A cautionary note on using the scale prior for the parameter N of a binomial distribution

Cristiano Villa and Stephen G. Walker

January 28, 2014

Abstract

Statistical analysis of ecological data may require the estimation of the size of a population, or of the number of species with a certain population. This task frequently reduces to estimating the discrete parameter N representing the number of trials in a binomial distribution. In Bayesian methods, there has been a substantial amount of discussion on how to select the prior for N. We propose a prior for N based on an objective measure of the *worth* that each value of N has in being included in the model space. This prior is compared (through the analysis of the popular snowshoe hare dataset) with the scale prior which, in our opinion, cannot be understood from solid objective considerations.

Keywords abundance, binomial, Kullback–Leibler divergence, loss function, objective prior

1 Introduction

In this paper we discuss objective prior distributions for the discrete parameter N of a binomial distribution, with specific applications to estimation of population or species sizes. In particular, we argue that in statistical applications the scale prior $\pi(N) \propto 1/N$ should not be emplyoed for it is lacking a probabilistic interpretation.

In the statistical analysis of ecological data it is frequent to deal with data that comes from binomial outcomes, such us the size of a population or the number of species. The capture-recapture models for closed population introduced by Otis et al. (1978) represent an example on how the estimation of N proceeds in wildlife data analysis.

A common choice of objective prior for N is the scale prior, that is $\pi(N) \propto 1/N$. Recently, (Link, 2013) has shown its support to the scale prior for N on the basis of its better performance in comparison to the uniform prior and, in addition, that it has been proposed by Berger et al. (2012). We argue that there is no real motivation in the use of the scale prior; on the countrary, it appears to be an*ad-hoc* solution rather than the result of specifi probabilistic considerations. In other words, $\pi(N) \propto 1/N$ has no "meaning". We believe that a way of defining an objective prior for N has to take into considerations the reason why a particular value of the parameter has been included in the parameter space $\mathcal{N} = \{1, 2, \ldots\}$. In particular, the objective approach defines losses instead of probabilities. This idea is discussed in Villa & Walker (2013a) and Villa & Walker (2013b).

It is noteworthy to point out that the scale prior has been used by Wang et al. (2007), King & Brooks (2008) (and the references therein), for applications in ecology, and by Basu & Ebrahimi (2001), for an example of an application in capture-recapture models in software reliability.

The organisation of the paper is as follows. In Section 2 we discuss some background on objective priors for N, and define the prior we propose. Section 3 shows a comparison of the scale prior with our by analysing the popular snowshoes hare data. Finally, Section 4 includes some discussion points and general considerations.

2 Objective priors for N

Consider $x \sim Bin(N, p)$, where $N \in \mathcal{N} = \{1, 2, \ldots\}$ represents the number of independent Bernoulli trials, and $p \in (0, 1)$ the probability of success at each trial. The aim is to make inference on the discrete parameter N, assuming p is unknown.

The task of assigning an objective prior to a discrete parameter is not a trivial and, in the past, has represented an interesting challenge. The main reason comes from the fact that common objective approaches such as Jeffreys' rule (Jeffreys, 1961) and reference analysis (Berger et al., 2009) are not suitable for discrete parameters and, when they are, they do not provide sensible results. Note that the uniform prior $\pi(N) \propto 1$, which may appear to be a natural choice to represent ignorance about N, is not suitable for inference as, for when p is unknown, leads to an improper posterior (Berger et al., 1999, 2012).

A motivation behind the choice of $1/N$ is that, although Jeffreys himself never discussed the prior for N when p is unknown, the choice of $\pi(N) \propto 1/N$ is assumed as natural (Berger et al., 2012), as it is the prior Jeffreys recommends for (continuous) scale parameters. Link (2013), in addition to the above motivation, recommends the scale prior as it solves estimation problems related to the use of the uniform prior (when $\mathcal{N}$ is finite).

