Computational Ecology and Software, 2011, 1(4):218-223

Article

Estimating survival rates in ecological studies with small unbalanced sample sizes: an alternative Bayesian point estimator

Christian Damgaard[1], Adeline Fayolle[2]
[1]Department of Bioscience, Aarhus University, Vejlsøvej 25. 8600 Silkeborg, Denmark
[2]Cirad, Environments and Societies Department, 'Goods and Services of Forest Ecosystems' Research Unit, Campus International de Baillarguet, TA 10C, BP 5035, Montpellier, 34035, France
E-mail: cfd@dmu.dk

Received 18 August 2011; *Accepted 23 September 2011*; *Published online 1 December 2011*
IAEES

Abstract
Increasingly, the survival rates in experimental ecology are presented using odds ratios or log response ratios, but the use of ratio metrics has a problem when all the individuals have either died or survived in only one replicate. In the empirical ecological literature, the problem often has been ignored or circumvented by different, more or less *ad hoc* approaches. Here, it is argued that the best summary statistic for communicating ecological results of frequency data in studies with small unbalanced samples may be the mean of the posterior distribution of the survival rate. The developed approach may be particularly useful when effect size indexes, such as odds ratios, are needed to compare frequency data between treatments, sites or studies.

Keywords frequency data; binomial distribution; small sample size; unbalanced design; effect size index; log response ratio.

1 Introduction

Individual survival as a vital rate is an important variable that often is measured in empirical studies in ecology or conservation biology, and, increasingly, the survival rates in experimental ecology are presented using odds ratios or log response ratios, which have been found to be valuable for communicating ecological results (Hedges et al., 1999). For example, ratio metrics are often used in studies of plant-plant interactions to quantify the proportionate change of the experimental manipulation, where a treatment effect (with interaction) is compared to a control (without interaction, removed vegetation) (e.g. Liancourt et al., 2005; Violle et al., 2006; Fayolle et al., 2009). Additionally, in a meta-analysis context, the ratio metrics allow comparing results of independent experiments (Hedges et al., 1999).

In most studies, the stochastic process of survival is assumed to be a binomially distributed random variable, and the survival rate of a population is usually reported by the maximum likelihood estimate, $\hat{p} = x/n$, where x is the number of survivors, and n is number of individuals at a fixed starting point.

In this paper, we focus on the use of ratio metrics with frequency data and the special problem that often arises when the response between the control and the treatment is extreme - that is all the individuals either died or survived. In the empirical ecological literature, the problem often has been ignored or circumvented by different, more or less *ad hoc* approaches. For example, in a meta-analysis concerning plant-plant interactions in arid environments, Maestre et al. (2005) used the odds ratio of survival between treated and control plants.

These authors decided to add one to the number of surviving individuals in order to avoid values that would require division by 0, whereas Hyatt et al. (2003) decided to add 0.5 to the number of surviving seeds and seedlings in their meta-analysis of the relationship between survival and parental distance. In other studies, ratios of zero are replaced by 1/(2N) and ratios of one are replaced by 1-1/(2N), which is an adjustment that was originally proposed by Berkson (1944). Other commonly used, and highly criticisable, practices are to aggregate data (in order to reduce the likelihood of observations of zero or one) or even to throw data away.

Additionally, the survival rate has some undesirable properties even without calculating ratios. i) Ecological studies may have relatively few replicates. This means that the uncertainty of the survival probability is relatively large, or, using Bayesian terminology, the density of the posterior distribution of the survival probability is relatively wide. Since the shape of the density of the posterior distribution is generally asymmetric, the maximum likelihood estimate is not necessarily the most informative point estimate of the posterior distribution, especially when either no or all individuals survive. ii) Ecological studies may be unbalanced. It is common to compare survival between treatments that differ in the number of individuals in the beginning because of uncontrolled differences in germination success or population size. For example, when comparing zero survivors out of ten with zero survivors out of five, our belief of a low survival is stronger in the case where ten individuals are followed, but this notion is not reflected in the maximum likelihood estimates, which both are zero.

