Information Processing & Management Vol 23, No 3, pp 171-182, 1987
Printed in Great Britain

0306-4573/87 $3.00 + .00

ANALYSIS OF ZIPF'S LAW: AN INDEX APPROACH

YE-SHO CHEN
Department of Quantitative Business Analysis, Louisiana State University,
Baton Rouge, LA 70803

and

FERDINAND F. LEIMKUHLER
School of Industrial Engineering, Purdue University, West Lafayette, IN 47907

(*Received* 27 *July* 1986; *in final form* 7 *November* 1986)

Abstract—A rigorous analysis of Zipf's law is made using an index for the sequence of observed values of the variables in a Zipf-type relationship. Three important properties relating rank, count, and frequency are identified. Using this approach, the shape of Zipf-type curves can be described in terms of three distinct regions and two parameters of the Mandelbrot-Zipf law. This result has considerable practical significance, since it provides rigorous foundations for the application of Zipf's law.

1. INTRODUCTION

One of the most widely cited laws of human behavior is based on Zipf's [1] observation that if one takes the words making up an extended body of text and ranks them by their number of occurrences, then the rank r multiplied by its corresponding frequency of occurrence, $g(r)$, will be approximately constant, that is,

$$g(r) = ar^{-1}, \quad r = 1,2,3,\ldots \tag{1}$$

where a is a positive constant. Zipf's law has been applied to the study of many different kinds of human phenomena. Recent applications include program complexity in software engineering by Shooman [2], keyword distribution in bibliographic data base design by Fedorowicz [3–6] and design of very large data bases by Wiederhold [7], data compacting in computer networking by Ting [8], information retrieval by Smith and Devine [9], and statistical analysis of text by Chen [10].

A major difficulty in using the law is that frequency-rank data typically fail to follow a simple linear log-log relationship, as shown in Fig. 1. In the figure, Latin words exhibit the typical concave frequency-rank pattern that has been found in English words [1, p. 123], German words [1, p. 117], and Norwegian words [1, p. 128]. In contrast to these findings, a convex frequency-rank pattern was found in Chinese characters [1, p. 91] and in Gothic root morphemes [1, p. 91]. More flexible models, such as those proposed by Simon [12], Mandelbrot [13], and Sichel [14–18], have assumed an a priori relationship between the count, rank, and frequency of words. In this paper, we propose a new formulation of Zipf's law that takes explicit account of the typical characteristics observed in Zipf-type data.

In particular, we use an index to designate sequential observations of the ranked data. This formulation of Zipf's law is shown to have three important rank-count-frequency properties that are subsequently used to study the common shapes of Zipf-type curves. The various possible curve forms are then classified in terms of three regions of rank and two parameters associated with an earlier formulation of the law by Mandelbrot [13].

As a prelude to examining the paper, Section 2 examines the traditional frequency-count and frequency-rank approaches and the associated problems. Section 3 discusses the index approach. Section 4 derives the slope of Zipf-type curves, and the shapes of the curves with respect to three important regions are discussed in Sections 5 and 6. Finally, Section 7 represents the conclusion.

