received December 26, 1979; revised June 6, 1980; accepted July 10, 1980.

0002-8231/81/0032-0113$01.00

TABLE 1. The conditions of the three parameters in each type of formulation.

type	parameter		
	a	b	c
1	unknown	1	0
2	unknown	unknown	0
3	$\frac{1}{\log(1 + 1/c)}$	$\frac{\log(1/c)}{\log(1 + 1/c)}$	unknown
4	unknown	$a \log(1/c)$	unknown
5	unknown	unknown	unknown

The second type of formulation is given by Brookes [3]. This model, called the graphical formulation, represents the linear section of Bradford's distribution. Drott and Griffith [17] point out that the slope k can be used to estimate the completeness of a particular search and that the intercept s on the x(log rank) axis is an indicator of subject breadth. By introducing an unknown parameter b, the straight line does not always pass the last point (N,R), and thus it is possible to obtain a better estimation. However, the formulation of the nucleus section is still left unsolved.

The third type of formulation is developed by Leimkuhler [5] and Brookes [4]. In Leimkuhler's model, the parameter z is related to the subject field and the completeness of the collection. The unsolved nucleus section can be analyzed by this model because of the shift parameter. When the value of the shift parameter is estimated, the value of the slope and intercept parameters can be obtained by simple calculation (see Table 1). However, it is a rigid constraint that the straight line always has to pass both the supposed point (0,0) according to Bradford's law and the last point (N,R).

Three models for the fourth type of formulation are proposed as follows:

(i) Fairthorne [7] presents an equation with continuous variables.

(ii) Wilkinson [8] deduces the verbal formulation in proof of the ambiguity between the graphical explanation and the verbal description described by Bradford.

(iii) Leimkuhler [6] proposes the model of a collection of journals and presents the various algebraic methods for estimating a scale parameter f and a dispersion parameter g.

By introducing an unknown slope parameter, the fourth type of formulation is more flexible than the third one. However, the straight line always passes the supposed point (0,0) according to Bradford's law. This puts a limitation on obtaining a straight line fitted to the observed data.

The fifth type of formulation is the yield formula deduced by Haspers [9]. The parameter u in his model is a measure of the subject breadth, and the parameter v is an indicator of the compactness of the yield of the most productive sources. By introducing an intercept parameter v, the limitation in the fourth type of formulation disappears. The fifth type of formulation is the most flexible of all because of the three unknown parameters. Equation (1) is algebraically equivalent to Haspers' model. The only difference is that his model is formulated using discrete variables, while the model proposed here is formulated using continuous variables.

It is clear that each of eight previously published mathematical models is a special case of a general formulation and is one of the five types of formulation. The statistical method for estimating the slope, intercept, and shift parameters of the five types of formulation is described below.

A Statistical Method for Estimating Parameters

There are three methods for estimating parameters: (i) the graphic method, which plots the relationship between the cumulative frequency of articles and the rank of journal on a semilogarithmic scale [2]; (ii) the algebraic method, which solves simultaneous algebraic equations, substituting a few typical data for the variables, under the assumption that the curve of Bradford's distribution is a straight line [6]; and (iii) the statistical method, which minimizes error sum of squares between observed and esti-

TABLE 2. The corresponding models of each type of formulation, including the relationships between parameters in the model and general formulation, and the source reference.

type	model*	relationships between parameters	source reference
1	$y = d \log x + 1$	$d = a$	Cole-62[2]
2	$R(n) = k \log(n/s)$	$k = aR$, $s = N\ 10^{-b/a}$	Brookes-69[3]
3	$y = \log(1 + zx)/\log(1 + z)$	$z = 1/c$	Leimkuhler-67[5]
	$y = \log_r [(m + n)/m]$	$m = cN$, $r = 1 + 1/c$	Brookes-78[4]
4	$y = p \log(1 + qx)$	$p = a$, $q = 1/c$	Fairhtorne-69[7]
	$R(n) = j \log(n/t + 1)$	$j = aR$, $t = cN$	Wilkinson-72[8]
	$R(n) = f \log(1+gn)/\log(1+g)$	$f = aR \log(1 + 1/cN)$, $g = 1/cN$	Leimkuhler-77[6]
5	$R(n) = h \log(n/u + 1) + v$	$h = aR$, $u = cN$, $v = bR + aR \log c$	Haspers-76[9]

*The original notation is altered in order to avoid confusion.

mated data [17]. In this article, statistical methods for estimating the slope, intercept, and shift parameters of five types of formulation are used.

