

# Objective 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 multivariate hypergeometric distribution, the binomial-beta distribution, the binomial distribution, and determination of prior model probabilities. The recommended objective priors for the first, third and fourth 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 a more sensible 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  $\theta$ , 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 which seeks to choose that prior distribution which maximizes the asymptotic missing information. Because of the nature of the asymptotics in discrete parameter spaces, however, the prior which will maximize the missing information is the constant prior. And we have been unable to find any modification of the approach – e.g., adding constraints on the problem – that yield anything other than 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  $\theta$  denote the unknown discrete parameter, taking values in a countable discrete space  $\Theta$ . 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  $\theta$  as a continuous variable in the original model. Unfortunately, this typically introduces a normalizing factor into the expression for non-integer  $\theta$ , 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  $\theta$  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  $\theta$  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, Jeffreys (1946, 1961) postulated that the unknown  $R$  arises as a Binomial random variable  $\text{Bi}(R | N, p)$  with  $p$  unknown. The problem can then be reduced to finding the reference prior  $\pi^R(p)$  for  $p$ , a continuous parameter, and this can be used

to determine the desired objective prior for  $R$  via  $\pi^*(R) = \int \text{Bi}(R | N, p) \pi^R(p) dp$ . (If the reference prior for the continuous hyperparameter is improper, one would proceed with the usual reference prior device of approximating it by a series of priors restricted to an increasing series of compact subsets, and taking the appropriate limit of the resulting marginal priors for the parameter of interest.)

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. Also, it may be possible to introduce more than one hierarchical model, resulting in different possible objective priors. This seems unavoidable; different “hierarchical origins” for the discrete parameter model may naturally result in different priors.

### 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  $\theta$  (based on the  $k$  replications), which effectively becomes continuous as  $k \rightarrow \infty$ ;
- finding the asymptotic sampling distribution of  $\hat{\theta}_k$  as  $k \rightarrow \infty$ ;
- pretending that  $\theta$  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  $\theta$  is continuous, and hence the desired objective prior is (cf. Theorem 9 of Bernardo, 2005)  $\pi^*(\theta) \propto \sigma(\theta)^{-1}$ .

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 an objective 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  $\theta$  to make the problem continuous:

- Let  $\theta \rightarrow \infty$  (or ensure this by forcing some other parameter to  $\infty$ ).
- In the limiting asymptotic distribution of  $x$  (or some related random variable), let  $\theta$  be continuous (if possible).
- Do the reference prior replications over  $k$  with this asymptotic distribution, to define an objective prior.

For instance, in the uniform problem mentioned in Section 1.2.1,  $x/\theta$  has a uniform distribution on the discrete set  $\{0, \frac{1}{\theta}, \dots, \frac{\theta-1}{\theta}, 1\}$ . As  $\theta$  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  $x/\theta$  being Uniform on  $(0,1)$ , or  $x$  being Uniform on  $(0, \theta)$ . Pretending that  $x$  and  $\theta$  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 an objective prior for ‘large  $\theta$ ,’ and this may not be right for small  $\theta$ . 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,  $\pi^*(\theta)$ , is infinite at the endpoints of the discrete parameter space  $\Theta$ . A natural (if ad hoc) solution is to then define the objective prior by discretizing  $\pi^*(\theta)$  on the continuous extension of the parameter space  $\Theta^*$ : write  $\Theta^* = \cup_j \Theta_j^*$ , where  $\Theta_j^*$  is a set containing the discrete parameter value  $\theta_j$ , and then define  $\pi^*(\theta_j) = \int_{\Theta_j^*} \pi^*(\theta) d\theta$ , 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 finite (indeed proper) discrete objective 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. Even though the resulting priors arise from reference priors in an extended model, we will not call them reference priors because they do not formally arise as priors that minimize asymptotic missing information in the original problem.

