

Reference Priors for Discrete Parameter Spaces

JAMES O. BERGER

Duke University, USA

JOSE M. BERNARDO

Universidad de Valencia, Spain

and

DONGCHU SUN

University of Missouri-Columbia, USA

Abstract

The development of objective prior distributions for discrete parameter spaces is considered. Formal approaches to such development – such as the *reference prior* approach – often result in a constant prior for a discrete parameter, which is questionable for problems that exhibit certain types of structure. To take advantage of structure, this article proposes embedding the original problem in a continuous problem that preserves the structure, and then using standard reference prior theory to determine the appropriate objective prior. Four different possibilities for this embedding are explored, and applied to a population-size model, the hypergeometric distribution, the binomial-beta distribution, the binomial distribution, and determination of prior model probabilities. The recommended objective priors for the first three problems are new.

AMS 2000 subject classification: Primary 62F15; secondary 62A01, 62B10, 62E20.

Some key words: Parameter-based asymptotics, Binomial, Discrete parameter space, Hypergeometric, Jeffreys-rule priors, Prior model probabilities, Reference priors.

1 Introduction

1.1 Background

Discrete parameter spaces have long posed a problem for objective Bayesian analysis, since the obvious objective prior for a discrete parameter is often the constant prior; for instance, applying standard reference prior theory (cf. Bernardo & Smith (1994), p. 307) will always yield a uniform prior on a finite parameter space. If the parameter space is indeed finite and has no structure, this is not a problem, it being hard to imagine how anything other than the uniform distribution could be labeled as objective. If the parameter space has structure, however, a non-constant objective prior is often desired, as the following two examples show.

Example 1.1 A simple example of a structured discrete parameter problem is provided by observing a binomial random variable x with PDF $\text{Bi}(x | n, p)$, where the sample size n is the unknown quantity of interest and the success probability p is known. The parameter space is thus the set $\mathcal{N} = \{1, 2, \dots\}$ of the natural numbers.

This problem is particularly interesting when p is also unknown since it is well known from the literature (see, e.g., Kahn (1987)) that use of a constant prior density for n is then inadequate, resulting in an improper posterior distribution for standard objective priors for p (such as the Jeffreys-rule prior or uniform prior). Thus we need to use the structure of the binomial model to develop something more sensible as an objective prior for n .

Example 1.2 Consider the hypergeometric distribution for a population of size N (known). The unknown parameter is R , the number of items in the population having a certain property, which can take values $\{0, 1, \dots, N\}$. The data is r , the number of items with the property out of a random sample of size n (taken without replacement).

The parameter space is clearly finite, but reflects an obvious structure through the

hypergeometric distribution. Indeed, for large N it is well known that the hypergeometric distribution is essentially equivalent to the binomial distribution, with parameter $p = R/N$. The objective prior for R should thus be compatible with the objective prior for p , which most schools of objective Bayesian analysis take to be the non-uniform Jeffreys-rule prior $\pi(p) \propto p^{-1/2}(1-p)^{-1/2}$.

There have been a number of proposals for objective priors for specific discrete parameter problems, and we will refer to these proposals when considering specific problems. There have been a few general proposals. Jeffreys (1961) simply suggests using $1/\theta$ for any unbounded positive parameter θ , including discrete. This choice has the appeal of having the largest polynomial decay that retains propriety. Rissanen (1983) proposed a prior for positive integers that in some sense is the vaguest proper prior. Barger & Bunge (2008) propose using the formal Jeffreys-rule method or reference prior method, based on an extension of the concept of an information matrix to discrete parameter problems that was developed in Lindsay & Roeder (1987). We discuss this interesting approach further in Section 1.2.3.

Our preferred method of constructing objective priors is the reference prior approach but, as mentioned above, direct application of the approach to discrete parameters seems to yield the constant prior. Motivated by the hypergeometric example, we thus instead approach the structured discrete parameter problems by embedding the original model into a model with a continuous parameter. In this continuous parameter problem, we can then apply the ordinary reference prior theory (Bernardo, 1979, 2005; Berger & Bernardo, 1992; Berger, Bernardo & Sun, 2009a), and appropriately discretize the resulting continuous reference prior (if necessary).

1.2 Possible Embeddings

There does not appear to be a generally successful single methodology for embedding a discrete parameter problem into a continuous parameter problem. In this section we discuss

four of the embedding methodologies that we have found to be useful in dealing with discrete, structured parameter spaces. In the remainder of Section 1, we will generically let θ denote the unknown discrete parameter, taking values in a countable discrete space Θ . In the later specific examples we will revert to using the natural notation for the example (e.g., n for the binomial problem and R for the hypergeometric problem).

1.2.1 Approach 1: Assuming parameters are continuous

The simplest possibility is just to treat θ as a continuous variable in the original model. Unfortunately, this typically introduces a normalizing factor into the expression for non-integer θ , and the behavior of this normalizing factor can cause problems. Sometimes, however, one can also treat the available data x as continuous and analyze the resulting completely continuous problem. For example, suppose that x is uniform on the integers $\{0, 1, \dots, \theta\}$. Treating both θ and x as continuous variables, we have that x is uniform on $(0, \theta)$, and we know that the reference prior for this continuous problem is $\pi(\theta) = 1/\theta$.

Even if making both x and θ continuous in the original density does introduce an additional normalizing factor, it is often a smooth function that can be handled within ordinary reference prior analysis. The solution, however, is usually not available in closed form. Furthermore, when a different normalization factor is introduced, it is hard to argue that the new continuous model has the same structure as the original discrete model, and hence the utilization of the continuous reference prior for the discrete model is suspect. We thus do not recommend using this simple embedding if a new (nonconstant) normalization factor is introduced.

1.2.2 Approach 2: Introducing a continuous hierarchical hyperparameter

In some problems, it is natural to add a hierarchical level of modeling to create a continuous hyperparameter that can be analyzed with usual reference prior methods. As an example,

in the hypergeometric problem, one can postulate that the unknown R arises as a Binomial random variable from the $\text{Bi}(R|N, p)$ distribution, with p unknown. The problem can then be reduced to finding the reference prior for p , a continuous parameter, and this can be used to determine the implied reference prior for R .

When such hierarchical modeling is possible and natural, the case for the ensuing reference prior for the discrete parameter seems rather compelling. Unfortunately, this situation is rather uncommon.

1.2.3 Approach 3: Applying reference prior theory with a consistent estimator

This approach uses the usual methodology of reference prior theory, based on replication of the experiment under study. In other words, one considers $\mathbf{x}^{(k)} = (x_1, \dots, x_k)$, where the x_i are k independent replications of the data x from the model under consideration. (As usual in reference prior theory, x will, itself, typically be a vector of observations; the replication being considered is imaginary replication of the entire original experiment.) In the reference prior approach, one considers the asymptotic behavior of an information-based criterion as $k \rightarrow \infty$ (cf. Berger & Bernardo (1992)).

Approach 3 proceeds by

- choosing a consistent linear (or some other simple) estimate $\hat{\theta}_k = \hat{\theta}_k(\mathbf{x}^{(k)})$ of θ (based on the replications), which effectively becomes continuous as the number of replications $k \rightarrow \infty$;
- finding the asymptotic sampling distribution of $\hat{\theta}_k$ as $k \rightarrow \infty$;
- pretending that θ is continuous in this asymptotic distribution and finding the corresponding reference prior.

Notice that using fully efficient estimators, which only take values on the original discrete parameter space, cannot work here, since there is then no possible continuous embedding.

Suppose that, as $k \rightarrow \infty$ and for some series of constants c_k (typically $c_k = \sqrt{k}$), $c_k(\hat{\theta}_k - \theta)$ has a limiting normal distribution with mean zero and variance $\sigma^2(\theta)$. In this limiting normal distribution, one now simply assumes that θ is continuous, and hence the reference prior is (cf. Theorem 9 of Bernardo, 2005) $\pi^R(\theta) \propto \sigma(\theta)^{-1}$.

That this approach requires use of inefficient estimators is both philosophically problematic and also practically ambiguous, because use of different inefficient estimators can result in different reference priors (as we shall see). Hence this approach might be best used to suggest a reference prior, which is then validated by other criteria.

