Zipf's law, the central limit theorem, and the random division of the unit interval

Richard Perline*
Flexible Logic Software, 34-50 80th Street, Suite 22, Queens, New York, 11372
(Received 30 August 1995)

It is shown that a version of Mandelbrot's monkey-at-the-typewriter model of Zipf's inverse power law is directly related to two classical areas in probability theory: the central limit theorem and the ''broken stick'' problem, i.e., the random division of the unit interval. The connection to the central limit theorem is proved using a theorem on randomly indexed sums of random variables [A. Gut, *Stopped Random Walks: Limit Theorems and Applications* (Springer, New York, 1987)]. This reveals an underlying log-normal structure of pseudoword probabilities with an inverse power upper tail that clarifies a point of confusion in Mandelbrot's work. An explicit asymptotic formula for the slope of the log-linear rank-size law in the upper tail of this distribution is also obtained. This formula relates to known asymptotic results concerning the random division of the unit interval that imply a slope value approaching -1 under quite general conditions. The role of size-biased sampling in obscuring the bottom part of the distribution is explained and connections to related work are noted. [S1063-651X(96)01007-0]

PACS number(s): 05.40.+j, 02.50.−r, 87.10.+e

I. INTRODUCTION

Mandelbrot's creative and influential work on Zipf's inverse power law of word frequency distributions contains two potential points of confusion. One of these concerns what Mandelbrot [1] now acknowledges as the ''linguistic shallowness'' of the law. The possible confusion here stems from the fact that Mandelbrot [2] has shown that an inverse power distribution of pseudoword frequencies can be generated from random text, although in contrast his earlier work [3,4] aimed at information-theoretic models that might shed light on deeper properties of language or thought. A second point of confusion concerns Mandelbrot's, [5, see pp. 209–211] extended argument opposing the log-normal distribution as a model for word frequency distributions, although a version of his random text model has a natural underlying log-normal structure. The first issue, linguistic shallowness, was addressed decades ago by the prominent language researchers Miller and Chomsky [6,7]. It is the purpose of this article to address the second issue, log-normal structure, by proving the previously unrecognized—but direct—connection between a version of Mandelbrot's monkey-at-the-typewriter model and a special case of Anscombe's [8] important generalization of the central limit theorem. We also show that the upper tail of this log-normal structure is an approximate inverse power law and discuss asymptotic results from the classical ''broken stick'' problem that explain why a log-log rank-size plot of the upper tail will tend to have a slope close to -1 under quite general conditions.

The confusion on both points—linguistic shallowness and log-normal structure—is intertwined. Mathematically oriented researchers in psychology and linguistics were at first quite interested in Mandelbrot's use of information theory to derive Zipf's law, but then also realized the significance of his derivation of an inverse power law using nothing but a Markov random text model. Miller [6] and Miller and Chomsky [7] greatly clarified the situation by discussing an illustrative, special case of Mandelbrot's model. Miller [6] showed that one can straightforwardly derive a step function approximating an inverse power law with the simplifying assumptions of *equiprobable* and independently combined letters. He concludes with the clear words—which are implied, but not stated, in the Appendix to [2]—that Zipf's law can be derived ''without appeal to least effort, least cost, maximal information, or any branch of the calculus of variations.''

Physical scientists interested in Zipf's law through Mandelbrot's work may be unaware of Miller's clarification, probably because it was published in the psychology literature. For example, Li [9] writes that ''Miller did not give a proof of his statement''—referring to a comment by Miller [10] that Zipf's law could be generated by random text—and then gives the same proof of the special equiprobable letters case given by Miller [6] and Miller and Chomsky [7]. And recently Mantegna *et al.* [11] have come under criticism (see [12]) for arguing that noncoding DNA sequences may be transmitting biological information based on an analysis which they assert shows that noncoding DNA pseudoword frequencies conform approximately to a Zipf-like law. They might have been less likely to make such an argument had they been familiar with Miller's clarification of Mandelbrot's results.