The choice of $1/N$ as an objective prior for N is questionable for the following reasons. The motivation for Jeffreys prior in a discrete setting is obsolete. Jeffreys rule is based on invariance property under one-to-one transoformations of the parameter of interest, and this notion has no meaning for a discrete parameter space. Furthermore, Kahn (1987), shows that if we assign a Beta prior to p, $\pi(p) \sim Be(a, b)$, and assume the parameters of the binomial independent a prior, then $\pi(N) \propto 1/N^c$ yields a proper posterior for N if $a + c > 1$. It is therefore legitimate to wonder why c has to be chosen as equal to one. Why not, for example, $\pi(N) \propto 1/N^2$ or $\pi(N) \propto 1/N^3$? This fact adds a level of subjectivity and arbitrariness to the whole procedure, making the process not as objective as intended.

One may argue that the scale prior is the result of a different objective procedure as well. Berger et al. (2012) use an approach which consists in embedding the discrete problem into a continuous one and then apply reference analysis. However, as there exist more than one embedding procedure, they obtain two different priors: the scale prior and

$$\pi(N) \propto \sqrt{N\{N + 4/(n+3)\}}$$

where n is the size of indepdent and identically distributed random variables: $X_i \sim Bin(N,p)$, $i = 1, \ldots, n$. As both priors have similar properties, the recommendation of $1/N$ lays in the simplicity of its functional form. Again, the choice of the scale prior does not appear to be truly objective.

It is fundamental to highlight that in an applied (statistical) setting, such as in ecology, an objective prior needs an idea which is well supported. Unlike academic statisticians, who can discourse on objective priors on theoretical grounds, applied statisticians have to put the motive first: an objective prior needs to have a meaning. In fact, the derivation of $\pi(N)$ should be the result of a process where there is a clear explanation on why a particular prior is chosen and what it represents; we find, for example, that in Link (2013) this explanation is missing, and that the justification in adopting the scale prior is just a reminder to someone else's work.

The prior we propose is based on the idea of assigning a *worth* to each element $N \in \mathcal{N}$. The *worth* is objectively measured by assessing what is lost if that parameter value is removed from $\mathcal{N}$, and it is the true one. Once the *worth* has been determined, this will be linked to the prior probability by means of the self-information loss function (Merhav

& Feder, 1998) $-\log \pi(N)$. A detailed illustration of the idea can be found in Villa & Walker (2013a) and Villa & Walker (2013b), but here is an overview.

Let us indicate by f_N the binomial distribution with parameters N, give p (for the moment assumed to be known). The utility (i.e. *worth)* to be assigned to f_N is a function of the Kullback–Leibler divergence (Kullback & Leibler, 1951) measured from the model to the nearest one; where the nearest model is the one defined by $N' \neq N$ such that $D_{KL}(f_N \| f_{N'})$ is minimised. In fact (see Berk (1966)) N' is where the posterior asymptotically accumulates if N is excluded from $\mathcal{N}$. The objectivity of how the utility of f_N is measured is obvious, as it depends on the choice of the model only.

Let us now write $u_1(N) = \log \pi(N)$ and let the minimum divergence from f_N be represented by $u_2(N)$. Note that $u_1(N)$ is the utility associated with the prior probability for model f_N, and $u_2(N)$ is the utility in keeping N in $\mathcal{N}$. We want $u_1(N)$ and $u_2(N)$ to be matching utility functions, as they are two different ways to measure the same utility in N. As it stands, $-\infty < u_1 \leq 0$ and $0 \leq u_2 < \infty$, while we actually want $u_1 = -\infty$ when $u_2 = 0$. The scales are matched by taking exponential transformations; so $\exp(u_1)$ and $\exp(u_2) - 1$ are on the same scale. Hence, we have

$$e^{u_1(N)} = \pi(N) \propto e^{g\{u_2(N)\}}, \tag{1}$$

where

$$g(u) = \log(e^u - 1). \tag{2}$$

By setting the functional form of g in (1), as it is defined in (2), we derive the proposed objective prior for the discrete parameter N

$$\pi(N) \propto \exp\left\{\min_{N\neq N'\in\mathcal{N}} D_{KL}(f_N\|f_{N'})\right\} - 1. \tag{3}$$