For discrete models that have *linear difference score* (see Lindsay & Roeder (1987) for definition) it is possible to define an analogue of the expected Fisher information matrix and apply the Jeffreys-prior formula or the reference prior formula to obtain an objective prior. This was proposed in Barger & Bunge (2008), who analyze several examples. The only overlap with the examples in this paper (i.e., the only example we consider that has a linear difference score) is the binomial problem with p known. For this problem and, indeed, any problem with a linear difference score, the objective prior obtained from their approach will be the same as that obtained using our Approach 3 with a linear estimator. Barger & Bunge (2008) thus provides an extrinsic justification for Approach 3 with linear estimators, in problems with a linear difference score.

1.2.4 Approach 4: Using parameter-based asymptotics

Following Lindsay & Roeder (1987) and Sweeting (1992), this approach uses a formal limiting operation in θ to make the problem continuous:

- Let $\theta \rightarrow \infty$ (or ensure this by forcing some other parameter to ∞).
- In the limiting asymptotic distribution of x (or some related random variable), let θ be continuous (if possible).

- Do the reference prior replications over k with this asymptotic distribution, to define a reference prior.

For instance, in the uniform problem mentioned in Section 1.2.1, x/θ has a uniform distribution on the discrete set $\{0, \frac{1}{\theta}, \dots, \frac{\theta-1}{\theta}, 1\}$. As θ gets large, it seems reasonable to replace the discrete values on the unit interval by the unit interval itself, so we end up with the limiting distribution of x/θ being Uniform on $(0,1)$, or x being Uniform on $(0, \theta)$. Pretending that x and θ are continuous in this problem results in $\pi(\theta) = 1/\theta$, as before.

This approach has the big advantage of being well-defined, as it is based on the full (or an asymptotically efficient) posterior, but it only gives the reference prior for ‘large θ ,’ and this may not be right for small θ . Hence this is perhaps best used in conjunction with one of the other approaches, as a validation of the answer from that approach.

1.2.5 Technical aside: discretization of the prior

It can happen in Approaches 1, 3 and 4 that the resulting prior, $\pi^R(\theta)$, is infinite at the endpoints of the discrete parameter space Θ . A natural (if ad hoc) solution is to then define the discrete prior by, instead, discretizing the continuous reference prior, $\pi^*(\theta)$, on the continuous extension of the parameter space Θ^* : write $\Theta^* = \cup_j \Theta_j^*$, where Θ_j^* is a set containing discrete parameter value θ_j , and then define

$$\pi^R(\theta_j) = \int_{\Theta_j^*} \pi^*(\theta) d\theta,$$

assuming these probabilities are finite.

As an example, for the hypergeometric problem in Example 1.2, we will see that Approaches 3 and 4 lead to an objective prior of the form $\pi(R) = 1/\sqrt{R(N-R)}$, which is infinite at the endpoints $R = 0$ and $R = N$, so that a discretization of this prior would be needed (and the resulting prior probabilities at the endpoints would be finite).

We do not pursue this issue, however, because it is not strictly needed in the examples of

the paper; for the hypergeometric problem, we can utilize Approach 2, which directly yields a discrete reference prior.

1.3 Overview

In the remainder of the paper, we consider a variety of discrete parameter problems and explore the application of the above four approaches to development of an objective prior for these problems. We will call the resulting prior a reference prior, because it arises as a standard reference prior in the extended model.

When Approach 1 (without renormalization) or Approach 2 (when there is a single natural hierarchical model) are applicable, we view the resulting reference prior as being the natural reference prior, and look no further. For problems where neither approach applies, we utilize some combination of Approaches 3 and 4 to propose a reference prior. Unfortunately, it is not the case that the resulting prior is necessarily unique, although the possible priors seem to result in essentially the same answer for large θ . This is probably unavoidable; unless there is enough structure to use Approaches 1 or 2, unequivocal determination of an objective prior for small θ seems not to be possible. That all the methods seem to provide essentially the same objective prior for large θ is, however, gratifying, in that it is typically for large θ that there are problems with the likelihood.

The particular discrete problems that are considered in the paper (besides the discrete uniform distribution already dealt with above) are a *population-size model* (in Section 2), the *hypergeometric distribution* (in Section 3), the *binomial-beta distribution* (in Section 4), the *binomial distribution* (in Section 5), and *determination of prior model probabilities* (in Section 6). The first three applications involve derivation of reference priors that have not previously been considered for the problems, while the last two choices validate priors that had been previously proposed via more ad hoc arguments.

2 Estimating a Population Size

Consider an experiment with Type II censoring (cf. Lawless (1982), Section 1.4), in which there is a sample of N units whose lifetimes follow an exponential distribution with mean $1/\lambda$. The experiment is stopped as soon as R units have failed; let $t_1 \leq \dots \leq t_R$ denote the resulting failure times. Here $N \geq R$ and λ are both unknown, while R is pre-specified.

The problem of estimating N has many interesting applications. For example, Starr (1974) and Kramer & Starr (1990) desired to estimate the number of fish in a lake, assuming that the time to catch any particular fish is exponentially distributed with mean $1/\lambda$. Goudie & Goldie (1981) considered a linear pure death process, where the data consisted only of the lifetimes of those who had died, assumed to be iid exponentially distributed with mean $1/\lambda$; of interest was an estimate of the initial population size N . This model can also arise in software reliability, where the number N of bugs is unknown, and the lengths of time to discover the first R bugs are assumed to be independently exponential with mean $1/\lambda$. See Basu & Ebrahimi (2001).

The density of (t_1, \dots, t_R) is

$$p(t_1, \dots, t_R | N, \lambda) = \frac{N!}{(N-R)!} \lambda^R \exp \left\{ -\lambda [t_1 + t_2 + \dots + t_R + (N-R)t_R] \right\}.$$

The pair $(t_1 + \dots + t_R, t_R)$ is minimal sufficient for (N, λ) and, hence, so is the pair $(V, W) = (t_1 + \dots + t_R)/t_R, t_R$. For any given $c > 0$, the transformation (t_1, \dots, t_R) to (ct_1, \dots, ct_R) induces a transformation of the parameter (N, λ) to $(N, c\lambda)$ and the sufficient statistic (V, W) to (V, cW) . So a maximal invariant statistic is V .

Goudie & Goldie (1981) (formula (4)) derived the joint density of (V, W) as follows,

$$p(V, W | N, \lambda) = \frac{R}{(R-2)!} \binom{N}{R} \lambda^R W^{R-1} \exp \left\{ -\lambda(V + N - R)W \right\} g_R(V), \quad (1)$$

for $1 < V < R, W > 0$, where

$$g_R(V) = \sum_{i=1}^{[V]} (-1)^{i-1} \binom{R-1}{i-1} (V-i)^{R-2}, \quad 1 < V < R. \quad (2)$$

The marginal density of V depends only on N and is (formula (5) of Goudie & Goldie (1981))

$$p(V | N) = \frac{1}{(R-2)!} \frac{N!}{(N-R)!} \frac{1}{(V+N-R)^R} g_R(V), \quad 1 < V < R. \quad (3)$$

Because V is a maximal invariant statistic whose distribution depends only on N , frequentist inference concerning N would be based on (3). Because of the invariance, this marginal density also arises from the Bayesian perspective, when the Haar density (also the Jeffreys-rule prior and reference prior) $\pi(\lambda) = 1/\lambda$ is used as the conditional prior for λ given N . Indeed, it can be directly shown that (3) results, up to a proportionality constant, from integrating out λ in (1), with respect to the Haar density. Thus, from either a frequentist or objective Bayesian perspective, dealing with unknown N reduces to analysis with the density (3).

Part of the interest in this problem is that it is difficult to address by standard methods. For instance, Goudie & Goldie (1981) showed that there is no unbiased estimate of N , and that moment estimates and maximum likelihood estimates do not exist with probability roughly 1/2. For objective Bayesian analysis, note that the likelihood $p(V | N)$ in (3) tends to one as $N \rightarrow \infty$, so that the posterior would not exist if either a constant prior or a prior proportional to $1/N$ (two common objective choices for integer N) were used. Similar problems were encountered in Raftery (1988b) for a situation involving Type I censoring.

