Ecological Applications, 3(2), 1993, pp. 351–356

Green
Young
1993

SAMPLING TO DETECT RARE SPECIES[1]

ROGER H. GREEN
Department of Zoology, University of Western Ontario, London, Ontario, Canada N6A 5B7

RICHARD C. YOUNG
Woodward-Clyde Consultants, P.O. Box 680925, Franklin, Tennessee 37068 USA

Abstract. Often a sampling program has the objective of detecting the presence of one or more species. One might wish to obtain a species list for the habitat, or to detect the presence of a rare and possibly endangered species. How can the sampling effort necessary for the detection of a rare species can be determined?

The Poisson and the negative binomial are two possible spatial distributions that could be assumed. The Poisson assumption leads to the simple relationship $n = -(1/m)\log \beta$, where n is the number of quadrats needed to detect the presence of a species having density m, with a chance β (the Type 2 error probability) that the species will not be collected in any of the n quadrats. Even if the animals are not randomly distributed the Poisson distribution will be adequate if the mean density is very low (i.e., the species is rare, which we arbitrarily define as a true mean density of <0.1 individuals per sample unit), and the spatial distribution is not highly aggregated. Otherwise a more complicated relationship based on the negative binomial distribution would have to be used.

Published sampling distributions of 37 unionid mollusc species over river miles (distance measured along the path of the river; 1 mile = 1.609347 km) in two southern Appalachian rivers were evaluated to determine the appropriateness of the simple Poisson-based formula for estimation of necessary sample size to detect species presence. For each of 273 species × river mile combinations we estimated the mean, the variance, and the negative binomial parameter k, and then estimated "necessary n" from both the Poisson- and the negative-binomial-based formulae. We defined "Poisson adequacy" to be the proportion that the Poisson estimate is of the negative binomial estimate of necessary sample size, and stated the requirement that it be >0.95. Only 8 of the 273 cases represented rare species that failed this requirement. Thus we conclude that a Poisson-based estimate of necessary sample size will generally be adequate and appropriate.

Key words: detection; endangered species; molluscs; Poisson; power; presence; rare species; sample size; sampling; unionid.

INTRODUCTION

What should be the basis for allocation of sampling effort when the objective is to detect the presence of a rare species? Kovalak et al. (1986) discuss this question with emphasis on sampling methods, on species vs. number of individuals models, and on number of quadrats needed to obtain specified confidence limits on estimates of mean density. Here we will focus on the number of quadrats needed to detect the presence of (i.e., collect at least one of) a rare species (given some target density, i.e., degree of rareness), with some specified probability of detection.

Intuitively the sampling distribution should approximate the Poisson, which can be derived from other common distributions (e.g., binomial, negative binomial) by simply assuming that the event is rare—that the probability of collecting an individual in any given sample is low. Even aggregated (clumped, patchy) distributions, such as the negative binomial ($s^2 = m + m^2/k$, where s^2 is the variance in number per unit area and m is the mean) with k small, approach the Poisson ($s^2/m \to 1$) as the event becomes rare and thus the mean m becomes small. Since $s^2/m = 1 + m/k$ in the negative binomial, it is obvious that $s^2/m \to 1$ as $m \to 0$ (so long as k is not also approaching 0 at the same or a faster rate). Obviously "rarity" is a relative thing. If there are fewer individuals in a given size area then m will decrease, but so will it decrease if there is the same density of individuals and smaller quadrats are used. It is easy to show that when sampling aggregated distributions of organisms, for a given coverage ($\Sigma X = mn$) it is better to sample with many small quadrats (large n and small m) than to use a few large quadrats (small n and large m). See Green (1979) for discussion.

For our purposes here we wish to maintain generality, and try to show that our results apply to any rare-species sampling situation. Somewhat arbitrarily we will declare that any species having true density exceeding 0.1 per sample unit size (e.g., quadrat size) is *not* rare. What densities <0.1 are to be considered rare we leave to individual taste; we assume the burden of demonstrating valid results for any density <0.1.

We approach the problem in two steps. First we

[1] Manuscript received 20 November 1991; revised and accepted 2 September 1992.

derive appropriate formulae, based on the Poisson and negative binomial distributions, for a sample number adequate to detect the presence of rare species with some specified probability of detection. This is in essence a power analysis. Power = $1 - \beta$, and β = the probability of the Type 2 error = the probability of allocating n quadrats and failing to collect a species that is actually present in that habitat and has some mean density m. (See Green 1989 and references cited therein for more discussion of power analysis in biological applications.) Then we evaluate the sampling distributions of the unionid species in two data sets to assess the adequacy of the Poisson formula relative to the negative binomial. The latter describes most any spatial distribution of organisms, from random to highly aggregated, quite well (Green 1979 and references cited therein), and there is no reason it should not do so for our purposes. Thus we use the negative binomial estimate of necessary sample size (number of quadrats) n as the standard against which the estimate based on the Poisson approximation will be judged.