The aim of this paper is to examine frequency data in small ecological samples from a Bayesian perspective in an attempt to rationalise the previous *ad hoc* approaches to avoid the problem of calculating ratios of sparse frequency data, where sometimes either no or all individuals survive. The discussion of sparse frequency data dates back to Laplace in the 18th century and the presented results are not new, but since the results have not been used by ecologists, we find it appropriate to discuss them again in this paper. It is argued that it may be preferable to estimate the probability parameter by the mean of the posterior distribution of the probability rather than the maximum likelihood estimate, when a point estimate of the probability is needed in cases with unbalanced and small sample sizes. The mean of the posterior distribution of a binomially distributed variable is not more complicated to calculate compared to the maximum likelihood estimate, and it does not suffer from the above-mentioned problems in studies of frequency data with small sample sizes, unbalanced designs and when effect size indexes are needed. It is important to note that the focus of this paper is only on the most relevant point estimate of frequency data and *not* on the representation of statistical uncertainty or testing of hypotheses.

2 A Bayesian Approach

The Bayesian posterior distribution of the probability parameter p in the binomial distribution given the number of individuals that fulfil a certain criteria (e.g. alive or germinated), x, out of a total number of individuals, n, is calculated as:

$$\pi(p \mid x,n) = \frac{l(x,n \mid p)\ \pi(p)}{\int l(x,n \mid p)\ \pi(p)\ dp} \qquad (1),$$

where $l(x,n \mid p)$ is the likelihood function of the binomial distribution, and $\pi(p)$ is the prior distribution of the probability p. A convenient prior distribution for probabilities is the beta distribution: $\pi(p) = (1-p)^{\beta-1}\ p^{\alpha-1} / \text{Beta}(\alpha,\beta)$, where α, $\beta > 0$, which is a flexible distribution in the domain between zero and one. If the beta distribution is chosen as the prior distribution, then the mean of the posterior distribution is:

$$\bar{p} = E\left[\pi(p \mid x, n)\right] = \frac{x+\alpha}{n+\alpha+\beta} \quad (2),$$

(Chew, 1971) and in the simplifying case of an uniformly distributed (uninformative) prior distribution, i.e. $\alpha = \beta = 1$, then

$$\bar{p} = \frac{x+1}{n+2} \quad (3).$$

The posterior distribution depends on the assumed prior distribution, and consequently, it is critical to find objective, or at least rational, methods for specifying the prior distribution of the frequency data; otherwise the criticised *ad hoc* approaches have only been replaced by different *ad hoc* approaches. The specification of prior distributions in Bayesian methods has received a lot of attention (e.g. Chew, 1971; Carlin and Louis, 1996; Stafford and Lloyd, 2011) and it is outside the scope of this paper to review this topic. Generally, we suggest to use the uninformative prior, i.e. the beta distribution with $\alpha = \beta = 1$, in cases when there is no prior information on the survival rates at the specific environment. Otherwise, if data from previous experiments or observations exist, we suggest to fit such data to the beta-distribution and, thus, to integrate existing knowledge into the prior distribution.

The use of the maximum likelihood estimate as the summary statistic of the probability is motivated by its superior asymptotic estimation properties, i.e. $\hat{p}$ is the minimum variance unbiased estimator of the probability parameter in the binomial distribution (Johnson et al., 1993). In contrast, the mean of the posterior distribution, $\bar{p}$, is a biased, although consistent, (i.e. the bias becomes negligible for large sample sizes) estimator of the probability parameter *p* in the binomial distribution. In the case of an uninformative prior distribution, the bias is $\bar{p} - \hat{p} = (1 - 2\hat{p})/(n+2)$. However, the reason for calculating summary statistics in ecological studies is to provide a good representation of our knowledge of the process, and it is not necessarily the summary statistic with the best asymptotic estimation properties that is best at communicating our knowledge.

