Ecology (1977) **58**: pp. 693–696

ON FITTING THE TRUNCATED LOGNORMAL DISTRIBUTION TO SPECIES-ABUNDANCE DATA USING MAXIMUM LIKELIHOOD ESTIMATION[1]

JOHN SLOCOMB
Department of Biology and Center for Environmental Studies

BARBARA STAUFFER
Department of Statistics

KENNETH L. DICKSON
Department of Biology and Center for Environmental Studies,
Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061 USA

Abstract. The truncated lognormal distribution can be used to graduate certain species-abundance data, provided that estimates of the location and scale parameters are obtained. A computer program has been written which groups the data on a $\log_2$ scale and numerically solves the maximum likelihood equations for this type of distribution. Results show that the estimates obtained by this method compare well with those of Hald and Cohen. Examples are presented using the diatom data of Hohn and Hellerman, and it is shown that a better fit is obtained by using the entire data set instead of selectively disregarding the most abundant tail intervals. Other published techniques for this type of analysis are also discussed.

Key words: Computer methods; lognormal distribution; maximum likelihood estimation; numerical analysis; species abundance.

INTRODUCTION

Since Preston (1948) first used the lognormal distribution to graduate species-abundance data, the practice of fitting this distribution to various types of ecological data has become important as one means of describing community structure. Provided that the sampling interval is complete, that is, includes the complete lognormal "universe," no difficulty arises in the estimation of the location and scale parameters. Unfortunately, it is rare in ecological work for an entire universe to be collected, and, thus, the distribution is nearly always truncated at the left side (Preston 1962), making estimation of the parameters more difficult.

For such truncated lognormal distributions, Hald (1949) published tables from which maximum likelihood estimates could be obtained. Cohen (1959, 1961) published simplified tables, and Patrick et al. (1954) gave an account of a graphical method which could be used for the same purpose. Gauch and Chase (1974) have proposed a variation of the parameters algorithm which can be used to get least squares estimates of the required parameters and which alleviates the need for cumbersome tables and graphs. Bulmer (1974) has written a program in ALGOL to calculate maximum likelihood estimates for compound Poisson lognormal distributions. In our investigations, we felt it desirable to obtain rapid and accurate solutions to the maximum likelihood equations for the lognormal distribution, and for this purpose, we have written a FORTRAN IV program. Copies of this program along with the appropriate documentation are available upon request. The results presented in this paper were obtained using an IBM 370-158 computer employing double precision arithmetic.

THE LOGNORMAL DISTRIBUTION

Species-abundance data are commonly presented in the form of a histogram with the number of species, $f(x)$, containing x members as ordinate ($x = 1,2,3\ldots$) and the number of individuals (usually grouped in some way) as abscissa. Presented in this manner, the $f(x)$ are thus frequencies of frequencies. Customarily, this frequency distribution represents species abundances from a relatively narrow taxonomic range which the investigator feels constitutes an important ecological entity. Attempts to obtain a representative sample from every taxon in the community would be impractical, and so it is often the case that the sample range is restricted by taxonomic rank (Pielou 1969).

In an attempt to find a mathematical expression to summarize data of this type, Fisher et al. (1943) proposed the logarithmic series as an interpretation of the observed frequency distribution. Subsequently, various distributions have been proposed to describe species-abundance data, but Preston (1948) was the first to test the idea that the species-abundance distribution might well be described by theoretical lognormal frequencies. He found that for a sufficiently large aggregation of individuals of many species, the distribution often conformed to a normal law after the individuals were grouped on a logarithmic scale. That is (after Pielou 1969) it is assumed that the probability density function for the distribution is

[1] Manuscript received 15 September 1976; accepted 15 November 1976.

$$P(\log_2\lambda) = (\sigma\sqrt{2\pi})^{-1}\exp[-\log_2(\lambda/m)^2/2\sigma^2]. \quad (1)$$

Letting $x = \log_2\lambda$ and $\alpha = \log_2 m$, Eq. 1 becomes

$$P(x) = (\sigma\sqrt{2\pi})^{-1}\exp[-(x-\alpha)^2/2\sigma^2],\ -\infty < x < \infty, \quad (2)$$

where σ is the logarithmic standard deviation, α is the position of the mode, and x is the position of an observed number of species. Preston's original method of grouping was to group the individuals into "octaves," i.e., intervals on a base 2 logarithmic scale, with boundaries $r = 1,2,4,8 \ldots$, so that the midpoints of the octaves were at $r = 1.5,3,6,12\ldots$ The matter of labeling the octaves is arbitrary; however, it is convenient to label the endpoints $x = 1$ through R where R is the number of octaves required to account for the sample distribution. Those species that fall on a group boundary are split equally between that octave and the next higher or lower octave. Plotting the observed distribution then becomes a matter of noting into which octave the observed $f(x)$ fall, splitting these $f(x)$ as needed.