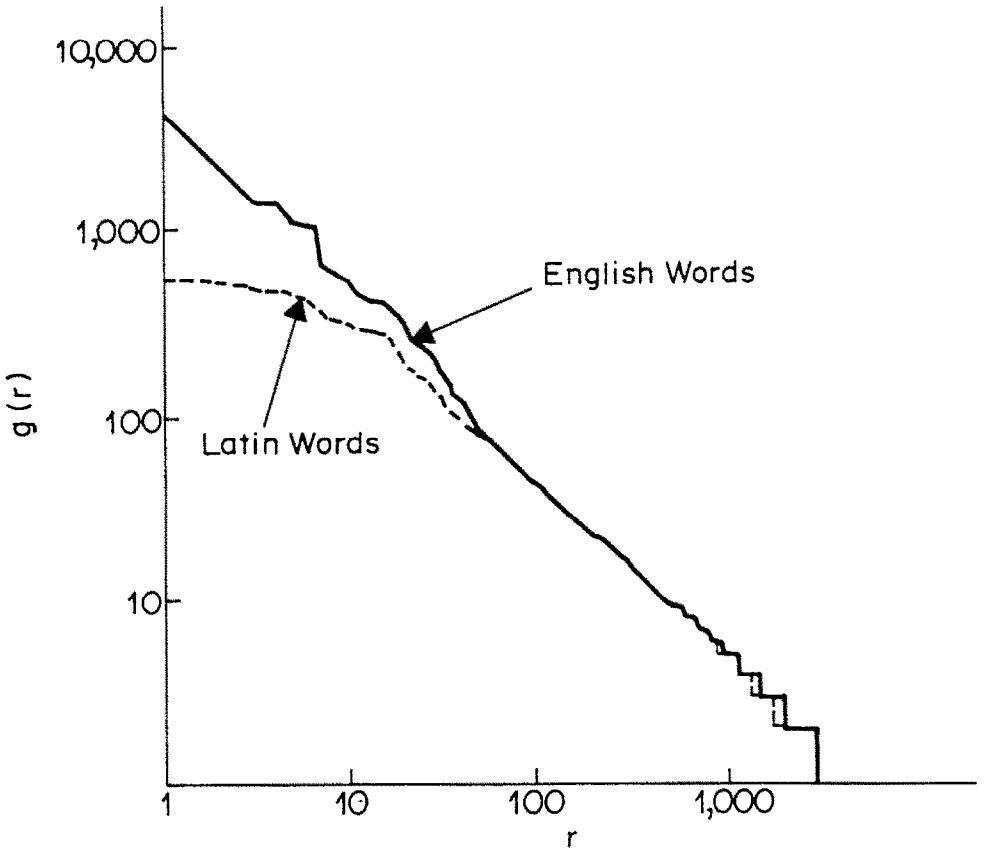

Fig. 1. Zipf-type (or frequency-rank) plot for English and Latin Words [11, p. 44]. Note that ranks one and two of English words have been modified to fit the data. Zipf's original figure suggests that they have the same frequency of occurrence and is a mistake. The authors thank one of the referees for pointing this out.

2. FREQUENCY-COUNT AND FREQUENCY-RANK DISTRIBUTIONS

Zipf-type data on word occurrences are based on observations of four entities: (a) word count, n, that is, the number of occurrences of a certain word contained in a text; (b) the count frequency, $f(n)$, or the number of words of each count; (c) the word rank, r, that is, the cumulative frequency of words of the same or greater count; and (d) the rank frequency, $g(r)$, that is, the number of words of the same rank. Note that when several words have the same count, they are assumed to have the same maximal-rank, which is the largest possible rank. Other ranking methods use the minimum-rank, average-rank, and random-rank [19]. Zipf [1] assigned the random-rank to all words with the same frequency of occurrence. The differences between Zipf's random-rank and our maximal-rank methods can be seen most clearly when viewed in terms of the index approach introduced in Section 4, as discussed later.

Two approaches are taken with Zipf's law: (a) frequency-rank and (b) frequency-count. When the different words in a large text are ranked in the order of decreasing frequency of occurence, a frequency-rank distribution, $g(r)$ versus r, is obtained, as shown in eqn (1). An alternative approach is to arrange the words in increasing order of occurrence or count to obtain a frequency-count distribution [20], $f(n)$ versus n, for example,

$$f(n) = bn^{-2}, \quad n = 1,2,3,\ldots \tag{2}$$

This equation is the same as Lotka's law of scientific productivity [21].

Two drawbacks are associated with most formulations proposed in the literature. First, they assume the independent variables r and n run from one to infinity and, second, they

assume that r and n are consecutive without any "jump" or gaps. Typically, the observed values of count and rank, beyond the first smaller values, will "jump" to larger values in progressively larger steps; that is, they contain "gaps" and do not run consequently from one to infinity. The nature of the gaps has a significant impact on the shapes of Zipf-type curves and cannot be ignored. One common problem with this assumption is to invalidate goodness-of-fit tests.

For example, consider the zero-truncation generalized inverse Gaussian-Poisson (GIGP) distribution proposed by Sichel [14–18]:

$$f(n/t) = [(1-\theta_t)^{-\gamma_t/2} K_{\gamma_t}\{\alpha_{\gamma_t}(1-\theta_t)^{1/2}\}\,]^{-1} \frac{\left(\frac{1}{2}\alpha_t\theta_t\right)^n}{n!} K_{n+\gamma_t}(\alpha_t)$$

$$n = 0,1,2,3,\ldots$$

where the three parameters are $-\infty < \gamma_t < \infty$, $0 \leqq \theta_t \leqq 1$ and $\alpha_t \geqq 0$. $K_\nu(z)$ is the modified Bessel function of the second kind of order ν with argument z, and t stands for the length of the time period to be considered. In Table 1 of Sichel's paper [15], $f(89) = f(255) = 1$ and $f(n) = 0$ for $n = 90,91,\ldots,254$. However, the GIGP distribution ignored the gap (between 89 and 255) and "predicts" $f(n) > 0$ for $n = 90,91,\ldots,254$.

A more realistic approach is to associate an index, $i = 1,2,\ldots,m$, with each step or observed value of count and rank, as discussed below.

3. THE INDEX APPROACH

We introduce the notion of an index $i = 1,2,\ldots,m$ and let n_i and r_i denote the ith different observed value of count and rank, respectively, so that $n_{i+1} > n_i$ and $r_{i+1} > r_i$. Let $f(n_i)$ and $F(n_i)$, respectively, denote the number of words having a count of exactly n_i and no less than n_i. Also, let $g(r_i)$ denote the frequency of occurrence of words with the rank r_i. The data in Table 1 are taken from Good [22], who had analyzed a sample of English nouns in Macaulay's essay on Bacon. The frequency-rank or Zipf-type plot of the data is shown in Fig. 2.