Because of log rank, the use of error sum of squares has two results: the droop section is emphasized and the deviation in the nucleus section increases. This defeats our purpose of focusing on the nucleus and linear sections. This difficulty may be solved by introducing a measure of weight. Therefore, the statistical criterion can be expressed by

$$Z = \left[\sum_{n=1}^{N} w_n (y_n - \bar{y}_n)^2 \right]^{1/2}, \quad (2)$$

where Z is root-weighted square error; w_n is the normalized weight to the square error in rank n of journal; y_n is the normalized cumulative frequency of articles in rank n of journal; and $\bar{y}_n$ is the estimate of y_n.

The density of data is higher in low rank than in high rank because of log rank. In order to avoid the gravitation of data to low rank, it is necessary to assign high-ranking data a larger weight. Therefore, the normalized weight can be given by

$$w_n = y_n - y_{n-1}, \quad (3)$$

where $y_0 = 0$.

As Eq. (1) is nonlinear to variable x, the method of nonlinear regression analysis can be applied [24]. If good starting estimators had been obtained, the Gauss-Newton method could have been used for estimating parameters. Any method of nonlinear regression analysis, however, is in need of iteration procedure in which the value of the parameters is reset according to a rational rule which improves the criterion.

The method for estimating three parameters of various types of formulation consists of the estimating procedure and the searching procedure. The estimating procedure, which minimizes the statistical criterion Z for a given value of shift parameter, is as follows: (i) give a value for the shift parameter, (ii) compute the weight using Eq. (3), (iii) decide the value of slope and intercept parameters according to Table 1, in which the unknown parameters are obtained using the method of regression analysis on Eq. (2), and (iv) obtain the value of Z using the estimated data.

The searching procedure in which the optimal value of shift parameter is obtained by introducing an increment parameter δ is as follows:

(a) set $\delta = 0.1$ and $c_1 = 0$,

(b) set $\delta = 0.1\delta$, and if $\delta < 0.00001$, then go to (f),

(c) set $c_2 = c_1 + \delta$, and obtain Z_2 for c_2 using the estimating procedure,

(d) set $c_3 = c_2 + \delta$, and obtain Z_3 for c_3 using the estimating procedure,

(e) if $Z_2 < Z_3$, then go to (b), otherwise, set $c_1 = c_2$, $c_2 = c_3$, and $Z_2 = Z_3$, and go to (d),

(f) c_2 is optimal solution, and end.

By altering 0.00001 in step (b) to a smaller value, a more accurate value of shift parameter can be found.

The estimation of parameters in the first and second types of formulation starts from step (i) in the estimating procedure and ends on step (iv). However, the estimation of parameters in the third, fourth and fifth types of formulation starts from step (a) in the searching procedure, and ends on step (f), though in steps (c) and (d) the routine jumps to the estimating procedure and returns. It is possible to program these procedures.

A Comparative Experiment

In order to investigate the characteristics of each type of formulation, a comparative experiment will be performed using 11 databases collected from published articles. These sources are shown in Table 3. This sample of databases is selected arbitrarily, but it provides us with an example of

TABLE 3. The sources of 11 databases used in a comparative experiment.

data-base	source reference	subject topic	N*1	R*2
1	Bradford-34[1]	Applied Geophysics	326	1332
2	Bradford-34[1]	Lubrication	164	395
3	Kendall-60[18]	Operations Research	370	1763
4	Cole-62[2]	Petroleum Industry	197	903
5	Goffman-69[19]	Mast Cell	587	2378
6	Goffman-69[19]	Schistosomialsis	1738	9914
7	Goffman-70[20]	Allen Memorial Medical Lib.	371	876
8	Goffman-70[20]	Transplantation Immunology	272	1120
9	Lawani-73[21]	Tropical Agriculture	374	2294
10	Saracevic-73[22]	Library Literature	242	3420
11	Pope-75[23]	Information and Technology	1011	7368

*1 N is the total number of journals.
*2 R is the total number of articles.

the many differences among the various types of formulation.