We note that in this way the Bayesian approach is conceptually consistent, as we update a prior utility assigned to N, through the application of Bayes theorem, to obtain the resulting posterior utility expressed by $\log\pi(N|x)$. Indeed, there is an elegant procedure akin to Bayes which works from a utility point of view, namely that

$$\log\pi(N|x) = K + \log f_N(x|N) + \log\pi(N),$$

which has the interpretation of

$$\text{Utility}(N|x,\pi) = K + \text{Utility}(N|x) + \text{Utility}(N|\pi),$$

where K does not depend on N. There is then a retention of meaning between the prior and the posterior information (here represented as utilities). This property is not shared by the usual interpretation of Bayes theorem when priors are objectively obtained; in fact, the prior would usually be improper, hence not representing probabilities, whilst the posterior is (and has to be) a proper probability distribution.

In Villa & Walker (2013a) we show that the nearest model to f_N is at $N' = N+1$. Thus, the prior for N is given by

$$\pi(N) \propto \frac{1}{(N+1)(1-p)}\exp\left\{\sum_{x=0}^{N}\log(N+1-x)\binom{N}{x}p^x(1-p)^{N-2}\right\} - 1. \tag{4}$$

The prior in (4) is improper but, with just one observation, yields a proper posterior.

If p is unknown, the joint prior distribution for the parameters of the binomial is given

Figure 1: Snowshoes hare in its natural habitat.

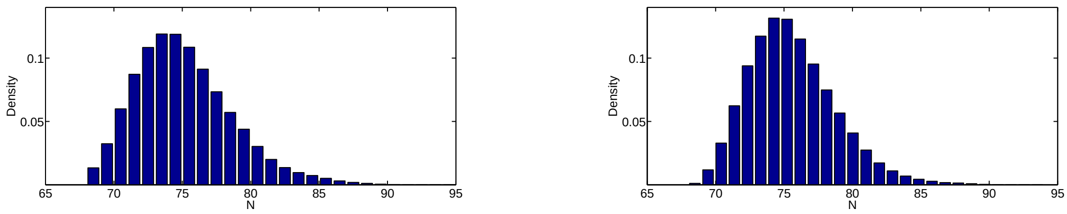
by

$$\pi(N, p) = \pi(N|p)\pi(p), \tag{5}$$

where $\pi(N|p)$ is the prior in (4) above, and $\pi(p)$ a suitable prior for the probability of success at each trial.

3 Snowshoes hares analysis

To illustrate the objective prior we propose, and to compare it with the scale prior, we analyse a popular capture-recapture data set. The problem has been originally discussed in Otis et al. (1978) and, from a Bayesian perspective, for example in Royle et al. (2007) and Link (2013). In particular, Link (2013) has analysed the data using a scale prior for N (although using a data augmentation approach).



(a) (b)

Figure 2: Histogram of the posterior distribution for the parameter N for the hare data using scale prior (a) and our prior (b).

The data consists of a sample of $n = 68$ hares captured-recaptured, over $T = 6$ days. The encounter frequencies, over the 6 days, gives the set $\{25, 22, 13, 5, 1, 2\}$, that is for a total of 145 capture-recapture occurrences. For this illustration, we consider model for closed populations M_0, as defined in Otis et al. (1978), which assumes that the capture-recapture probabilities are constant for all the animals and across the 6 days. Thus, indicating by y_i the detection frequency of animal i, with $i = 1, \ldots, N$, the likelihood function is given by

$$L(N, p|y) \propto \frac{N!}{(N-n)!} p^{\sum_i y_i} (1-p)^{T \cdot N - \sum_i y_i}. \tag{6}$$

We analyse the data by considering both the scale prior and our prior for N. For the scale prior, we have $\pi(N, p) = \pi(N)\pi(p)$, assuming prior independence of the parameters. When we use the prior (4), the joint prior has the form of (5). In both circumstances we set $\pi(p) \sim Be(1/2, 1/2)$, that is Jeffreys' prior. As the posterior distributions are analytically intractable, we obtain the marginal distribution for N through MCMC methods.