When Approach 1 (without renormalization) or Approach 2 (when there is a single natural hierarchical model) are applicable, we view the resulting prior as being the natural objective prior, and look no further. For problems where neither approach applies, we utilize some combination of Approaches 3 and 4 to propose an objective 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  $\theta$ . This is probably unavoidable; unless there is enough structure to use Approaches 1 or 2, unequivocal determination of an objective prior for small  $\theta$  seems not to be possible. That all the methods seem to provide essentially the same objective prior for large  $\theta$  is, however, gratifying, in that it is typically for large  $\theta$  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 *multivariate hypergeometric distribution* (in Section 4), the *binomial-beta distribution* (in Section 5), the *binomial distribution* (in Section 6), and *determination of prior model probabilities* (in Section 7). The objective priors derived in Sections 2, 4, and 5 are new. The analyses in Sections 6 and 7 validate objective 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  $\lambda$  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  $(N, \lambda)$  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  $(N, \lambda)$  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  $\lambda$  given  $N$ .

Indeed, it can be directly shown that (3) results, up to a proportionality constant, from integrating out  $\lambda$  in (1), with respect to the Haar density. Thus, from either a frequentist or objective Bayesian perspective, dealing with unknown  $N$  reduces to analysis of (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 an objective 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 Appendix A).

**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 objective prior, it is convenient to consider the reparameterization  $\theta = N - R + 1$ , since the range of  $\theta$  is then the positive integers  $\mathcal{N} = \{1, 2, \dots\}$ . The objective prior for  $\theta$  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 objective prior for any  $R$ . (Of course, the exact prior is not difficult to program and work with.)

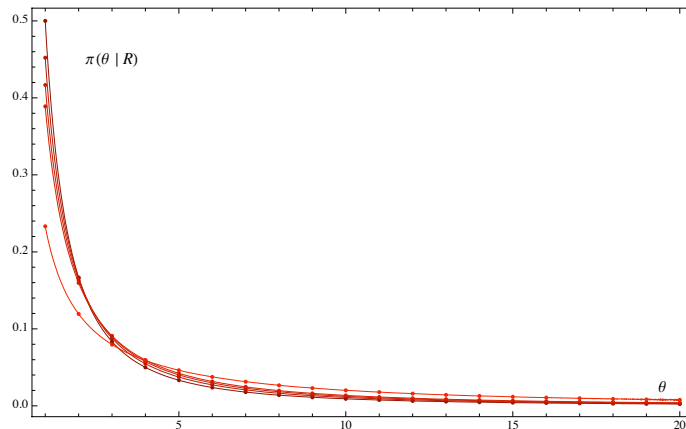


Figure 1: Objective priors for  $\theta$  for  $R = 2, 3, 4, 5$ , and 20 (top to bottom at  $\theta = 1$ ).

### 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

$$\begin{aligned} \Pr(r | n, N, p) &= \sum_{R=0}^N \Pr(r | n, R, N) \Pr(R | N, p) \\ &= \sum_{R=0}^N \frac{\binom{R}{r} \binom{N-R}{n-r}}{\binom{N}{n}} \binom{N}{R} p^R (1-p)^{N-R} \\ &= \binom{n}{r} p^r (1-p)^{n-r}, \end{aligned} \quad (7)$$

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 objective prior for  $R$  in the hypergeometric model

$$\begin{aligned} \pi^*(R | N) &= \int_0^1 \text{Bi}(R | N, p) \text{Be}(p | \frac{1}{2}, \frac{1}{2}) dp \\ &= \frac{1}{\pi} \frac{\Gamma(R + \frac{1}{2}) \Gamma(N - R + \frac{1}{2})}{\Gamma(R + 1) \Gamma(N - R + 1)}, \end{aligned} \quad (8)$$

for  $R \in \{0, 1, \dots, N\}$ , as suggested in Jeffreys (1946, 1961). Note that, by construction, this is a proper prior.

