

## MOMENTS OF PARTICLE SIZE DISTRIBUTIONS UNDER SEQUENTIAL BREAKAGE WITH APPLICATIONS TO SPECIES ABUNDANCE

ANDREW F. SIEGEL\* AND  
GEORGE SUGIHARA,\*\* *Princeton University*

### Abstract

The sequential broken stick model has appeared in numerous contexts, including biology, physics, engineering and geology. Kolmogorov showed that under appropriate conditions, sequential breakage processes often yield a lognormal distribution of particle sizes. Of particular interest to ecologists is the observed variance of the logarithms of the sizes, which characterizes the evenness of an assemblage of species. We derive the first two moments for the logarithms of the sizes in terms of the underlying distribution used to determine the successive breakages. In particular, for a process yielding  $n$  pieces, the expected sample variance behaves asymptotically as  $\log(n)$ . These results also yield a new identity for moments of path lengths in random binary trees.

BROKEN STICK MODEL; LOGNORMAL DISTRIBUTION; ECOLOGY; NICHE HIERARCHY MODEL; RANDOM BINARY TREES

### 1. Introduction

When a distributional pattern is generated by an unobservable process, insight into the mechanism of genesis can sometimes be inferred with the aid of a suitable model. Our focus here will be on lognormal distributions, with the aim of studying a class of processes that give rise to empirical families of lognormal curves. The importance of this pattern rests largely on its ubiquity and the broad spectrum of contexts in which it appears.

In engineering and geology lognormal distributions have been used to describe quantities produced by natural and mechanical processes, such as frequencies of particle sizes and life lengths of materials and machines before failure (Epstein (1947); Herdan (1953)). In economics and sociology these distributions have been fitted to data on incomes (Gibrat (1931); Davies (1946); Kapteyn (1916)) and numbers of people per occupation (Clark (1964)). Applications in biology include characterizing data on body sizes (Yuan (1933); Camp

---

Received 26 November 1981; revision received 24 March 1982.

\* Postal address: Department of Statistics, Princeton University, Princeton, NJ 08544, U.S.A.

\*\* Postal address: Department of Biology, Princeton University, Princeton, NJ 08544, U.S.A.

pieces. To go from stage  $n$  to stage  $n + 1$ , first choose a piece at random (uniformly without regard to size, so that each piece has probability  $1/n$  of being chosen) and then break it in two according to a proportion chosen independently from  $F$ . Because the piece to be broken is chosen randomly in this way, we lose no generality by requiring  $F$  to be symmetric, in order to simplify the mathematical treatment.

If  $W$  is an observation from  $F$ , define moments  $\mu_m = E[\log(W)]^m$  and  $\nu = E[\log(W)\log(1 - W)]$ . We assume that  $\mu_1$ ,  $\mu_2$ , and  $\nu$  are finite. At stage  $n$ , let the pieces have sizes  $X_{1n}, \dots, X_{nn}$  in some random ordering so that these are exchangeable (but not independent) random variables. The logarithms of the sizes,  $U_{in} = \log(X_{in})$ , are of interest in many applications, as are the sample mean  $\bar{U}_n = (\sum_{i=1}^n U_{in})/n$  and the sample variance  $S_n^2 = (\sum_{i=1}^n (U_{in} - \bar{U}_n)^2)/(n - 1)$ . Note that each  $X_{in}$  is, marginally, the product of a random number of independent proportions chosen from  $F$ , and each  $U_{in}$  is the sum of the corresponding logarithms.

*Theorem 1.* The mean and variance of the logarithm,  $U_{in}$ , of the size of a single random piece at stage  $n$  are

$$(1) \quad E(U_{in}) = 2\mu_1 \sum_{k=2}^n \frac{1}{k}$$

$$(2) \quad \text{Var}(U_{in}) = 2\mu_2 \sum_{k=2}^n \frac{1}{k} - 4\mu_1^2 \sum_{k=2}^n \frac{1}{k^2}.$$

*Proof.* By exchangeability, it will suffice to compute  $E(U_{1n})$ . Condition according to whether this piece was or was not involved in the most recent breakage, events with probabilities  $2/n$  and  $(n - 2)/n$  respectively. Because the conditional distributions of the log lengths are  $U_{1,n-1} + \log(W)$  and  $U_{1,n-1}$  respectively, where  $W$  has distribution  $F$  and is independent of  $U_{1,n-1}$ , we obtain the recurrence

$$(3) \quad E(U_{1n}) = E(U_{1,n-1}) + \frac{2}{n} \mu_1$$

whose solution with initial condition  $E(U_{1,1}) = 0$  is (1). For the variance,  $\text{Var}(U_{1n}) = E(U_{1n}^2) - [E(U_{1n})]^2$ , condition as before for each term of this difference, to establish

$$(4) \quad E(U_{1n}^2) = E(U_{1,n-1}^2) + \frac{4}{n} \mu_1 E(U_{1,n-1}) + \frac{2}{n} \mu_2$$

and

$$(5) \quad [E(U_{1n})]^2 = [E(U_{1,n-1})]^2 + \frac{4}{n} \mu_1 E(U_{1,n-1}) + \frac{4}{n^2} \mu_1^2.$$