The indices $i = 1,2,\ldots,m$, can be divided into three groups: where i is small, where i is close to m, and otherwise. For small i, usually $n_i = i$. For instance, in Table 1, $n_i = i$, for $i = 1,2,\ldots,41$. For i close to m, usually $f(n_i) = 1$. For instance, in Table 1, $f(n_i) = 1$, for $i = 43,\ldots,50$. Let i_l be the maximum i such that $n_i = i$ and let i_u be the minimum i such that $f(n_i) = 1$ and $f(n_{i-1}) \neq 1$. Then we have the following three important properties:

$$n_i = i, \quad 1 \leqq i \leqq i_l. \tag{3}$$

(For example, in Table 1, $n_i = i$ for $1 \leqq i \leqq 41$.)

$$f(n_1) = 1, \quad i_u \leqq i \leqq m. \tag{4}$$

(For example, in Table 1, $f(n_i) = 1$ for $43 \leqq i \leqq 50$.)

$$n_i \cong i \text{ and } f(n_i) \cong 1, \; i_l + 1 \leqq i \leqq i_u - 1. \tag{5}$$

(For example, in Table 1, $n_i \cong i$ and $f(n_i) \cong 1$ for $i = 42$.)

In terms of the index notations, one also sees the following relationships between r, n, F, and g. Proofs are given by Chen [10] and Hubert [23]. For $i = 1,2,\ldots,m$,

$$r_i = F(n_{m-i+1}). \tag{6}$$

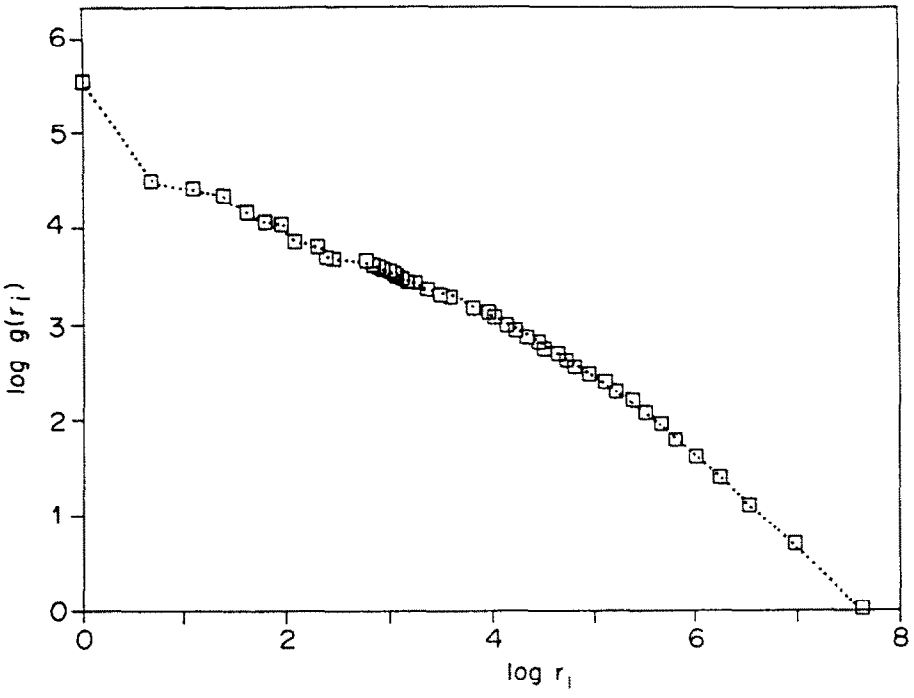

Fig. 2. Zipf-type plot of Good's data [22].

(For example, $r_{10} = 11$ and $F(n_{41}) = 11$ [both from Table 1].)

$$g(r_i) = n_{m-i+1}. \tag{7}$$

(For example, $g(r_{10}) = 41$ and $n_{41} = 41$ [both from Table 1].)

The empirical relationships in eqns (6) and (7) can be used to prove the following:

THEOREM 1: *For* $i = 1,2,\ldots,m$,

$$g(r_i) = a(r_i + b)^c \tag{8}$$

$$\text{iff } F(n_i) = dn_i^e - b \tag{9}$$

where a, b, c, d, and e, *are constants, and* $a, d > 0$, $c, e < 0$, $ce = 1$, $da^e = 1$, *and* $b > -1$.