Our goal, in the end, is to have a generally applicable formula for estimating necessary sample size to collect (i.e., detect the presence of) a species having a specified degree of rareness and a specified Type 2 error probability. Any formula based on the negative binomial would also require an estimate of the parameter k, and thus a much greater sampling effort than that needed to simply collect an individual of the species. A Poisson-based formula, on the other hand, would only require specification of a degree of rareness and a Type 2 error probability. Therefore we hope to demonstrate the adequacy of a Poisson-based formula.

Derivation of Power Formulae

The probability of obtaining X individuals in a single sample from a Poisson distribution with mean m is

$$p_X = m^X \frac{e^{-m}}{X!},$$

and the probability of obtaining 0 individuals (an empty sample) is

$$P_0 = e^{-m}.$$

In n independent samples the probability of obtaining 0 individuals (all n samples are empty) is

$$P_0{}^n = (e^{-m})^n = e^{-mn} = e^{-\Sigma X}.$$

The probability of obtaining at least one individual in at least one of n independent samples (=the probability of detecting that species) is

$$P_{>0,n} = 1 - P_0{}^n = 1 - e^{-mn}.$$

By rearrangement we obtain

$$n = -\frac{1}{m}\log(1 - p_{>0,n}), \tag{1}$$

which is the number of samples needed to detect the presence of a rare species with power $1 - \beta = p_{>0,n}$. Thus Eq. 1 can be written

$$n = -\frac{1}{m}\log\beta. \tag{2}$$

For $1 - \beta = 0.95$ this reduces to the simple relationship

$$n = 3/m.$$

Thus the necessary sample size for a 0.95 chance of detecting a species is equal to 3 divided by the mean density of that species.

Now let us repeat this derivation using the negative binomial distribution

$$p_X = \left(1 + \frac{m}{k}\right)^{-k} \frac{(k + X - 1)!}{X!(k - 1)!}\left(\frac{m}{m + k}\right)^X$$

$$p_0 = \left(1 + \frac{m}{k}\right)^{-k}$$

$$p_0{}^n = \left[\left(1 + \frac{m}{k}\right)^{-k}\right]^n = \left(1 + \frac{m}{k}\right)^{-kn}$$

$$p_{>0,n} = 1 - p_0{}^n = 1 - \left(1 + \frac{m}{k}\right)^{-kn}.$$

By rearrangement we obtain

$$n = -\frac{1}{k}\frac{\log(1 - p_{>0,n})}{\log\left(1 + \frac{m}{k}\right)} = -\frac{1}{k}\frac{\log\beta}{\log\left(1 + \frac{m}{k}\right)}. \tag{3}$$

In all of the above we have used the same symbols (e.g., m and k) interchangeably for population parameters and for sampling estimates of these parameters. Which applies should be clear from the context. This keeps the notation simple, and in any case the development of our argument is more heuristic than formal.

When estimates of parameters are used as predictors in equations, e.g., the sample mean in Eq. 3, there will be an added component of variance associated with the equation's estimate (of n in this case). This could be variance due to sampling or counting error, or both. (For our data there would be negligible counting error, relative to sampling error.) There should be no bias unless regression slope parameters are being estimated. When they are (see next section), the effect of a predictor variable estimated with error is to bias the estimate of the slope of the ordinary least squares regression toward zero. Thus the strength of a relationship will be underestimated. See McArdle (1988) for a review of this problem.

We define the adequacy of the Poisson assumption as the ratio $R_1 = n_p/n_{nb}$. For example, the Poisson-based estimate of sample size (from Eq. 2) needed to have a 95% chance of detecting the presence of a species having mean density 0.1 is $n_p = 29.96$. If k is 1, then the negative binomial estimate (from Eq. 3) is $n_{nb} = 31.43$. Thus the adequacy of the Poisson assumption

TABLE 1. The 37 unionid mollusc species that were collected.

