Modeling Count Data of Rare Species: Some Statistical Issues
Author(s): Ross B. Cunningham and David B. Lindenmayer
Source: *Ecology*, Vol. 86, No. 5 (May, 2005), pp. 1135-1142
Published by: Ecological Society of America
Stable URL: http://www.jstor.org/stable/3450874
Accessed: 20/08/2009 22:02

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/page/info/about/policies/terms.jsp. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/action/showPublisher?publisherCode=esa.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

JSTOR is a not-for-profit organization founded in 1995 to build trusted digital archives for scholarship. We work with the scholarly community to preserve their work and the materials they rely upon, and to build a common research platform that promotes the discovery and use of these resources. For more information about JSTOR, please contact support@jstor.org.

*Ecological Society of America* is collaborating with JSTOR to digitize, preserve and extend access to *Ecology*.

http://www.jstor.org

*Ecology*, 86(5), 2005, pp. 1135–1142
© 2005 by the Ecological Society of America

# MODELING COUNT DATA OF RARE SPECIES: SOME STATISTICAL ISSUES

ROSS B. CUNNINGHAM[1] AND DAVID B. LINDENMAYER

*Centre for Resource and Environmental Studies, The Australian National University, Canberra, A.C.T., 0200, Australia*

*Abstract.* Most species abundance data show that a small number of species contribute the vast majority of individuals to a community. Thus, most taxa in a community are uncommon or rare. Yet such species will frequently be of ecological, conservation, or management interest. Data for uncommon or rare species will be presence/absence data or counts of abundance that contain a greater number of zero observations than would be predicted using standard, unimodal statistical distributions. Such data are generally referred to as zero-inflated data and require specialized methods for statistical analysis. Statistical approaches to modeling zero-inflated data include nonstandard mixture models; two-part, conditional models; and birth process models. In this paper, we briefly summarize two of these methods and illustrate the two-part, conditional approach to the problem of modeling count data with extra zeros. An advantage of this approach includes separate fits and separate interpretations of both components of count data; that is to say, the presence/absence component and the abundance component (given presence) can be analyzed separately. This can be valuable not only for simplicity, but also such a two-step method may assist ecological understanding in cases where the basis for species presence might be separated from the underlying reasons affecting the population size of that species at those sites where it is present.

We present two case studies of the application of the two-part conditional model for modeling count data with extra zeros. One deals with modeling relationships between counts of the rare and endangered arboreal marsupial, Leadbeater's possum (*Gymnobelideus leadbeateri*) and habitat variables in the wet eucalypt forests of southeastern Australia. The other is an analysis of data obtained from a monitoring study of seabird nesting from the Coral Sea off northeastern Australia. Finally, we briefly discuss some inferential and practical issues in developing designs and models for presence/absence data (which is the first component in the two-part conditional approach) when observed occurrences are low (e.g., <5%).

*Key words: count data; habitat analysis; Leadbeater's possum; monitoring; over dispersion; rare species; sea birds; statistical modeling; zero-inflated data.*

## INTRODUCTION

Species occurrence can be broadly described by three key qualitative factors: abundance of a species within a community (common or rare), habitat specificity (common or specialized), and geographic range (widespread or restricted). The full factorial expansion of these factors gives eight combinations (Table 1). Seven of these describe the various types of rarity as defined by Rabinowitz et al. (1986). Note that species in cell 1 (the top row) are widespread and common and do not qualify as rare species.

Many of the forms of rarity in Table 1 indicate that in most cases, a "rare species" will typically not be numerically abundant in the majority of communities in which it is a member. This proposition has some congruence with the ideas of Preston (1962) and MacArthur and Wilson (1967) both of whom show that within almost all ecological communities, a few species are common and the vast majority are uncommon or rare. This phenomenon has been known in ecology for a long time (see Williams 1944, reviewed by Gaston 1994). Yet, rare and uncommon species will often be those of substantial management and conservation interest (Meffe and Carroll 1997, Fagan et al. 2002, Hartley and Kunin 2003), in part because they may be among the most extinction-prone taxa in an assemblage.

Frequency data arising from studies of the abundance of rare phenomena will often have special nonstandard features. For example, a common characteristic of count data of rare species is that there are many more zeros than would be expected on the basis of the nonzero data (see Fig. 1 for an example). Indeed, depending on the scale and type of a given field study, this could be true of all classifications in Table 1 except those in cell 1 (i.e., common and widespread taxa that would not be considered rare). Data with many zeros

Manuscript received 30 March 2004; revised 17 June 2004; accepted 23 June 2004; final version received 21 July 2004. Corresponding Editor: A. M. Ellison. For reprints of this Special Feature, see footnote 1, p. 1079.