(1938); Cramér (1946)) and species abundance (Preston (1962); Aitchison and Brown (1968); Patrick (1968); Bulmer (1974); May (1975); Pielou (1975); Sugihara (1980)). Brown and Sanders (1981) have shown that the lognormal distribution arises in a large variety of classification procedures.

Some of these physical and biological contexts, where the natural method of genesis involves repeated breakages, produce special families of lognormal distributions. Kolmogorov (1941) has shown that when the frequency of breakage is independent of the size of each particle, the asymptotic distribution of particle sizes should tend to be lognormal. Of interest here is that the mean and variance will depend on the number of breakages applied. These parameters will therefore be coupled to the number of particles generated, producing families of lognormal distributions in which the variance of log sizes will increase with the application of additional breakage events.

When such coupling is actually observed, it may suggest sequential breakage as a possible method of genesis. This argument was used, for example, in a recent model of species abundance (Sugihara (1980)) in which sequential binary breakages of niche space was proposed to explain a particular coupling of parameters observed in the lognormal species abundance distribution (Preston (1962)). Investigating sequential breakage mechanisms may be useful not only for clarifying the predictions of this species abundance model, but also in general, for understanding the genesis of empirical families of lognormal curves having coupled parameters.

It should be noted that knowledge of some moments does not entirely specify a distribution. In fact, there exist other distributions with the exact same sequence of moments as a given lognormal (Feller (1971), p. 227). Nonetheless, even the first two moments clearly provide useful partial information about an otherwise unknown probability distribution.

Monte Carlo estimates have been available (Sugihara (1980)) for the relationship between the expected variance and the number of particles (or species) in some special cases of repeated binary breakage. Our aim here is to provide exact and asymptotic formulae for this relationship. In addition to simplifying computation, these results will also yield further insight into the nature of breakage processes. In particular, we shall show how the expected mean and variance of the logarithmic sizes can be expressed in terms of auxiliary moments of the distribution of breakage applied at each step. Underlying these results is a somewhat surprising identity involving cross-moments of path lengths in random binary trees.

## **2. Expected means and variances**

Begin with a stick of unit length and a breakage distribution  $F$  that is symmetric on  $(0, 1)$ . This is stage 1; at stage  $n$  the stick will be broken into  $n$

Subtracting (5) from (4) and simplifying, we obtain

$$(6) \quad \text{Var}(U_{1n}) = \text{Var}(U_{1,n-1}) + \frac{2}{n} \mu_2 - \frac{4}{n^2} \mu_1^2$$

whose solution with initial condition  $\text{Var}(U_{11}) = 0$  is (2).

In a real situation, care must be taken with regard to the variance term. Due to the dependence among  $U_{1n}, \dots, U_{nn}$ , the sample variance  $S_n^2$  should not be compared to (2), which is the expected sample variance for an *independent* sample with the same marginals. Instead, the following quantities should be used.

*Theorem 2.* The expected sample mean and variance at stage  $n$  are

$$(7) \quad E(\bar{U}_n) = 2\mu_1 \sum_{k=2}^n \frac{1}{k}$$

$$(8) \quad E(S_n^2) = \left\{ 2 \left( 1 + \frac{2}{n-1} \right) \sum_{k=2}^n \frac{1}{k} - 2 \right\} \mu_2 - \nu.$$

*Proof.* Equation (7) follows by linearity from (1). For (8), expand and use exchangeability to obtain

$$(9) \quad E(S_n^2) = E(U_{1n}^2 - U_{1n}U_{2n}).$$

Next, condition according to the four possibilities of involvement of  $U_{1n}$  and  $U_{2n}$  in the most recent breakage (neither,  $U_{1n}$  only,  $U_{2n}$  only, or both). For example, with probability  $2/(n(n-1))$  they were both involved in the most recent breakage, and  $U_{1n}^2 - U_{1n}U_{2n}$  has the conditional distribution

$$(10) \quad [U_{1,n-1} + \log(W)]^2 - \{[U_{1,n-1} + \log(W)][U_{1,n-1} + \log(1-W)]\}$$

where  $W$  is a random variable with distribution  $F$  and is independent of  $(U_{1n}, U_{2n})$ . Combining this with the results from the other three possibilities, then simplifying, we find with some effort that

$$(11) \quad E(S_n^2) = \frac{(n-2)(n+1)}{n(n-1)} E(S_{n-1}^2) + \frac{2\mu_2}{n} - \frac{2\nu}{n(n-1)}.$$

With some patience, it can be shown by induction that (8) is the solution to the recurrence (11) with initial condition  $E(S_2^2) = \mu_2 - \nu$ .