To find a reference prior here, we can directly apply Approach 1, embedding the discrete parameter space for N , $\{R, R+1, \dots\}$ into the continuous space $(R-0.5, \infty)$, since, for each $N \geq R-0.5$, $p(v | N)$ is still a probability density for v (with the same normalization). The reference prior for N in this continuous problem is simply the Jeffreys-rule prior

$$\pi(N | R) \propto \sqrt{i_R(N)}, \quad N \geq R-0.5,$$

where $i_R(N)$ is the Fisher information given in the following lemma (the proof of which is given in the Appendix).

Lemma 2.1 *The Fisher information of N in the continuous parameter problem is*

$$i_R(N) = -\frac{RN!}{(R-2)!(N-R)!} J_{R,N} + \sum_{j=0}^{R-1} \frac{1}{(N-j)^2}, \quad (4)$$

where

$$J_{R,N} = \frac{2}{R^3 - R} \sum_{i=0}^{R-1} (-1)^i \binom{R-1}{i} \frac{1}{(N-R+1+i)^3}. \quad (5)$$

To understand some of the properties of this reference prior, it is convenient to consider the reparameterization $\theta = N - R + 1$, since the range of θ is then the positive integers $\mathcal{N} = \{1, 2, \dots\}$. The reference prior for θ is then

$$\pi(\theta | R) \propto \sqrt{i_R(\theta + R - 1)}, \quad \theta \in \mathcal{N}.$$

Special cases, when $R = 2, 3, 4$, are as follows

$$\pi(\theta | R) \propto \begin{cases} \frac{1}{\theta(\theta + 1)} & \text{if } R = 2, \\ 1.3036 \frac{\sqrt{(\theta + 2)\theta + 4/3}}{\theta(\theta + 1)(\theta + 2)} & \text{if } R = 3, \\ 1.6017 \frac{\sqrt{[(\theta + 3)\theta + 22/5](\theta + 3)\theta + 27/5}}{\theta(\theta + 1)(\theta + 2)(\theta + 3)} & \text{if } R = 4. \end{cases}$$

Note that these are proper priors, and so their normalization constants are included. It appears that the tail of $\pi(\theta | R)$ is always of order $1/\theta^2$ (verified numerically for R up to 100), and hence the prior seems always to be proper. It is, of course, necessary for the prior to be proper in order for the posterior to be proper, since the likelihood is constant as $N \rightarrow \infty$; this is thus another example of the rather remarkable tendency of the Jeffreys-rule (reference) prior to yield a proper posterior for challenging likelihoods.

For $R = 2, 3, 4, 5$, and 20, these priors are plotted in Figure 1. While they do differ somewhat at $\theta = 1$, they otherwise are remarkably similar. Indeed, the differences between the priors will not greatly affect the posterior, since the main effect of the prior on the posterior is in the tails, where all the priors are very similar. Hence one could reasonably just use $1/[\theta(1 + \theta)]$ as the reference prior for any R . (Of course, the exact reference prior is not difficult to program and work with.)

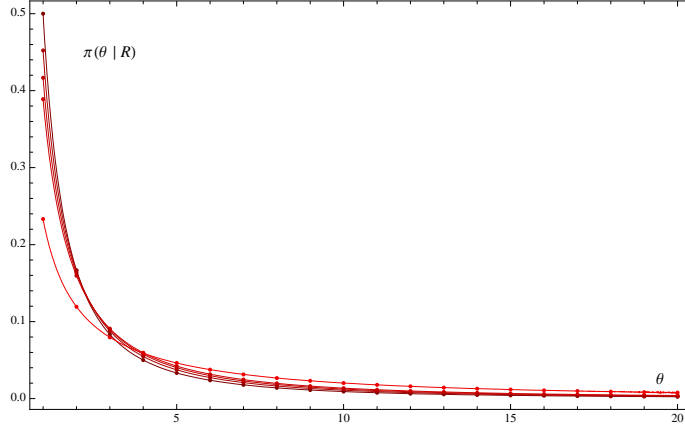


Figure 1: Reference priors for θ for $R = 2, 3, 4, 5,$ and 20 (top to bottom at $\theta = 1$).

3 The Hypergeometric Distribution

Consider the hypergeometric distribution,

$$p(r | n, R, N) = \frac{\binom{R}{r} \binom{N-R}{n-r}}{\binom{N}{n}}, \text{ for } r \in \{0, \dots, R\}, \quad (6)$$

where R is unknown and $R \in \{0, 1, \dots, N\}$. Approach 1 would introduce a complicated non-constant normalization factor here, so we, instead, turn to the other approaches.

3.1 Approach 2

As mentioned in the introduction, a natural hierarchical model for the unknown R is to assume that it is $\text{Bi}(R | N, p)$, with unknown p . The problem then reduces to finding the reference prior for the continuous hyperparameter p .

The appropriate reference prior for a hyperparameter in a hierarchical setting (cf. Bernardo and Smith, 1994, p. 339) is found by first marginalizing out the lower level parameters having specified distributions. Here, that would mean marginalizing over R , resulting in

$$\Pr(r | n, N, p) = \sum_{R=0}^N \Pr(r | n, R, N) \Pr(R | N, p) \quad (7)$$

$$= \sum_{R=0}^N \frac{\binom{R}{r} \binom{N-R}{n-r}}{\binom{N}{n}} \binom{N}{R} p^R (1-p)^{N-R} \quad (8)$$

$$= \binom{n}{r} p^r (1-p)^{n-r}, \quad (9)$$

which, as one would expect, is simply the binomial model $\text{Bi}(r | n, p)$. But the reference prior for the binomial model is known to be the corresponding Jeffreys-rule prior, which is the Beta distribution $\pi^R(p) = \text{Be}(p | \frac{1}{2}, \frac{1}{2})$ and, therefore, integrating out p in the $\text{Bi}(R | N, p)$ hierarchical prior for R yields the induced reference prior for R in the hypergeometric model

$$\begin{aligned} \pi^R(R | N) &= \int_0^1 \text{Bi}(R | N, p) \text{Be}(p | \frac{1}{2}, \frac{1}{2}) dp \\ &= \frac{1}{\pi} \frac{\Gamma(R + \frac{1}{2}) \Gamma(N - R + \frac{1}{2})}{\Gamma(R + 1) \Gamma(N - R + 1)}, \end{aligned} \quad (10)$$

for $R \in \{0, 1, \dots, N\}$. Note that, by construction, this is a proper prior.

For $N = 1$, one obtains the uniform prior

$$\pi^R(R | N = 1) = \{1/2, 1/2\}, \quad R \in \{0, 1\}. \quad (11)$$

However, this is the only case where the reference prior for the hypergeometric is uniform; for instance, for $N = 2$ one gets

$$\pi^R(R | N = 2) = \{3/8, 1/4, 3/8\} \quad R \in \{0, 1, 2\}. \quad (12)$$

The upper panel of Figure 2 represents a set of reference priors $\pi^R(\theta | N)$ for the proportion $\theta = R/N$, $\theta \in \{1, 1/N, \dots, 1\}$, for several values of the population size N .

The behavior of the reference prior (10) for large N is, as expected, compatible with the continuous reference prior $\pi^R(p) = \text{Be}(p | \frac{1}{2}, \frac{1}{2})$ for the binomial probability model. Indeed, using Stirling's approximation for the Gamma functions in (10) one obtains, in terms of $\theta = R/N$,

$$\pi^R(\theta | N) \approx \frac{1}{N + \frac{2}{\pi}} \text{Be}\left(\frac{N\theta + \frac{1}{\pi}}{N + \frac{2}{\pi}} \mid \frac{1}{2}, \frac{1}{2}\right), \quad \theta = 0, 1/N, \dots, N, \quad (13)$$

which is basically proportional to $\text{Be}(\theta | \frac{1}{2}, \frac{1}{2})$. As illustrated in the lower panel of Figure 2, where the solid line represents (13) as a continuous function of θ , the approximation is very good, even if N is moderate.

