Information Processing & Management Vol 23, No 3, pp 155–170, 1987
Printed in Great Britain

0306–4573/87 $3.00 + .00

THE MAXIMAL VALUE OF A ZIPF SIZE VARIABLE: SAMPLING PROPERTIES AND RELATIONSHIP TO OTHER PARAMETERS

JEAN TAGUE and PAUL NICHOLLS
School of Library and Information Science, University of Western Ontario, London, Ontario N6G 1H1, Canada

(*Received* 29 *April* 1986; *in revised form* 2 *October* 1986)

Abstract—Because the Zipf size-frequency distribution is used so often as a mathematical model for bibliometric variables, it is important that the relationships among its parameters and its sampling properties be understood by investigators in this field. This paper examines these relationships and properties. In addition, it provides tables for the sampling distribution of the maximal value of a finite Zipf distribution and an approximation formula for confidence intervals. Confidence limits for the maximal value in a number of previous studies are determined.

The Zipf distribution plays a central role in the modeling of human activities, particularly of the variables studied in bibliometrics and scientometrics: productivity of researchers in a discipline, impact of authors or publications, use of words in a text or keys in a data base, and dispersion of a subject literature among sources. In general, it may be described as representing the distribution of a set of tokens over a set of types, for example, publications over authors, citations over authors or publications, word occurrences in a text over word forms, data base accesses over keys, and so forth. It has been represented in a number of functional forms, which may be distinguished by the number of parameters and by the nature of the property or variable described, whether a size (frequency) or a rank.

In its most general form, the Zipf function describes the distribution of a set of m tokens over a set of t types using one of the following expressions:

$$g(x) = \frac{a}{(x+c)^b}, \qquad x = 1,2,\ldots,x_{\max},\ a,b > 0,\ c \geq 0$$

$$f(r) = \frac{a'}{(r+c')^{b'}}, \qquad r = 1,2,\ldots,t,\ a',b' > 0,\ c' \geq 0$$

where $g(x)$ is the number of types with exactly x tokens and $f(r)$ is the number of tokens for the rth ranking type when types are arranged in descending order of number of tokens. The function $g(x)$ is commonly called a size-frequency distribution, as opposed to $f(r)$, which is a rank-frequency distribution.

This study is concerned with the parameters of the size-frequency Zipf distribution. The parameter $x_{\max}$ represents the maximum number of tokens for a type, or the maximal size or value of the productivity variable x. Note that $x_{\max} = f(1)$, that is, the frequency of the highest ranked type. In most applications, c is assumed to be 0, that is

$$g(x) = \frac{a}{x^b}, \qquad x = 1,2,\ldots,x_{\max},\ a,b > 0.$$

In this case, the parameter a will represent the number of types with exactly one token. The larger the exponent b, the larger will be this number relative to the total number of types.

The Zipf size-frequency distribution can be expressed as a relative frequency or probability distribution by dividing by a suitable constant. If X represents the number of tokens assigned to a random type, $p(x)$ the probability X assumes a specific value x, and t the total number of types, then

$$p(x) = \frac{g(x)}{t} = \frac{a}{tx^b}, \qquad x = 1,2,\ldots,x_{\max}.$$

The size variable X can be generalized to a continuous productivity variable, that is, productivity of a type rather than number of tokens of a type. The discrete Zipf distribution is then replaced by its continuous analog, the Pareto distribution. In some applications, either when the number of tokens assumes a very large number of values or when fractional assignment of tokens to types can be made, the use of this continuous analog is appropriate and simplifies the derivation of some sampling properties, as shall be seen later. Such a replacement might arise, for example, in an author productivity distribution in which pages or words of text were counted rather than papers, or in which multiple authors were assigned fractions of papers.

When a continuous productivity variable is used, the density function has the form

$$g(x) = \frac{a}{x^b}, \qquad 0 \le x \le x_{\max}$$

and the cumulative distribution has the form

$$G(x) = \int_1^x \frac{a}{v^b}\,dv = \frac{a}{b-1}\left[1 - \frac{1}{x^{b-1}}\right], \qquad b \neq 1$$

$$G(x) = a \log x, \qquad b = 1.$$

A number of studies of the Zipf distribution have appeared since the initial presentations by Zipf, Estoup, and Lotka, the most extensive being that of Haitun [1]. In these early studies, and in many later ones, it was assumed that there was no upper limit to the number of types, that is, that $x_{\max} = \infty$. In any particular data set, of course, the maximal value is finite. However, in most discussions of the size-frequency model, it is assumed that no upper limit exists for the number of tokens. In many cases, this is a reasonable assumption; journals may be published and papers may be cited forever. In other cases, this is not a reasonable assumption; authors' publications are limited by their lifespans.

Four different views of the data have been taken in earlier studies in determining the particular Zipf model that is most appropriate in a given circumstance. The data are regarded as being one of:

1. a complete set of tokens from a finite population
2. an incomplete set of tokens from a finite population
3. a random sample of tokens from a finite population
4. a random sample of tokens from an infinite population

The number of tokens for each type is then determined. In a size-frequency analysis, the number of types for each of the possible numbers of tokens or sizes is tabulated. From this tabulation, the values of the parameters of the Zipf model are determined using one or more of the following techniques:

1. visual scanning of a plot of the tabulation
2. least squares estimation of the parameters
3. maximum likelihood estimation of the parameters
4. minimum chi-square estimation of the parameters
5. moment estimation of the parameters or estimates involving a subset of the frequencies

Visual scanning is useful as a preliminary step, but for more reliable estimators of the parameters, one of the other methods is usually employed. Frequently the parameters a and x_{max} are fixed and c as indicated earlier assumed to be 0. Estimation is then concerned with the parameter b, because, for fixed a, x_{max}, and b, t will be completely determined. Maximum likelihood estimates of b in general produce distributions with good fits to the empirical data [2]. Strictly speaking, when complete or incomplete populations of tokens are involved, the parameter values obtained by these techniques are more properly considered as approximations, rather than estimates in the statistical sense, because they are not based on random samples. It is the purpose of this paper to examine the sampling distribution of the maximal sample value x_{max}. Our reason for doing so is to indicate the errors that may be made in using a sample value of the maximum number of tokens assigned to a type as an estimate of the population value. The magnitude of the error in using this estimate will be shown to depend on the other parameters and on the sample size.