[1] E-mail: rbc@cres.anu.edu.au

TABLE 1. Table of species distributions and patterns of abundance classified by abundance, habitat specificity, and geographic range (modified from concepts outlined in Cody [1986], Rabinowitz et al. [1986], and New [2000]).

| Abundance of species within a community | Habitat specificity | Geographic range | Description |
|---|---|---|---|
| Common | common | widespread | widespread, occurs in a wide range of habitats and is abundant in those habitats (and therefore cannot be considered rare) |
| Common | common | restricted/localized | highly localized distribution but occurs across a range of habitats and is abundant in places where it occurs |
| Common | rare/specialized | widespread | widespread, but occurs in few habitats, and is common in places where it occurs |
| Common | rare/specialized | restricted/localized | highly localized distribution and occurs in few habitats, but is common in places where it occurs |
| Rare | common | widespread | widespread and occurs across a range of habitats but is scarce in places where it occurs |
| Rare | common | restricted/localized | highly localized distribution, occurs across a range of habitats but is scarce in places where it occurs |
| Rare | rare/specialized | widespread | widespread, but occurs in few habitats, and is scarce in places where it occurs |
| Rare | rare/specialized | restricted/localized | highly localized distribution, occurs in few habitats, and is scarce in places where it occurs |

pose statistical challenges because key distributional assumptions (e.g., normality, homoscedasticity, and others) are not fulfilled for standard statistical analyses.

Zero counts can arise in ecological data for two reasons; either they are inevitable, (known as structural or necessary zeros), which arise when presence is not tenable (an example might be the occurrence of the Lion [*Panthera leo*] in Australia), or they are random or accidental zeros, which arise due to sampling (e.g., see Green and Young 1993), where conditions are potentially suitable but absence is observed. There are many possible reasons why absence may be observed ranging from non-detection when present, to absent due to several unidentifiable factors, even though all identifiable factors would suggest presence. A simple example might be that of a target species being removed by a predator (Elton 1927, Kavanagh 1988). In most applications, there is usually insufficient knowledge to distinguish the two types of zeros.

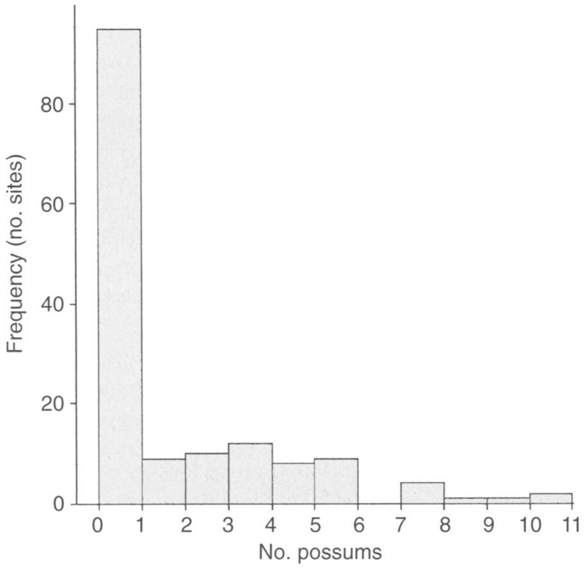

FIG. 1. Frequency graph of the abundance of Leadbeater's possum from 151 field survey sites, each measuring 3 ha in size (see Lindenmayer et al. [1991] for further details).

Statistical modeling provides a powerful framework for modeling presence and/or abundance in terms of a set of possible set of explanators or covariates (Morrison et al. 1992, Welsh et al. 1996, Burgman and Lindenmayer 1998). Such models are commonly used to identify important environmental variables that may explain patterns of distribution and abundance, to provide a concise description of the data, to allow the study of individual cases, and provide a means for constructing valid prediction intervals for new cases (Guisan and Zimmerman 2000). In recent years, there has been considerable activity by statistical scientists and others to develop methodologies for modeling count data with many zeros (Lambert 1992, Ridout and Demetrio 1992, Heilbron 1994, Welsh et al. 1996, Faddy 1998; M. S. Ridout, C. G. B. Demetrio, and J. Hinde, *unpublished manuscript*). These methods are not known or only poorly known by the majority of ecologists. Given this, in this paper we:

1) Briefly outline some of the current approaches to modeling count data with extra zeros. A more detailed review of that approach is given by Ridout et al. (M. S. Ridout, C. G. B. Demetrio, and J. Hinde, *unpublished manuscript*).

2) Provide two illustrations of the use of one of these methods—the two-part approach:

a) a statistical model for count data of Leadbeater's possum (*Gymnobelideus leadbeateri*), a species with restricted distribution, which has limited habitat specificity, but can be relatively common where it does occur.

b) an analysis of data arising from a long-term monitoring program of seabird populations (particularly Frigatebirds [*Fregata minor* and *F. ariel*] and the Red-footed Booby [*Sula sula*]) in the Coral Sea off north-eastern Australia.

3) Discuss some statistical estimation and inference issues that arise when studying rare species and outline some of the practical problems when attempting to model data which exhibit a very low frequency of occurrence.