Proof: Using equations (6), (7), and (8), we have

$$n_{m-i+1} = a(F(n_{m-i+1}) + b)^c, \quad i = 1,2,\ldots,m$$

which is equivalent to eqn (9). A different proof of Theorem 1 can be found in Chen and Leimkuhler [24]. □

Equations (8) and (9) are realistic formulations of Zipf's law and Lotka's law in terms of index levels of rank and count. Equation (8), without the index notation, was first derived by Mandelbrot [12] and is often called the Mandelbrot-Zipf law. Equation (9), on the other hand, was derived through the indexed Mandelbrot-Zipf law and could be called the Mandelbrot-Lotka law. For further details on the formulations of the laws, see Chen and Leimkuhler [25].

4. THE SLOPE OF ZIPF-TYPE CURVES

There are several variations of Zipf-type curves. Figures 1 and 2 show three classes of the curves. Each figure represents a typical example(s) of variation and is well known in the literature. Figure 1 shows the Zipf's plot of the data taken from Zipf [11]. It shows

Table 1. Zipfian approach of Good's data [22].

A	B	C	D	E	F	G
i	n_i	$f(n_i)$	$n_i f(n_i)$	$F(n_i)$	r_i	$g(r_i)$
1	1	990	990	2,048	1	255
2	2	367	734	1,058	2	89
3	3	173	519	691	3	81
4	4	112	448	518	4	76
5	5	72	360	406	5	65
6	6	47	282	334	6	58
7	7	41	287	287	7	57
8	8	31	248	246	8	48
9	9	34	306	215	10	45
10	10	17	170	181	11	41
11	11	24	264	164	12	40
12	12	19	228	140	16	39
13	13	10	130	121	18	38
14	14	10	140	111	19	37
15	15	13	195	101	20	36
16	16	3	48	88	21	35
17	17	10	170	85	22	34
18	18	7	126	75	23	33
19	19	6	114	68	24	32
20	20	5	100	62	26	31
21	21	1	21	57	29	30
22	22	4	88	56	30	29
23	23	7	161	52	34	28
24	24	2	48	45	37	27
25	25	1	25	43	42	26
26	26	5	130	42	43	25
27	27	3	81	37	45	24
28	28	4	112	34	52	23
29	29	1	29	30	56	22
30	30	3	90	29	57	21
31	31	2	62	26	62	20
32	32	1	32	24	68	19
33	33	1	33	23	75	18
34	34	1	34	22	85	17
35	35	1	35	21	88	16
36	36	1	36	20	101	15
37	37	1	37	19	111	14
38	38	2	76	18	121	13
39	39	4	156	16	140	12
40	40	1	40	12	164	11
41	41	1	41	11	181	10
42	45	2	90	10	215	9
43	48	1	48	8	246	8
44	57	1	57	7	287	7
45	58	1	58	6	334	6
46	65	1	65	5	406	5
47	76	1	76	4	518	4
48	81	1	81	3	691	3
49	89	1	89	2	1,058	2
50	255	1	255	1	2,048	1
Sum		2,048	8,045			

Column A = index i, $i = 1, 2, \ldots, m$; $m = 50$ in this case.
Column B = number of occurrences.
Column C = number of words $f(n_i)$.
Column D = Column B * Column C.
Column E = $F(n_i) = \sum_{k=i}^{m} f(n_k)$ = number of words with occurrences no less than n_i.
Column F = cumulation of Column C from the bottom.
= rank r_i of words having a corresponding given number of occurrences.
Column G = $g(r_i) = n_{m-i+1}$
= the frequency of occurrence of words with rank r_i.

a linear curve (English words) and a curve (Latin words) with the concavity to the origin. Figure 2, on the other hand, shows a curve with the convexity to the origin. The three classes of curves do show a common characteristic, that is, a linear decreasing pattern to the right of the plots.

Several general formulations of Zipf's distribution, for example, Good [22], Sichel [14–18], and Simon [12], have been proposed to model the various shapes of Zipf-type curves. However, because they ignored the gaps, their formulations have failed to model the variations. As we will see later, the nature of gaps plays an important role in the shape of a Zipf-type curve. This fact is identified through the study of slopes by using the index approach to analyze the classes of Zipf-type curves reported in the literature. Consider the slopes of the Zipf-type curve

$$s_j = \frac{\log g(r_{j+1}) - \log g(r_j)}{\log r_{j+1} - \log r_j}, \quad j = 1,2,\ldots,m-1$$

The following lemma shows an equivalent form of the slopes.

LEMMA 1: *For* $j = 1,2,\ldots,m-1$,

$$s_j = \frac{\log\left(\dfrac{n_{m-j}}{n_{m-j+1}}\right)}{\log \dfrac{F(n_{m-j})}{F(n_{m-j+1})}} \tag{10}$$

Proof: Apply eqns (6) and (7), one has eqn (10). □

From Lemma 1 and the three properties of eqns (3), (4), and (5), we can derive the following theorem. The proof can be found in the Appendix.