Random samples are relatively unusual in bibliometrics, at least at the present time. However, Allison [3] uses random samples of chemists and biochemists and tracks their publishing productivity over a number of years using Lotka's law and the negative binomial as a model; Potter [4,5] describes studies by the Library of Congress and University of Illinois using random samples from the library catalogs and modeling with Lotka's law. Richardson [6] does the same thing with an Australian academic library. Subramanyam [7,8] uses random samples from abstracting services to study the productivity of computer scientists. From the Samson and Bendell study [9] it does not seem unusual for random samples to be employed in the design and investigation of information retrieval systems. Even if random samples are not utilized frequently at the present time, it is obviously desirable to use them in the case where large catalogs, author populations, and data bases are involved. They may not now be used because not enough is known about sampling from such populations.

The ability to estimate the population maximal value from its sample counterpart is useful in designing the file structure for an online index or catalog. Whether an inverted file or hashed structure is used, it is important, in the design phase, to estimate the maximal number of postings for an author or a keyword. Where such systems are developed from existing printed indexes or card catalogs, this parameter may be estimated from a random sample of entries in the manual system, using the procedures developed in the following paragraphs. The use of Zipf-type distributions in file design is discussed in Tague, Nelson, and Wu [10] and Nelson and Tague [11].

Because the Zipf distribution assumes a central role in bibliometrics, it is important that the magnitude of estimation errors be known when using sample data to draw conclusions about a Zipf model. Bibliometricians are increasingly engaged in developing precise fitting techniques for Zipf-type distributions. It is hoped that this paper will be a contribution to this development.

Following this introduction and brief review of the Zipf distribution, the meaning of and the relations between the parameters for the simple Zipf distribution will be discussed. Previously suggested methods for estimating the parameters will be reviewed. Finally, the results of a computer derivation of several quantiles of the maximal size distribution will be presented. We seek to determine what relationships characterize this distribution and the extent to which these are dependent on the sample size. The application of these results in setting up a confidence interval for the maximal population value will then be indicated.

PARAMETERS OF A SIMPLE SIZE-FREQUENCY ZIPF DISTRIBUTION

The simple Zipf size-frequency distribution, then, has three parameters: a, b, and x_{max}. When this function is used as a model for the distribution of tokens over types, x_{max} represents the maximal size, that is, the largest number of tokens that can be assigned to a type. The parameter a represents the number of types with a single token. The parameter b, to some extent, represents the dispersion of the distribution of tokens over types. The larger the b, the larger the number of types with only one token and the more rapid the decline, with increasing x, of the frequencies $g(x)$. If b is small, the numbers $g(x)$ decline very slowly with x, and the distribution has a very long tail. Hence a higher proportion of

the tokens are concentrated among a few highly productive types. As b approaches 0, all sizes approach an equiprobable state. In a compilation of 105 distributions of human activities, using least squares to estimate b, Haitun [1] found that 33% had $b \leq 1$, 62% had $b \leq 2$, and only 5% $b > 10$. With regard to the c parameter in the more general Zipf distribution, Samson and Bendell [9] note that the effect of increasing this parameter is almost indistinguishable from the effect of reducing the b parameter.

Two other parameters that characterize empirical distributions of tokens over types, namely the total number of types t and the total number of tokens m, are related to these three fundamental parameters. If the continuous approximation to the Zipf is used, these relationships can be described by a simple function. The total number of types t is given by

$$t = \sum_{x=1}^{x_{\max}} g(x) = a \sum_{x=1}^{x_{\max}} 1/x^b.$$

If x is treated as a continuous variable and the summation replaced by an integration, then

$$t = \int_1^{x_{\max}} g(x)\, dx.$$

Thus,

$$t = \frac{a}{1-b}\,[x_{\max}^{1-b} - 1], \qquad \text{for } b < 1$$

$$t = a \log_e x_{\max}, \qquad \text{for } b = 1$$

$$t = \frac{a}{b-1}\left[1 - \frac{1}{x_{\max}}\, b - 1\right], \qquad \text{for } b > 1.$$

If b and $x_{\max}$ are fixed, the total number of types t varies directly with a, the number of types with one token. If a and $x_{\max}$ are fixed, t varies inversely with b, that is, the larger the b, the smaller the total number of types. As $b \to 0$, t approaches its maximum $ax_{\max} - 1$ value. If a and b are fixed, then t will increase with increasing $x_{\max}$, so that the larger the maximum number of tokens per type, the larger the total number of types. For $b > 1$, if $x_{\max}$ is large, the total number of types will be approximately $t = a/(b-1)$.

Since we must have $g(x_{\max}) \geq 1$ and therefore $a \geq x_{\max}^b$, t must satisfy the inequality

$$t \geq x_{\max}^b \sum_{x=1}^{x_{\max}} 1/x^b.$$

The total number of tokens is given by

$$m = \sum_{x=1}^{x_{\max}} xg(x) = a \sum_{x=1}^{x_{\max}} 1/x^{b-1}.$$

If the variable X can be approximated by a continuous productivity variable, the total productivity m will be as follows:

$$m = \frac{a}{2-b}\,[x_{\max}^{2-b} - 1], \qquad \text{if } b < 2$$

$$m = a \log_e x_{\max}, \qquad \text{if } b = 2$$

$$m = \frac{a}{b-2}\left[1 - \frac{1}{x_{\max}^{b-2}}\right], \qquad \text{if } b > 2.$$

In all cases, the total number of tokens increases with x_{max} if a and b are fixed, and with a if b and x_{max} are fixed. If x_{max} is very large and $b > 2$, m is approximated by $a/(b-2)$. If a and x_{max} are fixed, the total number of tokens varies inversely with b. Thus, if the number of types with lowest productivity and the maximum productivity for a type are known, a large population of tokens will be associated with low b values and a small population of tokens with high b values.

