Research Technical Completion Report
A-085-IDA

STATISTICS OF SPECIES ABUNDANCE DISTRIBUTIONS IN MONITORING AQUATIC COMMUNITIES

by

Brian Dennis
Assistant Professor

Department of Forest Resources,
College of Forestry, Wildlife and Range Sciences
University of Idaho

Idaho Water and Energy Resources Research Institute
University of Idaho
Moscow, ID 83843

December 1982

The research on which this report is based was financed in part by the U.S. Department of the Interior, as authorized by the Water Research and Development Act of 1978 (P.L. 95-467).

Research Technical Completion Report

A-085-IDA

STATISTICS OF SPECIES ABUNDANCE DISTRIBUTIONS IN MONITORING AQUATIC COMMUNITIES

by

Brian Dennis
Assistant Professor

Department of Forest Resources,
College of Forestry, Wildlife and Range Sciences
University of Idaho

Submitted to
Bureau of Reclamation
United States Department of the Interior
Washington, D.C. 20242

The work on which this report is based was supported in part by funds provided by the United States Department of the Interior as authorized under the Water Research and Development Act of 1978.

Idaho Water and Energy Resources Research Institute
University of Idaho
Moscow, ID 83843

December 1982

ACKNOWLEDGEMENTS

This research was funded by the United States Department of the Interior through a grant from the Idaho Water and Energy Resources Research Institute (Contract No. A-085-IDA).

TABLE OF CONTENTS

ABSTRACT

This report documents statistical procedures for using species abundance distributions for monitoring pollution impacts in aquatic biological communities. The species abundance distributions are also appropriate for use in other areas of applied ecology as well. Previous use of these tools by ecologists has relied on largely ad hoc data analysis procedures having little basis in statistical theory. This report is a synthesis of the statistical theory that applies to species abundance distributions. The report discusses appropriate statistical interpretation of the distributions, sampling distributions, types of abundance models such as the lognormal and the gamma, the canonical hypothesis of species abundance, and methods of testing hypotheses. It also identifies a number of important questions needing further research.

Species Abundance Distributions. The typical ecological data set consists of a list of S species and their abundances, Λ_1, Λ_2, . . ., Λ_S. Abundance is usually measured in number of individuals, or more infrequently, biomass. The relative abundances are given by Λ_1/T, Λ_2/T, . . ., Λ_S/T, where T is the total community abundance found by summing all the Λ_i values. Ecological diversity indices are used to detect changes in the pattern of relative abundance. Diversity indices normally contain two kinds of information about relative abundance patterns: (1) Species richness, or the number of species, positively affects most diversity indices. (2) Evenness, or the degree to which the relative abundances approach uniformity, is a factor measured by most diversity indices. Patil and Taillie (1982) have recently given a comprehensive account of the statistical theory behind these indices.

Aquatic communities sampled in pollution monitoring studies typically contain large numbers of species, particularly in invertebrate communities. Diatoms and other phytoplankton, zooplankton, and benthic invertebrates are frequently used in such studies; samples usually have large numbers of species. Ties are common among species abundances: many species will be represented by only one individual, many others by two, etc. Such data are traditionally tabulated in frequency form. Thus N_1 = number of species with one representative in the sample, N_2 = number of species with two representatives, and so on. This report is concerned with these species abundance data, rather than the relative abundance data.

The use of species abundance distributions is an alternative to using diversity indices. This distribution method involves fitting a traditional probability distribution to the data tabulated in species abundance form. The data are thereby summarized in compact form, and the parameter values of the distribution may be used to compare data from different environments. The parameters, in fact, often are related to species richness, evenness, and diversity, as will be seen. Fitting species abundance distributions works best in communities with larger numbers of species (over 20 or so) such as are encountered in phytoplankton, zooplankton, or benthic invertebrate studies.

Many assemblages of species display similar patterns of abundances (Fisher et al. 1943; Preston 1948; Williams 1964; Patrick 1968; Whittaker 1972; Kempton and Taylor 1974; May 1975; Pielou 1975). Simple probability distributions successfully describe species abundance data from wide varieties of communities. Two distributions in particular, the gamma and the lognormal, have recieved wide attention. Both distributions have similar shapes (unimodal and skewed on the right), the lognormal having a somewhat "heavier" right tail. The widespread empirical success of these distributions has spawned a number of hypotheses about their ecological causes (Whittaker 1972; May 1975; Dennis and Patil 1979; Sugihara 1980).

Ecologists have known for nearly 30 years that species abundance patterns respond to environmental changes (Patrick et al. 1954). In the lognormal and gamma models, certain parameters characterize the evenness and species richness of the communities (Taillie 1979), which are the components of ecological diversity. Typically, the numbers of rare species in disturbed communities decline, causing shifts in the

diversity or parameter values, and corresponding shifts in the shapes of the species abundance distributions. The shapes become long-tailed, reflecting the greater proportion of total community abundance residing in common species. These trends, though often documented, are not frequently utilized in biomonitoring studies.