II. EXHIBITING LOG-NORMAL STRUCTURE

Unfortunately, Miller's simplified model is degenerate in the sense that it does not reveal the natural log-normal structure of word probabilities that exists for the case of *unequal* letter probabilities. For this case, Mandelbrot, [5, p. 210] did note that the logarithmic probabilities of randomly generated pseudowords of a *fixed* number n of letters will be approximately normal; however, he overlooked a stronger result, following directly from Anscombe's theorem on random sums [8], which shows that the logarithmic probabilities of *all words of n or fewer letters* will be approximately normal.

*Electronic address: rkper@acm.org

Therefore the probabilities themselves will be approximately log-normal.

Mandelbrot's derivation of Zipf's law from his Markov random text model is based on the ensemble of all the word probabilities that can be generated by random sequences of all possible lengths, i.e., an infinite vocabulary. For proving the log-normal structure, our analysis will be based on a scheme assuming (1) unequal and independent letter probabilities (i.e., we drop the Markov assumption) and (2) a maximum word length of n letters, although we study asymptotic behavior as $n\rightarrow\infty$.

Assume an alphabet consisting of $K\geqslant 2$ nonspace characters $L_1,L_2,\ldots,L_K$ and the space character L_{K+1}. Let the letter keys be struck independently with probabilities $a_1,a_2,\ldots,a_{K+1}$, where $\sum_{j=1}^{K+1}a_j=1$ and $0<a_j<1$ for $1\leqslant j\leqslant K+1$. For the purposes of proving the log-normal structure of the word probabilities, it is necessary to assume that $a_i\neq a_j$ for at least one case where $i\neq j$ and $i,j\leqslant K$. Choose an integer n and define a trial as the outcome after $n+1$ or fewer letters have been selected. The outcome will result in either (a) a ''word'' consisting of n or fewer nonspace characters plus the ending space character or (b) a ''nonword'' consisting of $n+1$ consecutive nonspace characters with no ending space character. As many trials are run as desired, but each is limited to a maximum of $n+1$ characters and ends when either outcome (a) or (b) occurs.

A finite vocabulary model introduces the difficulty of the K^{n+1} nonwords, which have a positive aggregate probability of occurrence on each trial of $(\sum_{j=1}^{K}a_j)^{n+1}=(1-a_{K+1})^{n+1}$. We exclude the nonwords and analyze the distributional characteristics of the probabilities for the $N_n=\sum_{j=0}^{n}K^j$ ''legitimate'' words.

It simplifies matters to factor out a_{K+1} and refer to the resulting values as *base values*. The largest base value is always equal to 1, corresponding to the word of 0 nonspace letters. Let B_j denote the multiset (i.e., a set that can contain repetitions of its elements) of the K^j base values for each of the words of exactly j letters. Let $U_n=B_0\cup\cdots\cup B_n$ be the multiset of all base values for words from 0 to n letters. A generic element of the multisets B_j and U_n will be written b. Write the ranked values of U_n as b_r, where r represents rank from the top. Define a probability space on U_n by the natural counting measure, i.e., each element $b\in U_n$ is assigned an atom of probability equal to $1/N_n$. Denote the random variable defined in this way as Y_n.

By this definition, Y_n can be represented as the product of a random number R_n of independent and identically distributed random variables: $Y_n=X_1X_2\cdots X_{R_n}$, where $0\leqslant R_n\leqslant n$. Each X_i is a multinomial random variable taking on the values a_j, $1\leqslant j\leqslant K$, with equal probability. The case $R_n=0$, when $Y_n=1$, corresponds to the single 0-letter word. Since there are K^j words of exactly j letters, the probability that Y_n will have the representation $Y_n=X_1X_2\cdots X_j$ involving exactly j factors is K^j/N_n. Therefore, letting $P\{\ \}$ be the probability of the expression inside the braces, the probability mass function of R_n is $P\{R_n=j\}=K^j/N_n$ for $j=0,1,\ldots,n$.