## BRIEF OUTLINE OF APPROACHES FOR MODELING COUNT DATA WITH EXTRA ZEROS

In the section below, we consider several approaches to the relatively widespread problem of modeling data characterized by extra zeros and where extra-Poisson variation may be present. The more accessible approaches are: (1) use standard unimodal distributions for discrete data with extra dispersion; (2) use nonstandard mixture models to account for the extra zeros; and (3) employ two-part models. Another somewhat different approach, not presented here, has been developed and promoted by Faddy (1997), uses ideas associated with birth process models.

### *Standard unimodal distributions for discrete data with extra dispersion*

The baseline model for count data is the Poisson model. Allowing for extra-dispersion (i.e., variance > mean) in a standard Poisson model or fitting a negative binomial model may be a simple way of dealing with data with moderate numbers of extra zeros. These approaches are widely known and the Poisson regression model is a special case of the generalized linear model (McCullagh and Nelder 1989).

For illustrative purposes, let us assume we are dealing with counts of a given animal for a number of sites. Then from the baseline Poisson model, the predicted number of sites with no animals for a total of $n$ sites is $ne^{-\lambda(z)}$, where $\lambda(z)$ is the Poisson mean for the set of covariates $z$. However, the fit is usually poor because, as is usual in studies of rare species, there are typically many more observations with no animals than would be expected from this model.

The negative binomial distribution can be derived as a standard mixture of Poisson distributions. However as it only has only has one mode, as does the Poisson, it does not deal with the extra-zero problem even though it allows for extra dispersion. Other distributions arising as a mixture of Poisson distributions, such as the Neyman Type A (see Dobbie and Welsh 2001*a*, *b*) can have more than one mode, including a mode at zero, provide a more elegant solution to the problem.

### *Nonstandard mixture models to account for the extra zeros*

Another approach to the problem is to model the response variable as a mixture of a Bernoulli distribution and a Poisson or negative binomial distribution. For the Poisson case, this mixture model, with covariates, is defined as follows:

Given a response vector of counts $y_i$, $i = 1, \ldots, n$ are independent and

$$\begin{aligned} y_i &= 0 && \text{with probability } 1 - \pi(x) \\ y_i &\sim \text{Poisson } [\lambda(z)] && \text{with probability } \pi(x) \end{aligned}$$

so that

$$P(Y = 0 \mid x, z) = 1 - \pi(x) + \pi(x)e^{-\lambda(z)}$$

$$P(Y = r \mid x, z) = \frac{\pi(x)e^{-\lambda(z)}\lambda(z)^r}{r!} \qquad r = 1, 2, \ldots .$$

Here $\pi(x)$ is the probability that the number of animals on a site has a Poisson distribution and, given that the number of animals on a site has a Poisson distribution, $\lambda(z)$ is the mean number of animals on the site. Both $\pi$ and $\lambda$ may depend on the same, or possibly a different, set of covariates $x$ and $z$, respectively. This has become known as the zero-inflated (ZIP) Poisson model (Lambert 1992, Welsh et al. 1996). The simplest form of the model occurs if the covariates $x$ and $z$ coincide. Modification of the standard ZIP by replacing the Poisson distribution with a negative binomial distribution is relatively straightforward and is discussed by Lambert (1992) and Welsh et al. (1996).

ZIPs may provide insight into processes or mechanisms that may have generated the zero data i.e., distinguish between structural and random zeros. However, it will usually be unknown as to whether they distinguish the two types correctly.

### *Two-part, conditional models*

The previous models are based on single distributions, a mixture of distributions or, in the case of ZIPS, an extreme form of mixture. Parameters of the resulting distributions will usually not be independent of each other and so interpretation will be difficult. In the most interesting case, that is regression modeling, parameters of the distribution will typically depend on covariates. This makes interpretation even more complicated.

Here, we consider an alternative to the mixture of distributions idea. Consider a response that has two states: one in which no animals occur and another in which animals occur with varying levels of abundance. If we are only concerned about modeling state 1 (that is, whether any animals occur at a site), then linear logistic modeling is commonly applied. Given that animals are observed, the number of animals recorded can be modeled by a truncated discrete distribution such as the Poisson or negative binomial distribution (Grogger and Carson 1991). We refer to the complete model as the two-part, conditional model. In this case, the components are orthogonal and so the model separates processes that determine whether or not an animal is present from the processes determining the number of animals, given they are present. In the Poisson case, this is defined as follows:

Suppose the counts $y_i$, $i = 1, \ldots, n$ are independent and

$$y_i = 0 \qquad \text{with probability } 1 - p(x)$$

$$y_i \sim \text{truncated Poisson } [\lambda(z)] \qquad \text{with probability } p(x)$$

so that

$$P(Y = 0 \mid x) = 1 - p(x)$$

$$P(Y = r \mid x, z) = \frac{p(x)e^{-\lambda(z)}\lambda(z)^r}{r![1 - e^{-\lambda(z)}]} \qquad r = 1, 2, \ldots .$$

Here $p(x)$ is the probability of observing at least one animal on a site and, given that there is at least one animal, $\lambda(z)$ is the parameter of the truncated Poisson distribution which describes the number of animals observed.