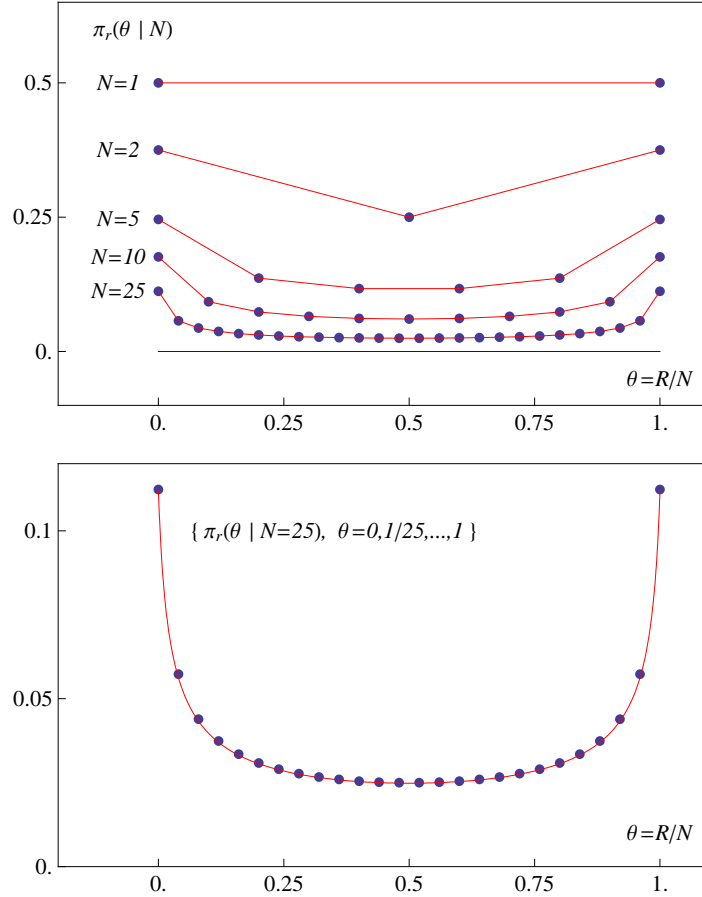


Figure 2: Structured reference priors $\pi^R(\theta | N)$ for the proportion $\theta = R/N$ of conforming items in a population of size N , for several N values (upper panel), and its continuous approximation (lower panel).

3.2 Approaches 3 and 4

Since Approach 2 is a preferred approach, we do not formally present Approaches 3 and 4. It is worth mentioning, however, that an application of Approach 3 with a simple linear estimator of R yields a reference prior proportional to $1/\sqrt{R(N-R)}$, which is very similar to (10). Likewise, applying Approach 4 by letting $N \rightarrow \infty$ results in utilizing the binomial approximation to the hypergeometric directly, and again suggests using $1/\sqrt{R(N-R)}$ as the reference prior. Both methods thus give essentially the right answer (if N is not small and R is not 0 or N) although, as discussed in the introduction, we view Approach 2 as being

superior when it can be applied. Also, as discussed in Section 1.2.5, Approach 2 avoids the technical issue of dealing with the infinite endpoints of $1/\sqrt{R(N-R)}$.

3.3 Laplace's Rule of Succession

A philosophical curiosity is that of determining the probability that the next element in a randomly selected sequence of elements will have a specified property, given that all previous n elements possessed the property. This is usually phrased in the context of a potentially infinite number of elements, but can also be asked when it is known that there are only N elements in the population. In this case, one can view the problem from the hypergeometric viewpoint: of n sampled elements, $r = n$ have the specified property, with the total number R that have the property being unknown. The probability that the next randomly selected element (from the $N - n$ unsampled elements) has the property (call this $+$) is clearly $\Pr(+ | N, n, R, r = n) = (R - n)/(N - n)$. Marginalizing over the posterior distribution of R , given $r = n$, will yield the desired $\Pr(+ | N, n, r = n)$.

If one uses a uniform prior for R , Broad (1918) shows that the resultant probability is

$$\Pr(+ | N, n, r = n) = \frac{n + 1}{n + 2}, \quad (14)$$

which is independent of N . This is usually known as Laplace's rule of succession, although Laplace (1774) did not consider the case of finite N and only derived (14) for its continuous binomial approximation.

Given that we argue for $\pi^R(R | N)$ in (10) as the appropriate objective prior for R , as opposed to the uniform prior, it is natural to ask what the resulting 'reference prior' law of succession would be. The analog of (14), as computed in Berger et al. (2009b), is

$$\Pr^R(+ | N, n, r = n) = \frac{n + 1/2}{n + 1}. \quad (15)$$

As one would require, for $n = 0$ (and hence with no initial information), both (14) and (15) yield $1/2$. For $n = 1$, Laplace yields $2/3$ while the reference probability is $3/4$. Table 1 lists

Table 1: Probability of the next random element having the property, given that all n previously observed elements have the property.

	<i>Laplace</i>	<i>Reference</i>
n	$\frac{n+1}{n+2}$	$\frac{n+1/2}{n+1}$
0	1/2	1/2
1	2/3	3/4
2	3/4	5/6
5	0.8571	0.9167
10	0.9167	0.9545
25	0.9630	0.9808
100	0.9900	0.9950
1000	0.9990	0.9995

other values. As illustrated in that table, the reference law of succession has an appreciably faster convergence to one as n increases.

4 The Binomial-Beta Distribution

Consider the binomial-beta distribution formed as the marginal distribution from

$$x \mid (n, p) \sim \text{Bi}(x \mid n, p) \text{ and } p \sim \text{Be}(p \mid a, b), \quad (16)$$

where (a, b) are known positive constants. Clearly, the marginal density of x is

$$\begin{aligned} p(x \mid n) &= \frac{n!}{x!(n-x)!} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 p^{x+a-1} (1-p)^{b-1} dp \\ &= \frac{n!}{x!(n-x)!} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(x+a)\Gamma(n-x+b)}{\Gamma(n+a+b)}, \end{aligned} \quad (17)$$

for $x = 0, 1, \dots, n$. For fixed x , the tail (when n is large) of this marginal likelihood for n has the form

$$p(x \mid n) \propto \frac{n!}{(n-x)!} \frac{\Gamma(n-x+b)}{\Gamma(n+a+b)} \approx \frac{1}{n^{a+b-1} (n-x)^{1-b}} \approx \frac{1}{n^a}, \quad (18)$$

which depends on a only. Note that, for $a \leq 1$, this is not integrable and hence, for instance, a constant prior for n would not yield a proper posterior.

To find the reference prior for n from this discrete distribution, application of Approach 1 would result in a non-constant normalization factor, and there is no natural hierarchical embedding for Approach 2. Hence we consider Approaches 3 and 4.

4.1 Approach 3

Going to the asymptotics of reference prior theory, let $\{x_1, \dots, x_k\}$ be k independent replications from (17). The mean and variance of the x_i are easily found to be

$$\mathbb{E}(x_i | n) = \mathbb{E}(np | n) = n \frac{a}{a+b}, \quad (19)$$

$$\text{Var}(x_i | n) = \frac{n(n+a+b)}{(a+b)^2(a+b+1)}. \quad (20)$$

A simple estimate of n is the linear estimate

$$\hat{n} = \frac{a+b}{a} \bar{x} = \frac{a+b}{ak} \sum_{j=1}^k x_j. \quad (21)$$

Note that this is not an efficient estimate but, as mentioned in the introduction, efficient estimates cannot be used in direct reference prior asymptotics for discrete distributions. This estimator has mean and variance

$$\mathbb{E}(\hat{n} | n) = n \quad \text{and} \quad \text{Var}(\hat{n} | n) = \frac{n(n+a+b)}{a^2(a+b+1)k}.$$

It follows from the central limit theorem that

$$p(\hat{n} | n) \approx N \left(\hat{n} \mid n, \frac{n(n+a+b)}{a^2(a+b+1)k} \right). \quad (22)$$

We extend the asymptotic distribution to a continuous parameter model by pretending that n is a continuous parameter. Then the reference prior for n is clearly the Jeffreys-rule prior,

$$\pi_1(n) \propto \left(\frac{n(n+a+b)}{a^2(a+b+1)k} \right)^{-\frac{1}{2}} \propto \frac{1}{\sqrt{n(n+a+b)}}. \quad (23)$$

4.2 Approach 4

We use the fact that for fixed $p \in (0, 1)$, when $n \rightarrow \infty$, x/n is asymptotically normal with mean p and variance $p(1-p)/n$. For any $y \in (0, 1)$, it follows that, as $n \rightarrow \infty$,

$$\begin{aligned} P(x/n \leq y) &= \int_0^1 P(X/n \leq y \mid p) \text{Be}(p \mid a, b) dp \\ &\rightarrow \int_0^1 1_{(0,y)}(p) \text{Be}(p \mid a, b) dp \\ &= \int_0^y \text{Be}(p \mid a, b) dp, \end{aligned}$$

so that the limiting density of $y = y/n$ is $\text{Be}(y \mid a, b)$. The parameter-based asymptotic distribution of x thus has density $n^{-1} \text{Be}(x/n \mid a, b, n)$. In this density we can treat $n > 0$ as a continuous parameter (as well as x), in which case it is clearly a scale parameter, and the reference prior for a scale parameter is $\pi_2(n) \propto 1/n$. (This is only a first-order parameter-based asymptotic argument. We will see an example of a second order argument in Section 5.)