The subject topic is quite varied. The material used in it includes references of articles on a given topic (databases 1, 2, and 4), references of articles in a given journal (database 3), references of articles in a given review book (database 11), articles in bibliography (databases 5, 6, and 9), indexed articles (databases 8 and 10), and journal usage records in a library (database 7). All of these materials have been used to examine Bradford's distribution.

Each database is made of a series of the productivity of articles and the frequency of journals arranged in descending order of productivity (see Table B1 in Appendix B). Because of the summarized data, we need the summation for generating the frequency distribution. That is, the cumulation of the productivity of articles equals the cumulative frequency of articles, and the cumulation of the frequency of journals equals the rank of journal. Dividing by the total number of articles or journals, these data are normalized in the range from zero to one. The complete frequency distribution will be used in which the total number of the pairs of data equals the total number of journals [16], and the logarithms have a base 10.

The comparative experiment is performed using a computer with the performance of 1.6 μsec in Gibson mix. The programming language is FORTRAN, and the memory needs 45K words. The data set of productivity and frequency from each database is punched and read, and the cumulative frequency distribution is generated. Then, the various values such as error and parameters in each type of formulation are computed using the statistical method described above. The execution time in this section is about 145 sec.

Table 4 shows the minimal value of root-weighted square error according to the type of formulation and the database. This gives us a clear answer to finding the best fit of five types of formulation adapted to the observed data. In any database, the error of the fifth type of formulation is the least of all, and thus it is concluded, from a statistical viewpoint, that the fifth type of formulation with three unknown parameters is the best fit to the observed data. This is in accordance with intuition.

TABLE 4. The minimal value of root-weighted square error according to the type of formulation and the database.

data-base	type-1	type-2	type-3	type-4	type-5
1	0.0438	0.0365	0.0113	0.0111	0.0106
2	0.0668	0.0470	0.0114	0.0069	0.0066
3	0.0189	0.0164	0.0084	0.0084	0.0083
4	0.0378	0.0354	0.0160	0.0136	0.0115
5	0.0480	0.0459	0.0247	0.0197	0.0173
6	0.0337	0.0335	0.0207	0.0137	0.0121
7	0.1174	0.0877	0.0073	0.0073	0.0072
8	0.0228	0.0192	0.0070	0.0066	0.0054
9	0.0818	0.0735	0.0265	0.0134	0.0131
10	0.0483	0.0482	0.0341	0.0273	0.0136
11	0.0415	0.0333	0.0480	0.0320	0.0269
mean	0.0510	0.0433	0.0196	0.0145	0.0121

TABLE 5. The mean, maximum, and minimum values of parameters in each type of formulation.

parameter		type-1	type-2	type-3	type-4	type-5
a	mean	0.418	0.392	0.491	0.516	0.521
	max	0.563	0.458	0.851	0.858	0.863
	min	0.340	0.337	0.359	0.370	0.368
b	mean	1.000	0.965	0.996	1.023	1.024
	max	———	1.054	1.000	1.091	1.090
	min	———	0.870	0.974	0.973	0.972
c	mean	0.000	0.000	0.014	0.015	0.016
	max	———	———	0.072	0.073	0.074
	min	———	———	0.002	0.002	0.001

Looking at the mean errors in Table 4, the mean error of the third type of formulation drops sharply to less than half of the mean error of the second type of formulation. The decrease can be attributed to the shift parameter, which alters the rising curve in the nucleus section to a straight line.

The characteristics of estimated parameters in each type of formulation are given in Table 5. This shows an outline of the static structure of scatter. The range of value of parameters is presented clearly, though it is obtained from only 11 databases. If the distribution function of parameters on a given field is confirmed by using many databases, many of the frequency distributions similar to the actual Bradford's distribution could be generated by the technique of computer simulation. This may be useful for studying the dynamic structure of scatter.

The correlation matrix in the fifth type of formulation is shown in Table 6. The correlation of 0.963 between parameters a and c is the greatest of all. This is caused by the fact that the value of the slope parameter on a semilogarithmic sheet increases with the increased value of the shift parameter.