Take logarithms (base-K logarithms prove convenient; ln will be used for natural logarithms) to obtain $\log_K Y_n=\sum_{j=1}^{R_n}\log_K X_j$. So $\log_K Y_n$ is the sum of a *random number* of independent and identically distributed random variables with finite variance. Anscombe's theorem [8] assures the asymptotic normality of $\log_K Y_n$ if it can be shown that the ratio R_n/n converges in probability to a constant $c>0$ as $n\rightarrow\infty$. Since $P\{R_n=n-j\}=K^{n-j}/N_n>(K-1)/K^{j+1}$, it is not difficult to show that $P\{|R_n/n-1|\leqslant j/n\}>1-1/K^{j+1}$. Letting $j=[\sqrt{n}]$, the largest integer in $\sqrt{n}$, and $n\rightarrow\infty$ proves that R_n/n converges in probability to $c=1>0$. Consequently, $\log_K Y_n$ is asymptotically normal and so Y_n will be approximately log-normal for sufficiently large n.

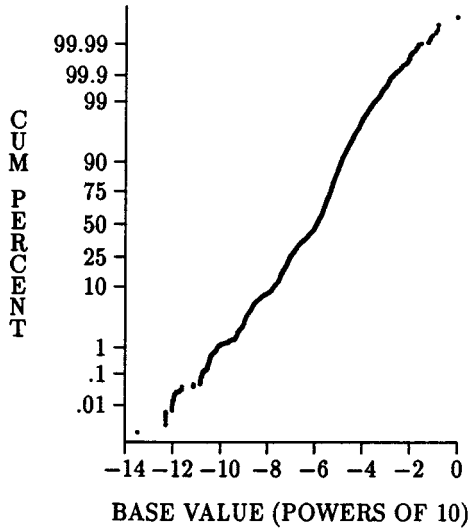

FIG. 1. Log-normal probability plot of base values for the random text model. The approximate linearity of the curve shows the approximate log-normality of the values.

The approximate log-normality of Y_n is illustrated in the approximately linear log-normal probability plot of Fig. 1, where we take $n=5$, $K=10$, and use nonspace letter probabilities a_1 through a_{10} with values 0.002 04, 0.0305, 0.0575, 0.06, 0.0668, 0.0715, 0.0837, 0.12, 0.144, and 0.148. All the values in U_5 have been generated and plotted. The points (x,y) plotted in the graph are $x=\log_{10}b_r$ and $y=\Phi^{-1}[(N_n-r+1)/(N_n+1)]$ for $1\leqslant r\leqslant N_n$, where Φ^{-1} is the inverse of the standard normal distribution function.

III. AN INVERSE POWER UPPER TAIL AND THE ''BROKEN STICK'' PROBLEM

Significantly, the upper tail of the distribution of base values in U_n conforms approximately to an inverse power law, as can been seen in Fig. 2. This figure shows the same base values previously displayed in Fig. 1 now represented in log-log coordinates. The vertical axis represents $\log_K b_r$ and the horizontal axis represents the logarithm of the associated rank, $\log_K r$. The evident linearity of the graph over most of the top range of base values is graphic proof of an approximate inverse power law in this range.

Mandelbrot's [2] combinatorial proof explains this phenomenon, but his derivation does not lead to an explicit formula for the slope of the linear part of this graph. We obtain an asymptotic estimate for it here and show that it connects up in a natural way with the ''broken stick'' problem. Consider the least squares regression of $\log_K b_r$ onto $\log_K r$ con-