One reason for this is a widespread misunderstanding, particularly among North American ecologists, of the statistical methods associated with using species abundance distributions. Two approaches to analyzing data are through Preston's (1948) lognormal model or through Fisher's (Fisher et al. 1943) gamma model. Not only do the distributions differ between Preston's and Fisher's approaches, but the statistical role of the probability distributions differs. North American ecologists generally follow Preston's interpretation of the distributions, which unfortunately has little basis in statistical theory. In Preston's approach, a species abundance distribution has no probabilistic content. Rather, the distribution is simply a function or curve to be fit (via ad hoc methods) to data. Contemporary writings continue to perpetuate this approach (May 1975; Preston 1980). In the approach of Fisher, by contrast, the role of a species abundance distribution is a probabilistic one. Random variables are explicitly defined, sampling variation is modeled, and parameter estimation methods have a sound statistical basis. This approach has been largely confined to the statistics literature in papers written by European statistical ecologists (Kempton 1975; Bulmer 1974; Engen 1978). With extensions to this statistical approach to be documented here, formal hypothesis testing can be conducted. Since such extensions and applications are important to a biomonitoring program, the statistical approach to species abundance distributions should be the preferred approach.

The statistical approach to species abundance distributions. Consider a given species with an average abundance of λ in a sample. It is reasonable to assume that the number of individuals of this species in a sample, X, has a Poisson distribution:

$$Pr[X = x \mid \lambda] = e^{-\lambda}\lambda^x/x!, \quad x = 0,1,2,... \tag{1}$$

A species abundance distribution takes the abundances, $\Lambda_1, \Lambda_2, \ldots \Lambda_s$, of all the species in the community to be independent, identically distributed random variables with some probability density function, $f(\lambda)$. The total number of species, S, is itself a random variable. Let

$$E[S] = s.$$

Then the expected total abundance in the community is

$$E[T] = E[\Lambda_1 + \Lambda_2 + \ldots + \Lambda_s] = E[S]E[\Lambda] = s\int_0^\infty \lambda f(\lambda)d\lambda, \tag{2}$$

where T is the total abundance in the community (a random variable also).

Observe that the expected number of species with abundances greater than λ is

$$sPr[\Lambda > \lambda] = s\int_\lambda^\infty f(u)du, \tag{3}$$

and that the expected abundance of a single species, given that its abundance is greater than λ, is

$$\int_\lambda^\infty uf(u)du \Big/ \int_\lambda^\infty f(u)du\ . \tag{4}$$

The expected total abundance of all the species with abundances greater than λ therefore has the form

$$[s\int_\lambda^\infty f(u)du][\int_\lambda^\infty uf(u)du \Big/ \int_\lambda^\infty f(u)du] = s\int_\lambda^\infty uf(u)du\ . \tag{5}$$

Because of (3) and (5), $sf(\lambda)$ is called the species curve and $s\lambda f(\lambda)$ is called the individuals curve.

In Preston's approach, $sf(\lambda)$ is a function to be fit through species abundance data, with the probabilistic content of $f(\lambda)$ not explicitly defined. Fisher's approach, outlined here, explicitly treats $f(\lambda)$ as a probability density function. Note that the fact that Fisher used a gamma distribution for $f(\lambda)$ whereas Preston used a lognormal distribution is incidental. Either probability distribution could serve in either approach.

The advantage to Fisher's approach is the opportunity to model sampling variation through traditional statistical methods. If the abundance of a given species has the probability density function $f(\lambda)$, then the number of individuals of this species in a sample has a "mixed" Poisson distribution (which follows from (1)):

$$Pr[X = x] = \int_0^\infty [e^{-\lambda}\lambda^x/x!]f(\lambda)d\ , \quad x = 0, 1, 2, \ldots \quad . \qquad (6)$$

Let N_x = the number of species with x representatives in the sample. The expected value of N_x, denoted m_x, is given by

$$E[N_x] = m_x = E[S]Pr[X=x] = s \int_0^\infty [e^{-\lambda}\lambda^x/x!]f(\lambda)d\lambda. \qquad (7)$$

Usually S is assumed to have a Poisson distribution, and the N_x, $x = 0, 1, 2, \ldots$, are assumed to be independent Poisson random variables with expected values m_x.

The gamma model. Fisher suggested using the gamma distribution as a form for $f(\lambda)$(Fisher et al. 1943):

$$f(\lambda) = \mu^k \lambda^{k-1} e^{-\mu\lambda}/\Gamma(k),\ 0<\lambda<\infty\ . \qquad (8)$$

Here μ and k are positive-valued parameters, and $\Gamma(\cdot)$ is the gamma function (Abramowitz and Stegun 1965).