As discussed above, Eq. 2 is nearly always truncated on the left, reflecting the fact that there are a number of rarer species that have escaped collection. Consequently, the total number of species in the universe, N, is unknown and must be estimated. The point of truncation of the lognormal curve is called the "veil line"·(Preston 1948) and is assumed to occur at a count of one individual. That is, the part of the normal curve lying to the left of one individual cannot be observed without further sampling. In handling the data, distances along the x axis are measured using the veil line as the y axis. Thus, for example, the first x value of 0.5 corresponds to the midpoint of the octave containing those species represented by one or two individuals. The corresponding $f(x)$ value is the sum of one half the singleton species plus one half the species with two individuals. The parameter α is the distance from the veil line to the mode.

The effects of truncation on Eq. 2 are to replace the constant, $(\sigma\sqrt{2\pi})^{-1}$ by a new constant, A, and to restrict the domain of the function to the positive x axis, so that the probability density function becomes

$$P(x) = A\exp[-(x-\alpha)^2/2\sigma^2],\ x > 0, \quad (3)$$

where α and σ are given above and x is the distance from the origin (point of truncation). The constant A is a function of α and σ^2.

Maximum Likelihood Estimates

In the following, the x values are the distances of the appropriate octaves from the origin, with the first x value being 0.5. The $f(x)$ values are the species frequencies for the appropriate octaves, with the end point frequency count split as previously described. The value n is the total number of species observed less one half of the number of singletons (i.e., species represented by one individual), so that $n = \Sigma f(x)$. The probability density function of the truncated normal distribution is given by Eq. 3. Since the total area under a probability density function is always unity, we can write

$$A\int_0^\infty \exp[-(x-\alpha)^2/2\sigma^2]dx = 1,$$

and thus

$$A^{-1} = \int_0^\infty \exp[-(x-\alpha)^2/2\sigma^2]dx = \sigma\int_{-\alpha/\sigma}^\infty \exp(-y^2/2)dy.$$

The logarithm of the likelihood function is then given by

$$L = n\log_e A - (½\sigma^2)\,\Sigma f(x)(x-\alpha)^2.$$

For ease of computation, let $b = \alpha/\sigma$. Then maximizing L with respect to b and σ will also maximize L with respect to α and σ. Since σ is the standard deviation of the untruncated normal distribution, it must be non-negative, and, therefore, a non-negativity constraint was imposed on its estimate. Thus, the actual function maximized was

$$L = n\log_e A - (½\sigma^2)\,\Sigma f(x)(x - b\sigma)^2 \quad (4)$$

where

$$A^{-1} = \sigma\int_{-b}^\infty \exp(-y^2/2)dy \quad (5)$$

subject to the constraint $\sigma \geqslant 0$.

Finding the first partial derivatives of L with respect to b and σ and setting these derivatives equal to zero gives as the likelihood equations

$$\Sigma f(x)x - bn\sigma - n\sigma^2 A\exp(-b^2/2) = 0 \quad (6)$$

and

$$\Sigma f(x)x^2 - n\sigma^2 - b\sigma\,\Sigma f(x)x = 0. \quad (7)$$

These equations were solved on the computer, using a modified Newton-Raphson procedure (Stark 1970, Section 4.5) with the above constraint imposed on σ. The solutions are the same as those obtained using Hald's and Cohen's tables.

The program output includes several items in addition to the estimates of α and σ. The raw data and the data grouped into octaves are both listed. Estimates are given for the height of the mode and the number of species in the population. The expected frequency in each octave (using the estimates as the population parameters) is listed, and a chi-square statistic for goodness of fit is calculated. Finally, the estimated asymptotic variance-covariance matrix for the estimates of α and σ is printed.