Actinonaias carinata	*Lasmigona costata*
Actinonaias pectorosa	*Lastena lata*
Amblema plicata	*Leptodea fragilaris*
Conradilla caelata	*Lexingtonia dolabelloides*
Cumberlandia monodonta	*Ligumia recta*
Cyclonaias tuberculata	*Medionidus conradicus*
Cyprogenia irrorata	*Megalonaias gigantea*
Dromus dromas	*Obliquaria reflexa*
Elliptio dilatatus	*Plethobasus cyphyus*
Epioblasma brevidens	*Pleurobema cordatum*
Epioblasma capsaeformis	*Pleurobema oviforme*
Epioblasma triquetra	*Potamilus alata*
Fusconaia barnesiana	*Ptychobranchus fasciolaris*
Fusconaia cuneolus	*Ptychobranchus subtentum*
Fusconaia edgariana	*Quadrula cylindrica*
Fusconaia subrotunda	*Quadrua pustulosa*
Lampsilis fasciola	*Tritongonia verrucosa*
Lampsilis ovata	*Truncilla truncata*
	Villosa iris

is $R_1 = n_p/n_{nb} = 0.953$. The bias of the Poisson estimate (relative to the assumed unbiased negative binomial estimate) is $R_1 - 1 = -0.047$, or -4.7%. In general, combining Eqs. 2 and 3,

$$R_1 = \frac{n_R}{n_{nb}} = \frac{k}{m}\log\left(1 + \frac{m}{k}\right).$$

Now we define $R_2 = m/k$, and, substituting R_2 for m/k, we obtain

$$R_1 = \frac{\log(1 + R_2)}{R_2}.$$

Thus we see that the ratio $R_2 = m/k$ determines the adequacy of the Poisson assumption $R_1 = n_p/n_{nb}$. For example the ratio $R_2 = m/k = 0.107$ corresponds to an adequacy of $R_1 = 0.95$. In a scatterplot of m vs. k for real data, the boundary for a Poisson adequacy of 0.95 would be the line $m = 0.107\ k$. In a log–log plot of m vs. k, the boundary would be the line $\log m = \log 0.107 + \log k$ (as in Fig. 2, which presents results described below).

THE UNIONID SAMPLING DISTRIBUTIONS

Two data sets (Ahlstedt 1986) were used to evaluate the sampling distributions of the 37 unionid species that were collected (Table 1). The first data set was obtained from 345 0.25-m² quadrats allocated within 11 "river miles" (17.6 river kilometres; distance measured along the path of the river) in the Clinch River (Virginia–Tennessee) in 1979. The number of quadrats per river mile varied between 16 and 40. The second

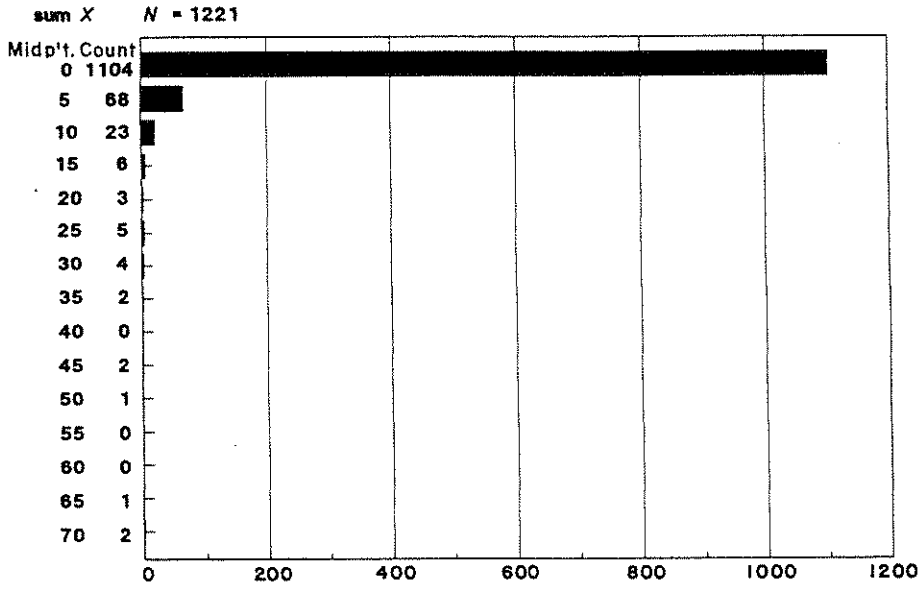

FIG. 1. The frequency distribution of number of individuals collected ΣX, in both unionid mollusc species data sets, for all 1221 species × river mile combinations (of which 948 have $\Sigma X = 0$).