4.3 Comparison

Approach 3 led to $\pi_1(n) \propto 1/\sqrt{n(n+a+b)}$, while Approach 4 led to $\pi_2(n) = 1/n$, as candidates for the reference prior. These are both clearly compatible for large n , which is reassuring. The interesting question is thus whether the dependence of π_1 on a and b results in a superior objective prior.

The following argument suggests that π_1 is better. Suppose that a and b are very large, in which case the $\text{Be}(p \mid a, b)$ distribution in (16) will be very concentrated around $p = a/(a+b)$. It follows that $p(x \mid n)$ should behave like the $\text{Bi}(x \mid n, a/(a+b))$ distribution, in this situation. In the next section, we will see that the recommended reference priors for a binomial n , when p is known, behave like $1/\sqrt{n}$, which is the behavior of $1/\sqrt{n(n+a+b)}$ for n of interest when a and b are very large. This points to choosing π_1 as the reference prior.

To further study this, we utilize the common device of studying the frequentist coverage of

credible sets arising from the priors, to assist in the determination of a good choice. Suppose we have a sample $\mathbf{x} = (x_1, \dots, x_m)$ from (17) for given n , and construct the one-sided credible sets that have posterior probability exceeding a desired threshold $1 - \alpha$, namely

$$C_i(\mathbf{x}) = \{1, \dots, n_i^*\}, \quad n_i^* = \inf_{n^* \in \mathcal{N}} \Pr_i[n \leq n^* \mid \mathbf{x}] \geq 1 - \alpha,$$

corresponding to the two priors $\pi_i, i = 1, 2$. We then consider the frequentist coverage of these credible sets, namely $\Pr(C_i(\mathbf{X}) \ni n \mid n)$, with the goal being to have frequentist coverage similar to the stated posterior probability. Because of problems caused by the discreteness here, we choose the target not be $1 - \alpha$ itself, but rather the frequentist expectation (given n) of the posterior coverage. (A Bayesian, in repeated use of the prior to construct credible sets in different applications, will, on average, be quoting this as the coverage, and it is reasonable to compare average stated coverage with the average actual coverage of the credible sets.)

The straightforward implementation of this idea, for given true parameter value n and sample size m , is to generate, say, 25,000 sample m -vectors, $\mathbf{x}_{(j)}$, with each coordinate being drawn independently from (17), compute $C_i(\mathbf{x}_{(j)})$ for each sample, and approximate the frequentist coverage and the average posterior coverage (APC) as, respectively,

$$\text{Coverage} \cong \frac{\#C_i(\mathbf{x}_{(j)}) \text{ that contain } n}{25,000}, \quad \text{APC} \cong \frac{1}{25,000} \sum_{j=1}^{25,000} \Pr(n \in C_i(\mathbf{x}_{(j)}) \mid \mathbf{x}_{(j)}). \quad (24)$$

In Tables 2 and 3, we present some of the results from the simulation study, for true $n = 10$ and sample size $m = 1$. (Small sample sizes provide more revealing differences between the priors.) We report Coverage and APC, as well as Difference = Coverage – APC, which is best if close to zero, with negative errors being better than positive errors (since the stated accuracy of the credible sets is then at least conservative). As a secondary criterion, we also report the average size (AS) of the posterior credibility sets, defined as $\text{AS} = \text{E}[\text{length of } C_i(\mathbf{x}_{(j)}) \mid n]$, and approximated by $\sum_j \text{length of } C_i(\mathbf{x}_{(j)})/25,000$.

It is clear that π_2 can significantly overstate the coverage probability of the credible sets. In contrast, π_1 is not only much more accurate in terms of its coverage statement, but also

Table 2: Average posterior coverage, frequentist coverage, and average size of one-sided 50% credible intervals based on 25,000 simulations of sample size $m = 1$ from the Binomial-Beta distribution with true $n = 10$ and various values of $a = b$.

$a = b$	Prior	APC	Coverage	Difference	AS
5	π_1	0.549	0.579	-0.030	9.972
	π_2	0.502	0.408	0.094	8.983
20	π_1	0.559	0.608	-0.049	9.996
	π_2	0.501	0.390	0.110	8.999
50	π_1	0.562	0.618	-0.055	10.033
	π_2	0.500	0.385	0.115	9.035

Table 3: Average posterior coverage, frequentist coverage, and average size of one-sided 90% credible intervals based on 25,000 simulations of sample size $m = 1$ from the Binomial-Beta distribution with true $n = 10$ and various values of $a = b$.

$a = b$	Prior	APC	Coverage	Difference	AS
5	π_1	0.910	0.872	0.038	19.618
	π_2	0.908	0.872	0.037	18.328
20	π_1	0.914	0.927	-0.012	16.046
	π_2	0.916	0.927	-0.011	15.254
50	π_1	0.914	0.937	-0.022	15.202
	π_2	0.917	0.813	0.105	14.556

tends to err on the side of conservatism, which is desirable. And, as expected, π_2 significantly degrades as $a = b$ grows larger, while the performance of π_1 remains stable as $a = b$ grows. The performance of π_1 is actually rather remarkable, considering that the study is only for a sample of size $m = 1$.

In conclusion, all arguments point to choosing $\pi^R(n) = 1/\sqrt{n(n+a+b)}$ as the reference prior for the Binomial-Beta distribution.

5 Reference Prior for the Binomial Sample Size

Assume that $x \sim \text{Bi}(x | n, p)$. Estimating n has been a challenging problem for over half a century, with literature dating from Haldane (1941). Recent articles (which have many other references) include Berger et al. (1999) and DasGupta & Rubin (2005). Basu & Ebrahimi (2001) develop objective priors for a related problem in software reliability.

We are interested in finding a reference prior for n , first for the case of known p and then for the situation in which p is unknown.

5.1 Known p

To find the reference prior for n from this discrete distribution, application of Approach 1 would again result in a non-constant normalization factor. There is no single hierarchical extension that is natural for Approach 2, but a reasonable extension – first used by Raftery (1988a) (see also Moreno & Giron (1998))– is to assume a Poisson $\text{Ps}(n-1 | \lambda)$ distribution for $n-1$. One then derives the reference prior $\pi(\lambda)$ for the corresponding integrated model $p(x | \lambda, p) = \sum_{n=1}^{\infty} \text{Bi}(x | n, p) \text{Ps}(n-1 | \lambda)$, and uses this to obtain the corresponding reference prior $\pi(n | p) \propto \int_0^{\infty} \text{Ps}(n-1 | \lambda) \pi(\lambda | p) d\lambda$. As this prior is not available in closed form and the choice of the Poisson hierarchical extension is rather arbitrary, we do not pursue this further, but it is worth noting that the result is approximately proportional to our eventually

recommended prior: $\pi(n | p) \propto n^{-1/2}$.