The sampling distribution for the number of individuals of a species in a sample is a negative binomial distribution. This results from substituting (8) into (6):

$$\Pr[X = x] = \int_0^\infty [e^{-\lambda}\lambda^x/x!][\mu^k\lambda^{k-1}\, e^{-\mu\lambda}/\Gamma(k)]d\lambda$$

$$= \{\mu^k/[x!\,\Gamma(k)]\} \int_0^\infty \lambda^{k+x-1}\, e^{-(1+\mu)\lambda} d\lambda$$

$$= \frac{\Gamma(k+x)}{x!\,\Gamma(k)} \left(\frac{1}{1+\mu}\right)^x \left(\frac{\mu}{1+\mu}\right)^k$$

$$= \binom{k+x-1}{x} q^x p^k, \quad x = 0, 1, 2, \ldots, \tag{9}$$

where $p = \mu/(1+\mu) = 1 - q$. Values of m_x are therefore given by

$$m_x = s \binom{k+x-1}{x} q^x p^k, \quad x = 0, 1, 2, \ldots . \tag{10}$$

The parameter k is intuitively seen as a measure of the evenness with which the species' abundances are apportioned in the community. Observe that the coefficient of variation for X ($= \sqrt{\mathrm{Var}(X)}/E[X]$) in (9) is given by $1/\sqrt{kq}$. This quantity decreases as k increases. Thus, the species' abundances are more likely to be close to the mean abundance of a single species when k is large, resulting in greater evenness.

Fisher noted that values of k tended to be quite small when (10) was fitted to Lepidoptera data. In other words, the distribution of relative abundance was very uneven, though large numbers of species were present. In effect, k was a nuisance parameter for those data, continually taking values close to zero. The model (10) contains three parameters: s, k, and p. For small k values, s is inextricably bound up with k, prompting Fisher to combine s and k into a single parameter α. The concept is easiest to see by noting that the generating function (z - transform) for the terms m_x in (10) is

$$\sum_{x=0}^{\infty} m_x z^x = s[p/(1 - qz)]^k. \qquad (11)$$

The term m_0 is not represented in the sample data, being the expected number of species not present. The generating function for the terms $m_1, m_2, \ldots$ is then

$$\sum_{x=1}^{\infty} m_x z^x = s[p/(1 - qz)]^k - sp^k$$

$$= s \exp \{k \log[p/(1 - qz)]\} - s \exp [k \log p]$$

$$= s\{ 1 + k \log[p/(1 - qz)] + o(k) \}$$

$$- s[1 + kp + o(k)], \qquad (12)$$

where $o(k)$ denotes terms of order k^2 or higher. Therefore, for small k,

$$\sum_{x=1}^{\infty} m_x z^x \approx - sk \log (1 - qz). \qquad (13)$$

The parameters k and s are thus confounded into a single parameter, α, defined by

$$\alpha = sk \quad . \qquad (14)$$

This parameter is a diversity index, combining properties of species richness and evenness into a single number. The generating function (13) is that of the logarithmic series:

$$M_x = \alpha q^x/x. \qquad (15)$$

Extensions to the gamma model. The small values of k found in the Lepidoptera data resulted from communities with vary low evenness. Species abundance distributions from uneven communities have long right tails, representing concentration of abundance in a few species. The

long tail of the logarithmic series (15) is partly responsible for the astonishing success of this model in describing Lepidoptera data (Kempton and Taylor 1974). For some data though, even the log series tail is not heavy enough.

This fact led Kempton (1975) to propose an extension to the log series. The extension results from considering the underlying gamma abundance model (8). Kempton assumed that heterogeneity was present in the data, and he modeled this by assigning a probability distribution to the parameter α. The distribution used was a gamma distribution: $g(\mu) = \beta^{l}\mu^{l-1}e^{-\beta\mu}/\Gamma(l)$. (16)

The abundance distribution (8) then becomes

$$h(\lambda) = \int_0^\infty f(\lambda)g(\mu)d\mu$$

$$= \frac{\beta^{l}\lambda^{k-1}}{\Gamma(k)\Gamma(l)}\int_0^\infty \mu^{l+k-1}e^{-(\lambda+\beta)\mu}d\mu$$

$$= \frac{\Gamma(k+l)}{\Gamma(k)\Gamma(l)}\ \frac{\beta^{l}\lambda^{k-1}}{(\lambda+\beta)^{l+k}}\ , \quad 0<\lambda<\infty \qquad (17)$$