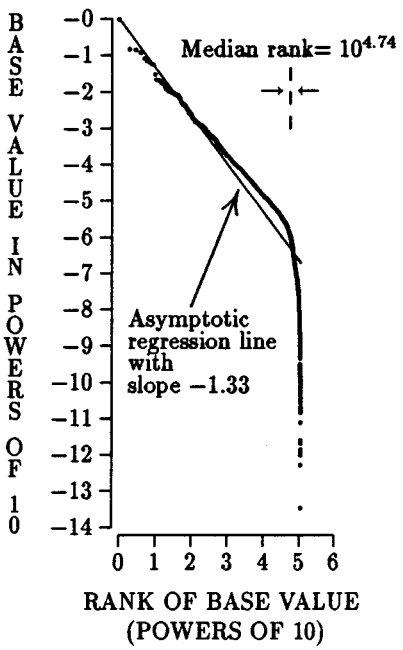

FIG. 2. Log-log plot of base values by rank for the random text model. The linear trend is evident for approximately the top half of the data. In the random text model, observed word frequencies would be those of a sample drawn from a multinomial distribution with probabilities proportional to these base values. Zipf's inverse power law would tend to be observed because the bottom part of the distribution (i.e., words of low probability) would tend not to be represented in the sample data.

strained to go through the origin, $(0,0)=(\log_K 1,\log_K b_1)$. This constrained, least squares regression line can be proven to be a very good fit in the sense that the R^2 (coefficient of determination) associated with the model approaches 1 as $n\to\infty$. (The fact that the regression line is *not* a good fit for the bottom of the distribution of base values is addressed below.) The following calculations show that the slope β of this line is asymptotically equal to $(\Sigma_{j=1}^{K}\log_K a_j)/K$.

Let μ_j be the mean and σ_j^2 the variance of the logarithms of the K^j base values in B_j. For example, $\mu_0=\sigma_0^2=0$, $\mu_1=(\Sigma_{j=1}^{K}\log_K a_j)/K$, and $\sigma_1^2=[\Sigma_{j=1}^{K}(\log_K a_j-\mu_1)^2]/K$. Write $x_n\sim y_n$ whenever $\lim_{n\to\infty}x_n/y_n=1$. The following required intermediate results are stated without proof.

(R1) The equations $\mu_j=j\mu_1$ and $\sigma_j^2=j\sigma_1^2$ hold for $j=0,1,\ldots,n$.

(R2) The least squares line of $\log_K b_r$ regressed onto $\log_K r$ and constrained to pass through the origin has the slope $\beta=(\Sigma_{r=1}^{N_n}\log_K b_r\log_K r)/\Sigma_{r=1}^{N_n}\log_K^2 r$.

(R3) (i) $\Sigma_{j=1}^{n}\log_K j\sim n\log_K n$ and (ii) $\Sigma_{j=1}^{n}\log_K^2 j\sim n\log_K^2 n$.

(R4) $\Sigma_{j=1}^{n}j^\alpha K^j\sim n^\alpha K^{n+1}/(K-1)$ for $\alpha\geqslant 0$.

(R5) (i) $\Sigma_{r=1}^{N_n}|\log_K b_r|\sim|\mu_1|nK^{n+1}/(K-1)$ and (ii) $\Sigma_{r=1}^{N_n}\log_K^2 b_r\sim\mu_1^2 n^2K^{n+1}/(K-1)$.

(R1) through (R4) are straightforward to prove, and (R5) follows from (R1) and (R4).

It is convenient to use $|\beta|$ in our calculations, and we can obtain an asymptotic estimate of it by giving an upper and lower bound that are asymptotically equivalent. The key inequalities are

$$\frac{1}{N_n}\sum_{r=1}^{N_n}|\log_K b_r|\sum_{r=1}^{N_n}\log_K r\leqslant\sum_{r=1}^{N_n}|\log_K b_r|\log_K r\leqslant\left(\sum_{r=1}^{N_n}\log_K^2 b_r\right)^{1/2}\times\left(\sum_{r=1}^{N_n}\log_K^2 r\right)^{1/2},\qquad(1)$$

where the upper bound follows from the Cauchy-Schwartz inequality and the lower follows from Chebychev's monotonic inequality [13]. The latter is applicable here because $|\log_K b_r|$ and $\log_K r$ are both monotonically increasing in r.

After dividing through by $\Sigma_{r=1}^{N_n}\log_K^2 r$ in (1), the middle expression in the chain of inequalities is just $|\beta|$. Then, using the asymptotic relations in (R3)–(R5), routine calculations show that both the upper and lower bounds of $|\beta|$ are $\sim|\mu_1|$. Consequently, for $n\to\infty$ we must have $|\beta|\sim|\mu_1|$, as well. The inequalities of (1) are also the crux of similar manipulations that prove R^2 approaches 1 as a limit as $n\to\infty$.