Note that if we substitute $1 - \pi(x) + \pi(x)e^{-\lambda(z)}$ for $1 - p(x)$ in the above formula we have the ZIP parameterization. However, in the regression context, $\pi$ and $p$ are different parameters and so the two-part, conditional model and the ZIP model are not equivalent.

A major advantage of two-part, conditional models over ZIPs is that both model fitting and model interpretation of the components can be done separately. This also leads to computational advantages. These models treat both structural and random zeros together.

## CASE STUDIES IN MODELING OF COUNT DATA WITH EXTRA ZEROS USING TWO-PART MODELS

### *Case study #1. The relationships between the abundance of Leadbeater's possum (*G. leadbeateri*) and key habitat variables*

Our first case study highlights an application of two-part, conditional modeling for use in analysis of the habitat requirements of a rare species. The target taxon is Leadbeater's possum (*Gymnobelideus leadbeateri*), which is a rare and endangered species virtually restricted to the montane ash forests of the Central Highlands of Victoria, southeastern Australia (Lindenmayer et al. 1991). As the species is of considerable management concern because of its occurrence in some of the most valuable wood production forest in Australia (Lindenmayer 2000), it is important for conservation as well as ecological reasons to identify factors which explain occurrence and/or local population sizes. *G. leadbeateri* conforms to cell four (in Table 1) in terms of the type of rarity it exhibits. The species' distribution is highly restricted and virtually confined to the Central Highlands of Victoria—an area of 60 × 80 km. Within this region, particular structural and floristic conditions of wet montane ash eucalypt forests provide suitable habitat for *G. leadbeateri*. Where such suitable habitat does occur, it may support sets of loosely linked colonies of two to 12 animals (Lindenmayer 2000).

As outlined above, the study of species with rarity characteristics like *G. leadbeateri* (Rabinowitz et al. 1986; see Table 1) will often lead to the collection and analysis of data which consists of counts with a high frequency at zero (Gaston 1994). Indeed, a histogram of counts of *G. leadbeateri* from 151 survey sites showed there are many more zeros than would be expected from standard statistical distributions for count data (Fig. 1).

Two-part, conditional Poisson and negative binomial models, as well as a Poisson mixture model (ZIP), were fitted to the data gathered on *G. leadbeateri* (Welsh et al. 1996). Covariates considered included forest age, slope, aspect, tree canopy height, crown cover, a score for degree of decorticating bark, basal area of acacia, the number of shrubs, and the number of trees with hollows (log transformed) (see Lindenmayer et al. 1991 for further details). These data are given in the Appendix.

The selection of significant explanatory covariates ($P < 0.05$) for each model was undertaken by assessing the magnitude of the changes in deviance for both components, and the magnitude of the ratio of parameter estimates divided by their standard errors. These statistics are distributed (approximately) as chi-square and Student's *t*, respectively. Methods for model selection are well known (e.g., see Nicholls 1989, 1991) and have been described in detail by McCullagh and Nelder (1989).

The two-part conditional Poisson model for data on *G. leadbeateri* restricting the linear predictors to have common covariates was:

Component 1: $p = P\{\text{at least one animal present}\}$ and $\text{logit}(p) = -2.178 + 0.857$ lstags (SE = 0.251)

Component 2: $\lambda$ = mean abundance of animals (given presence) and $\log(\lambda) = 0.572 + 0.321$ lstags (SE = 0.103)

where lstags is the $\log_e$(number of trees with hollows on the site + 1).

As the parameters for this model are orthogonal, interpretation of the components of the model can be made separately. Thus, where animals occur, abundance increases by approximately 0.32% for a 1% increase in the number of trees with hollows and the odds of recording at least one animal at a site is increased by approximately 0.86% for a 1% increase in the number of trees with hollows.

Fitting the conditional negative binomial model to explore the possibility of extra-Poisson variation gave the following results:

Component 2: $\lambda$ = mean abundance of possums and $\log(\lambda) = 0.485 + 0.344$ lstags (SE $\approx$ 0.181).

Our estimate of the extra variation parameter was 0.143, with an estimated standard error of 0.161. Thus there was no evidence of extra-Poisson variation so the truncated Poisson model was considered appropriate.

The ZIP model for the data on *G. leadbeateri* restricting the linear predictors to have common covariates only was:

Component 1: $p = P\{\text{at least one animal present}\}$ and $\text{logit}(\pi) = -1.912 + 0.772$ lstags (SE = 0.257)

Component 2: $\lambda$ = mean abundance of animals and $\log(\lambda) = 0.573 + 0.320$ lstags (SE = 0.102).