This is a beta II distribution (of which the F distribution is a special case). The resulting sampling distribution (7) is an integral with no closed form:

$$m_x = \int_0^\infty s\,[e^{-\lambda}\lambda^{x}/x!]h(\lambda)d\lambda$$

$$= \frac{s\beta^{l}}{x!}\ \frac{\Gamma(k+l)}{\Gamma(k)\Gamma(l)}\int_0^\infty \frac{\lambda^{x+k-1}e^{-\lambda}}{(\beta+\lambda)^{l+k}}d\lambda \qquad (18)$$

Setting $sk = \alpha$ and taking the limit as $s\to\infty$, $k\to 0$ provides the generalized log series model:

$$m_x = \frac{\alpha\beta}{x!}\int_0^{\infty}\frac{\lambda^{x-1}e^{-\lambda}}{(\beta+\lambda)^{1}}\,d\lambda \quad . \tag{19}$$

This, too, is an integral with no closed form. Fitting this model to data requires repeated numerical integration. Nonetheless, it worked quite successfully for the Lepidoptera data examples tested by Kempton.

Engen (1978) proposed a different approach to the tail-length problem. He pointed out that the m_x values in (10) can be computed for all values of k such that $-1 \leq k$, provided $x \geq 1$:

$$m_x = \alpha\frac{\Gamma(k+x)}{\Gamma(k+1)x!}q^x p^k, \quad x=1, 2, \ldots \quad . \tag{20}$$

This is called the extended negative binomial. The underlying species abundance distribution though, is no longer a gamma distribution, as (8) is only valid for positive k values.

The lognormal model. Preston (1948) introduced the lognormal distribution as a model of species abundance:

$$f(\lambda) = (\lambda\sigma\sqrt{2\pi})^{-1}\exp[-(\log\lambda-\mu)^2/(2\sigma^2)], \quad 0<\lambda<\infty \quad . \tag{21}$$

Here μ and σ^2 are parameters; $\log\Lambda$ has a normal distribution with mean μ and variance σ^2. Preston observed that the species abundances of bird census data resembled a normal curve when plotted on a logarithmic scale. Preston's estimation method was to fit the normal curve directly to the logged abundance data, rather than develop a sampling distribution. It is preferable to incorporate (21) into the sampling framework developed for the gamma model. From (7), the expected species frequencies are:

$$m_x = \frac{s}{x!\,\sigma\sqrt{2\pi}} \int_0^\infty \lambda^{x-1} \exp\{-\lambda - (\log\lambda - \mu)^2/(2\sigma^2)\}\, d\lambda$$

$$x = 0, 1, 2, \ldots . \quad (22)$$

These frequencies form the Poisson-lognormal distribution. The integral has no closed form, requiring numerical integration in data analysis.

The parameter σ^2 is a measure of the unevenness of the species abundances. The coefficient of variation for the Poisson-lognormal distribution is $[\exp(-\mu - \sigma^2/2) + \exp(\sigma^2) - 1]^{1/2}$, an increasing function of σ^2. Thus, $1/\sigma$ is a measure of evenness. This suggests the possibility of finding a limiting sequence of m_x values analogous to the log series. The log series had $sk = \alpha$, suggesting for the Poisson lognormal the limit $s/\sigma \to \gamma$. The resulting sequence is proportional to a harmonic series:

$$\lim_{\substack{s \to \infty \\ \sigma \to \infty \\ s/\sigma \to \gamma}} m_x = \frac{\gamma}{x!\sqrt{2\pi}} \int_0^\infty \lambda^{x-1} e^{-\lambda}\, d\lambda$$

$$= \frac{\lambda\Gamma(x)}{x\sqrt{2\pi}} = \left(\frac{\gamma}{\sqrt{2\pi}}\right)\left(\frac{1}{x}\right). \quad (23)$$

In the above, m_x is given by (22). The limit does not exist for the value x=o.

Canonical hypothesis. Preston (1962), in fitting a number of lognormal curves to ecological data sets, noticed a curious pattern. When the abundances were plotted on a logarithmic scale, the most common species often had an abundance near the mode of the individuals curve (see (5)). Preston hypothesized an empirical relationship between the parameters of the lognormal which fixes the dominant species to the

mode of the individuals curve. This hypothesis produces a "canonical" lognormal distribution with only two parameters instead of three. The canonical hypothesis has attracted considerable attention in theoretical ecology (May 1975, Sugihara 1980, Preston 1962). Conjectured causes of the canonical lognormal to date have failed to account for the fact that the canonical hypothesis implies an inverse relationship between species richness and evenness. The inverse relationship was pointed out by Patil and Taillie (1980), using a statistical definition of the canonical hypothesis for the lognormal and gamma models. The statistical definition permits actual statistical testing of the hypothesis on data sets. The definition will be generalized here to allow use of any species abundance distribution.