The ratio m/t represents the average number of tokens per type or average productivity of the population, that is, a type-token ratio. When x_{max} is large and X is represented by a continuous productivity variable, this average productivity will be approximately given by

$$m/t = \frac{1-b}{2-b} x_{max}^{1-b}, \qquad \text{if } b < 1$$

$$m/t = x_{max}/\log_e x_{max}, \qquad \text{if } b = 1$$

$$m/t = \frac{b-1}{2-b} x_{max}^{2-b}, \qquad \text{if } 1 < b < 2$$

$$m/t = \log_e x_{max}, \qquad \text{if } b = 2$$

$$m/t = \frac{b-1}{b-2}, \qquad \text{if } b > 2.$$

Thus, for values of b greater than 2, the average productivity is approximately independent of the maximal productivity when this is large. However, the average is dependent on the maximal value when b is less than or equal to 2.

We have seen that the maximal size and exponent parameters of a Zipf distribution are important in characterizing the nature of a Zipf population. In general, high values for average productivity will be found in populations where the exponent is small and where, consequently, many of the tokens are distributed among the highly productive types. Low values of average productivity will be found in populations where the exponent is large and where, consequently, most tokens are distributed among types of low productivity.

ESTIMATION OF THE EXPONENT AND THE MAXIMAL SIZE FOR A ZIPF POPULATION

Estimation of the exponent and maximal size of a Zipf population from a random sample of values will now be discussed. As indicated earlier, in many cases, bibliometric sets cannot be considered to be random samples but, rather, incomplete populations. However, this paper will limit itself to the random sample situation. Unfortunately, most incomplete populations cannot be regarded as populations with random omissions, because most often the omissions are of a particular nature, for example, involving types of low productivity.

Previously suggested methods for estimating the parameters will first be reviewed. In the following presentation, $x_1, x_2, \ldots, x_n$ represent a random sample of sizes from a population that can be described by the simple Zipf distribution. We let $g'(x)$ represent the observed number of tokens of size x.

Johnson and Kotz [12] present three methods for estimating b when x_{max} is infinite and these may also be applied when x_{max} is known. The methods are as follows:

1. The maximum likelihood estimator of b is the sample statistic b' that satisfies the following equation:

$$\frac{\sum_{i=1}^{n} \log_e x_i}{n} = \frac{\sum_{x=1}^{x_{max}} (\log_e x)/x^{b'}}{\sum_{x=1}^{x_{max}} 1/x^{b'}}.$$

2. An estimator b' based on the first two empirical frequencies:

$$b' = \frac{\log_e(g'(1)/g'(2))}{\log_e 2} .$$

3. The moment estimator b' obtained by equating the empirical and theoretical means:

$$\frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{x=1}^{x_{\max}} 1/x^{b-1}}{\sum_{x=1}^{x_{\max}} 1/x^b} .$$

If $x_{\max}$ is assumed to be infinite, then we must have $b > 2$; otherwise the right-hand side cannot be determined.

In general, a Zipf variable has moments of order $k > b$ if $x_{\max}$ is infinite. The equations in 1 and 3 cannot be solved exactly, unless $b = 2$ or $b = 4$ and $x_{\max}$ is infinite, so that iterative solutions are necessary. A table in Johnson and Kotz [12] provides solutions to one decimal place for the maximum likelihood equation for $1.1 \leq b \leq 5.0$, assuming infinitely large sizes. Johnson and Kotz also give the expression for the variance of the maximum likelihood estimator:

$$\frac{d}{db}\left[\frac{n\sum_{x=1}^{x_{\max}} (\log_e x)/x^b}{\sum_{x=1}^{x_{\max}} 1/x^b}\right]^{-1}$$

Again, a table is provided to determine this value for $1.5 \leq b \leq 4.0$ when $x_{\max}$ is infinitely large.

Linear least squares approximations of a and b in the frequency function may be determined by taking logarithms to linearize the Zipf frequency function. Pao [13] has reviewed this approach when $x_{\max}$ is infinite. The expressions are then

$$b' = n' \frac{\sum_{x=1}^{x_{\max}} \log_e x \log_e g'(x) - \sum_{x=1}^{x_{\max}} \log_e x \sum_{x=1}^{x_{\max}} \log_e g'(x)}{n' \sum_{x=1}^{x_{\max}} (\log_e x)^2 - \left(\sum_{x=1}^{x_{\max}} \log_e x\right)^2}$$

where n' is the number of distinct values of $x_{\max}$ in the sample.

$$\log a' = \frac{\sum_{x=1}^{x_{\max}} \log_e g'(x) - b' \sum_{x=1}^{x_{\max}} \log_e x}{n'} .$$

If $x_{\max}$ is infinite, then the summation $1/x^b$ in the above two expressions is given by the Riemann zeta function and will converge only for $b > 1$. In this latter case, exact values of the summation exist only for $b = 2$ (i.e., $\pi^2/6$) and $b = 4$ (i.e., $\pi^4/90$). Pao [13] gives a method for estimating the sum to a desired degree of accuracy.

In one of the few papers to consider simultaneous estimation of all four parameters of the generalized Zipf distribution, Samson and Bendell [9] suggest that minimum chi-square estimates be used, that is, the values of a, b, c, and $x_{\max}$ that minimize

$$\chi^2 = \sum_{x=1}^{x_{\max}} \frac{(g'(x) - a/(x+c)^b)^2}{a/(x+c)^b} \qquad a, b > 0.$$

Again, exact solutions for a, b, c, and x_{max} cannot be determined. A time-consuming pattern search method and a simpler heuristic method for determining these minimizing values are suggested in the paper.

Little attention has been paid apart from the Bendell and Samson [9] paper to the problem of the estimation of the maximal population value from its sample counterpart. The maximum likelihood estimate is the maximal sample value x'_{max}. However, no analytical results on the reliability of this estimator are known to the authors. Knowledge of the extent to which the maximal sample value may deviate from the population value is important in bibliometric discussion. Frequently, the sample value is assumed to be the same as that in the population, with no concern for the extent of the error of the estimate.