Empirical studies of natural language showing plots of word frequencies f_r against their rank r in log-log coordinates almost always have slopes very near -1, as a glance through Zipf's work will show [14]. The value -1 has some significance for the slope estimate $\beta\sim(\Sigma_{j=1}^{K}\log_K a_j)/K$ of this random text model through its connection to the classical problem in probability theory concerned with the random division of the unit interval. Consider the letter probabilities $a_1,a_2,\ldots,a_{K+1}$ as a random division of the interval [0,1]. We represent this in the standard way. Let $X_1,X_2,\ldots,X_K$ be K independently and identically distributed random variables defined on [0,1]. Write the corresponding order statistics as $X_{(1)}\leqslant X_{(2)}\cdots\leqslant X_{(K)}$. The interval [0,1] is subdivided into $K+1$ mutually exclusive and exhaustive segments by the spacings D_j defined as $D_1=X_{(1)},D_2=X_{(2)}-X_{(1)},\ldots,D_j=X_{(j)}-X_{(j-1)},\ldots,D_{K+1}=1-X_{(K)}$. Think of the letter probabilities a_j as realized values of the spacings D_j.

Darling [15] has shown that for uniformly distributed X_j and $K\to\infty$, $\Sigma_{j=1}^{K+1}\ln D_j$ has the asymptotic mean $-K\ln K$. Blumenthal [16] has extended this result to a broader family of nonuniform distributions. Dropping any single term, say $\ln D_{K+1}$, from the sum has no effect on this asymptotic result, and replacing ln with $\log_K$ we see that $(\Sigma_{j=1}^{K}\log_K D_j)/K$ must have the asymptotic mean $(-K\log_K K)/K=-1$. The values of a_j used in the two figures of this paper were obtained by sampling the X_j from the uniform distribution on [0,1] and computing the associated spacings D_j. For this sample, $(\Sigma_{j=1}^{10}\log_{10}D_j)/10=-1.33$, which is the slope of the asymptotic regression line graphed in Fig. 2.

IV. THE ROLE OF SIZE-BIASED SAMPLING AND CONCLUSION

It is clear from visual inspection of Fig. 2 that the asymptotic regression line with slope μ_1 does not fit the smallest base values well, and this will not change as $n\to\infty$. However, for the toy model defined here, the observed word frequencies of an experiment would be those of a sample drawn from a multinomial parent distribution with N_n categories

(words) with probabilities proportional to the base values plotted in Fig. 2. The smallest probability words would tend not to occur unless the sample size was very large. Consequently, an approximate inverse power law for the observed frequencies would be seen. *The true log-normal structure of the parent distribution would be obscured or distorted for most practical sample sizes*. In fact, Li [9] carries out a simulation that demonstrates this phenomenon. One of his simulations illustrates how Zipf's law can be generated in a sample of random text based on the independent, unequal letter probability model with a maximum word length (see his Fig. 2). However, he did not recognize that the underlying parent distribution is actually approximately log-normal—not a power law. With a sufficiently large sample size, his sample power law would have to break down as more and more of the low-probability random word sequences appear.

The significance of *size-biased* sampling in relation to this problem and many others should be emphasized. To a remarkable extent, what we observe are extreme, upper tail events. We see the brightest stars, not the dimmest; we record seismic occurrences only when they are sufficiently large to be detected by our instruments; we collect and report statistics on the heights of the tallest buildings, the areas of the world's largest lakes, and so on. In each of these cases, there is an obvious *visibility bias* in observing events that makes it difficult to assess the nature of the underlying parent distribution from which the sample was drawn. Statisticians [17] have remarked that this pervasive issue needs greater recognition and have developed analytic methods for attacking it.