The species curve is given by $sf(\lambda)$, and the individuals curve by $s\lambda f(\lambda)$. Let $r = \log \lambda$. On a logarithmic scale these curves become $se^{r}f(e^{r})$ and $se^{2r}f(e^{r})$, respectively. The mode of the log-individuals curve, $\tilde{r}$, is found by setting the derivative of the curve equal to zero:

$$f'(e^{\tilde{r}}) + 2e^{-\tilde{r}}f(e^{\tilde{r}}) = 0.$$

For the gamma (8) and lognormal (21) distributions, we find that

$$\tilde{r} = \log[(k+1)/\mu] \quad \text{(gamma)}, \tag{24}$$

$$\tilde{r} = \mu + \sigma^2 \quad \text{(lognormal)}.$$

The cumulative distribution function of λ is defined by

$$F(\lambda) = \int_0^{\lambda} f(u)du. \tag{25}$$

Let Λ_{max} be the abundance of the largest of s species. By a well-known result from statistical theory (Patel et al. 1976),

$$E[F(\Lambda_{max})] = s/(s+1). \tag{26}$$

This suggests using the probability transform of s/(s+1) as a definition for λ_{max}:

$$\lambda_{max} = F^{-1}\left(\frac{s}{s+1}\right); \tag{27}$$

$$r_{max} = \log \lambda_{max}. \tag{28}$$

The quantity r_{max} is some measure of central tendency for the random variable $\log \Lambda_{max}$, the logged abundance of the largest species.

The canonical hypothesis asserts that $r_{max} \approx \tilde{r}$, or,

$$F^{-1}\left(\frac{s}{s+1}\right) \approx e^{\tilde{r}}. \tag{29}$$

The relationship constrains the species number, s, and the parameters in $f(\lambda)$ to a contour. For the lognormal and gamma models, this contour is the inverse relationship between species richness and evenness.

The canonical hypothesis (29) for the lognormal model reduces to

$$\sigma^2 = \left[\Phi^{-1}\left(\frac{s}{s+1}\right)\right]^2 \tag{30}$$

where $\Phi(\cdot)$ is the cumulative distribution function for the standard normal distribution. For large s the following asymptotic formula gives a good approximation to (30) (Patil and Taillie 1980):

$$\sigma^2 = 2 \log s - \log \log s - \log 4\pi. \tag{31}$$

As s (species richness) increases, σ^2 (unevenness) must increase.

With the gamma model, (29) becomes

$$\Gamma_k(k+1) = \frac{s}{s+1} \tag{32}$$

Here $\Gamma_k(\cdot)$ is the cumulative distribution function of a standard gamma distribution with index k. Alternatively, letting $\bar{\Gamma}_k(v) = 1 - \Gamma_k(v)$ be the right tail of the standard gamma, one sees that

$$\bar{\Gamma}_k(k+1) = \frac{1}{s+1} \tag{33}$$

Patil and Taillie (1980) note that (33) is approximately a linear relationship between s and k for large s and small k. Specifically, $sk \approx \alpha$, where α is a constant. This is just Fisher's limit given earlier (14). The value of α is fixed here, however, by the canonical relationship (33). Dividing (33) by k and letting $s \to \infty$, $k \to 0$, $sk \to \alpha$, one sees that

$$\int_1^\infty \frac{t^{-1}e^{-t}}{\Gamma(1)}\,dt = \frac{1}{\alpha} \tag{34}$$

which works out to $\alpha \approx 4.56$. Thus, the canonical hypothesis for the log series distribution (15) fixes a value of the parameter α. The hypothesis (33) for the gamma model constrains s (species richness) to be an inverse function of k (evenness).

Methodology for data analysis. The statistical approach to species abundance distributions discussed here provides meaningful ways of data analysis. The ad hoc approach of Preston (1948), by contrast, offers little opportunity for drawing statistical inferences. Statistical inferences made possible by Fisher's approach include point estimation, interval estimation, and hypothesis testing.

Let $m_1, m_2, \ldots,$ be a set of data or realized values of the random variables $N_1, N_2, \ldots$. The distribution of N_x, the number of species with x representatives in the sample, is Poisson, with mean m_x given by (7). Let $\underset{\sim}{\theta}$ be the (vector of) unknown parameters in m_x. For the gamma model, $\underset{\sim}{\theta} = [s, k, \mu]$. For the lognormal model, $\underset{\sim}{\theta} = [s, \sigma, \mu]$. The likelihood function, $l(\underset{\sim}{\theta})$, is the product of the probabilities $\Pr[N_1 = n_1] \Pr[N_2 = n_2] \ldots$:

$$\Pr[N_x = n_x] = [m_x(\underset{\sim}{\theta})]^{n_x} \exp[-m_x(\underset{\sim}{\theta})]/n_x!, \tag{35}$$

so that

$$l(\underset{\sim}{\theta}) = \exp[-\sum_{x=1}^{\infty} m_x(\underset{\sim}{\theta})] \prod_{x=1}^{\infty} [m_x(\underset{\sim}{\theta})]^{n_x}/n_x! \tag{36}$$