In any correlation to error Z in Table 6, the relationship between the error Z and the intercept parameter b is the strongest of all. The intercept parameter represents the value of y to $x = 1 - c$. As the value of c is near zero, the value of x is near one. Therefore, the error is affected by the last part of the data from the graph and the correlation. A discussion of this phenomenon follows in the next section.

Deletion of the Droop Data

The existence of the droop section in Bradford's distribution was first recognized by Groos in 1967 [14]. He indicates that the droop data lies in greater than 10% of the total number of journals in the Keenan-Atherton data. Drott and Griffith [17] estimate the value of parameters in the second type of formulation using the data

TABLE 6. Correlation matrix of the total number of journals, N, the total number of articles, R, the minimal value of root-weighted square error, Z, and parameters a, b, and c in the fifth type of formulation.

	articles R	error Z	parameter a	parameter b	parameter c
journals N	0.945	0.438	-0.379	0.290	-0.293
articles R		0.585	-0.436	0.547	-0.377
error Z			-0.263	0.758	-0.325
parameter a				-0.265	0.963
parameter b					-0.365

deleted on all journals contributing only one article. In this article, instead of formulating the complicated portion, the deletion of the droop data will be attempted for better estimation.

In order to identify the droop data, a further experiment is performed using the same 11 databases as in the previous section. The last part of the data is deleted successively, 2% at a time, 25 times, up to 50% of the total number of articles. In every estimation, the parameters and the error in each type of formulation are obtained. The results of the mean error are given in Figure 1.

As shown in Figure 1, the mean error in all types of formulation except the first decreases with the increased deletion rate of articles, and the relationship between the mean errors in any two types of formulation remains unchanged. In the first type of formulation, as presented in the previous section, the nucleus section is left unsolved, and the straight line always passes the last point (N,R). With the increased deletion rate of articles, the relative weight in the unfitted nucleus section increases, the relative weight in the fitted droop section decreases, and thus the mean error increases.

As an objective criterion for deleting the droop data, the degree of fit to the observed data is adequate. In this article, the following criterion is adopted: the root-weighted square error is less than 0.01. The results of the fifth type of formulation are given in Table 7, including the rate of journals and articles, error, and the value of parameters a, b, and c. The databases with less than 0.01 of error from the start, i.e., databases 2, 3, 7, and 8, are not deleted at all. The mean rates of journals and articles are 0.740 and 0.938, respectively.

It can be concluded that the deletion of the droop data does indeed lead to better estimation of parameters and less error.

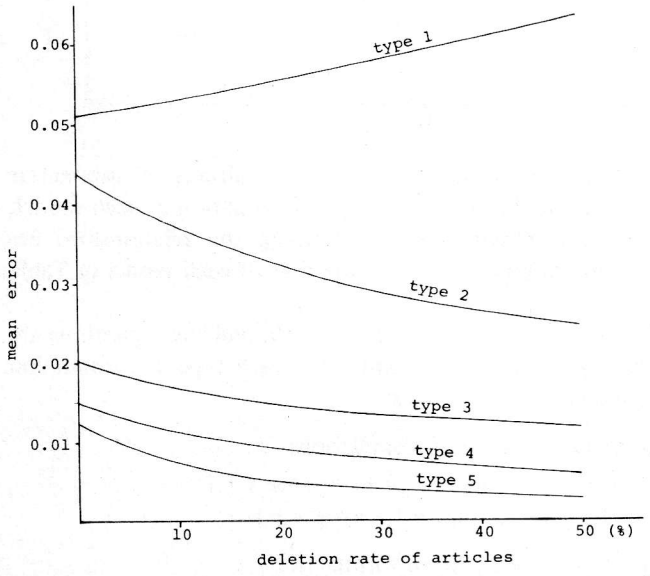

FIGURE 1. A plotting of the mean error against the deletion rate of articles in each type of formulation.

Conclusion

Bradford's distribution can be definitively explained by a general formulation deduced from the graphical analysis of eight previously published mathematical models. It is clear that each of the models is a special case of the general formulation and is one of the five types of formulation.