We conclude with very brief pointers to related topics. Empirical, approximately log-normal distributions characterized by upper tails with inverse power laws have been noted numerous times in many phenomena, as discussed, for example, by Montroll and Shlesinger [18,19] and Perline [20]. Empirical and theoretical arguments in support of log-normal models of word frequency distributions are given by Herdan [21]. (The similarity between our construction of U_n and the derivation of a hybrid lognormal-Pareto distribution in [18,19] should be noted.) The fractal character of the set U_n (as $n \to \infty$) seems intuitively evident, and the expression $(\sum_{j=1}^{K} \log_K a_j)/K$ pops up as what Evertsz and Mandelbrot [22] call the ''most probable Hölder exponent'' of a multifractal. Finally, Gut [8] has shown the importance of Anscombe's generalization of the central limit theorem for more realistic models of random walks, and we suggest that it can be extended in many ways for applications to a great variety of phenomena.

ACKNOWLEDGMENTS

I am deeply grateful to Ron Perline, Department of Mathematics and Computer Science, Drexel University, for encouraging the composition of this paper.

[1] B. Mandelbrot, *Fractals: Form, Chance, and Dimension* (Freeman, New York, 1977).
[2] B. Mandelbrot, in *Information Networks, the Brooklyn Polytechnic Institute Symposium*, edited by E. Weber (Interscience, New York, 1955), pp. 205–221.
[3] B. Mandelbrot, in *Proceedings of the Symposium on Applications of Information Theory*, edited by W. Jackson (Butterworth, London, 1953), pp. 486–499.
[4] B. Mandelbrot, Trans. IRE **3**, 124 (1954).
[5] B. Mandelbrot, Proc. Symp. Appl. Math. **12**, 190 (1961).
[6] G. Miller, Am. J. Psychol. **70**, 311 (1957).
[7] G. Miller and N. Chomsky, in *Handbook of Mathematical Psychology II*, edited by R. Luce, R. Bush, and E. Galanter (Wiley, New York, 1963), pp. 419–491.
[8] A. Gut, *Stopped Random Walks: Limit Theorems and Applications* (Springer, New York, 1987).
[9] W. Li, IEEE Trans. Inf. Theory **38**, 1842 (1992).
[10] G. Miller, in the preface to G. Zipf, *The Psycho-Biology of Language: An Introduction to Dynamic Psychology* (MIT Press, Cambridge, MA, 1965).
[11] R. Mantegna, S. Buldyuv, A. Goldberger, S. Havlin, C.-K. Peng, M. Simons, and H. Stanley, Phys. Rev. Lett. **73**, 3169 (1994).
[12] P. Niyogi and R. Berwick, CMP-LG@XXX.LANL.GOV, Document No. 9503012, 1995; see also Phys. Rev. Lett. **76**, 1976 (1996) for several Comments and a Reply.
[13] R. Graham, D. Knuth, and O. Patashnik, *Concrete Mathematics* (Addison-Wesley, Reading, MA, 1994).
[14] G. Zipf, *Human Behavior and the Principle of Least Effort* (Addison-Wesley, Cambridge, MA, 1949).
[15] D. Darling, Ann. Math. Stat. **23**, 450 (1952).
[16] S. Blumenthal, SIAM J. Appl. Math. **16**, 1184 (1968).
[17] G. P. Patil and C. Radhakrishna Rao, in *Applications of Statistics*, edited by P. Krishnaiah (North-Holland, Amsterdam, 1977), pp. 383–405.
[18] E. Montroll and M. Shlesinger, Proc. Natl. Acad. Sci. U.S.A. **79**, 3380 (1982).
[19] E. Montroll and M. Shlesinger, J. Stat. Phys. **32**, 209 (1983).
[20] R. Perline, Ph.D. dissertation, University of Chicago, 1982.
[21] G. Herdan, *Type-Token Mathematics: A Textbook of Mathematical Linguistics* (Mouton, 'S-Gravenhage, 1961).
[22] C. Evertsz and B. Mandelbrot, in *Chaos and Fractals: New Frontiers of Science*, edited by H.-O. Peitgen, H. Jürgens, and D. Saupe (Springer, New York, 1992).