Discussion

As examples we have analyzed the diatom collections published by Hohn and Hellerman (1963) which were graduated by theoretical lognormal frequencies. It should be noted that Hohn and Hellerman used only

TABLE 1. Summary of estimates and goodness-of-fit statistics obtained using the full data set (I) and only the first ten octaves (II)

		Estimates							
		α		σ		N		χ^2 (df)	
Data sets	Substrate	I	II	I	II	I	II	I	II
La Vase River	Glass slide	0.62	1.6	4.32	3.45	221.0	176.5	4.53 (10)	6.18 (8)
October 1957	Styrofoam®	−0.63	1.8	4.93	3.14	380.9	226.4	15.64 (11)	49.55 (8)
Potomac River	Glass slide	1.93	2.2	3.51	3.26	144.5	135.2	2.66 (9)	5.76 (8)
August 1957	Styrofoam®	1.64	2.1	3.70	3.20	148.4	132.1	2.36 (10)	2.59 (8)
Ridley Creek	Glass slide	1.46	2.0	3.52	3.06	257.4	226.1	3.43 (10)	4.58 (8)
May 1957	Styrofoam®	1.97	2.1	3.27	3.09	225.2	215.7	7.60 (9)	8.59 (8)
Ridley Creek	Glass slide	−5.88	1.4	6.70	3.20	560.5	149.3	25.36 (14)	63.14 (8)
March 1958	Styrofoam®	−0.87	1.6	5.00	3.26	416.7	248.7	11.48 (13)	28.05 (8)

the first 10 octaves of the sample distribution in order that the mode be more rapidly exposed, even though, in all eight data sets, the sample distribution extended beyond the tenth octave. As a consequence, their estimates of α and σ differ from ours, as was expected. Hohn and Hellerman suggested that the most abundant tail intervals may be disregarded in analyzing species-abundance data. Although this technique alleviates the need for further sampling (and hence further analysis), the fit so obtained (as determined by a chi-square statistic) is not as good as that obtained using the full data set. In three of the data sets considered, the modal octaves were somewhat arbitrarily exposed, although analysis of all of the observed octaves showed that, in each case, the mode was still behind the veil line. Table 1 shows the relationship between the two methods of analysis.

It may also be seen from the table that the estimates of N (total number of species in the population) obtained using the two methods are quite different. This discrepancy is not surprising, given the two types of analysis. More importantly, Pielou (1975) points out that currently available estimates of N are not satisfactory; this observation is certainly supported by the results presented here and in Bulmer (1974). This lack of confidence in estimates of N is unfortunate because reliable knowledge of total species size in a population is more interesting as a descriptive statistic than either the location or scale parameter estimates of the distribution. In addition, many applications of the Shannon-Weaver diversity index (H) require knowledge of N (Basharin 1959, Pielou 1969, and Hutcheson 1970), and recently Longuet-Higgins (1971) and Bulmer (1974) have derived estimates of H based on the estimates of N and σ^2. It is clear that a cautious attitude is warranted in these instances, especially when estimation of H is of interest.

At least two other computer routines are available for fitting lognormal curves to species-abundance data. Gauch and Chase (1974) use a curve fitting approach and obtain least squares estimates of the lognormal parameters. This technique assumes a model composed of a lognormal-type term plus random error and is thus a regression method. The estimates obtained are not generally the same as maximum likelihood estimates.

Use of the lognormal distribution involves the assumption that each species is represented in the collection by its expected number of individuals. Assuming instead that the number of individuals representing a species is a Poisson random variable leads to the use of the compound Poisson lognormal distribution. Bulmer (1974) has written an ALGOL program to find the maximum likelihood estimates of the parameters of this latter distribution, thus alleviating the computational difficulties of finding these estimates. According to Bliss (1966) and Pielou (1969, 1975), however, the lognormal distribution is adequate for describing species-abundance (discrete) data; in addition, the ALGOL program may be limited in its use in the United States.