The comparative experiment using 11 databases suggests that the minimal value of root-weighted square error decreases in ascending order of the type of formulation. This means that the fifth type of formulation with

TABLE 7. The results of deletion of the droop data in the fifth type of formulation.

data-base	rate of journals	rate of articles	error Z^*	para-meter a	para-meter b	para-meter c
1	0.252	0.739	0.0098	0.465	0.993	0.008
2	1.000	1.000	0.0066	0.604	0.972	0.024
3	1.000	1.000	0.0083	0.368	1.001	0.002
4	0.812	0.959	0.0095	0.502	1.037	0.011
5	0.513	0.880	0.0099	0.549	1.073	0.010
6	0.772	0.960	0.0098	0.390	1.042	0.002
7	1.000	1.000	0.0072	0.863	0.977	0.074
8	1.000	1.000	0.0054	0.414	1.001	0.003
9	0.877	0.980	0.0099	0.725	1.066	0.035
10	0.715	0.980	0.0083	0.553	1.106	0.016
11	0.196	0.820	0.0099	0.478	1.181	0.003
mean	0.740	0.938	0.0086	0.537	1.041	0.017

*Z is the minimal value of root-weighted square error.

TABLE B1. A list of the productivity of article and the frequency of the journal arranged in descending order of productivity to each data base.

data base	productivity-frequency
1	93-1,86-1,56-1,48-1,46-1,35-1,28-1,20-1,17-1,16-4,15-1,14-5,12-1,11-2,10-5,9-3, 8-8,7-7,6-11,5-12,4-17,3-23,2-49,1-169
2	22-1,18-1,15-1,13-2,10-2,9-1,8-3,7-3,6-1,5-7,4-2,3-13,2-25,1-102
3	242-1,114-1,102-1,95-1,58-1,49-1,34-1,22-2,21-2,20-2,18-1,16-4,15-2,14-1,12-2, 11-5,10-3,9-4,8-8,7-8,6-6,5-10,4-17,3-29,2-54,1-203
4	122-1,51-1,43-1,29-2,26-1,24-1,20-2,17-2,16-1,15-1,14-2,11-1,10-1,9-8,8-5,7-2, 5-9,4-14,3-14,2-28,1-100
5	66-1,58-1,57-1,55-1,53-1,46-1,40-1,38-2,37-1,35-1,34-1,32-1,31-1,30-1,28-1,27-1, 23-2,22-1,21-1,20-2,19-2,18-2,17-1,16-1,15-3,14-6,13-3,12-5,11-8,10-6,9-11,8-6, 7-8,6-8,5-16,4-24,3-35,2-90,1-328
6	325-1,266-1,259-1,215-1,211-1,171-1,159-1,143-1,137-1,136-1,118-1,115-1,112-1,108-1, 105-2,94-1,90-1,80-1,74-1,72-2,70-2,68-1,66-1,64-1,56-1,55-2,51-2,50-1,47-1,45-1, 44-1,42-2,41-1,40-1,39-2,37-3,36-1,35-2,34-1,33-1,32-3,31-3,29-2,28-5,27-1,26-1, 25-2,24-3,23-4,22-2,21-4,20-3,19-4,18-10,17-8,16-10,15-9,14-10,13-10,12-6,11-11, 10-14,9-19,8-29,7-27,6-44,5-57,4-76,3-137,2-266,1-908
7	18-1,12-1,11-2,10-1,9-3,8-7,7-4,6-12,5-15,4-28,3-36,2-75,1-186
8	124-1,71-1,66-1,54-1,40-1,28-1,25-1,24-1,22-1,19-1,17-1,15-3,13-1,12-2,10-1,9-6, 8-4,7-6,6-1,5-9,4-19,3-20,2-34,1-155
9	80-1,70-1,51-1,41-1,33-1,32-1,31-2,30-2,29-1,28-2,27-2,26-1,25-1,24-1,22-1,21-1, 20-2,19-3,18-3,17-1,16-7,15-3,14-5,13-3,12-3,11-10,10-8,9-11,8-13,7-11,6-18,5-25, 4-26,3-40,2-49,1-113
10	575-1,149-1,131-1,94-1,88-1,86-1,75-1,74-1,71-1,63-1,60-2,57-2,45-1,43-1,42-1,41-1, 37-1,36-1,33-1,31-3,30-1,29-1,28-1,27-3,25-2,22-3,21-1,20-5,19-4,18-4,17-2,16-4, 15-2,14-1,13-3,12-3,11-4,10-7,9-9,8-9,7-7,6-8,5-12,4-11,3-18,2-25,1-68
11	261-1,259-1,220-1,211-1,205-1,176-1,168-1,164-1,155-1,134-1,120-2,115-1,105-1,102-1, 96-1,85-1,80-1,79-2,78-1,74-1,64-1,63-1,60-2,59-1,53-1,52-1,51-2,45-1,44-1,42-2, 40-1,38-2,36-1,33-2,32-1,31-5,30-1,29-1,28-1,27-1,25-1,24-3,23-1,22-6,21-2,20-5, 19-4,18-8,17-5,16-3,15-4,14-7,13-10,12-9,11-9,10-7,9-8,8-12,7-20,6-14,5-35,4-45, 3-68,2-140,1-534