The remainder of this paper is a development of the cumulative probability distribution of the maximal sample value given population values for b, x_{max}, and t (thus determining a). From these distributions, it is possible to construct confidence intervals for the population maximal value based on a sample value.

DISTRIBUTION OF THE MAXIMAL SAMPLE VALUE

If we let $Y = x'_{max}$ represent the largest number of tokens possessed by a type in a sample of n types from a Zipf population of t types, then the cumulative distribution function for Y represents the probability that the maximal sample value will be less than or equal to a specified value. This probability, that is, the probability that the random variable Y assumes a value less than or equal to y is given by the following expression:

$$\text{Prob}[Y \leq y] = F(y|n,t,b,x_{max}) = \prod_{i=1}^{n} \left\{ \frac{\sum_{x=1}^{y} a(t,b,x_{max})/x^b - i + 1}{\sum_{x=1}^{x_{max}} a(t,b,x_{max})/x^b - i + 1} \right\}$$

where n is the sample size, t is the number of types, x_{max} is the maximal population size, and

$$a(t,b,x_{max}) = t/\sum_{x=1}^{x_{max}} 1/x^b.$$

Notice that this distribution is independent of m, the total number of tokens. The derivation of this distribution is given in the Appendix. An example of the distribution for $b = 1.5$, $x_{max} = 100$, $t = 1000$, and $n = 10, 30, 50, 100, 500$ is shown in Fig. 1. By calculating this cumulative probability for each value of Y in turn, for fixed n, t, b, and x_{max}, it is possible to determine the appropriate pth quantile values of Y, that is, the smallest value y_p such that

$$\text{Prob}[Y \leq y_p] = F(y_p|n,t,b,x_{max}) \geq p.$$

Thus, if y_p is obtained in a sample and used as an estimate of x_{max}, the error of estimate is $e_p = x_{max} - y_p$. The probability the error will be less than or equal to this value is the probability the sample maximal value is greater than or equal to y_p, that is, approximately $1 - p$. Thus, if we obtain a sample maximal value y', not knowing the population maximal value x_{max}, we can be $100(1-p)\%$ confident the interval $y' + e_p$ contains the true population maximal value. In general, we will not know e_p. However, if e_p is seen to converge under certain conditions, it may be possible to set up approximate confidence intervals.

In order to examine the behavior of Y, particularly as the sample size n increases, several values of y_p were determined for the values of t in the body of Table 1, for each

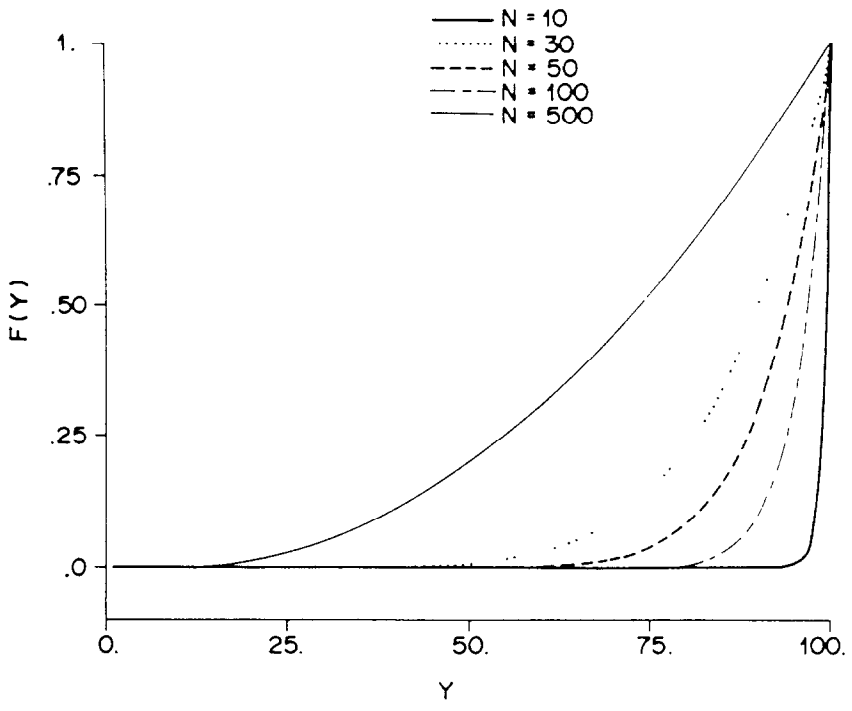

Fig. 1. Sampling distribution of the maximal sample value [F(Y)] for $b = 1.5$; $x_{max} = 100$; $t = 1{,}000$; $n = 10, 30, 50, 100, 500$.

value of b specified in the rows and each value of x_{max} specified in the columns. For each such value of t, the values of n are given in Table 2. Recall from the earlier discussion that

$$t \geq x_{max}^{b} \sum_{x=1}^{x_{max}} 1/x^{b}$$

and that, of course, $n \leq t$. Samples constituting a high proportion of the population of types were considered, as well as samples that were small compared to the population. The reason for this choice has been mentioned earlier: In many bibliometric studies, the data represent an incomplete population rather than a small sample. Of course, the results will be valid only if the omissions are random.

The results of these calculations are shown in Table 3, for p values of .001, .01, .05, and .1. For each combination of b, x_{max}, and t values, once a sample size has resulted in $Y_p = x_{max}$ for all p, no further rows are indicated, because all remaining Y_p will be equal to x_{max}.