The computer program described in this paper provides a rapid and convenient method of fitting the lognormal distribution to various types of species-abundance data. Moreover, the program is designed to sort through large amounts of data and group it on a $\log_2$ scale prior to the initiation of analysis. Although other methods of grouping the data are entirely feasible, we chose Preston's octave method since we felt it provided a much clearer way of graphically observing species abundance (it results in a doubling of the number of individuals in each of the octaves). Grouping on a natural logarithmic scale or any other arbitrarily chosen base, though mathematically acceptable, is less easily visualized in most cases. Furthermore, acceptable ecological hypotheses to account for the distribution of species abundance still remain to be advanced. The fitting of a lognormal distribution is, thus, entirely of heuristic benefit and should be made easily interpretable.

The development of the computer program was in response to the growing need for rapid data analysis techniques that could be interfaced with biological monitoring systems. Automated "early warning" devices (for example, see Almeida et al. 1972; Cairns et al. 1972, 1974) are currently being developed for monitoring pollution-related community stress in aquatic ecosystems and require rapid processing of biological (e.g., species-abundance) information.

Acknowledgments

We thank Peter Stauffer and Wolfgang Pelz for their advice and assistance in writing the computer program. We also express appreciation to Dr. Walter Pirie, Department of Statistics, VPI and SU for his suggestions during the progress of this work. It was supported by NSF/RANN/IT/GI-38357/PR/73/2.

Literature Cited

Almeida, S. P., D. R. Del Balzo, J. Cairns, Jr., K. L. Dickson, and G. R. Lanza. 1972. Holographic microscopy of diatoms. Trans. Kansas Acad. Sci. **74**:257–260.

Basharin, G. P. 1959. On a statistical estimate for the entropy of a sequence of independent random variables. Theory Probab. Its. Appl. **4**:333–336.

Bliss, C. I. 1966. An analysis of some insect trap records. Sankhya, Ser. A **28**:123–136.

Bulmer, M.G. 1974. On fitting the Poisson lognormal distribution to species-abundance data. Biometrics **30**:101–110.

Cairns, J., Jr., K. L. Dickson, G. R. Lanza, S. P. Almeida, and D. R. Del Balzo. 1972. Coherent optical spatial filtering of diatoms in water pollution monitoring. Arch. Mikrobiol. **83**:141–146.

Cairns, J., Jr., K. L. Dickson, J. Slocomb, S. P. Almeida, J. K. T. Eu, C. Y. C. Liu, and H. F. Smith. 1974. Microcosm pollution monitoring, p. 223–228. *In* D. D. Hemphill [ed.] Trace substances in environmental health-VIII. University of Missouri, Columbia, Missouri.

Cohen, A. C. 1959. Simplified estimators for the normal distribution when samples are singly censored or truncated. Technometrics **1**:217–237.

———. 1961. Tables for maximum likelihood estimates: singly truncated and singly censored samples. Technometrics **3**:535–541.

Fisher, R. A., A. S. Corbet, and C. B. Williams. 1943. The relation between the number of species and the number of individuals in a random sample of an animal population. J. Anim. Ecol. **12**:42–58.

Gauch, H. G., and G. B. Chase. 1974. Fitting the Gaussian curve to ecological data. Ecology **55**:1377–1381.

Hald, A. 1949. Maximum likelihood estimation of the parameters of a normal distribution which is truncated at a known point. Skandinavisk Aktuarietidskrift **32**:119–134.

Hohn, M. H., and J. Hellerman. 1963. The taxonomy and structure of diatom populations from three Eastern North American rivers using three sampling methods. Trans. Am. Microsc. Soc. **82**:250–329.

Hutcheson, K. 1970. A test for comparing diversities based on the Shannon formula. J. Theor. Biol. **29**:151–154.

Longuet-Higgins, M. S. 1971. On the Shannon-Weaver index of diversity, in relation to the distribution of species in bird censuses. Theor. Pop. Biol. **2**:271–289.

Patrick, R., M. H. Hohn, and J. H. Wallace. 1954. A new method for determining the pattern of the diatom flora. Notes Natl. Acad. Nat. Sci., Philadelphia. **259**:1–12.

Pielou, E. C. 1969. An introduction to mathematical ecology. John Wiley and Sons, New York. 286 p.

———. 1975. Ecological diversity. John Wiley and Sons, New York. 165 p.

Preston, F. W. 1948. The commonness and rarity of species. Ecology **29**:254–283.

——— 1962. The canonical distribution of commonness and rarity. Ecology **43**:185–215, 410–432.

Stark, P. A. 1970. Introduction to numerical methods. Macmillan, New York. 334 p.