Note that

$$s = \sum_{x=0}^{\infty} m_x(\underset{\sim}{\theta}). \tag{37}$$

Also note that if $n_x = 0$ for any x, then

$$[m_x(\underset{\sim}{\theta})]^{n_x}/n_x! = 1, \quad x \notin A, \tag{38}$$

where A is the set of x values where $n_x > 0$. From (37) and (38), the likelihood function (36) becomes

$$l(\underset{\sim}{\theta}) = \exp[m_o(\underset{\sim}{\theta}) - s] \prod_{x \in A} \{[m_x(\underset{\sim}{\theta})]^{n_x}/n_x!\} \tag{39}$$

The maximum likelihood (ML) estimates of the unknown parameters $\underset{\sim}{\theta}$ are the values maximizing $l(\underset{\sim}{\theta})$, or, equivalently, the values maximizing the log of $l(\underset{\sim}{\theta})$ given by

$$\log l(\underset{\sim}{\theta}) = m_o - s + \sum_{x \in A} [n_x \log m_x - \log(n_x!)] \tag{40}$$

One way of findng these estimates is by differentiating $\log l(\underset{\sim}{\theta})$ with respect to each of the parameters and equating the results to zero:

$$\frac{\partial \log l}{\partial \theta_j} = \frac{\partial}{\partial \theta_j}(m_o - s) + \sum_{x \in A} \frac{n_x}{m_x} \frac{\partial m_x}{\partial \theta_j} = 0, \tag{41}$$

$$j = 1, 2, \ldots,$$

The system of q equations (one equation for each unknown parameter) must then be solved for the roots, $\hat{\underset{\sim}{\theta}}$ say. The first equation for s

can be seen by noting that $m_x = sp_x$, where p_x (=Pr[X=x]) are the probabilities given in (6):

$$\frac{\partial \log l}{\partial s} = p_o - 1 + \sum_{x \in A} \frac{n_x}{s} = 0, \tag{42}$$

yielding the ML estimate

$$\hat{s} = \sum_{x \in A} n_x / (1 - \hat{p}_o). \tag{43}$$

Here $\hat{p}$ is p with the other ML parameter estimates substituted. The equations for the other parameter estimates are can be rephrased by solving (41) for s and equating to (43):

$$\frac{\partial p_o}{\partial \theta_j} \sum_{x \in A} n_x + (1 - p_o) \sum_{x \in A} \frac{n_x}{p_x} \frac{\partial p_x}{\partial \theta_j} = 0, \tag{44}$$

$$j = 1, \ldots, q.$$

For statistical inferences and computing, it is valuable to obtain the q x q information matrix of the parameters, $I(\underset{\sim}{\theta})$ say, in which the i, jth element is

$$-E\left[\frac{\partial^2 \log l}{\partial \theta_i \partial \theta_j}\right] = \frac{\partial^2}{\partial \theta_i \partial \theta_j}(s - m_o)$$

$$+ \sum_{x \in A} \left[\frac{1}{m_x}\left(\frac{\partial m_x}{\partial \theta_i}\right)\left(\frac{\partial m_x}{\partial \theta_j}\right) - \frac{\partial^2 m_x}{\partial \theta_i \partial \theta_j}\right]. \tag{45}$$

Several approaches to computing the ML estimates are possible. The first is to solve the likelihood equations (41) directly using a numerical technique such as the Newton-Raphson, or its statistical counterpart, the scoring method. For the scoring method, an initial guess at the parameter values, $\underset{\sim}{\theta}_1$, say, is required. One then evaluates the likelihood equations (41) at $\underset{\sim}{\theta}_1$, putting the results in

a 1 x q row vector, $\underset{\sim}{h}(\underset{\sim}{\theta}_1)$ say. Then, the parameter estimates are iteratively improved by computing

$$\underset{\sim}{\theta}_{k+1} = \underset{\sim}{\theta}_k + \underset{\sim}{h}(\underset{\sim}{\theta}_k) I^{-1}(\underset{\sim}{\theta}_k) \qquad (46)$$

for k = 1, 2, ..., until $\underset{\sim}{h}(\underset{\sim}{\theta}_k)$ is as close to zero as desired. The resulting roots are the ML estimates, $\hat{\underset{\sim}{\theta}}$. ML theory states that $\hat{\underset{\sim}{\theta}}$ has a multivariate normal distribution (asymptotically) with mean $\underset{\sim}{\theta}$ and variance-covariance matrix $I^{-1}(\underset{\sim}{\theta})$. The variance-covariance matrix may be estimated by computing $I^{-1}(\hat{\underset{\sim}{\theta}})$.