Table 1. Values of t used in exact calculations of y_p

	x_{max}		
b	10	100	1,000
1	100, 1,000, 5,000, 10,000	1,000, 5,000, 10,000	10,000
1.5	100, 1,000, 5,000, 10,000	5,000, 10,000	
2	1,000, 5,000, 10,000		
2.5	1,000, 5,000, 10,000		
3	5,000, 10,000		

Table 2. Values of n used in exact calculations of y_p

t	n
100	10, 30, 50, 90
1,000	10, 30, 50, 90, 500, 900
5,000	10, 30, 50, 90, 500, 900, 2,500, 4,500
10,000	10, 30, 50, 90, 500, 900, 2,500, 4,500, 5,000, 9,000

If the number of types t is large, then the sampling distribution may be approximated by

$$F(y|n,b,x_{\max},t) = \left[\frac{\sum_{x=1}^{y} 1/x^b}{\sum_{x=1}^{x_{\max}} 1/x^b}\right]^n = F(y|n,b,x_{\max})$$

In other words, the distribution is independent of the total number of types t.

If the number of tokens per type or productivity can be approximated by a continuous distribution, the expression is simplified even further:

$$F(y|n,b,x_{\max}) = \left[\frac{1 - 1/y^{b-1}}{1 - 1/x_{\max}^{b-1}}\right]^n \qquad b \neq 1$$

$$F(y|n,b,x_{\max}) = \left[\frac{\log_e y}{\log_e x_{\max}}\right]^n \qquad b = 1$$

This distribution, for $b = 1.5$, $x_{\max} = 100$, and $n = 10, 30, 50, 100, 500$ is shown in Fig. 2. If we set $F(y'_p|n,b,x_{\max}) = p$, using these two expressions, and then solve for y'_p, we get

$$y'_p = [1 - (p)^{1/n}(1 - 1/x_{\max}^{b-1})]^{-1/(b-1)}, \qquad b \neq 1 \tag{1}$$

$$\log_e y'_p = (p)^{1/n} \log_e x_{\max}, \qquad b = 1. \tag{2}$$

The values of y'_p determined from this approximation may be considered limiting values of the pth quantiles when the number of types and of productivity values is large. Figure 3 shows the difference $y_p - y'_p$, which represents the error introduced by the approximation, for $t = 10{,}000$, $p = .05$, and various values of b and $x_{\max}$. As expected, the error decreases as t increases. It also appears least for small values of b and $x_{\max}$. Its behavior with increasing n needs further study. For small n, the error is large and erratic; as n increases, the error declines to 0.

If $x_{\max}$ also becomes very large, then we have a second approximation:

$$y''_p = [1 - p^{1/n}]^{-1/(b-1)}, \qquad b \neq 1.$$

This approximation cannot be used if $b \leq 1$ since, in this case, y''_p will increase without limit. As the sample size n increases, y_p increases to $x_{\max}$, as would be expected. Similarly, as b increases, y_p will decrease to 1, another indication that large values of b are associated with a dispersion of the tokens among the types in the lower productivity values.

APPROXIMATE CONFIDENCE INTERVALS

The final topic to be considered is that of a confidence bound for the maximal population value given a maximal sample value. From the definition of y_p, we know that the

Table 3. pth quantile values y_p of the maximum sample value

t	n	$y_{.001}$	$y_{.01}$	$y_{.05}$	$y_{.1}$
$b = 1$, $x_{max} = 10$					
100	10	3	4	5	6
	30	6	7	8	9
	50	8	9	9	10
		10	10	10	10
1,000	10	2	4	5	6
	30	6	7	8	8
	50	7	8	9	9
	90	9	9	10	10
	500	10	10	10	10
5,000	10	2	4	5	6
	30	6	7	8	8
	50	7	8	9	9
	90	8	9	10	10
	500	10	10	10	10
10,000	10	2	4	5	6
	30	6	7	8	8
	50	7	8	9	9
	90	8	9	10	10
	500	10	10	10	10
$b = 1$, $x_{max} = 100$					
1,000	10	8	15	26	35
	30	35	49	62	69
	50	52	64	75	80
	90	70	79	85	89
	500	95	97	98	99
	900	99	99	100	100
5,000	10	8	15	26	35
	30	35	48	61	69
	50	52	64	74	80
	90	69	78	85	88
	500	94	96	98	98
	900	97	98	99	99
	2,500	99	100	100	100
	4,500	100	100	100	100
10,000	10	8	15	26	35
	30	35	48	61	69
	50	52	64	74	80
	90	69	78	85	88
	500	94	96	98	98
	900	97	98	99	99
	2,500	99	100	100	100
	4,500	100	100	100	100
$b = 1$, $x_{max} = 1000$					
10,000	10	24	63	144	215
	30	215	345	492	576
	50	382	519	648	715
	90	577	690	784	829
	500	905	936	958	968
	900	947	965	977	982
	2,500	983	989	993	995
	4,500	992	995	997	998
	5,000	993	996	997	998
	9,000	998	999	1,000	1,000
$b = 1.5$, $x_{max} = 10$					
100	10	2	2	3	4
	30	4	5	7	7
	50	6	7	8	9
	90	9	9	10	10
1,000	10	2	2	3	4
	30	4	5	6	7
	50	5	6	7	8
	90	7	8	9	9
	500	10	10	10	10

continued

Table 3. continued.

t	n	$y_{.001}$	$y_{.01}$	$y_{.05}$	$y_{.1}$
$b = 1.5$, $x_{max} = 10$ continued					
5,000	10	2	2	3	4
	30	4	5	6	7
	50	5	6	7	8
	90	7	8	9	9
	500	10	10	10	10
10,000	10	2	2	3	4
	30	4	5	6	7
	50	5	6	7	8
	90	7	8	9	9
	500	10	10	10	10
$b = 1.5$, $x_{max} = 100$					
5,000	10	2	3	6	8
	30	8	14	22	28
	50	15	24	35	42
	90	28	39	52	59
	500	75	82	88	91
	900	86	90	94	95
	2,500	96	97	98	99
	4,500	99	100	100	100
10,000	10	2	3	6	8
	30	8	14	22	28
	50	15	24	35	42
	90	28	39	52	59
	500	74	82	88	90
	900	85	90	93	95
	2,500	95	97	98	99
	4,500	98	99	99	100
	5,000	98	99	99	100
	9,000	100	100	100	100
$b = 2$, $x_{max} = 10$					
1,000	10	1	1	2	2
	30	2	3	4	5
	50	3	4	5	6
	90	5	6	7	8
	500	9	9	10	10
	900	10	10	10	10
5,000	10	1	1	2	2
	30	2	3	4	5
	50	3	4	5	6
	90	5	6	7	7
	500	9	9	10	10
	900	9	10	10	10
	2,500	10	10	10	10
10,000	10	1	1	2	2
	30	2	3	4	5
	50	3	4	5	6
	90	5	6	7	7
	500	9	9	10	10
	900	9	10	10	10
	2,500	10	10	10	10
$b = 2.5$, $x_{max} = 10$					
1,000	10	1	1	1	2
	30	2	2	3	3
	50	2	3	4	4
	90	3	4	5	5
	500	7	8	9	9
	900	9	10	10	10
5,000	10	1	1	1	2
	30	2	2	3	3
	50	2	3	4	4
	90	3	4	5	5
	500	7	8	8	9
	900	8	9	9	10
	2,500	10	10	10	10
	9,000	86	90	94	95

