SHORT COMMUNICATION

Calculating the probability of absence using the Credible Interval Value

Graham B. McBride[1,*] and Peter Johnstone[2]

[1]NIWA, PO Box 11 115, Hamilton 3216, New Zealand
[2]AgResearch, Invermay Agricultural Centre, Private Bag 50 034, Mosgiel 9053, New Zealand
[*]Author for correspondence (Email: g.mcbride@niwa.co.nz)

Abstract: A common question that arises when considering the results from a well-designed sampling programme for a rare or invasive species is: 'Sampling has failed to detect a species that could have been present, so can we calculate the probability that it truly was absent during the sampling period?' Noting that this invokes a Bayesian view of 'probability', which therefore must be accepted if the question is to be answered in the affirmative, we present a method of addressing it.

Keywords: absence; detection; disease; pests; probability

Introduction

Policy developers commonly express the need to be informed about a measure of probability that an invasive species was actually absent in some domain during a sampling period, given that sampling has failed to detect it (e.g. *Didymosphenia geminata* in the North Island of New Zealand; Biosecurity New Zealand 2009). In probability theory this is a Bayesian question in that it seeks to make a probability statement about a hypothetical state-of-nature given the data obtained.

Methods

Without loss of generality we keep matters simple by assuming that: (1) sampling methods have perfect sensitivity and specificity (no false negatives or false positives), and (2) sampling is random within the space-time domain of interest (i.e. in those parts of the domain where conditions could allow for the species' presence). The perfect sensitivity assumption can be relaxed, as discussed later.

The essence of Bayesian methods in this context can be stated as using new data to update an investigator's prior belief about the species' absence once new data come to hand. That updating uses Bayes' Rule, and the result takes the form of a 'posterior probability'. While the validity of the Rule is well accepted, its outcome in the context of hypothesis assessment is less so—posterior probabilities depend to some extent on the investigator's prior belief and so may vary between investigators. Nevertheless we suggest that if one uses a standard or 'reference' prior (Lee 1997), progress can be made. Also, we note that the influence of the prior decreases with the number of sample data, such that the data 'begin to speak for themselves'.

We propose the use of the Bayesian Credible Interval Value (*CIV*), which permits an inference statement of the form 'The probability is X that the species is present at less than p^* of all possible sites in the sampled domain, given that sampling in that domain has failed to detect it'. In this statement the value X is the calculated *CIV*, and the value p^* is a value that must be decided externally (e.g. by a committee of managers and scientist). The value of p^* needs to be 'small' (e.g. 0.01), but it cannot be zero (absolute guarantees of absence cannot be inferred).

Mathematical details of the derivation of *CIV* formulae are given in the Appendix. The key issue is the appropriate choice of the two parameters (α and β) of the beta prior distribution.

Results and Discussion

On choosing $p^* = 0.01$ we obtain the results shown on Fig. 1, in which two well-known alternative choices of a 'reference' beta prior distribution are shown: uniform ($\alpha = \beta = 1$) and Jeffreys' ($\alpha = \beta = \frac{1}{2}$). So for 100 samples all showing absence, under the uniform prior we could say with 64% confidence the species is present at less than 1% of all possible sites in the sampled domain, given that sampling in that domain has failed to detect it. Under the Jeffreys' prior that confidence rises to 84%. To achieve 99% confidence for this statement would require at least 458 sampled sites (for the uniform prior) and 330 sites (for Jeffreys' prior). So which prior should be used?

The former prior is invoked by 'Bayes' postulate' (Lee 1997), which is often proposed for situations of 'complete ignorance'. However, for two reasons, we propose that Jeffreys' prior be preferred: (1) because it has desirable 'invariance under scale transformation' properties that the uniform prior does not share (Lee 1997), and (2) because few would agree with the notion that, under a uniform prior, prior to any new data being obtained an investigator would hold that *any* value of prevalence and would be equally likely. Note however that Jeffreys' prior is 'cup-shaped', such that very low *or* very high prevalence values are held to be much more likely than more moderate values. But as site data come to hand, all failing to detect the species, the influence of the high end of the prior is quickly quenched. This seems a very desirable feature. It has been used to justify the choice of Jeffreys' prior for 'Confidence of Compliance' with percentile standards for environmental waters and for drinking waters (McBride & Ellis 2001; McBride 2005). Any choice of a prior other than
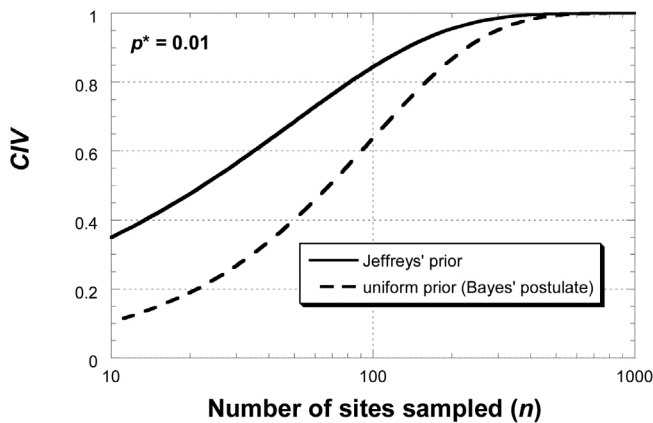
Figure 1. *CIV* (probability of absence, given failure to detect) versus sampling effort for prevalence limit *p** = 0.01.