THEOREM 2:

a. For $1 \leqq j \leqq m - i_u$,

$$s_j = \frac{\log \dfrac{n_{m-j}}{n_{m-j+1}}}{\log \dfrac{j+1}{j}} \tag{11}$$

b. For $m - i_u + 1 \leqq j \leqq m - i_l - 1$,

$$s_j \cong \frac{\log \dfrac{m-j}{m-j+1}}{\log \dfrac{j+1}{j}} \tag{12}$$

c. For $m - i_l \leqq j \leqq m - 1$,

$$s_j = \frac{\log \dfrac{m-j}{m-j+1}}{\log \dfrac{F(m-j)}{F(m-j+1)}}. \tag{13}$$

We define: $1 \leqq j \leqq m - i_u$, $m - i_u + 1 \leqq j \leqq m - i_l - 1$, and $m - i_l \leqq j \leqq m - 1$, as region I, region II, and region III, respectively, and study the underlying mechanisms for

the shape of a Zipf-type curve within each region. Before continuing, we discuss the differences between Zipf's random-rank and our maximal-rank methods as follows.

Consider Table 1, which shows two words contributing 45 occurrences each, with each assigned the maximal-rank 10. In Zipf's random-rank approach, the assigned ranks for the two words will be 9 and 10, respectively. Basically, there are no differences between the two approaches, except for region III. In Zipf's approach, there are steps for large values of rank as shown in Figure 1, while in our approach, there are only connected dot points. The advantages in using maximal-rank versus random-rank are the following:

1. The maximal-rank approach is reversible. That is, we can convert the frequency-rank distribution into the frequency-count distribution and vice versa through eqns (6) and (7). Reversibility is very important for the comparison of those theoretical explanations of Zipf's law, because some researchers used the frequency-count approach (e.g., Simon [12] and Sichel [14–18]) and some used the frequency-rank approach (e.g., Mandelbrot [13]). The random-rank method makes the reversibility impossible.
2. The maximal-rank approach shows a simple functional relationship between rank and frequency. Zipf's original intention was to find a simple equation to fit the phenomenon he observed. However, the use of random-rank makes things more complicated, because step functions are not simple functions.
3. The maximal-rank approach avoids the controversy of assigning rank in the case of ties. This may happen in some applications of Zipf's law where the order of ranks is sensitive.

5. REGION I AND THE MANDELBROT-ZIPF LAW

Equation (11) indicates that the shape of a Zipf-type curve in the region $1 \leqq j \leqq m - i_u$ depends on the scattering pattern of n_i/n_{i+1}, $i_u \leqq i \leqq m - 1$. To analyze further, consider eqns (7) and (11). We have

$$s_j = \frac{\log \dfrac{n_{m-j}}{n_{m-j+1}}}{\log \dfrac{j+1}{j}} = \frac{\log \dfrac{g(r_{j+1})}{g(r_j)}}{\log \dfrac{j+1}{j}}, \qquad 1 \leqq j \leqq m - i_u$$

Equation (4) shows $F(n_i) = m - i + 1$ for $i_u \leqq i \leqq m$, which is equivalent to $F(n_{m-j+1}) = j$ for $1 \leqq j \leqq m - i_u + 1$. From equation (6), we obtain $r_j = j$ for $1 \leqq j \leqq m - i_u + 1$. Thus,

$$s_j = \frac{\log \dfrac{g(j+1)}{g(j)}}{\log \dfrac{j+1}{j}}, \qquad 1 \leqq j \leqq m - i_u \tag{14}$$

The last equation shows the distribution of $g(j)$, $1 \leqq j \leqq m - i_u + 1$, plays the crucial role in determining the shape of the first region. A reasonable choice of the function is to assume eqn (8) is true, because it is a general form of the Mandelbrot-Zipf law and is flexible enough to model the data of $g(j)$, especially for $1 \leqq j \leqq m - i_u + 1$.

By using the index approach, we deduce the following corollary:

COROLLARY 1: *In region I, the Zipf-type curve is*

a. *Concavely decreasing iff* $b < 0$,

b. *Linearly decreasing iff* $b = 0$,

c. *Convexly decreasing iff* $b > 0$,

where $b > -1$ *is the shift coefficient of the Mandelbrot-Zipf law of eqn (8).*

Proof. From eqns (8) and (14), we obtain

$$s_j = c \frac{\log\left(1 + \frac{1}{j+b}\right)}{\log\left(1 + \frac{1}{j}\right)}, \qquad 1 \leqq j \leqq m - i_u \tag{15}$$

If $b = 0$, *then* $s_j = c < 0$ is linearly decreasing. If $b \neq 0$, then we apply Taylor expansion of the natural logarithm (p. 51, Feller [26]) to the numerator of eqn (15) and obtain

$$\log\left(1 + \frac{1}{j+b}\right) = (j+b)^{-1} - \frac{1}{2}(j+b)^{-2} + \frac{1}{3}(j+b)^{-3} - \ldots \cong (j+b)^{-1}$$