TABLE 2. Analysis of covariance results for models predicting $1/k$, where k is the negative binomial parameter.

a. $1/k$ predicted by species, river, and density m; weighted by total sampling area ΣX

Source	df	F	P
Species	36	0.31	>.05 NS
River	1	0.71	>.05 NS
m	1	0.15	>.05 NS
Error	234		
Total	272		

b. Rank $1/k$ predicted by species, river, and rank density m; weighted by ΣX

Source	df	F	P
Species	36	2.33	<.01**
River	1	0.86	>.05 NS
m	1	63.59	<.01**
Error	234		
Total	272		

** $P < .01$.

data set was obtained from 509 0.25-m² quadrats allocated over 22 river miles in the Duck River (Tennessee) in 1980. The number of quadrats per river mile varied between 8 and 40. Thus for both rivers the proportion of the total area that was actually sampled was small; we can assume an "infinite population" sampling situation, and no finite population correction is needed.

For each combination of species and river mile we calculated the frequency distribution, the sample variance and mean, an estimate of the negative binomial parameter k (by solving the relationship $k = m^2/(s^2 - m)$, and the chi square test of the equality of variance and mean (H_0: Poisson distribution, which is a negative binomial distribution with $k \to \infty$). See Elliott (1977) for discussion and examples of testing H_0: Poisson and estimating parameters of the negative binomial distribution. Note that these data consist of independent observations: different species obtained from sets of quadrats collected from different river miles.

When a species was not collected at all in a given river mile, no calculations were possible. This was the case for 948 out of the 1221 species × river mile combinations ($37 \times 33 = 1221$) in the two rivers. For many of the remaining $152 + 121 = 273$ species × river mile combinations there were only one or two individuals collected. This is something of a catch-22 situation in that it *is* the instances of rarity that are of greatest interest, and yet when few individuals are collected the parameter estimates are poor and the power of the test against H_0: Poisson is low. Therefore we must estimate the variance, mean, and k, putting the greatest weighting on the instances where there are more quadrats and where there are more mussels, and then back-extrapolate to the condition of rarity. The value of ΣX (the total number of mussels in all quadrats) seems an appropriate weighting function for summary statistical analyses (Fig. 1). It simply weights each species × river mile combination by how many mussels were present, and thus roughly according to how reliable the parameter estimates are.

An analysis of covariance (ANCOVA) was performed on the combined data from both rivers, with the dependent variable $1/k$ used as an index of patchiness ($1/k$ near zero indicating randomness and $1/k$ large indicating patchiness; see Elliott 1977). Category predictor variables were river and species, and the covariate was the mean density, m. The purpose of this statistical analysis is to determine whether there is a "patchiness vs. mean density" relationship, and if there is whether it differs among species or between the two rivers. Because the data contain several extreme outliers ($1/k$ having five values of 5×10^5 while 97% of the values were <20), no relationship is found (Table 2a). Conover and Iman (1981) showed that the robustness of nonparametric methods can be obtained by performing parametric statistical analyses on rank data, so we repeated this ANCOVA using ranks of the $1/k$ values and ranks of the means (Table 2b) and found that rank $1/k$ is significantly predicted by rank m, that the rank $1/k$ vs. rank m relationship significantly differs among species, and that there is no difference between the two rivers. The unbalanced design did not permit tests for interactions (e.g., between species and rank m), but plots of the data do not suggest any interactions. The relationship between rank $1/k$ and rank m is positive—a within species × river slope of +0.510 from this ANCOVA—and only slightly different at +0.522 in a simple regression of rank $1/k$ on rank m (species and river not included as predictor variables). This indicates that sampling distributions become more random as the mean density decreases, which agrees with the theoretical rationale stated in the Introduction—that the Poisson distribution is approached as the mean becomes small.

Any bias in the slope estimate caused by m, or rank

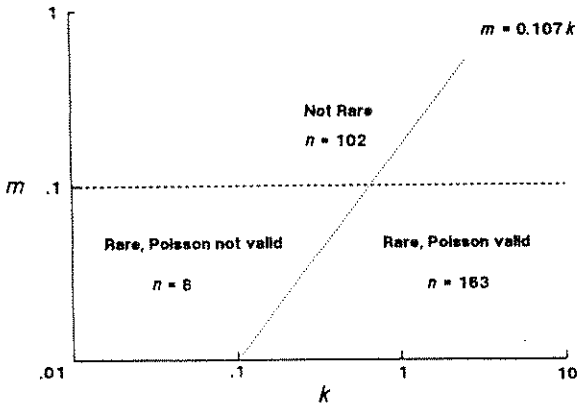

FIG. 2. A log–log plot of the mean density m against the Poisson parameter k, divided into three regions: $m > 0.1$, and the $m < 0.1$ region divided into two parts by the line $m = 0.107\ k$ (representing $m/k = 0.107$). n = the number of cases from the two unionid mollusc data sets that fall into each region.