continued

Table 3. continued.

t	n	$y_{.001}$	$y_{.01}$	$y_{.05}$	$y_{.1}$
$b = 2.5$, $x_{max} = 10$ continued					
10,000	10	1	1	1	2
	30	2	2	3	3
	50	2	3	4	4
	90	3	4	5	5
	500	7	8	8	9
	900	8	9	9	9
	2,500	9	10	10	10
	4,500	10	10	10	10
$b = 3$, $x_{max} = 10$					
5,000	10	1	1	1	1
	30	1	2	2	2
	50	2	2	3	3
	90	2	3	3	4
	500	5	6	7	7
	900	6	7	8	8
	2,500	8	9	9	10
	4,500	10	10	10	10
10,000	10	1	1	1	1
	30	1	2	2	2
	50	2	2	3	3
	90	2	3	3	4
	500	5	6	7	7
	900	6	7	8	8
	2,500	8	9	9	10
	4,500	9	10	10	10
	5,000	9	10	10	10
	9,000	10	10	10	10

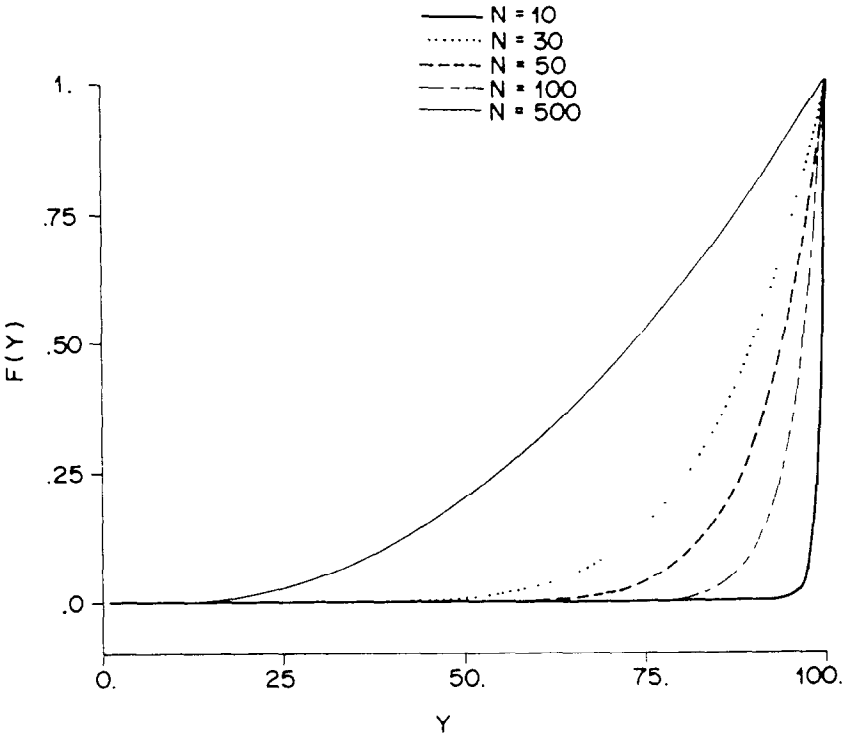

Fig. 2. Approximate sampling distribution of the maximal sample value [F(Y)] for $b = 1.5$; $x_{max} = 100$; $t = 1{,}000$; $n = 10, 30, 50, 100, 500$.