Similar technique applies to the denominator of equation (15) and we have

$$s_j \cong c \frac{1}{1 + \frac{b}{j}}$$

When $-1 < b < 0$, the last equation shows that s_j is a decreasing function of j, and thus the curve in this region is concave. When $b > 0$, the last equation indicates that s_j is an increasing function of j and therefore the curve in this region is convex. □

Corollary 1 makes a significant modeling contribution by identifying parameter b as the crucial factor in determining the shape of a Zipf-type curve in region I. The Zipf-type curve looks linear if $b = 0$. If $-1 < b < 0$, the curve in region I is concave and if $b > 0$, the curve is convex.

6. REGIONS II AND III: LINEARITY

Equation (12) indicates that the shape of a Zipf-type curve in region II, where $m - i_u + 1 \leqq j \leqq m - i_l - 1$, is approximately linear. We show this as follows.

COROLLARY 2: *In region II, the Zipf-type curve is approximately linear.*

Proof: Consider $m - i_u + 1 \leqq j \leqq m - i_l - 2$, □

$$\frac{s_{j+1}}{s_j} \cong \frac{\log\frac{m-j-1}{m-j}\log\frac{j+1}{j}}{\log\frac{m-j}{m-j+1}\log\frac{j+2}{j+1}}$$

$$= \frac{\log\left(1 - \frac{1}{m-j}\right)}{\log\left(1 - \frac{1}{m-j+1}\right)} \frac{\log\left(1 + \frac{1}{j}\right)}{\log\left(1 + \frac{1}{j+1}\right)} \cong 1$$

Equation (13) indicates that the shape of a Zipf-type curve in region III, where $m - i_l \leqq j \leqq m - 1$, depends heavily on the values of $F(i)$, $1 \leqq i \leqq i_l + 1$. To analyze further, we need some assumption on $F(i)$, $1 \leqq i \leqq i_l + 1$. A reasonable choice is to assume eqn (9) is true, since it is a general form of the Mandelbrot-Lotka law and is flexible enough to model the data of $F(i)$, especially for $1 \leqq j \leqq i_l + 1$.

From eqns (9) and (13), we derive that for $m - i_l \leqq i \leqq m - 1$,

$$s_j = \frac{\log \dfrac{m-j}{m-j+1}}{\log \dfrac{(m-j)^e - b/d}{(m-j+1)^e - b/d}} \tag{16}$$

Evaluating eqn (16) on a computer, we find that $s_j \cong e^{-1} = c$. This is because b/d is relatively small and eqn (16) becomes an identity equation. We summarize the result as follows:

COROLLARY 3: *In region III, the Mandelbrot-Lotka law implies an approximately linear Zipf-type curve with slope equal to c, the exponent coefficient of the Mandelbrot-Zipf law of eqn (8).*

7. CONCLUSION

In this analysis, we take explicit account of the sequence of observed values of the variables in a Zipf-type relationship by means of an index. This approach identifies three important properties relating count, rank, and frequency, as shown in eqns (3), (4), and (5). Using these properties, an analysis is made of the shapes of three classes of Zipf-type curves. Three significant regions are identified. Fig. 3 summarizes the findings in the regions in terms of two shape parameters.

1. The first region is concavely decreasing, linearly decreasing, or convexly decreasing. The deciding factor for the shape is the distribution gaps between n_i's. The shift coefficient b in the Mandelbrot-Zipf law provides a parametric explanation of the concavity, linearity, and convexity.
2. The second region is approximately linear. This property is robust; that is, in all of the Zipf-type curves, the middle region is linear.
3. The third region is approximately linear with slope equal to c, the exponent coefficient of the Mandelbrot-Zipf law.

This formulation of Zipf's law makes it possible to account for the variations normally encountered with Zipf-type data. This result has considerable practical significance, since it provides rigorous foundations for the application of Zipf's law, for example, in computer science [2–10], communications [27], language structure [28], anthropology [29], psychology [30], and information science [31].

APPENDIX

Proof of Theorem 2

From eqn (10), for $i = 1,2,\ldots,m-1$,

$$s_{m-i} = \frac{\log \dfrac{n_i}{n_i+1}}{\log \dfrac{F(n_i)}{F(n_{i+1})}} \tag{17}$$

The three groups of indices identified in eqns (3), (4), and (5) play an important role in the following proof.