3 Motivating Example

Imagine two plots, where the survival of five and ten individuals, respectively, is followed. In both plots, there are zero survivors. The densities of the posterior distribution of the survival probability, assuming that the number of survivors is binomially distributed and there is no prior information, are shown in Fig. 1. The density of the posterior distribution may be interpreted as our ignorance or uncertainty about the probability of survival after we have made our observations (Clark, 2007). Clearly, we are more certain that the survival probability is low in the plot with ten individuals compared to the plot with five individuals, and this knowledge is reflected in the shape of the densities of the posterior distribution, i.e., the density in the case of $n = 5$ is more flat and, consequently, has a relatively larger variance compared to $n = 10$. However, this information is not reflected in the maximum likelihood estimate of the survival probability, which is zero in both cases. On the other hand, the means of the densities of the posterior distribution of the survival probability are $1/12$ and $1/7$, respectively, which is a better representation of the difference between the two

densities and, consequently, our uncertainty of the survival probability. Furthermore, note that the maximum likelihood estimates of the survival probability have relatively little support in the data, i.e. the mode of the asymmetric density at $\hat{p} = 0$ is a poor representation of the survival data as represented by their posterior distributions.

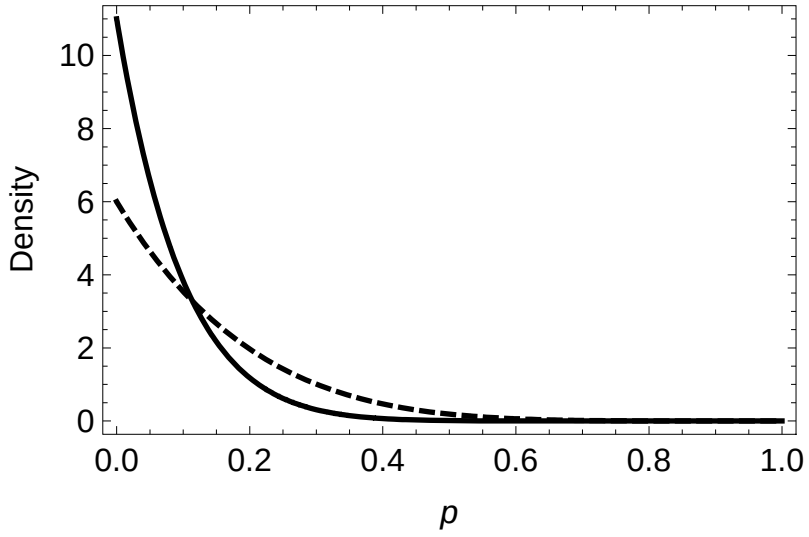

Fig. 1 The density of the posterior distribution of the survival probability (the parameter *p*) for *x* = 0 (zero survivors) when *n* = 10 (full line) and *n* = 5 (dashed line), assuming an uninformative prior distribution ($\alpha = \beta = 1$). The maximum likelihood estimates of the two cases are equal ($\hat{p}(x = 0, n = 10) = \hat{p}(x = 0, n = 5) = 0$), whereas the means of the posterior distributions differ, ($\overline{p}(x = 0, n = 10) = 1/12 < \overline{p}(x = 0, n = 5) = 1/7$), in accordance with our increased belief in a low survival in the case of ten individuals. Note that the variance of the posterior distribution is larger in the case of *n* = 5 compared to *n* = 10, which reflects our relatively larger uncertainty of the survival probability.

4 Discussion

In the sometimes difficult task of communicating complicated ecological results, it is often desirable to construct indices or graphs that summarise key ecological messages, and for this, point estimates are indispensible. For example, a common research question in ecological studies is to compare survival between two treatments along an environmental gradient or through time. In order to account for the effect of unexplained spatial heterogeneity in a natural environment, the study is often arranged as a number of randomly positioned paired plots, which may be unbalanced and have relatively small sample sizes. A popular choice for depicting and analysing the effects of a treatment is to quantify the proportionate change that result from the experimental treatment by odds ratios or log response ratios. The latter ratio, also called lnRR index, is the logarithm of the ratio of the mean response between the modified and control treatment (Hedges et al., 1999). The lnRR index was developed in a meta-analyse context and is now widely used in the plant interaction literature (see Armas et al. (2004), and Oksanen et al. (2006) for a comparison of commonly used index quantifying plant interactions). The lnRR index presents two desirable properties, which make it relatively easy to analyse the effects of the treatment using standard statistical methods: (1) the lnRR index linearise the metric, the log ratio is affected equally by changes in either nominator or denominator, which means symmetry for changes in either the modified or the control treatments and (2) normalizes the sampling distribution, which is originally skewed (Hedges et al., 1999). Furthermore, the key ecological messages are