5.1.1 Approach 3

Let x_1, \dots, x_k be k independent replications from $\text{Bi}(x | n, p)$. A simple linear estimate of n is

$$\hat{n} = \frac{\bar{x}_k}{p} = \frac{1}{pk} \sum_{j=1}^k x_j. \quad (25)$$

Note that

$$\text{E}(\hat{n} | n, p) = n, \text{ and } \text{Var}(\hat{n} | n) = \frac{n(1-p)}{kp}.$$

It follows from the central limit theorem that

$$p(\hat{n} | n, p) \approx N\left(\hat{n} \mid n, \sqrt{\frac{n(1-p)}{kp}}\right). \quad (26)$$

Extending this to the model continuous in n , the Jeffreys-rule prior is

$$\pi_1(n | p) \propto \frac{1}{\sqrt{n}}. \quad (27)$$

As mentioned in Section 1.2.3, this is (necessarily) the same result derived in Barger & Bunge (2008), since we utilized a linear estimator and the problem can be shown to have a linear difference score.

Of course, (25) is not the only inefficient (but consistent) estimate of p . Another possible estimate of n here is

$$\hat{n} = \frac{1}{2}(\sqrt{1 + 16S^2} - 1), \quad (28)$$

where $S^2 = \sum x_i^2/k$. A laborious application of Approach 3 with this estimate yields the reference prior

$$\pi_2(n | p) \propto \frac{1}{\sqrt{n}} \times \frac{2pn + 1 - p}{\sqrt{4p^2n^2 + 2pn(3 - 5p) + 1 - 6p(1 - p)}}. \quad (29)$$

Since this differs from the prior in (27), it can be concluded that the choice of (inefficient) statistic in Approach 3 does matter, making the approach less attractive. However, the actual difference between these two priors is relatively minor. Indeed, the ratio $\pi_2(n | p)/\pi_1(n | p)$ is maximized at $n = 1$, with the value $1 + p$, and decreases monotonically to 1 as $n \rightarrow \infty$.

5.1.2 Approach 4

From the fact that x/n is asymptotically normal with mean p and variance $p(1 - p)/n$, it follows that the parameter-based asymptotic distribution of x as $n \rightarrow \infty$ is (treating x as continuous) $N(x | np, np(1 - p))$. Treating n as also being continuous, the appropriate reference prior would be the Jeffreys-rule prior for n from this distribution, which can be computed to be

$$\pi_3(n | p) \propto \frac{1}{\sqrt{n}} \times \sqrt{1 + \frac{(1 - p)}{2pn}}. \quad (30)$$

Again, the large n behavior of this prior is the same as that of the previously determined priors to first order (in n). The behavior of this prior can be quite different, however. In particular, for small p , the prior behaves like $1/n$ instead of $1/\sqrt{n}$ when n is moderate, and this can lead to different answers. Note that there is no assurance that the performance of this prior for moderate n is adequate, since it was derived based on a “large n ” argument.

5.1.3 Comparison

We have three candidate reference priors. From the viewpoint of simplicity, $\pi_1(n) = 1/\sqrt{n}$ is clearly the most attractive prior, but it may be that the dependence on p of $\pi_2(n)$ and $\pi_3(n)$ results in superior performance. We again study this issue by looking at the frequentist performance of credible sets, following exactly the methods discussed in Section 4.3. We include in the comparison two other priors that have been considered for this problem, namely $\pi_4(n) = n^{-1}$ and $\pi_5(n) = 1$. Table 4 is a representative sample of the many simulation results

Table 4: Average posterior coverage, frequentist coverage, and average size of one-sided 50% credible intervals based on 20,000 simulations of sample size m from the $\text{Bi}(x \mid 10, p)$ distribution.

(m, p)	Prior	APC	Coverage	Difference	AS
(5, 0.5)	π_1	0.639	0.638	0.001	10.015
	π_2	0.639	0.638	0.001	10.011
	π_3	0.640	0.638	0.002	10.014
	π_4	0.637	0.615	0.022	9.908
	π_5	0.635	0.662	-0.027	10.103
(5, 0.1)	π_1	0.545	0.578	-0.033	10.275
	π_2	0.541	0.578	-0.037	10.218
	π_3	0.562	0.569	-0.008	10.105
	π_4	0.544	0.432	0.112	9.334
	π_5	0.539	0.598	-0.059	11.123
(50, 0.01)	π_1	0.527	0.555	-0.028	10.059
	π_2	0.531	0.555	-0.024	10.026
	π_3	0.523	0.410	0.113	9.214
	π_4	0.529	0.410	0.120	9.102
	π_5	0.523	0.573	-0.051	11.025

that were examined. The following conclusions can be reached from this (and the many other simulation results):

- The prior $\pi_4(n) = 1/n$ systematically overstates the actual coverage, and by a large amount; hence it can be eliminated from consideration.
- The prior $\pi_5(n) = 1$ systematically understates the actual coverage by a rather large amount, and yields too large credible sets; hence it can be eliminated from consideration.
- The prior $\pi_3(n)$ sometimes performed best, but seriously overstated the actual coverage for the case $m = 50$ and $p = 0.01$. It is for small p that this prior significantly differs from the other candidate reference priors; hence, the indication is that this difference is harmful.
- The reference priors $\pi_1(n)$ and $\pi_2(n)$ had very similar good performances, even for the large p cases (not shown here) where they can be somewhat different as priors. Given the fact that $\pi_1(n)$ is much simpler, it emerges as the clear choice.

The overall conclusion thus seems to be that the recommended reference prior for the binomial distribution with known p is $\pi^R(n) = 1/\sqrt{n}$.

5.2 Unknown p

With p unknown as well as n , suppose there are $m \geq 2$ observations x_1, \dots, x_m from the $\text{Bi}(x | n, p)$ distribution. (It actually also works to take $m = 1$ in the following, but using one observation when there are two unknown parameters would be rather unusual.) The likelihood function of (n, p) , based on (x_1, \dots, x_m) , is

$$p(x_1, \dots, x_m | n, p) = \prod_{j=1}^m \binom{n}{x_j} p^{x_j} (1-p)^{n-x_j}, \quad (31)$$

where $s = \sum_{j=1}^m x_j$. Because the Jeffreys-rule or reference prior for p given n is $\text{Be}(p \mid 1/2, 1/2)$, the marginal likelihood of n is thus

$$p(x_1, \dots, x_m \mid n) = \frac{1}{\pi} \prod_{j=1}^m \binom{n}{x_j} \frac{\Gamma(s + 1/2)\Gamma(mn - s + 1/2)}{\Gamma(mn + 1)}. \quad (32)$$

To finish, we need to find a reasonable objective prior for n for this discrete parameter model.

5.2.1 Approach 3

We first compute the mean and variance of $m^{-1} \sum_{i=1}^m x_i$, from the distribution (32). From (19) and (20), it is clear that $E(x_i) = E(x_1) = n/2$ and $\text{Var}(x_i) = n(n + 1)/2$. It is also easy to verify that, for $i \neq j$,

$$\text{Cov}(x_i, x_j) = E[\text{Cov}(x_i, x_j \mid p)] = E[(np - n/2)(np - n/2)] = n^2 \text{Var}(p) = n^2/8.$$

Thus

$$\begin{aligned} E\left(\frac{1}{m} \sum_{i=1}^m x_i\right) &= \frac{n}{2}, \\ \text{Var}\left(\frac{1}{m} \sum_{i=1}^m x_i\right) &= \frac{1}{m^2} \left\{ m \text{Var}(x_1) + m(m-1) \text{Cov}(x_1, x_2) \right\} \\ &= \frac{n}{2m} \left\{ \frac{(m+3)n}{4} + 1 \right\}. \end{aligned}$$

Now going to k independent replications $(x_{i1}, \dots, x_{im}), i = 1, \dots, k$, from (32) for the reference prior asymptotics, a simple consistent linear estimate of n is

$$\hat{n} = \frac{2}{km} \sum_{i=1}^k \sum_{j=1}^m x_{ij}. \quad (33)$$

It follows from the central limit theorem that

$$p(\hat{n} \mid n, p) \approx N\left(\hat{n} \mid n, \sqrt{\frac{2n}{km} \left\{ \frac{(m+3)n}{4} + 1 \right\}}\right). \quad (34)$$

Extending this to the model continuous in n , the reference prior for n is

$$\pi_1(n) \propto \left(\frac{n}{m} \left\{ \frac{(m+3)n}{4} + 1 \right\} \right)^{-\frac{1}{2}} \propto \frac{1}{\sqrt{n(n + \frac{4}{m+3})}}. \quad (35)$$

Note that, when $m = 1$, this is (necessarily) the same prior as that in (23) for $a = b = 1/2$, since the two models are then the same.