The Poisson mean abundance for the state where animals occur, increases by 0.32% for a 1% increase in the number of trees with hollows, as for the truncated model. In addition, the odds of a site being classified as state 2 (the Poisson model) increase by 0.77% for each 1% change in the number of trees with hollows.

If the case is considered where the probability of absence does not depend on covariates (i.e., $\text{logit}(\pi) = 0.422$), then the parameters for the Poisson mean model are 0.453 and 0.37, respectively. This differs from the values given above where the probability of absence depends on the number of trees with hollows. Thus, for the ZIP model, the parameters must to be interpreted together, and hence interpretation can be difficult when the covariates affect $\pi$ and $\lambda$ differently. Combining the two components leads to a change in abundance, which is not simply proportional to lstags.

When the covariates are allowed to differ in the two components of the model the significant factors include lstags, basal area of wattle, slope, degree of decorticating bark and number of shrubs (Welsh et al. 1996). These results were similar to those obtained using the ZIP model.

Our models showed the species was most likely to occur on sites with numerous trees with hollows and a high basal area of *Acacia* spp. Trees with hollows contain nest sites for the species (Lindenmayer et al. 1991). Sap produced by *Acacia* spp. trees is an important food source for *G. leadbeateri* (Lindenmayer et al. 1994). Mean abundance of *G. leadbeateri* was highest on areas of flatter topography supporting numerous trees with hollows, few shrubs and large quantities of decorticating bark (Welsh et al. 1996). Thus, there were some major differences in the explanatory variables for the models for presence/absence and models for abundance, given presence.

### *Case study #2. Monitoring the abundance of the Frigatebird nests on North East Herald Cay*

In contrast to the previous case study on *G. leadbeateri* that focused on the application of two-part conditional modeling to habitat analysis, our second one below highlights the use of the same approach, but in the context of ecological monitoring as illustrated through ongoing work on seabird nesting.

Australia's Coral Sea Island Territory supports extensive seabird rookeries of great ecological significance, with 13 seabird species recorded breeding in the area. While some of these species such as the Red-footed Booby (*Sula sula*), Lesser Frigatebird (*Fregata ariel*), Great Frigatebird (*Fregata minor*), and Red-tailed Tropicbird (*Phaethon rubricauda*) have an extensive distribution outside of Australian waters, they are uncommon within Australia. The islands and cays of the Coral Sea are important in that they contain a significant proportion of the region's breeding populations (Baker et al. 2004). The species of seabirds conform to cell 3 in Table 1. That is, they are widespread but they have very specialized nesting requirements. However, where they do occur, the numbers of birds can be large.

Given the status of Frigatebirds and other seabirds within Australian waters, a long-term monitoring program of nesting success was established (Baker et al. 2004). One of the areas chosen for study is the Coringa-Herald National Nature Reserve on North East Herald Island where 11 transects set at 100-m intervals were established in 1992. For each transect, quadrats measuring 10 × 10 m were marked. A total of 415 quadrats was established across all habitats, and vegetation mapped for all transects and quadrats. The monitoring program has been designed to allow detection of "shifts" in nesting patterns on the island, and permits estimates of change in nest density from year to year (see Welsh et al. 2000).

An essential step in estimating the number of nests constructed by different species of seabirds on North East Herald Cay in each year is to relate the nest counts to other relevant variables such as the number of nests in the previous year, the transect identity, and the quadrat number. The data on seabird nests contains a large number of quadrats with zero counts and is possibly also characterized by extra dispersion (Fig. 2). A two-part, conditional model based on the truncated negative binomial distribution accounted for both the extra zeros and possible extra-Poisson variation was found to be compatible with the data. The key covariate was the number of nests in the previous year. Welsh et al. (2000) provides a detailed account of the complete analysis.

Given that the seabird nest data were collected in a regular spatial pattern, it is possible that counts were spatially correlated. In the initial analysis, this problem was dealt with in a sequential way by examining residuals for spatial dependence. Since then, Dobbie and

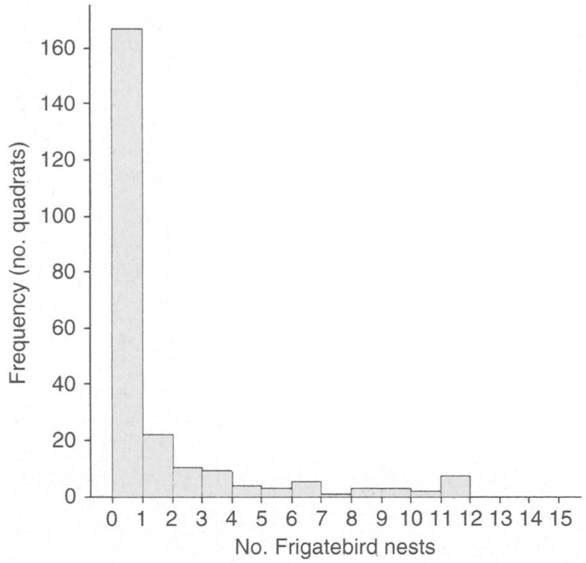

FIG. 2. Frequency graph of the number of Frigatebird (*F. ariel* and *F. minor*) nests on 236 100-m$^2$ quadrats established in the Coringa-Herald Nature Reserve on North East Herald Island in the Coral Sea.