three unknown parameters is the best fit to the observed data. A further experiment shows that the deletion of the droop data leads to a more accurate value of parameters and less error. These applications prove that the statistical method proposed here for estimating parameters is feasible. While the above findings apply only to the 11 databases discussed in this article, they can be said to hold in general.

Appendix A

Here we show proof of the relationships between parameters in each model and in the general formulation in Table 2. It is completed by showing that the following are exactly alike: (a) the equation in each type of formulation, which is deduced by substituting the conditions of parameters (see Table 1) in Eq. (1); (b) the equation in each model, which is obtained by substituting the relationships between parameters in the equation of each model in Table 2.

Note that $y = R(n)/R$, $x = n/N$, and the logarithms are base 10. The Eqs. (a) and (b) in each type of formulation are presented as follows:

(1) The first type of formulation:

(a) $y = a\log(x + c) + b = a\log x + 1$.

(b) $y = d\log x + 1 = a\log x + 1$.

(2) The second type of formulation:

(a) $y = a\log(x + c) + b = a\log x + b$.

(b) $R(n) = k \log(n/s) = aR \log [n/(N\,10^{-b/a})]$
$= aR\ [\log(n/N) + b/a], \therefore R(n)/R = a \log(n/N) + b, \ \therefore y = a \log x + b.$

(3) The third type of formulation:

(a) $y = a \log(x + c) + b$
$= \dfrac{\log(x + c)}{\log(1 + 1/c)} + \dfrac{\log(1/c)}{\log(1 + 1/c)} = \dfrac{\log(x/c + 1)}{\log(1 + 1/c)}$.

(b) * Leimkuhler's model:
$y = \dfrac{\log(1 + zx)}{\log(1 + z)} = \dfrac{\log(1 + x/c)}{\log(1 + 1/c)}$.

* Brookes' model:
$y = \log_r [(m + n)/m] = \log_{(1+1/c)} [(cN + n)/cN]$
$= \dfrac{\log [1 + (n/N)/c]}{\log(1 + 1/c)} = \dfrac{\log(1 + x/c)}{\log(1 + 1/c)}$.

(4) The fourth type of formulation:

(a) $y = a \log(x + c) + b = a \log(x + c) + a \log(1/c)$
$= a \log(x/c + 1).$

(b) * Fairthorne's model:
$y = p \log(1 + qx) = a \log(1 + x/c).$

* Wilkinson's model:
$R(n) = j \log(n/t + 1) = aR \log(n/cN + 1),$
$\therefore R(n)/R = a \log [(n/N)/c + 1],$
$\therefore y = a \log(x/c + 1).$

* Leimkuhler's model:
$R(n) = \dfrac{f \log(1 + gn)}{\log(1 + g)} = \dfrac{aR \log(1 + 1/cN) \log(1 + n/cN)}{\log\ 1 + 1/cN}$,
$\therefore R(n)/R = a \log [1 + (n/N)/c],$
$\therefore y = a \log(1 + x/c).$

(5) The fifth type of formulation:

(a) $y = a \log(x + c) + b.$

(b) $R(n) = h \log(n/u + 1) + v = aR \log(n/cN + 1) + bR + aR \log c$
$= aR \log(n/N + c) + bR,$
$\therefore R(n)/R = a \log(n/N + c) + b,$
$\therefore y = a \log(x + c) + b.$

Appendix B

The raw data used in the comparative experiment are shown in Table B1. The pair which includes the productivity of articles and the frequency of journals is expressed using the notation "-" for the sake of simplification. Further information can be had by going back to the source reference.