5.2.2 Approach 4

For fixed $p \in (0, 1)$ and as $n \rightarrow \infty$, the x_i/n are asymptotically independent and normally distributed with mean p and variance $p(1-p)/n$. Thus, for any $y_i \in (0, 1)$ and as $n \rightarrow \infty$,

$$\begin{aligned} P\left(\frac{x_1}{n} \leq y_1, \dots, \frac{x_m}{n} \leq y_m\right) &= \int_0^1 P\left(\frac{x_1}{n} \leq y_1, \dots, \frac{x_m}{n} \leq y_m \mid p\right) \text{Be}(p \mid \tfrac{1}{2}, \tfrac{1}{2}) dp \\ &\rightarrow \int_0^1 \prod_{i=1}^m 1_{(0, y_i)}(p) \text{Be}(p \mid \tfrac{1}{2}, \tfrac{1}{2}) dp \\ &= \int_0^{\min\{y_1, \dots, y_m\}} \text{Be}(p \mid \tfrac{1}{2}, \tfrac{1}{2}) dp. \end{aligned}$$

The limiting distribution of $(x_1/n, \dots, x_m/n)$ does not depend on n , and $n > 0$ is the scale parameter of the limiting distribution of (x_1, \dots, x_m) . The prior for n is thus the usual scale reference prior $\pi_2(n) \propto 1/n$.

5.2.3 Comparison

The ratio $\pi_2(n)/\pi_1(n)$ is $\sqrt{1 + 4/[n(m+3)]}$, which is small enough that there will not be an appreciable difference in the answers produced by the two priors. Hence we do not utilize a simulation study to select between them, but simply suggest – on the basis of simplicity – that $\pi_2(n)$ be used. The recommended reference prior for the binomial problem with both p and n unknown is thus

$$\pi^R(p, n) \propto \frac{1}{n} \times \text{Be}(p \mid \tfrac{1}{2}, \tfrac{1}{2}). \quad (36)$$

Although Jeffreys never explicitly studied this problem, $\pi^R(p, n)$ is presumably the prior he would have used, since he recommended the $\text{Be}(p \mid \frac{1}{2}, \frac{1}{2})$ prior for p and recommended $1/n$ for an infinite positive variable.

6 Prior Model Probabilities

Suppose we have m potential models \mathcal{M}_i , $i = 1, \dots, m$, for the data, and consider the problem of determining objective prior model probabilities, $\pi(\mathcal{M}_i)$.

6.1 Finite Number of Models, No Structure

If the models being considered have no structure, we would simply use the usual reference prior $\pi(\mathcal{M}_i) = 1/m$. This would occur for instance if there were just two models, or if the models had no relationships to each other. For instance, in selecting from among three different location-scale models – normal, Cauchy, and exponential – with only the location and scale parameters of each model being unknown, there is no structure that could lead to any objective assignment of prior probabilities other than $\pi(\mathcal{M}_i) = 1/3$.

6.2 Variable Selection

Consider the variable selection problem where there are m variables (or graphical nodes or links) that can make up a model. Let γ be the vector of zeroes and ones that indicates which variables are in the model, and denote the resulting model as \mathcal{M}_γ . Note that there are 2^m possible models so that, if structure were ignored, the reference prior would assign each model probability $\pi(\mathcal{M}_\gamma) = 2^{-m}$. It is being increasingly realized, however (cf. Ley & Steel (2007) and Scott & Berger (2008)) that this assignment of prior model probabilities has serious deficiencies (as first recognized by Jeffreys (1961)), such as not allowing for a multiplicity correction based on the number of variables being considered. Hence the use of reference priors based on structure is appealing.

Structure 1. One obvious structure here is that each variable can be in or out of the model. One can thus implement the hierarchical Approach 2, and assume that each variable is in the model with unknown probability p . Then the probability of any model with r variables is $\pi(\mathcal{M}_\gamma | p) = p^r(1-p)^{m-r}$. Formal application of the reference prior approach would now require marginalizing over γ , and then finding the reference (Jeffreys-rule) prior for p in the ensuing marginal model.

Unfortunately, this computation is extremely difficult and problem dependent, and cannot

be carried out in closed form. A rather ad hoc way forward is to, instead, simply use the usual reference prior for p in a $\text{Bi}(r \mid m, p)$ model, which is $\text{Be}(p \mid \frac{1}{2}, \frac{1}{2})$. Utilizing this, it follows that the structured reference model prior probabilities are

$$\pi(\mathcal{M}_\gamma) = \int_0^1 p^r (1-p)^{m-r} \text{Be}(p \mid \frac{1}{2}, \frac{1}{2}) dp = \frac{\Gamma[r + \frac{1}{2}] \Gamma[m - r + \frac{1}{2}]}{\pi \Gamma[m + 1]}. \quad (37)$$

Structure 2. Suppose we were instead to structure the problem (following Jeffreys (1961)) by saying that model size r is the parameter of interest and that, given r , there is no model structure. One might then assign the $m + 1$ possible values of r equal prior probabilities, i.e. $P(r) = (m + 1)^{-1}$ and assign all models of a given size r equal prior probability $1/\binom{m}{r}$. This leads to the prior probability assignment

$$\pi(\mathcal{M}_\gamma) = \frac{1}{(m + 1) \binom{m}{r}}. \quad (38)$$

This argument is somewhat less satisfying than that for Structure 1, in that the assignment of equal prior probabilities to each model size r seems rather arbitrary. In contrast, the main assumption going into Structure 1 was an exchangeability assumption. Interestingly, however, (38) can also be derived as in (37) if, instead, the uniform prior is used for p ; the choice of Jeffreys is thus reasonable from either perspective.

7 Appendix: Proof of Lemma 2.1

Proof. It is easy to see that

$$\frac{\partial^2}{\partial N^2} \log(f(V \mid N)) = \frac{R}{(V + N - 3)^2} - \sum_{j=0}^{R-1} \frac{1}{(N - j)^2}.$$

Then the Fisher information of N is (4), where

$$J_{R,N} = \int_1^R \frac{1}{(v + N - 3)^{R+2}} g_R(v) dv. \quad (39)$$

Here $g_R(v)$ is given by (2). We can rewrite (2) as

$$g_R(v) = \begin{cases} \binom{R-1}{0}(v-1)^{R-2}, & \text{if } 1 < v < 2, \\ \binom{R-1}{0}(v-1)^{R-2} - \binom{R-1}{2}(v-2)^{R-2}, & \text{if } 2 < v < 3, \\ \dots & \dots \\ \binom{R-1}{0}(v-1)^{R-2} - \dots + (-1)^R \binom{R-1}{R-2}(v-j)^{R-2}, & \text{if } R-1 < v < R. \end{cases}$$