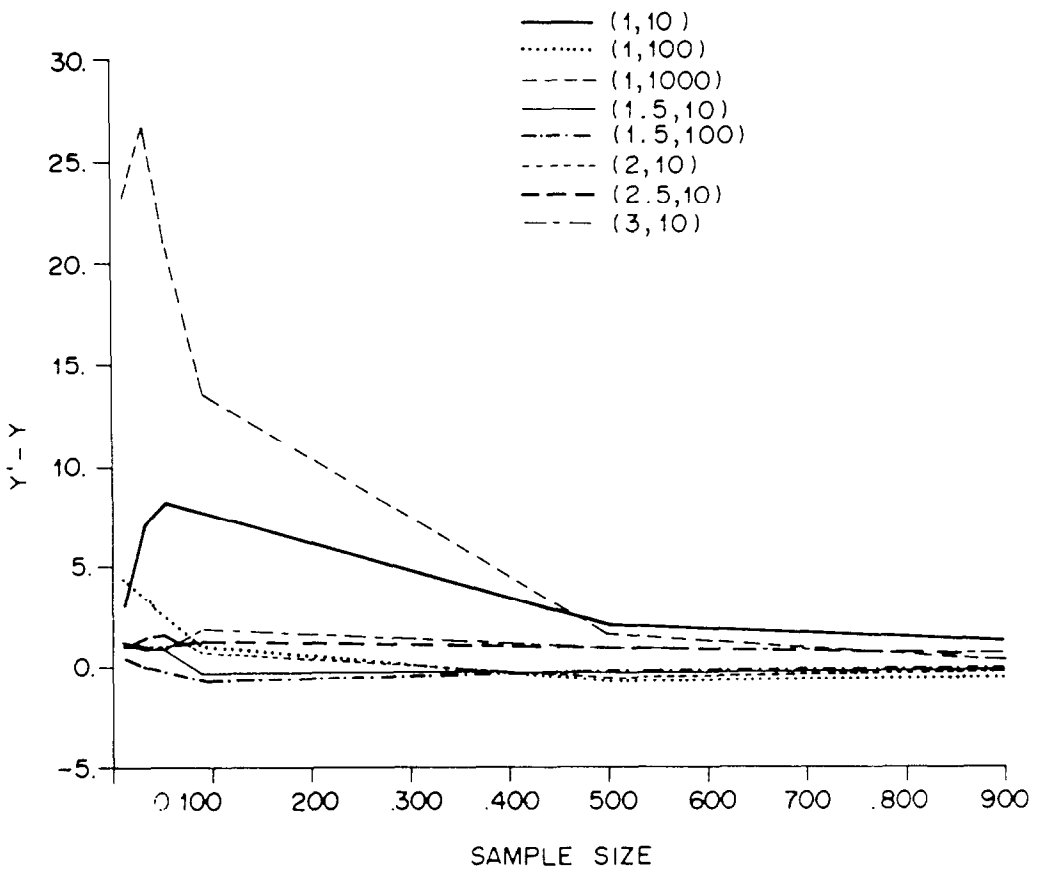

Fig. 3. Error $y' - y$ for $p = .05$; $t = 10{,}000$; various (b, x_{max}).

probability the maximal sample value will be greater than y_p is $1 - p$. Thus, in repeated sampling we would expect that $1 - p$ proportion of the sample maximal values will be within the interval $[y'_p, x_{max}]$. So, if we assume $y = y_p$ and solve for x_{max} in eqns (1) and (2) in the previous section, the interval from the observed y to the calculated x'_{max} will be a $100(1 - p)\%$ confidence interval when the number of types and of values for the number of types per token are both large. The expressions for x'_{max} are then as follows:

$$x'_{max} = \left[\frac{p^{1/n}}{p^{1/n} - 1 + 1/y^{b-1}}\right]^{1/(b-1)}, \qquad b \neq 1$$

$$\log_e x_{max} = \log_e y/p^{1/n}, \qquad b = 1.$$

Table 4 shows the upper limit for a 95% confidence interval based on the previous values of b and n and with y values of 10, 100, and 1,000. Particularly marked in this table is the unreliability of sample estimates of x_{max} when $b = 1$ and the sample size is small. This wide range of possible x_{max} values results from the concentration of productivity at the high end of the scale when $b = 1$ and, of course, the low reliability of small samples.

Finally, Table 5 shows the upper bound for x_{max} for some data sets published in connection with studies of the distribution of scientific productivity assuming different values of b. Of course, not all these data sets were intended to be random samples; however, it is instructive to see the extent of the difference between the empirical value and the theoretical limit. Whether the interval is realistic depends, of course, on the particular kind of bias the sampling procedure exhibits and, indeed, whether they are samples at all from a large population of authors.

CONCLUSIONS

This paper has discussed some of the properties of Zipf-type distributions, in particular, those relating to the estimation of the parameters of the distribution. The maximal size or productivity is one such parameter that has not, in the past, been studied extensively. The ability to estimate this parameter reliably is important is data base design.

Table 4. Upper limit for 95% confidence interval for x_{max} based on approximate sample maximal value y

	n					
b	10	30	50	100	500	1000
$y' = 10$						
1	22.3	12.7	11.5	10.7	10.1	10.1
1.5	11.6	10.5	10.3	10.1	10.0	10.0
2	13.5	11.0	10.6	10.3	10.1	10.0
2.5	15.7	11.6	10.9	10.4	10.1	10.0
3	18.2	12.2	11.3	10.6	10.1	10.0
$y' = 100$						
1	499.5	162.2	132.9	115.0	102.8	101.4
1.5	116.6	105.1	103.0	101.5	100.3	100.2
2	134.9	110.5	106.2	103.0	100.6	100.3
2.5	156.7	116.2	109.4	104.6	100.9	100.5
3	182.1	122.1	112.7	106.2	101.2	100.6
$y' = 1{,}000$						
1	1164.7	2065.6	1531.9	1233.8	1042.4	1020.9
1.5	1161.6	1051.2	1030.4	1015.1	1003.0	1001.5
2	1349.3	1105.0	1061.7	1030.4	1006.0	1003.0
2.5	1567.3	1161.6	1094.0	1045.9	1009.0	1004.5
3	1820.6	1221.1	1127.3	1061.7	1012.1	1006.0

Table 5. .95 upper confidence limit for population x_{max}

Study	Total No. Authors (t)	Sample Maximum (y)	Exponent (b)	.95 Upper Confidence Limit*
Dresden (1922) [14]	278	70	1.80	112
Lotka (1926) [15]	1,325	48	2.05	55
Lotka (1926) [15]	6,890	346	1.95	392
Dufrenoy (1938) [16]	1,529	8	2.55	8
Hersh (1942) [17]	826	131	1.85	177
Williams (1944) [18]	411	10	2.45	12
Williams (1944) [18]	1,537	11	2.45	11
Leavens (1953) [19]	721	46	2.10	62
Mantell (1966) [20]	2,255	16	2.75	18
Mantell (1966) [20]	97	9	2.25	15
Windsor (1975) [21]	93	4	2.50	5
Windsor (1975) [21]	71	8	2.35	18
Coile (1975) [22]	1,282	7	3.50	8
Coile (1975) [22]	1,339	8	3.45	10
Coile (1975) [22]	1,666	10	3.40	13
Coile (1975) [22]	3,206	10	3.50	11
Coile (1975) [22]	3,512	10	3.30	11
Radhakrishnan/Kernizan (1979) [23]	301	7	3.05	10
Radhakrishnan/Kernizan (1979) [23]	599	5	3.40	6
Radhakrishnan/Kernizan (1979) [23]	1,021	7	3.10	8
Radhakrishnan/Kernizan (1979) [23]	851	7	3.00	8
Rao (1980) [24]	1,111	23	2.15	25
Hubert (1981) [25]	754	19	2.35	23
Hubert (1981) [25]	1,630	21	2.45	23

*Numbers rounded up to the next higher integer value.