Welsh (2001*a*) have developed direct methods for modeling spatially correlated zero-inflated count data using the conditional, two-part approach.

## SOME OTHER STATISTICAL ISSUES IN THE DESIGN AND ANALYSIS OF STUDIES OF THE OCCURRENCE OF RARE SPECIES

It is well known that when detection probabilities of a species are very low, the ability to be able to establish the statistical significance of an effect tends to be low. That is, there is low statistical power. Fig. 3 (Nicholls and Cunningham 1995) shows the effect of increasing rarity (i.e., low probability of detection) on the standard error of log (relative risk) while maintaining a constant sample size. Here relative risk is a measure of the extent to which a site having a particular attribute is more (or less) likely to have a species present than a site without the attribute.

The effect is that for data on rare species, a change in a factor of interest (e.g., a "treatment") may result in a large relative change in odds of a given species being present, but this may not translate into a statistically significant effect. This is particularly evident as occurrence falls below 5%. Nicholls and Cunningham [1995] provide an example in the context of predicting the distribution of the koala and give details of these calculations.

In many instances when data are scant, particularly in terms of a very low frequency of occurrences (e.g., <5%), numerical computation problems arise in model estimation and the fitting of data. This arises because, on scales appropriate for analysis such as the linear logit scale, an attempt is being made to estimate parameters that are extremely large and negative, i.e., zero, on the natural scale. In essence, there is insufficient information for estimation and this is reflected in large standard errors and/or computational difficulties.

## DISCUSSION

Although the various types of rarity have been well discussed in ecology (Rabinowitz et al. 1986; see Table 1), the term "rarity" is used very loosely in the ecological literature. For example, Gaston (1994) listed many studies where the concept of rarity was used but its definition was different in almost all cases. From a statistical perspective, the problems and definitions of rarity and rare species need to be clearly formulated and stated so that progress can be made in solving relevant problems in ecology. This paper has discussed some approaches to modeling data having an excess of zeros; data that commonly arise in studies of rare and uncommon species. Recent advances in statistical methods, as briefly illustrated and discussed in this paper, can assist in the modeling of data on rare and uncommon species, thus providing a powerful, general framework for estimation and inference. Moreover, we believe that, depending on the scale and type of field study in question and the biology of the taxon in question, the modeling methods outlined in this paper would have potentially useful application for a wide range of the types of rare species discussed in detail by Rabinowitz et al. (1986) and outlined in the various cells in Table 1. Such flexibility is illustrated in this paper by both the differences in type of rarity between *G. leadbeateri* and *F. ariel* and *F. minor* and the different applications of modeling to data gathered for them—habitat analysis (i.e., comparative inference) vs. estimating population size (i.e., point estimation)

The methods outlined in this paper for the study of rare and uncommon species are increasingly important for two reasons. First, they extend the number of species in an assemblage which can be subject to the same

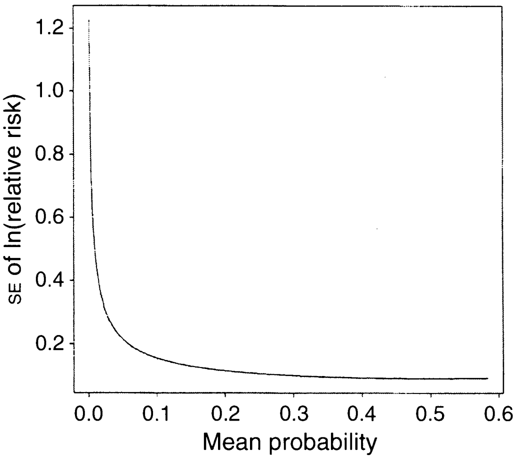

FIG. 3. Relationship between the standard error of ln(relative risk) and rarity.

sound, and flexible statistical analyses available for taxa characterized by extensive presence and abundance data. This is valuable because uncommon and rare taxa can comprise a substantial proportion of the species in most assemblages. In this context, flexible statistical analyses may help quantify what makes some species rare (e.g., quantifying habitat conditions that are themselves rare; as in the case of *G. leadbeateri*) which could make a valuable contribution to biodiversity conservation. Moreover, resource management practices are most often concerned with the conservation and persistence of rare and uncommon species. Hence, for example, the methods outlined here may assist in the quantification of the effects of an experimental treatment or environmental impact on rare or uncommon species.