It follows that

$$J_{R,N} = \sum_{i=1}^{R-1} (-1)^{i-1} \binom{R-1}{i-1} J_{i,R,N}, \quad (40)$$

where, for $i < R \leq N$,

$$J_{i,R,N} = \int_i^R \frac{(v-i)^{R-2}}{(v+N-R)^{R+2}} dv. \quad (41)$$

Making transformations $y = v - i$ and then $u = j - 1$,

$$\begin{aligned} J_{i,R,N} &= \sum_{j=1}^{R-i} \int_{j-1}^j \frac{y^{R-2}}{(y+N-R+i)^{R+2}} dy \\ &= \sum_{j=1}^{R-i} \int_0^1 \frac{(u+j-1)^{R-2}}{(u+N-R+i+j-1)^{R+2}} du \\ &= \sum_{j=1}^{R-i} \int_0^1 \frac{[(u+N-R+i+j-1) - (N-R+i)]^{R-2}}{(u+N-R+i+j-1)^{R+2}} du \\ &= \sum_{j=1}^{R-i} \sum_{k=0}^{R-2} (-1)^{R-2-k} \binom{R-2}{k} (N-R+i)^{R-2-k} \int_0^1 \frac{1}{(u+N-R+i+j-1)^{R+2-k}} du \\ &= \sum_{j=1}^{R-i} \sum_{k=0}^{R-2} (-1)^{R-2-k} \binom{R-2}{k} \frac{(N-R+i)^{R-2-k}}{R+1-k} \frac{1}{(u+N-R+i+j-1)^{R+1-k}} \\ &+ \sum_{j=1}^{R-i} \sum_{k=0}^{R-2} (-1)^{R-2-k} \binom{R-2}{k} \frac{(N-R+i)^{R-2-k}}{R+1-k} \frac{1}{(u+N-R+i+j)^{R+1-k}} \\ &= \frac{1}{(N-R+i)^3} \sum_{k=0}^{R-2} (-1)^{R-2-k} \binom{R-2}{k} \frac{1}{R+1-k} \\ &- \frac{1}{N^3} \sum_{k=0}^{R-2} (-1)^{R-2-k} \binom{R-2}{k} \left(\frac{N-R+i}{N} \right)^{R-2-k} \frac{1}{R+1-k} \\ &\equiv \frac{1}{(N-R+i)^3} G_R + \frac{1}{N^3} J_{i,R,N,2}, \end{aligned}$$

where $G_R = \sum_{k=0}^{R-2} (-1)^{R-2-k} \binom{R-2}{k} \frac{1}{R+1-k}$. Using the equality

$$\frac{1}{R+1-k} = \frac{1}{R-1-k} - \frac{2}{(R-1-k)(R-k)} + \frac{2}{(R-1-k)(R-k)(R+1-k)},$$

we can evaluate

$$\begin{aligned}
G_R &= \sum_{k=0}^{R-2} (-1)^{R-2-k} \binom{R-2}{k} \frac{1}{R-1-k} \\
&- 2 \sum_{k=0}^{R-2} (-1)^{R-2-k} \binom{R-2}{k} \frac{1}{(R-1-k)(R-k)} \\
&+ 2 \sum_{k=0}^{R-2} (-1)^{R-2-k} \binom{R-2}{k} \frac{1}{(R-1-k)(R-k)(R+1-k)} \\
&= (1-1)^{R-1} \frac{1}{R-1} - (-1)^{R-1} \frac{1}{R-1} \\
&- 2(1-1)^R \frac{1}{(R-1)R} + 2(-1)^R \binom{R}{R-1} \frac{1}{(R-1)R} + 2(-1)^R \frac{1}{(R-1)R} \\
&+ 2(1-1)^{R+1} \frac{1}{(R-1)R(R+1)} - 2 \sum_{k=R-1}^{R+1} (-1)^k \binom{R+1}{k} \frac{1}{(R-1)R(R+1)} \\
&= \frac{2(-1)^R}{(R-1)R(R+1)} = \frac{2(-1)^R}{R^3 - R}.
\end{aligned}$$

By a similar argument, one can obtain

$$\sum_{i=1}^{R-1} (-1)^{i-1} \binom{R-1}{i-1} J_{i,R,N,2} = \frac{2(-1)^R}{R^3 - R}. \quad (42)$$

Substituting these expressions into (40), results in

$$\begin{aligned}
J_{R,N} &= \frac{2}{R^3 - R} \sum_{i=1}^{R-1} (-1)^{R-i-1} \binom{R-1}{i-1} \frac{1}{(N-R+i)^3} + \frac{2(-1)^R}{(R^3 - R)N^3} \\
&= \frac{2}{R^3 - R} \sum_{i=0}^{R-1} (-1)^i \binom{R-1}{i} \frac{1}{(N-R+1+i)^3},
\end{aligned}$$

proving the lemma. □

References

- Barger, K. & Bunge, J. (2008), ‘Bayesian estimation of the number of species using noninformative priors’, *Biometrical Journal* **50**, 1064–1076.
- Basu, S. & Ebrahimi, N. (2001), ‘Bayesian capture-recapture methods for error detection and estimation of population size: heterogeneity and dependence’, *Biometrika* **2001**, 269–279.

- Berger, J. O. & Bernardo, J. M. (1992), On the development of reference priors, *in* ‘Bayesian Statistics 4. (J. M. Bernardo, J. O. Berger, D. V. Lindley and A. F. M. Smith, eds)’, Oxford University Press, London, pp. 35–60 (with discussion).
- Berger, J. O., Bernardo, J. M. & Sun, D. (2009a), ‘The formal definition of reference priors’, *The Annals of Statistics* **37**, in press.
- Berger, J. O., Bernardo, J. M. & Sun, D. (2009b), ‘Natural induction: an objective Bayesian approach’, *Rev. R. Acad. Cienc. Exactas Fis. Nat., Ser. A Mat.* **103**, in press.
- Berger, J. O., Liseo, B. & Wolpert, R. L. (1999), ‘Integrated likelihood methods for eliminating nuisance parameters’, *Statistical Science* **14**, 1–28 (with discussion).
- Bernardo, J. M. (1979), ‘Reference posterior distributions for Bayesian inference’, *Journal of the Royal Statistical Society, Series B* **41**, 113–147 (with discussion). Reprinted in *Bayesian Inference* **1** (G. C. Tiao and N. G. Polson, eds.), Oxford: Edward Elgar, 1995, 229–263.
- Bernardo, J. M. (2005), Reference analysis, *in* ‘*Handbook of Statistics* **25** (D. K. Dey and C. R. Rao eds.)’, Amsterdam: Elsevier, pp. 17–90.
- Bernardo, J. M. & Smith, A. F. M. (1994), *Bayesian Theory*, John Wiley & Sons.
- Broad, C. D. (1918), ‘On the relation between induction and probability’, *Mind* **27**, 389–404.
- DasGupta, A. & Rubin, H. (2005), ‘Estimation of binomial parameters when both n, p are unknown’, *Journal of Statistical Planning and Inference* **130**, 391–404.
- Goudie, I. B. J. & Goldie, C. M. (1981), ‘Initial size estimation for the pure death process’, *Biometrika* **68**, 543–550.
- Jeffreys, H. (1961), *Theory of Probability*, Oxford University Press, London.
- Kahn, W. D. (1987), ‘A cautionary note for Bayesian estimation of the binomial parameter n ’, *Amer Statistician* **41**, 38–39.

- Kramer, M. & Starr, N. (1990), ‘Optimal stopping in a size dependent search’, *Sequential Anal.* **9**, 59–80.
- Laplace, P. S. (1774), Mémoire sur la probabilité des causes par les événements, in ‘*Oeuvres Complètes* **8**’, Paris: Gauthier-Villars, 1891, pp. 27–68.
- Lawless, J. (1982), *Statistical Models and Methods for Lifetime Data*, John Wiley & Sons, New York.
- Ley, E. & Steel, M. F. (2007), On the effect of prior assumptions in Bayesian model averaging with applications to growth regression, in ‘*Policy Research Working Paper Series*’, **4238**, World Bank.
- Lindsay, B. G. & Roeder, K. (1987), ‘A unified treatment of integer parameter models’, *Journal of the American Statistical Association* **82**, 758–764.
- Moreno, E. & Giron, J. (1998), ‘Estimating with incomplete count data: A Bayesian approach’, *Journal of Statistical Planning and Inference* **66**, 147–159.
- Raftery, A. E. (1988a), ‘Inference for the binomial n parameter: a hierarchical Bayes approach’, *Biometrika* **75**, 223–228.
- Raftery, A. E. (1988b), ‘Analysis of a simple debugging model’, *Appl. Statist.* **37**, 12–22.
- Rissanen, J. (1983), ‘A universal prior for integers and estimation by minimum description length’, *Annals of Statistics* **11**, 416–431.
- Scott, J. & Berger, J. (2008), Bayes and empirical Bayes multiplicity adjustment in the variable-selection problem, Discussion Paper 2008-10, Duke University Department of Statistical Science.
- Starr, N. (1974), ‘Optimal and adaptive stopping based on capture times’, *J. Appl. Prob.* **11**, 294–301.
- Sweeting, T. J. (1992), ‘Parameter-based asymptotics’, *Biometrika* **79**, 219–232.