The tables of the sample maximal size distribution derived here serve as a warning that only large samples are likely to provide a maximal size that is close to the population value. The tabulated results also illustrate the significant part played by the exponent parameter as an indicator of inequality and in determining the extent to which the sample maximal value may be used as the population maximal value. It is hoped that the tables, approximation formulas, and confidence intervals will prove useful in assessing the reliability of empirically derived models.

REFERENCES

1. Haitun, S. D. Stationary scientometric distributions. Scientometrics 4(1):5-25; 4(2):89-104; 4(3):181-194; 1982.
2. Nicholls, P. T. Empirical validation of Lotka's law. Info. Proc. & Manag. 22(5):417-419; 1986.
3. Allison, P. D. Inequality and scientific productivity. Soc. Stud. Sci. 10:163-179; 1980.
4. Potter, W. G. When names collide: conflict in the catalog and AACR2. Lib. Res. Tech. Serv. 24(1):3-16; 1980.
5. Potter, W. G. Lotka's law revisited. Lib. Trends 31(2):21-39; 1981.
6. Richardson, V. L. Lotka's law and the catalogue? Aust. Academ. & Res. Lib. 12(3):185-190; 1981.
7. Subramanyam, K. Lotka's law and the literature of computer science. IEEE Trans. Prof. Commun. 22(4):187-189; 1979.
8. Subramanyam, K. Research productivity and breadth of interest of computer scientists. J. Am. Soc. Info. Sci. 35(6):369-371; 1985.
9. Samson, W. B.; Bendell, A. Rank order distributions and secondary key indexing. Comp. J. 28(3):309-312; 1985.
10. Tague, J. M.; Nelson, M. J.; Wu, H. Problems in the simulation of bibliographic retrieval systems. Information Retrieval Research: Proceedings of the Symposium; 1980; Cambridge. London: Butterworth's; 1981: (p. 236-255).
11. Nelson, M. J.; Tague, J. M. Split size-rank models for the distribution of index terms. J. Am. Soc. Info. Sci. 36(5):283-296; 1985.
12. Johnson, N. I.; Kotz, S. Discrete distributions. Boston: Houghton Mifflin; 1969.
13. Pao, M. L. Lotka's law: a test in procedure. Info. Proc. & Manag. 21(4):305-320; 1985.
14. Dresden, A. A. A report on the scientific work of the Chicago Section, 1897-1919. Bull. Am. Math. Soc. 28:303-307; 1922.
15. Lotka, A. J. The frequency distribution of scientific productivity. J. Wash. Acad. Sci. 16(12):317-323; 1926
16. Dufrenoy, J. The publishing behaviour of biologists. Quart. Rev. Biol. 13:207-210; 1938.
17. Hersh, A. H. *Drosophila* and the course of research. Ohio J. Sci. 42(5):198-200; 1942.
18. Williams, C. B. The number of publications written by biologists. Ann. Eugen. 12:143-145; 1944.
19. Leavens, D. H. Letter. Econometrica 21:620-622; 1953.
20. Mantell, L. H. On laws of special abilities and the production of scientific literature. Am. Document. 17(1):8-16; 1966
21. Windsor, D. A. Developing drug literatures, 1. Bibliometrics of baclofen and dantrolene sodium. J. Chem. Infor. Comp. Sci. 15(4):237-241; 1975.
22. Coile, R. C. Letter. J. Am. Soc. Info. Sci. 26(2):133-134; 1975.
23. Radhakrishnan, T.; Kernizan, R. Lotka's law and computer science literature. J. Am. Soc. Info. Sci. 30(1):51-54; 1979.
24. Rao, I. K. R. The distribution of scientific productivity and social change. J. Am. Soc. Info. Sci. 31(2):111-122; 1980.
25. Hubert, J. J. A rank-frequency model for scientific productivity. Scientometrics 3(3):191-202; 1981.

APPENDIX

Derivation of the distribution function for the maximal sample value

To prove:

$$\text{Prob}[Y \le y] = F(y|n,t,b,x_{\max}) = \prod_{i=1}^{n} \left\{ \frac{\sum_{x=1}^{y} a(t,b,x_{\max})/x^b - i + 1}{\sum_{x=1}^{x_{\max}} a(t,b,x_{\max})/x^b - i + 1} \right\}, \quad y = 1,2,\ldots,x_{\max}$$

where

$$a(t,b,x_{\max}) = t/\sum_{x=1}^{x_{\max}} 1/x^b.$$

Given a random sample of n types from a population of t types, where the number of types with exactly x tokens, $x = 1,2,\ldots,x_{\max}$ is given by

$$g(x) = \frac{a(t,b,x_{\max})}{x^b},$$

the probability that the largest value in the sample, Y, assumes a value less than or equal to y is equal to the probability of selecting n types from t types such that all n types have y or fewer tokens. The number of types t_y in the population of t types with y or fewer tokens is

$$t_y = \sum_{x=1}^{y} \frac{a(t,b,x_{\max})}{x^b} \tag{1}$$

Thus, the required probability is a hypergeometric probability equal to the ratio of the number of ways of selecting n types from t_y types to the number of ways of selecting n types from t types. This means

$$\begin{aligned} P[Y \leq y] &= \binom{t_y}{n} \Big/ \binom{t}{n} \\ &= \frac{t_y!\,n!\,(t-n)!}{n!\,(t_y-n)!\,t!} \\ &= \prod_{i=1}^{n} \left\{ \frac{t_y - i + 1}{t - i + 1} \right\} \end{aligned} \tag{2}$$

Since $t = t_{x_{\max}}$, from (1) and (2), we have

$$P[Y \leq y] = \prod_{i=1}^{n} \left\{ \frac{\sum_{x=1}^{y} a(t,b,x_{\max})/x^b - i + 1}{\sum_{x=1}^{x_{\max}} a(t,b,x_{\max})/x^b - i + 1} \right\}.$$