A second reason why the statistical methods outlined in this paper may be important is that some approaches for modeling zero-inflated data may provide insight into ecological processes/mechanisms, which may have generated the data. For example, through being able to separately model presence-only and conditional abundance (given presence) components of count data, it may (in some circumstances) be possible to identify separate factors affecting the size of a colony of a rare species, given presence, from those factors affecting presence of a species. This was certainly the case for *G. leadbeateri* where several explanatory variables were important in accounting for variation in colony size once conditions (nest trees) are suitable for the presence of the species on a site (Lindenmayer 1996). Indeed, recent efforts in forest management have (1) established a zoning system to ensure that areas that are suitable for the presence of the species are not logged (Macfarlane et al. 1998) and, (2) aimed to harvest forests in new and more environmentally sensitive ways that recreate forest structural conditions which maximize on-site population sizes (D. B. Lindenmayer and R. B. Cunningham, *unpublished data*). These new silvicultural methods are targeted toward flat terrain where it is more operationally feasible to do them and also where our two-part modeling work has shown that population sizes of *G. leadbeateri* will be higher if other suitable habitat attributes can be maintained or created.

Both ZIPs and two-part conditional models tend to be sensitive to the choice of underlying distribution. Gurmu (1997) has suggested some robust, semi-parametric alternatives. Where the nonzero part is not easily modeled by a specific probability distribution, a standard ordinal regression model (see McCullagh 1980) may be a suitable alternative. However, in that case, the covariates for the zero and non-zero parts of the model must be the same.

Our final comment relates to the importance of interdisciplinary work in scientific research. We believe the collaborative approach we have taken to our work, whereby we have combined expertise in ecology and statistics has resulted in outcomes (both in quality and quantity) that exceed the sum of what could have been achieved individually. Such collaborative approaches are increasingly important because both the science of ecology and the science of statistics are extensive and complex and are changing rapidly making it impossible for any one person to keep abreast of new innovations in both fields. Perhaps this is neatly illustrated by the early origins of the ZIP regression approaches applied in the ecological case studies in this paper. The methods were originally developed for problems in econometrics and manufacturing and appeared in a literature read by few (if any) ecologists.

## ACKNOWLEDGMENTS

The authors thank A. Welsh, M. Dobbie, A. O. Nicholls, and C. Donnelly for past collaborative work on which this paper was based. We also thank A. Ellison for encouraging this paper to be written and coordinating a compilation of studies on the statistics of rarity. Comments by A. Ellison, A. O. Nicholls, and an anonymous referee greatly improved earlier versions of this paper.

## LITERATURE CITED

Baker, G. B., M. Double, M. Holdsworth, and M. Hallam. 2004. Seabird monitoring program—Coral Sea Island territory. Report on 2003 field season and update of Herald Cays longitudinal data sets for the period 1992 to 2003. Unpublished report to Department of Environment and Heritage, Canberra, Australia.

Burgman, M. A., and D. B. Lindenmayer. 1998. Conservation biology for the Australian environment. Surrey Beatty, and Sons, Chipping Norton, Sydney, Australia.

Cody, M. L. 1986. Diversity, rarity, and conservation in Mediterranean climate regions. Pages 122–152 *in* M. Soulé, editor. Conservation biology: the science of scarcity, and diversity. Sinauer Associates, Sunderland, Massachusetts, USA.

Dobbie, M. J., and A. H. Welsh. 2001*a*. Modeling correlated zero-inflated count data. Australian and New Zealand Journal of Statistics **43**:431–444.

Dobbie, M. J., and A. H. Welsh. 2001*b*. Models for zero-inflated count data using the Neyman type A distribution. Statistical Modeling **1**:65–80.

Elton, C. S. 1927. Animal ecology. Methuen, London, UK.

Faddy, M. 1997. Extended Poisson process modeling, and analysis of count data. Biometrical Journal **39**:431–440.

Faddy, M. 1998. Stochastic models for analysis of species abundance data. Pages 33–40 *in* D. J. Fletcher, L. Kavalieris, and B. F. Manly, editors. Statistics in ecology, and environmental monitoring 2: decision-making, and risk assessment in biology. University of Otago Press, Dunedin, New Zealand.

Fagan, W. F., P. J. Unmack, C. Burgess, and W. L. Minckley. 2002. Rarity, fragmentation, and extinction risk in desert fishes. Ecology **83**:3250–3256.

Gaston, K. J. 1994. Rarity. Chapman and Hall, London, UK.

Green, R. H., and R. C. Young. 1993. Sampling to detect rare species. Ecological Applications **3**:351–366.

Grogger, J., and R. Carson. 1991. Models for truncated counts. Journal of Applied Econometrics **6**:225–238.

Guisan, A., and N. E. Zimmerman. 2000. Predictive habitat distribution models in ecology. Ecological Modelling **135**: 147–186.