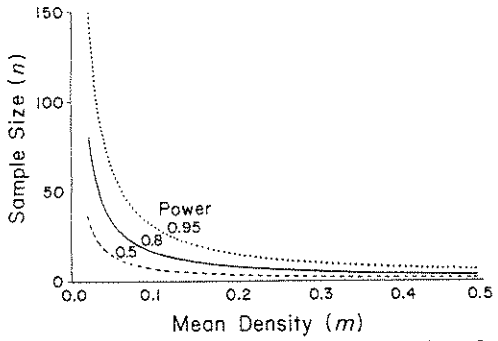

FIG. 3. The necessary sample size n as a function of mean density m, for various degrees of power $1 - \beta$, when sampling the Poisson distribution.

m, having been estimated with error is small. Geometric mean regression yields slope estimates little different from the ordinary least squares estimates. In any case the bias would always be toward a weaker "patchiness vs. mean density" relationship than actually exists.

As a check, a "resistant line" regression analysis (Minitab 1989), which ignores extent of deviation of points from the fitted line (and thus is unaffected by outliers), was also used to regress $1/k$ on m (the actual values, not the ranks). No ANOVA table or significance test is possible with this method, but a similar result was obtained, namely, a positive relationship between $1/k$ and m: $1/k = -0.039 + 0.792m$. Here we can actually predict how $1/k$ ("patchiness") will decrease as mean m decreases. It is obvious that as m goes to zero the value of $1/k$ goes to near zero. For a mean of 2 (slightly greater than the largest mean value of 1.71) the relationship predicts $k = 0.65$, i.e., strong patchiness. A mean of 1 predicts $k = 1.33$, moderate "typical" patchiness. A mean of 0.1 predicts $k = 24.8$, which is near random. Sampling to detect rare species is the topic here, and rarity by anybody's definition surely begins at densities of 0.1 per quadrat or less. In fact the relationship obtained by resistant line regression analysis predicts $m/k > 0.107$ (equivalent to Poisson adequacy $n_p/n_{nb} > 0.95$) for densities $m < 0.40$. Thus $m < 0.1$ would seem to satisfy both an intuitive definition of rarity and also Poisson adequacy in terms of the m/k ratio. Therefore these data support the proposition that the Poisson distribution can be assumed when sampling to detect the presence of rare species.

What about exceptions to this generality? In the ANCOVA on rank data (Table 2b) there were significant species differences in the patchiness vs. mean relationship. With so many species (37) the power of the test for species differences is high. The species differences are not large; the percentage variation among adjusted species means is only 25%. Also, there is no consistency of the order of the species with respect to their patchiness; the ANCOVA was done separately for each river and it was found that the correlation of species' adjusted mean patchiness between rivers was near zero. Thus there is no evidence of consistent species differences in the patchiness of their distributions.

How many of the 273 species × river mile combinations that yielded any mussels (hereafter referred to as "cases") have (a) low enough mean density to be considered rare (and thus be relevant to this paper's topic), and also have (b) m/k values too large for the Poisson assumption to be valid? Fig. 2 displays this information as a log–log plot of m vs. k, divided into three regions. Rarity is arbitrarily defined as $m < 0.1$, and then within that region a valid Poisson assumption is defined as $n_p/n_{nb} > 0.95$, corresponding to $m/k < 0.107$. (See previous section for discussion.) Only 8 of the 273 cases (171 of which constitute rarity by this definition) would be inappropriate for estimating the necessary n based on the Poisson assumption. In fact these estimates of Poisson adequacy are probably conservative. Some of the 948 "empty" species × river mile combinations undoubtedly did have that species present, but it was not collected in the sample quadrats. The number of such cases is unknown, but however many there are they would probably represent densities of $m < 0.1$ and $m/k < 0.107$. Thus they would fall into the "rare and Poisson valid" region in Fig. 2, and would increase the proportion of cases for which the Poisson assumption is adequate.