easily displayed in graphs, where log (1) = 0 is the point of interest corresponding to no difference between the modified and control treatment. However, a drawback of the index is that it is not defined for an extreme response, such as no survivors, in either the modified or control treatment. Our argument is that, when constructing such indices, it is better to summarise our uncertainty of the probability parameter after we have made our observations by the mean of the Bayesian posterior distribution rather than summarising the probability parameter by the maximum likelihood estimator. This allows a fair comparison of subplots with an unequal and low number of replicates and avoids the practical problem of undefined indices due to zero values, which have relatively little support by the data. Furthermore, when constructing indices or graphs using the mean of the posterior distribution of the probability, it is possible to include additional information in the prior distribution. For example, if background mortality is known to vary along an environmental gradient, then this knowledge may be integrated in the prior distribution. This integration of knowledge into the index will enable a more comprehensive graphical representation of the data, which, in some cases, may be useful when communicating ecological results.

The question of the relevant point estimator for the probability parameter in the binomial distribution is not a new problem. The problem has been studied since the 18th century, where it was introduced by Pierre-Simon Laplace as “the sunrise problem”, i.e. "What is the probability that the sun will rise tomorrow?". Laplace’s solution was identical to expression (3), although he derived it differently. Nevertheless, the solutions to the problem are generally unknown to ecologists, who consistently use the maximum likelihood estimator as the point estimator of the probability parameter, even at the price of discarding data, without realising that meaningful alternatives exist.

In the introduction, it was stressed that the focus of this paper is only on the most relevant point estimate of frequency data, and it is essential to keep in mind that the amount of statistical uncertainty is not reflected in any point estimate. In statistical analysis of frequency data, it is essential to utilize the collective properties of the assumed binomial distribution, for example by using generalized linear models (McCullagh and Nelder, 1999) and the present discussion on point estimates is irrelevant.

Acknowledgements

Thanks to Phillip Dixon for commenting on a previous version of this paper.

References

Armas C, Ordiales R, Pugnaire FI. 2004. Measuring plant interactions: a new comparative index. Ecology, 85: 2682-2886

Berkson J. 1944. Application of the logistic function to bio-assay. Journal of the American Statistical Association, 39: 357-365

Carlin BP, Louis TA. 1996. Bayes and Empirical Bayes Methods for Data Analysis. Chapman & Hall, London, UK

Chew V. 1971. Point estimation of the parameter of the binomial distribution. American Statistician, 25: 47-50

Clark JS. 2007. Models for Ecological Data. Princeton University Press, Princeton, USA

Fayolle A, Violle C, Navas ML. 2009. Differential impacts of plant interactions on herbaceous species recruitment: disentangling factors controlling emergence, survival and growth of seedlings. Oecologia, 159: 817-825

Hedges LV, Gurevitch J, Curtis PS. 1999. The meta-analysis of response ratios in experimental ecology. Ecology, 80: 1150-1156

Hyatt LA, Rosenberg MS, Howard TG, et al. 2003. The distance dependence prediction of the Janzen-Connell hypothesis: a meta-analysis. Oikos, 103: 590–602

Johnson NL, Kotz S, Kemp AW. 1993. Univariate Discrete Distributions. John Wiley, New York, USA

Liancourt P, Callaway RM, Michalet R. 2005. Stress tolerance and competitive-response ability determine the outcome of biotic interactions. Ecology, 86: 1611-1618

Maestre FT, Valladares F, Reynolds JF. 2005. Is the change of plant–plant interactions with abiotic stress predictable? A meta-analysis of field results in arid environments. Journal of Ecology, 93: 748–757

McCullagh P, Nelder JA. 1999. Generalized Linear Models. CRC, Boca Raton, USA

Oksanen L, Sammul M, Mägi M. 2006. On the indices of plant-plant competition and their pitfalls. Oikos, 112: 149-155

Stafford R, Lloyd JR. 2011. Evaluating a Bayesian approach to improve accuracy of individual photographic identification methods using ecological distribution data. Computational Ecology and Software, 1(1): 49-54

Violle C, Richarte J, Navas ML. 2006. Effects of litter and standing biomass on growth and reproduction of two annual species in a Mediterranean old-field. Journal of Ecology, 94: 196-205