Gurmu, S. 1997. Semi-parametric estimation of hurdle regression models with application to medical utilization. Journal of Applied Econometrics **12**:225–268.

Hartley, S., and W. E. Kunin. 2003. Scale dependency of rarity, extinction risk, and conservation priority. Conservation Biology **17**:1559–1570.

Heilbron, D. C. 1994. Zero-altered, and other regression models for count data with added zeros. Biometrical Journal **36**:531–547.

Kavanagh, R. P. 1988. The impact of predation by the Powerful Owl *Ninox strenua* on a population of the Greater Glider *Petauroides volans*. Australian Journal of Ecology **13**:445–450.

Lambert, D. 1992. Zero-inflated Poisson regression, with an application to defects in manufacturing. Technometrics **34**: 1–14.

Lindenmayer, D. B. 1996. Wildlife and woodchips: Leadbeater's possum as a testcase of sustainable forestry. University of New South Wales Press, Sydney, Australia.

Lindenmayer, D. B. 2000. Factors at multiple scales affecting distribution patterns, and its implications for animal conservation—Leadbeater's possum as a case study. Biodiversity and Conservation **9**:15–35.

Lindenmayer, D. B., R. B. Cunningham, and C. F. Donnelly. 1994. The conservation of arboreal marsupials in the montane ash forests of the Central Highlands of Victoria, southeastern Australia. VI. Tests of the performance of models of nest tree, and habitat requirements of arboreal marsupials. Biological Conservation **70**:143–147.

Lindenmayer, D. B., R. B. Cunningham, M. T. Tanton, H. A. Nix, and A. P. Smith. 1991. The conservation of arboreal marsupials in the montane ash forests of the Central Highlands of Victoria, south-east Australia. III. The habitat requirements of Leadbeater's possum, *Gymnobelideus leadbeateri* McCoy, and models of the diversity, and abundance of arboreal marsupials. Biological Conservation **56**:295–315.

MacArthur, R. H., and E. O. Wilson. 1967. The theory of island biogeography. Princeton University Press, Princeton, New Jersey, USA.

Macfarlane, M. A., J. Smith, and K. Lowe. 1998. Leadbeater's possum recovery plan, 1998–2002. Department of Natural Resources, and Environment, Government of Victoria, Melbourne, Australia.

McCullagh, P. 1980. Regression models for ordinal data. Journal of the Royal Statistical Society, Series B **42**:109–142.

McCullagh, P., and J. A. Nelder. 1989. Generalized linear models. Chapman, and Hall, London, UK.

Meffe, G., and C. R. Carroll. 1997. Principles of conservation biology. Sinauer, Sunderland, Massachusetts, USA.

Morrison, M. L., B. G. Marcot, and R. W. Mannan. 1992. Wildlife habitat relationships: concepts, and applications. University of Wisconsin Press, Madison, Wisconsin, USA.

New, T. R. 2000. Conservation biology. An introduction for southern Australia. Oxford University Press, Melbourne, Australia.

Nicholls, A. O. 1989. How to make biological surveys go further with generalised linear models. Biological Conservation **50**:51–75.

Nicholls, A. O. 1991. Examples of the use of generalised linear models in analysis of survey data for conservation evaluation. Pages 54–63 *in* C. R. Margules and M. P. Austin, editors. Nature conservation: cost effective biological surveys, and data analysis. CSIRO Australia, Melbourne, Australia.

Nicholls, A. O., and R. B. Cunningham. 1995. Strategies for evaluating wildlife management. Koala Conservation in the south-east forests. *In* S. Cork, S. Feary, and C. Mackowski, editors. Proceedings of an expert workshop. National Parks and Wildlife Service, Sydney, NSW, Australia.

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

Rabinowitz, D., S. Cairns, and T. Dillon. 1986. Seven forms of rarity, and their frequency in the flora of the British Isles. Pages 184–204 *in* M. E. Soulé, editor. Conservation biology: the science of scarcity, and diversity. Sinauer Associates, Sunderland, Massachusetts, USA.

Ridout, M. S., and C. G. B. Demetrio. 1992. Generalised linear models for positive count data. Revista de Metematica s Estatistica **10**:139–148.

Welsh, A. H., R. B. Cunningham, and R. L. Chambers. 2000. Methodology for estimating the abundance of rare animals: seabird nesting on North East Herald Cay. Biometrics **56**: 22–30.

Welsh, A. H., R. B. Cunningham, C. F. Donnelly, and D. B. Lindenmayer. 1996. Modeling the abundance of rare animals: statistical models for counts with extra zeros. Ecological Modeling **88**:297–308.

Williams, C. B. 1944. Some applications of the logarithmic series, and the index of diversity to ecological problems. Journal of Ecology **32**:1–44.

## APPENDIX

A table showing Leadbeater's possum abundance data and a selection of habitat covariates is presented in ESA's Electronic Data Archive: *Ecological Archives* E086-061-A1.