It is clear that the simple Poisson-based formula (Eq. 2) is usually adequate for estimating the necessary number of samples to detect the presence of a rare species. Fig. 3 shows how the necessary number of samples varies with the mean density, for various degrees of power, at $k = \infty$ (Poisson). A power of $1 - \beta = 0.95$ would balance the conventional $1 - \alpha = 0.95$ confidence level, with Type 1 and Type 2 error probabilities equal. Power equal to $1 - \beta = 0.8$ has become something of a standard in marine environmental

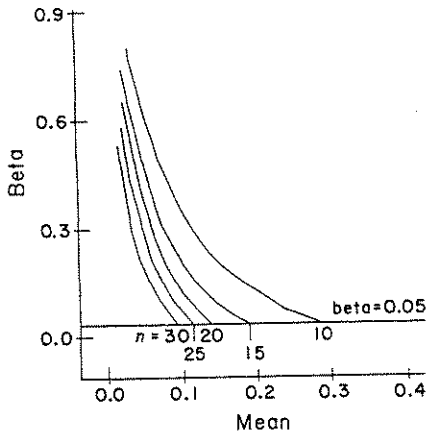

FIG. 4. The probability β of not collecting any individuals as a function of mean density m, for various values of n, when sampling the Poisson distribution. For example, if $n = 30$ quadrats yielded no individuals of a species, then one could conclude (with a risk of .05 of being wrong) that if the species occurs at all in that habitat it has a mean density of <0.100 per quadrat.

monitoring studies. A power of $1 - \beta = 0.5$ would imply settling for a 50:50 chance of detecting the presence of the species.

Discussion

Given some power to detect the presence of a rare species, what value of the mean density m should be used when determining the necessary number of quadrats n? Estimation of the mean (or the variance, or the negative binomial parameter k) would require even more samples than would detection of the species' presence. One can only decide how rare a species one wants to detect, and then allocate sampling effort accordingly. Kovalak et al. (1986) refer to this chosen value of m as the target density, and they ask "What should this density be?" They answer "This question must be addressed by regulatory agencies charged with protecting rare species." We agree.

If the species is not collected then the probability that any specified mean density exists can be stated (Fig. 4). This is based on a rearrangement of Eq. 2 to solve for β as a function of m, given the sample number n that was actually used:

$$\beta = e^{-mn} = e^{-\Sigma X}.$$

This provides a probabilistic solution to what Kovalak et al. (1986) refer to as "the corollary proposition: For a level of sampling effort, what population densities would have been detected?"

In conclusion, we wish to emphasize the generality of these results, which do not depend on species, habitat, sampling method, or sample unit size. The validity of the Poisson-based formula (Eq. 1) depends only on the assumption that the ratio m/k is less than some specified value, e.g., 0.107 for the estimate of n from the Poisson-based formula to be at least 95% of the correct value. Therefore it would be useful for some preliminary sampling to be done for any previously unstudied species and habitat, to estimate the distribution of m/k (for m at and below target densities, e.g., as described in Fig. 2), and thus estimate the likely proportion of cases for which the Poisson assumption will be inadequate.

Acknowledgments

We wish to thank our colleagues, students, and audiences at seminars and meeting presentations for their criticisms and suggestions as these ideas have evolved. The senior author was supported by an operating grant from the Natural Sciences and Engineering Research Council of Canada throughout the development of these ideas and the preparation of this paper.

Literature Cited

Ahlstedt, S. A. 1986. Cumberlandian mollusc conservation program, Activity 1: mussel distribution surveys. TVA/ONRED/AWR-86/15. Tennessee Valley Authority, Chatanooga, Tennessee, USA.

Conover, W. J., and R. L. Iman. 1981. Rank transformations as a bridge between parametric and nonparametric statistics. American Statistician **35**:124–130.

Elliott, J. M. 1977. Some methods for the statistical analysis of samples of benthic invertebrates. Scientific Publication Number 25. Freshwater Biological Association, The Ferry House, Ambleside, Cumbria, United Kingdom.

Green, R. H. 1979. Sampling design and statistical methods for environmental biologists. John Wiley & Sons, New York, New York, USA.

———. 1989. Power analysis and practical strategies for environmental monitoring. Environmental Research **50**: 195–205.

Kovalak, W. P. 1986. Sampling effort required to find rare species of freshwater mussels. Pages 34–45 *in* B. G. Isom, editor. Rationale for sampling and interpretation of ecological data in the assessment of freshwater ecosystems. ASTM STP 894. American Society for Testing and Materials, Philadelphia, Pennsylvania, USA.

McArdle, B. H. 1988. The structural relationship: regression in biology. Canadian Journal Zoology **66**:2329–2339.

Minitab Reference Manual Release 7. April 1989. Minitab, Inc., State College, Pennsylvania, USA.