uniform or Jeffreys' (e.g. a 'hockey-stick' beta distribution, with $\beta \gg \alpha$) may invite the very arbitrariness that can cause discomfort with Bayesian methods, although we note that in some circumstances a 'weakly informative prior' may be desirable (Gelman 2006).

　　Finally, the *CIV* can be simply modified to account for imperfect sensitivity of the sampling method, in that it may occasionally fail to detect the species when present in a sample. That is the assumption of perfect sensitivity can be relaxed. Denoting that probability as *q* (*q* < 1) the *CIV* can be approximately obtained by dividing *p** by *q*.

　　In proposing this approach it must be noted that the assumption of an adequate sample design must always be checked. The results of a *CIV* will be misleading were that programme to contain avoidable bias.

Acknowledgements

References

Biosecurity New Zealand 2009. Didymo *Didymosphenia geminata.* Wellington, Biosecurity New Zealand. www.biosecurity.govt.nz/pests/didymo/where-is-it. (accessed 2 July 2009).

Bolstad WM 2004. Introduction to Bayesian statistics. New York, John Wiley.

Gelman A 2006. Prior distributions for variance parameters in hierarchical models. Bayesian Analysis 1: 515–533.

Lee PM 1997. Bayesian statistics: an introduction. London, Arnold.

McBride GB 2005. Using statistical methods for water quality management: issues, problems and solutions. New York, John Wiley.

McBride GB, Ellis JC 2001. Confidence of Compliance: a Bayesian approach for percentile standards. Water Research 35: 1117–1124.

Wolfram S 2007. The Mathematica® book. Cambridge, Cambridge University Press.

Appendix: Mathematical development of the *CIV*

We define the following two 'events' in the context of having sampled *n* sites to make an assessment regarding a species' possible presence and failing to detect it:

A: the species was actually <u>a</u>bsent from the domain in which the sampled sites lie, and

F: <u>f</u>ailure to detect the species at any sampled site.

　　In Bayesian theory, on commencement of a study an investigator holds a 'prior probability' of the truth of a hypothesis (such as 'the species truly is absent from the domain'). After obtaining the study's data, Bayes' Rule is used to update that probability, so obtaining the 'posterior probability'. Under the Rule these probabilities are related, such that 'posterior ∝ prior x likelihood', where the likelihood is a description of the data (e.g. Lee 1997; Bolstad 2004). In the present case it is appropriate to describe the prior probability by a probability density function for prevalence, using the versatile beta distribution Be(α, β) with (positive) shape parameters α and β. This distribution, defined over the abscissa interval [0, 1], always encloses a unit area but can take a variety of shapes. In particular, Be(1, 1) is a flat uniform distribution while Be($\frac{1}{2}$, $\frac{1}{2}$) is the cup-shaped 'Jeffreys' prior', which is often taken as a 'reference prior' in Bayesian analyses (Lee 1997). We denote g(*p*) ~ Be(α, β) as the prior *probability density* of the species' prevalence (*p*), and $L(\mathbf{F} \mid p)$ as the 'likelihood' for all samples showing absence of the species. The posterior density function (McBride 2005) is

$$h(p \mid \mathbf{F}) = g(p) \left[\frac{L(\mathbf{F} \mid p)}{\int_0^1 L(\mathbf{F} \mid p)g(p)\,\mathrm{d}p} \right] \qquad (1)$$

　　Standard sampling theory shows that *L* is a binomial function involving the term $(1-p)^n$, a term also shared by the beta density function. Consequently, the likelihood and prior functions are 'conjugate' and so we obtain the simple algebraic result that the posterior density function is given by $h(p \mid \mathbf{F}) = $ Be(α, $\beta + n$). We can now integrate this function over a range of *p* to get lower and upper bounds (enclosing a 'credible interval' for *p*) to obtain the *CIV*. Since this interval does not have to be symmetric, we have included the possibility of true absence by setting the lower limit of the credible interval as $p_{\mathrm{lower}} = 0$. Therefore, defining $X = p^*$ as the upper limit, the desired probability [Pr(**A** | **F**)] can be obtained as Pr($p \le p^*$ | failure to detect the species in *n* sites) $= I_{p*}(\alpha, \beta + n)$, where I_{p*} is the 'incomplete beta function ratio' for $0 \le p^* \le 1$. Calculation of the incomplete beta function ratio results shown on Fig. 1 used the regularised incomplete beta function in Mathematica (Wolfram 2007). Other readily available statistical packages can perform these calculations.