An alternative computing procedure does not require evaluating the derivatives (45) at every iteration. The Nelder-Mead method instead computes a simplex on the likelihood surface (41) directly (see Olsson and Nelson 1975). Through a straightforward set of computational rules, the simplex "migrates" to the top of the likelihood surface, altering its shape each iteration to conform to the curvature of the surface. Experience has shown that the method is reliable but slow to converge.

The principal computing challenge with species abundance distributions is that many models require repeated numerical integrations to evaluate the m_x values (7) in the likelihood equations (41). Integration routines should be tested for their speed and efficiency. The gamma model provides a convenient vehicle for such testing, since its integral exists in closed form (10).

Once ML estimates are obtained, hypothesis testing is possible. Three types of possible tests are : 1) Goodness-of-fit, for evaluating the model's description of the data, 2) Multivariate two-sample, for

comparing two communities, 3) Likelihood ratio, for testing simple vs. complex models.

Goodness-of-fit tests are best performed on data pooled into abundance intervals. The reason is that investigators have found unpooled data to be rather variable, while the data quantiles remain rather stable (Kempton 1975). Logarithmic intervals to base 2 have become traditional, with abundance classes of x=1, 2-3, 4-7, 8-15, etc individuals. The observed n_x values, pooled into these classes, are compared with the expected $\hat{m}_x$ values. The usual chi-square statistic is computed, the degrees of freedom being the number of classes minus the number of parameters minus one. The fit of the model is rejected when the chi-square value exceeds some critical value.

Two communities can be compared by comparing the parameter estimates in the corresponding species abundance distributions. The two communities, A and B say, give rise to two sets of ML parameter estimates, $\hat{\underset{\sim}{\theta}}_A$ and $\hat{\underset{\sim}{\theta}}_B$. Since these estimates have multivariate normal distributions, a test of the hypothesis H_o: $\underset{\sim}{\theta}_A = \underset{\sim}{\theta}_B$ vs. the hypothesis H_1: $\underset{\sim}{\theta}_A \neq \underset{\sim}{\theta}_B$ can be conducted. Let

$$Q = (\hat{\underset{\sim}{\theta}}_A - \hat{\underset{\sim}{\theta}}_B)\ [I_A^{-1}(\hat{\underset{\sim}{\theta}}_A) + I_B^{-1}(\hat{\underset{\sim}{\theta}}_B)]^{-1}\ (\hat{\underset{\sim}{\theta}}_A - \hat{\underset{\sim}{\theta}}_B)'. \quad (47)$$

Under H_o, Q will have a chi-square distribution with degrees of freedom equal to the number of parameters in $\underset{\sim}{\theta}$.

Often an investigator will want to determine the level of complexity a model must have to adequately describe the data. For example, is the two parameter log series model (15) sufficient for a data set, or must the full, three-parameter negative binomial (10) be used? Also,

the canonical hypothesis would be an example. When a simple model is a special case of a more general model, likelihood ratio techniques can be used to test one against the other. One first computes the likelihood (36) under the simple model, l_o say, evaluated using the ML parameter estimates. Next, one computes l_1, the likelihood under the more complex model, here using the parameters estimated for the complex model. The parameter space under the simple model is assumed to be contained in (or a subset of) the parameter space of the complex model. As an example, the canonical hypothesis under the log series model would have the parameter α fixed at 4.56, while α is a free parameter under full log series. One finally computes the ratio of likelihoods from the two contending hypotheses:

$$R = l_o / l_1 \qquad (48)$$

Under H_o, ML theory states that -2 log R has a chi-square distribution (asymptotically). The degrees of freedom are the number of free parameters in the full model minus the number of free parameters in the reduced model.

Conclusions. On the basis of this synthesis of material on species abundance distributions, I have identified a number of important ecological questions that may be addressed with the statistical methodology documented in this report:

1) Species abundance distributions have shown good potential for evaluating water quality impacts on aquatic communities. However, their use for this purpose has been hampered by improper statistical methods. There have previously been no appropriate methods of hypothesis testing for evaluating the quality of the models or comparing

the abundance patterns between communities. Comparison is usually critical for biomontoring: studies are concerned with sampling before and after pollution, upstream and downstream of pollution, etc. Thus, previous studies using species abundance distributions for pollution impacts should be reanalyzed.

2) Aquatic ecologists have used the lognormal distribution almost exclusively. An assortment of other models are available and should be tried. These other models have met with considerable success in describing terrestrial communities.

3) The canonical hypothesis has provoked a considerable amount of excitement and speculation among ecologists. Large amounts of data are seemingly consistent with the hypothesis. Due to the ad hoc estimation methods used by these ecologists, however, the actual empirical status of the canonical hypothesis is uncertain. Existing data need to be reanalyzed using the statistical methods contained in this report. Furthermore, this report identified explicit techniques for conducting formal statistical tests of such hypotheses.