References

1. Bradford, S. C. "Sources of Information on Specific Subjects." *Engineering.* 137:85-86; 1934.
2. Cole, P. F. "A New Look at Reference Scattering." *Journal of Documentation.* 18(2):58-64; 1962.
3. Brookes, B. C. "Bradford's Law and the Bibliography of Science." *Nature.* 224:953-956; 1969.
4. Brookes, B. C.; Griffith, J. M. "Frequency-Rank Distributions." *Journal of the American Society for Information Science.* 29(1):5-13; 1978.
5. Leimkuhler, F. F. "The Bradford Distribution." *Journal of Documentation.* 23(3):197-207; 1967.
6. Leimkuhler, F. F. "Operational Analysis of Library Systems." *Information Processing and Management.* 13(2):79-93; 1977.
7. Fairthorne, R. A. "Empirical Hyperbolic Distributions (Bradford-Zipf-Mandelbrot) for Bibliometric Description and Prediction." *Journal of Documentation.* 25(4):319-343; 1969.
8. Wilkinson, E. A. "The Ambiguity of Bradford's Law." *Journal of Documentation.* 28(2):122-130; 1972.
9. Haspers, J. H. "The Yield Formula and Bradford's Law." *Journal of the American Society for Information Science.* 27(5/6):281-287; 1976.
10. Simon, H. A. "On a Class of Skew Distribution Functions." *Biometrika.* 42(3/4):425-440; 1955.
11. Naranan, S. "Power Law Relations in Science Bibliography. A Self-Consistent Interpretation." *Journal of Documentation.* 27(2):83-97; 1971.
12. de Solla Price, D. J. "A General Theory of Bibliometric and Other Cumulative Advantage Processes." *Journal of the American Society for Information Science.* 27(5/6):292-306; 1976.
13. Bookstein, A. "Patterns of Scientific Productivity and Social Change: A Discussion of Lotka's Law and Bibliometric Symmetry." *Journal of the American Society for Information Science.* 28(4):206-210; 1977.
14. Groos, O. V. "Bradford's Law and the Keenan-Atherton Data." *American Documentation.* 18(1):46; 1967.
15. Brookes, B. C. "The Complete Bradford-Zipf 'Bibliograph'." *Journal of Documentation.* 25(1):58-60; 1969.
16. Praunlich, P.; Kroll, M. "Bradford's Distribution: A New Formulation." *Journal of the American Society for Information Science.* 29(2):51-55; 1978.
17. Drott, M. C.; Griffith, B. C. "An Empirical Examination of Bradford's Law and the Scattering of Scientific Literature." *Journal of the American Society for Information Science.* 29(5):238-246; 1978.
18. Kendall, M. G. "The Bibliography of Operational Research." *Operational Research Quarterly.* 11(1):31-36; 1960.
19. Goffman, W.; Warren, K. S. "Dispersion of Papers among Journals Based on a Mathematical Analysis of Two Diverse Medical Literatures." *Nature.* 221:1205-1207; 1969.
20. Goffman, W.; Morris, T. G. "Bradford's Law and Library Acquisitions." *Nature.* 226:922-923; 1970.
21. Lawani, S. M. "Bradford's Law and the Literature of Agriculture." *International Library Review.* 5(3):341-350; 1973.
22. Saracevic, T.; Perk, L. J. "Ascertaining Activities in a Subject Area Through Bibliometric Analysis." *Journal of the American Society for Information Science.* 24(2):120-134; 1973.
23. Pope, A. "Bradford's Law and the Periodical Literature of Information Science." *Journal of the American Society for Information Science.* 26(4):207-213; 1975.
24. Draper, N. R.; Smith, H. *Applied Regression Analysis.* New York: Wiley; 1966.

A General Formulation of Bradford's Distribution: The Graph-Oriented Approach

Isao Asai

Reprinted from
JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE
Vol. 32, No. 2, March 1981, pp. 113–119