c. For $1 \leqq i \leqq i_l$, eqn (3) implies

$$s_{m-i} = \frac{\log \dfrac{i}{i+1}}{\log \dfrac{F(i)}{F(i+1)}} \tag{18}$$

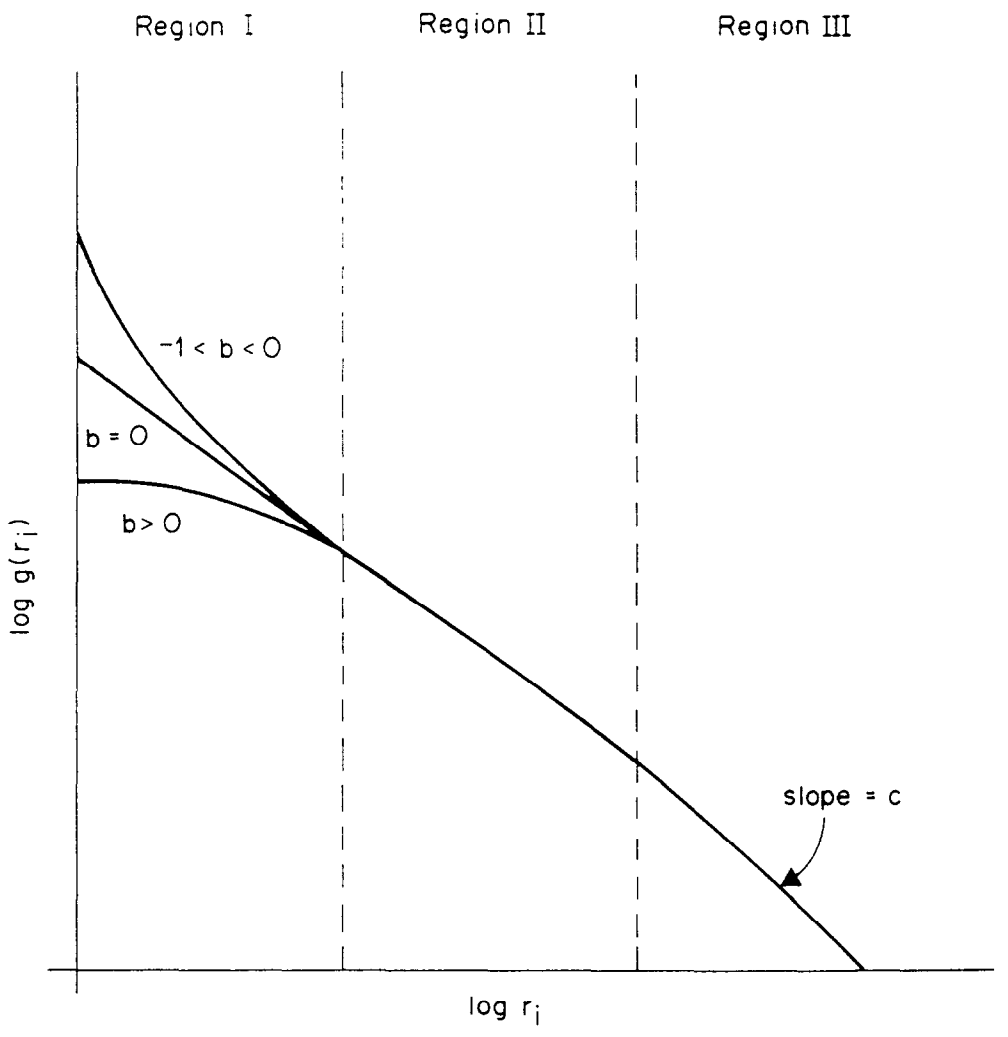

Fig. 3. Various variations of Zipf-type curves within the three regions.

Equation (18) is equivalent to

$$s_j = \frac{\log \dfrac{m-j}{m-j+1}}{\log \dfrac{F(m-j)}{F(m-j+1)}}, \qquad m - i_l \leqq j \leqq m-1 \tag{13}$$

b. For $i_l + 1 \leqq i \leqq i_u - 1$, eqn (5) implies

$$F(n_i) = \sum_{k=i}^{m} f(n_k) \cong m - i + 1$$

and

$$s_{m-i} \cong \frac{\log \dfrac{i}{i+1}}{\log \dfrac{m-i+1}{m-i}} \tag{19}$$

Equation (19) is equivalent to

$$s_j \cong \frac{\log \dfrac{m-j}{m-j+1}}{\log \dfrac{j+1}{j}}, \qquad m - i_u + 1 \leqq j \leqq m - i_\ell - 1 \tag{12}$$

a. For $i_u \leqq i \leqq m-1$, eqn (4) implies

$$s_{m-i} = \frac{\log \dfrac{n_i}{n_{i+1}}}{\log \dfrac{m-i+1}{m-i}} \tag{20}$$

Equation (20) is equivalent to

$$s_j = \frac{\log \dfrac{n_{m-j}}{n_{m-j+1}}}{\log \dfrac{j+1}{j}}, \qquad 1 \leqq j \leqq m - i_u \tag{11}$$

Acknowledgments—This research was supported in part by the National Science Foundation Grant IST-7911893A1 and by the Council on Research of Louisiana State University. The authors also thank the referees for their valuable comments.