LITERATURE CITED

Abramowitz, M. and I.A. Stegun, eds. 1965. **A handbook of mathematical functions.** Dover, New York.

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

Dennis, B. and G.P. Patil. 1979. Species abundance, diversity, and environmental predictability. In **Ecological diversity in theory and practice**, J.F. Grassle, G.P. Patil, W.K. Smith, and C. Taillie, eds. International Co-operative Publishing House, Fairland, Maryland.

Engen, S. 1978. **Stochastic abundance models**, Chapman and Hall, London.

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. **Journal of Animal Ecology**, 12, 42-58.

Kempton, R.A. and L.R. Taylor. 1974. Log-series and log-normal parameters as diversity discriminants. **Journal of Animal Ecology**, 3, 381-399.

Kempton, R.A. 1975. A generalized form of Fisher's logarithmic series. **Biometrika**, 62, 29-38.

May, R.M. 1975. Patterns of species abundance and diversity. In **Ecology and Evolution of Communities**, M.L. Cody and J.M. Diamond, eds. Belknap Press, Cambridge.

Olsson, D.M. and L.S. Nelson. 1975. The Nelder-Mead simplex method for function minimization. **Technometrics**, 17, 45-51.

Patel, J.K., C.H. Kapadia, and D.B. Owen. 1976. **Handbook of statistical distributions.** Marcel Dekker, New York.

Patil, G.P. and C. Taillie. 1980. Species abundance models, ecological diversity, and the canonical hypothesis. **International Statictical Institute Bulletin**, 42.

Patil, G.P. and C. Taillie. 1982. Diversity as a concept and its measurement. J. Amer. Stat. Assoc., 77, 548-561.

Patrick, R., M.H. Hohn, and J.H. Wallace. 1954. A new method for determining the pattern of diatom flora. **Notulae Naturae**, 259, 1-12.

Patrick, R. 1968. The structure of diatom communities in similar ecological conditions. **American Naturalist**, 102, 173-183.

Pielou, E.C. 1975. **Ecological diversity.** John Wiley and Sons, New York.

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

Preston, F.W. 1962. The canonical distribution of commonness and rarity. **Ecology**, 43, 185-215, 410-432.

Preston, F.W. Noncanonical distributions of commonness and rarity. **Ecology**, 61: 80-97.

Sugihara, G. 1980. Minimal community structure: an explanation of species abundance patterns. **American Naturalist**, 116, 770-787.

Taillie, C. 1979. Species equitability: a comparative approach. In **Ecological diversity in theory and practice**, J.F. Grassle, G.P. Patil, W.K. Smith, and C. Taillie, eds. International Co-operative Publishing House, Fairland, Maryland.

Whittaker, R.H. 1972. Evolution and measurement of species diversity. **Taxon**, 21, 213-251.

Williams, C.B. 1964. **Patterns in the balance of nature.** Academic Press, New York.

Selected Water Resources Abstracts

Input Transaction Form

1. Report No.

2.

3. Accession No. W

4. Title STATISTICS OF SPECIES ABUNDANCE DISTRIBUTIONS IN MONITORING AQUATIC COMMUNITIES

5. Report Date

6. March 1983

8. Performing Organization Report No.

7. Author(s) Dennis, B.

9. Organization Idaho University, Moscow, Forest Resources Dept.

10. Project No. A-085-IDA (1)

11. Contract/Grant No. 14-34-0001-2114

13. Type of Report and Period Covered

12. Sponsoring Organization

15. Supplementary Notes

Idaho Water and Energy Resources Research Institute Completion Report, Moscow, March 1983. 21 p., 22 ref.

16. Abstract

This report documents statistical procedures for using species abundance distributions for monitoring pollution impacts in aquatic biological communities. The species abundance distributions are also appropriate for use in other areas of applied ecology as well. Previous use of these tools by ecologists has relied on largely ad hoc data analysis procedures having little basis in statistical theory. This report is a synthesis of the statistical theory that applies to species abundance distributions. The report discusses appropriate statistical interpretation of the distributions, sampling distributions, types of abundance models such as the log-normal and the gamma, the canonical hypothesis of species abundance, and methods of testing hypotheses. It also identifies a number of important questions needing further research.

17a. Descriptors

*Species abundance distributions, pollution impacts, aquatic biological communities, ecology.

17c. COWRR Field & Group

05C

18. Availability IWERRI

19. Security Class. (Report)

20. Security Class. (Page)

21. No. of Pages 21

22. Price $

Send to:
Water Resources Scientific Information Center
OFFICE OF WATER RESEARCH AND TECHNOLOGY
U.S. DEPARTMENT OF THE INTERIOR
Washington, D.C. 20240

Abstractor IWERRI

Institution IWERRI