The means (1) and (7) are identical because the expectation operator is linear even under dependence. However, from (2) and (8) we see that the lack of independence has modified the variance. These differences can be studied in detail by examining an asymptotic expansion of each expression.

*Theorem 3.*

$$(12) \quad E(U_m) = E(\bar{U}_n) = \mu_1 \left\{ 2 \log(n) - 0.8456 + \frac{1}{n} - \frac{1}{6n^2} + \frac{1}{60n^4} + O\left(\frac{1}{n^6}\right) \right\}$$

$$(13) \quad \begin{aligned} \text{Var}(U_m) = & 2\mu_2 \log(n) - (2.5797\mu_1^2 + 0.8456\mu_2) + \frac{4\mu_1^2 + \mu_2}{n} \\ & - \frac{12\mu_1^2 + \mu_2}{6n^2} + \frac{2\mu_1^2}{3n^3} + \frac{\mu_2}{60n^4} - \frac{2\mu_1^2}{15n^5} + O\left(\frac{1}{n^6}\right) \end{aligned}$$

$$(14) \quad \begin{aligned} E(S_n^2) = & 2\mu_2 \log(n) - (\nu + 2.8456\mu_2) + 4\mu_2 \frac{\log(n)}{n-1} - \frac{0.6911\mu_2}{n} \\ & + \frac{0.1422\mu_2}{n^2} - \frac{0.02447\mu_2}{n^3} - \frac{0.007804\mu_2}{n^4} + \frac{0.008863\mu_2}{n^5} + O\left(\frac{1}{n^6}\right). \end{aligned}$$

*Proof.* These follow from two standard asymptotic expansions. Equation (12) depends on the expansion of the harmonic series (e.g. Knuth (1973), Vol. 1, p. 74):

$$(15) \quad \sum_{k=1}^n \frac{1}{k} = \log(n) + \gamma + \frac{1}{2n} - \frac{1}{12n^2} + \frac{1}{120n^4} + O\left(\frac{1}{n^6}\right)$$

where  $\gamma \cong 0.5772156649$  is Euler's constant. Equation (13) also uses (e.g. p. 61 of Hansen (1975))

$$(16) \quad \sum_{k=1}^n \frac{1}{k^2} = \frac{\pi^2}{6} - \frac{1}{n} + \frac{1}{2n^2} - \frac{1}{6n^3} + \frac{1}{30n^5} + O\left(\frac{1}{n^7}\right).$$

For (14), use (15) in (8) and multiply the non-logarithmic part by the expansion of  $2/(n-1)$  as a power series in  $1/n$ . Combine terms to find

$$(17) \quad \begin{aligned} E(S_n^2) = & \left\{ 2 \log(n) + (2\gamma - 4) + 4 \frac{\log(n)}{n-1} + \frac{4\gamma - 3}{n} + \frac{24\gamma - 13}{6n^2} \right. \\ & \left. + \frac{24\gamma - 14}{6n^3} + \frac{240\gamma - 139}{60n^4} + \frac{240\gamma - 138}{60n^5} \right\} \mu_2 - \nu + O\left(\frac{1}{n^6}\right) \end{aligned}$$

which evaluates to (14).

Although the leading terms in the variances (13) and (14) are identical, the differences in their second terms cannot be neglected even for moderately large  $n$  due to the slowly increasing behavior of  $\log(n)$ . For example, when  $F$  places mass  $1/2$  at  $1/4$  and at  $3/4$ , even with  $n$  as large as 50, we have  $\text{Var}(U_m) = 5.3$  while  $E(S_n^2) = 4.9$ .

### 3. An identity for random binary trees

Consider the class of random binary trees with  $n$  endpoints generated recursively by bifurcating an endpoint chosen uniformly at random from a tree with  $n - 1$  endpoints. These trees are responsible for part of the randomness of the sequential breakage model (the other component can be thought of as entering through the distribution  $F$ ). These trees are related to random binary search trees used in computer science (Knuth (1973), Vol. 3, p. 423–471). For a tree with  $n$  endpoints, let  $N_{1n}$  and  $N_{2n}$  denote the distances (in numbers of edges) from each of two randomly chosen endpoints to their nearest common ancestor, as illustrated in Figure 1. Although moments involving  $N_{1n}$  and  $N_{2n}$  generally increase with  $n$ , there is an expression for which this dependence cancels out.