The histogram of the posterior distributions are plotted in Figure 2. The posterior for N obtained by applying the scale prior $\pi(N) \propto 1/N$ is shown in (a), while the posterior obtained by applying the prior we propose in (4) is shown in (b). Both distributions are positively skewed and accumulate on the same values of N. When the scale prior is used, the median is $N = 81.5$, with 95% credible interval $(68.7, 94.3)$. When our prior is used, we have a median of $N = 81.0$ and 95% credible interval $(68.7, 93.4)$; note that prior (4) gives a smaller credible interval than the one obtained by adopting the scale prior.

For completeness, we note that for p we have medians $p = 0.33$ in both cases, with 95% credible intervals $(0.24, 0.43)$ and $(0.24, 0.42)$, for the scale and our prior respectively.

4 Discussion

The choice of an objective prior for N must be based not only on performance, but also on solid motivation. If this assumption is not met, it may appear that an objective approach is justifiable as long as the adopted prior leads to a posterior distribution that is suitable for inference (i.e. proper) and that has appealing performances. In the example of the hare data, we have shown that the prior based on losses results in a credible interval that is narrover that the one obtained by applying the scale prior for N. Additionally, while the latter prior has no probabilistic justification, the former one is the result of a clear objective motivation.

The prior for N can be applied to any of the remaining capture-recapture models (Otis et al., 1978), that is when either one or more effects (time effects, behavioral effects, heterogeneity effects) are considered. We have not included any example, either simulated or based on real data, for models including time, behavioral or heterogeneity effects. However, the implementation is similar to the one outlined.

References

BASU, S. & EBRAHIMI, N. (2001). Bayesian capture-recapture methods for error detection and estimation of population size: Heterogeneity and dependence. *Biometrika* **88**, 269–79.

Berger, J. O., Bernardo, J. M., & Sun, D. (2009). The formal definition of reference priors. *Annals of Statistics* **37**, 905–38.

Berger, J. O., Bernardo, J. M., & Sun, D. (2012). Objective priors for discrete parameter spaces. *Journal of the American Statistical Association* **107**, 636–48.

Berger, J. O., Liseo, B., & Wolpert, R. L. (1999). Integrated likelihood methods for eliminating nuisance parameters. *Statistical Science* **18**, 1–28.

Berk, R. H. (1966). Limiting behaviour of posterior distributions when the model is incorrect. *Annals of Mathematical Statistics* **37**, 51–8.

Jeffreys, H. (1961). *Theory of Probability.* University Press, Oxford.

Kahn, W. D. (1987). A cautionary note for Bayesian estimation of the binomial parameter n. *American Statistician* **41**, 38–39.

King, R. & Brooks, S. P. (2008). On the Bayesian estimation of a closed population size in the presence of heterogeneity and model uncertainty. *Biometrics* **64**, 816–24.

Kullback, S & Leibler, R. A. (1951). On information and sufficiency. *Annals of Mathematical Statistics* **22**, 79–86.

Link, W. A. (2013). A cautionary note on the discrete uniform prior for the binomial N. *Ecology* **94**, 2173–9.

Merhav, N. & Feder, M. (1998). Universal prediction. *IEEE Transactions on Information Theory* **44**, 2124–47.

Otis, D. L., Burnham, K. P., White, G. C., & Anderson, D. R. (1978). *Statistical inference from capture data on closed animal populations. Wildlife Monographs* **64**, 1–135.

Royle, J. A., Doraio, R. M. & Link, W. A. (2007). *Analysis of multinomial models with unknown index using data augmentation. Journal of Computational and Graphical Statistics* **16**, 67–85.

Villa, C. & Walker, S. G. (2013a). An objective approach to prior mass functions for discrete parameter spaces. *Journal of the American Statistical Association* Revision submitted.

Villa, C. & Walker, S. G. (2013b). Objective prior for the number of degrees of freedom of a t distribution. *Bayesian Analysis* To appear.

Wang, X, He, C. Z. & Sun, D. (2007). Bayesian population estimation for small sample capture-recapture data using noninformative priors. *Journal of Statistical Planning and Inference* **137**, 1099–118.