REFERENCES

1. Zipf, G. K. Human behavior and the principle of least effort. Reading, MA: Addison-Wesley; 1949.
2. Shooman, M. L. Software engineering: design, reliability, and management. New York: McGraw-Hill; 1983.
3. Fedorowicz, J. E. Modeling an automatic bibliographic system: a Zipfian approach. Pittsburgh, PA: Carnegie-Mellon University; 1981; Available from: University Microfilms, Ann Arbor, MI. Doctoral dissertation.
4. Fedorowicz, J. E. A Zipfian model of inverted file storage requirements. Vogt, W. G.; Mickie, M. H., eds Proc. Twelfth Annual Pittsburgh Conference on Modeling and Simulation; 1981 April 30-May 1. Research Triangle Park, NC: Instrument Society of America; 1981.
5. Fedorowicz, J. E. A Zipfian model of an automatic bibliographic system: an approach to MEDLINE. J. Am. Soc. Info. Sci. 33:223-232; July 1982.
6. Fedorowicz, J. E. The theoretical foundation of Zipf's law and its application to the bibliographic data base environment. J. Am. Soc. Info. Sci. 33:285-293; September 1982.
7. Wiederhold, G. Data base design. 2nd ed. New York: McGraw-Hill; 1983.
8. Ting, T. C. Compacting homogeneous text for minimizing storage space. Int. J. Comput. Info. Sci. 6:211-221; 1977.
9. Smith, F. J.; Devine, K. Storing and retrieving word phrases. Info. Process. & Manag. 21:215-225; 1985.
10. Chen, Y. S. Statistical models of text: a system theory approach. West Lafayette, IN: Purdue University; 1985. Doctoral dissertation.
11. Zipf, G. K. The psycho-biology of language: an introduction to dynamic biology. Boston: Houghton Mifflin; 1935 (paperback, Cambridge, MA: MIT Press, 1965).
12. Simon, H. A. On a class of skew distribution function. Biometrika. 52:425-446; 1955.
13. Mandelbrot, B. An information theory of the statistical structure of language. Proc. Symposium on Applications of Communication Theory; September 1952; London. London: Butterworths; 1953:486-500.
14. Sichel, H. S. On a distribution representing sentence-length in written prose. J. Roy. Stat. Soc. Series A. 137:25-34; 1974.
15. Sichel, H. S. On a distribution law for word frequencies. J. Am. Stat. Assoc. 70:542-547; 1975.
16. Sichel, H. S. Repeat-buying and the generalized inverse Gaussian-Poisson distribution. Appl. Stat. 31:193-204; 1982.
17. Sichel, H. S. Asymptotic efficiencies of three methods of estimation for the inverse Gaussian-Poisson distribution. Biometrika 69:467-472; 1982.
18. Sichel, H. S. On a distribution law for word frequency. J. Am. Stat. Assoc. 70:542-547; 1975.
19. Hubert, J. J. General bibliometric models. Libr. Trends. pp. 65-81; Summer 1981.
20. Johnson, N. L.; Kotz, S. Discrete distributions: univariate distributions—1. The Houghton Mifflin Series in Statistics; Boston: Houghton Mifflin; 1978.
21. Lotka, A. J The frequency of distribution of scientific productivity. J. Washington Acad. Sci. 16:317-323; 1926.
22. Good, I. J. The population frequencies of species and the estimation of population parameters. Biometrika 40; 1953.

23. Hubert, J. J. A relationship between two forms of Bradford's law. J. Am. Soc. Info. Sci. 29:159–161; 1978.
24. Chen, Y. S.; Leimkuhler, F. F. Bradford law: an index approach. Scientometrics. In press.
25 Chen, Y. S.; Leimkuhler, F. F. A relationship between Lotka's law, Bradford's law, and Zipf's law. J. Am. Soc. Info. Sci. 37:307–314; 1986.
26. Feller, W. An introduction to probability theory and its applications, V1. New York: John Wiley; 1968.
27. Benbasat, I.; Dexter, A. S.; Masulis, P. S. An experimental study of the human-computer interface. CACM 24:752–762; 1981.
28. Fenk, A.; Fenk, G. Correlation between short-term-memory and the flow of linguistic information. Z. Experimentelle und Argwandate Psychol. 27:400–414; 1980.
29. Ammermen, A. J Surveys and archaeological research Annual Rev. Anthropol. 10:63–88; 1981.
30. McCusker, L. Y.; Hillinger, M. L.; Bias, R. G. Phonological recoding and reading. Psychol. Bull. 89:217–245; 1981.
31. Leimkuhler, F. F.; Chen, Y. S.; Johnson, B. D. Analysis and application of information productivity models. 1-5 Progress Report, School of Industrial Engineering, Purdue University; NSF Grant IST-701189333A1; March 1983.