*Theorem 4.* Regardless of the value of  $n$ ,

$$(18) \quad E(N_{1n} + N_{1n}N_{2n} - N_{1n}^2) = 1.$$

*Proof.* Proceed by induction on  $n$ , conditioning on the four events describing which of  $N_{1n}$  and  $N_{2n}$  were involved in the most recent bifurcation. This is similar to the proof of Theorem 2.

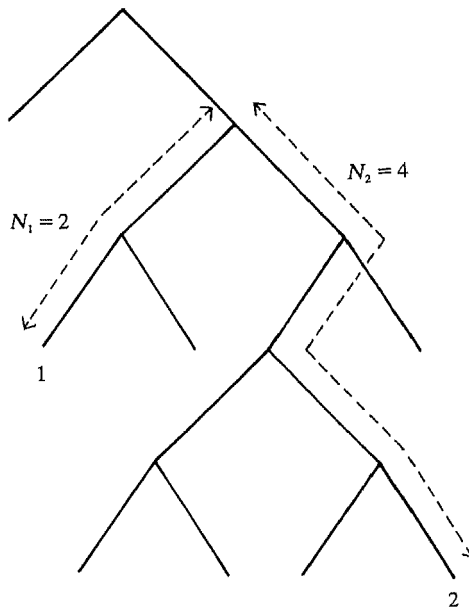


Figure 1. A random binary tree with  $n = 8$  endpoints. Distances  $N_1 = N_{1,8}$  and  $N_2 = N_{2,8}$  from two random endpoints to their nearest common ancestor are indicated

### Acknowledgements

This work was supported in part by U.S. Army Research Office Grant Number DAAG29-79-C-0205 and a Princeton University Prize Fellowship. We are grateful to Andrew Odlyzko, Stephen L. Teig, Vincent DellaPietra, and Clifford Hurvich for helpful conversations.

### References

- AITCHISON, J. AND BROWN, J. A. C. (1968) *The Lognormal Distribution, with Special Reference to its Uses in Economics*, 2nd edn. Cambridge University Press, London.
- BROWN, G. AND SANDERS, J. W. (1981) Lognormal genesis. *J. Appl. Prob.* **18**, 542-547.
- BULMER, M. G. (1974) On fitting the Poisson lognormal distribution to species-abundance data. *Biometrics* **30**, 101-110.
- CAMP, B. H. (1938) Notes on the distribution of the geometric mean. *Ann. Math. Statist.* **9**, 221-226.
- CLARK, P. J. (1964) On the number of individuals per occupation in a human society. *Ecology* **45**, 367-372.
- CRAMÉR, H. (1946) *Mathematical Methods of Statistics*. Princeton Mathematical Series **9**, Princeton University Press.
- DAVIES, G. R. (1946) Pricing and price levels. *Econometrica* **14**, 219-226.
- EPSTEIN, B. (1947) The mathematical description of certain breakage mechanisms leading to the logarithmico-normal distribution. *J. Franklin Inst.* **224**, 471-477.
- FELLER, W. (1971) *An Introduction to Probability Theory and its Applications*, Vol. 2, 2nd edn. Wiley, New York.
- GIBRAT, R. (1931) *Les inégalités économiques*. Librairie du Recueil Sirey, Paris.
- HANSEN, E. R. (1975) *A Table of Series and Products*. Prentice-Hall, Englewood Cliffs, NJ.
- HERDAN, G. (1953) *Small Particle Statistics*. Elsevier, Amsterdam.
- KAPTEYN, J. C. (1916) Skew frequency curves in biology and statistics. *Rec. Trav. Botaniques Néerlandais* **13**, 105-158.
- KNUTH, D. E. (1973) *The Art of Computer Programming* Vols. 1, 3. Addison-Wesley, Reading, MA.
- KOLMOGOROV, A. N. (1941) Über das logarithmisch normale Verteilungsgesetz der Dimensionen der Teilchen bei Zerstückelung. *C.R. Acad. Sci. U.R.S.S.* **31**, 99-101.
- MAY, R. M. (1975) Patterns of species abundance and diversity. In *Ecology and Evolution of Communities*, ed. M. L. Cody and J. M. Diamond, Harvard University Press, Cambridge, MA, 81-120.
- PATRICK, R. (1968) The structure of diatom communities in similar ecological conditions. *Amer. Natur.* **102**, 173-183.
- PIELOU, E. C. (1975) *Ecological Diversity*. Wiley, New York.
- PRESTON, F. W. (1962) The canonical distribution of commonness and rarity: Part I. *Ecology* **43**, 185-215.
- SUGIHARA, G. (1980) Minimal community structure: an explanation of species abundance patterns. *Amer. Natur.* **116**, 770-787.
- YUAN, P. T. (1933) On the logarithmic frequency distribution and the semilogarithmic correlation surface. *Ann. Math. Statist.* **4**, 30-74.