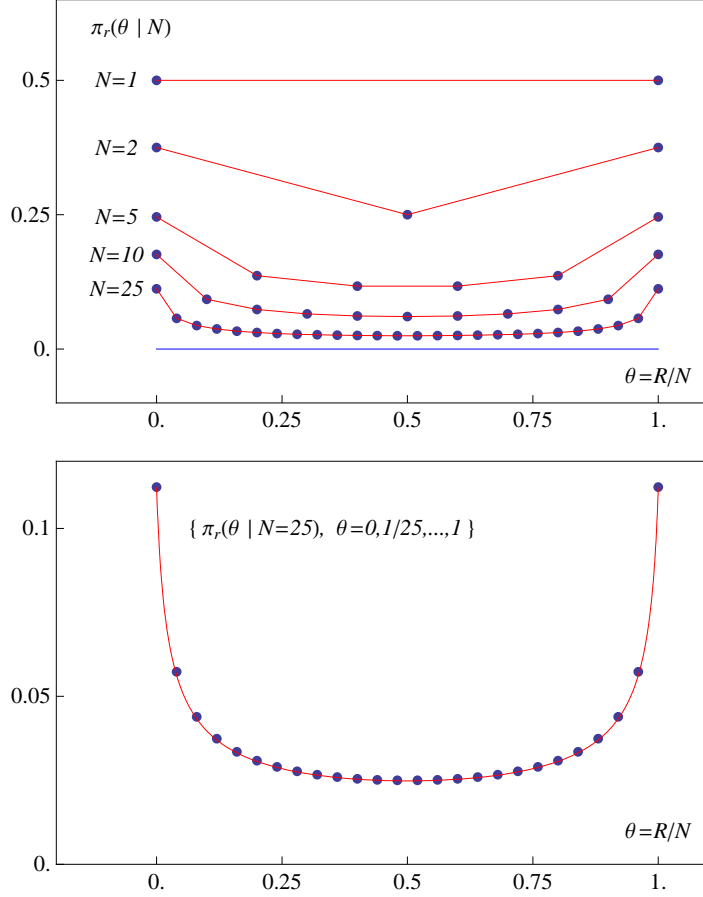


Figure 2: Objective priors  $\pi^*(\theta | N)$  for the proportion  $\theta = R/N$  of conforming items in a population of size  $N$ , for several  $N$  values (upper panel), and its continuous approximation (lower panel).

The upper panel of Figure 2 graphs the objective priors  $\pi^*(\theta | N)$  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 objective prior (8) 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 (8) one obtains, in terms of  $\theta = R/N$ ,

$$\pi^*(\theta | N) \approx \frac{1}{N + \frac{2}{\pi}} \text{Be}\left(\frac{N\theta + \frac{1}{\pi}}{N + \frac{2}{\pi}} \mid \frac{1}{2}, \frac{1}{2}\right), \quad \theta = 0, 1/N, \dots, N, \quad (9)$$

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 (9) as a continuous function of  $\theta$ , the approximation is very good, even if  $N$  is moderate.

## 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 an objective prior proportional to  $1/\sqrt{R(N-R)}$ , which is very similar to (8). 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 objective 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

$$\Pr(+ | N, n, r = n) = \frac{n + 1}{n + 2}, \quad (10)$$

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 (10) for its continuous binomial approximation.

Given that we argue for  $\pi^R(R|N)$  in (8) 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 (10), as computed in Berger et al. (2009b), is

$$\Pr^R(+ | N, n, r = n) = \frac{n + 1/2}{n + 1}. \quad (11)$$

## 4 Multivariate Hypergeometric Distribution

Let  $\mathcal{N}_+$  be the set of all nonnegative integers. Consider a multivariate hypergeometric distribution  $\text{Mu-Hy}(n, \mathbf{R}, N)$ , with the probability mass function on  $\mathcal{R}_{k,n} = \{\mathbf{r}_k = (r_1, \dots, r_k) : r_j \in \mathcal{N}_+, r_1 + \dots + r_k \leq n\}$ ,

$$\text{Hy}_k(\mathbf{r}_k | n, \mathbf{R}_k, N) = \frac{\binom{R_1}{r_1} \dots \binom{R_k}{r_k} \binom{R_{k+1}}{r_{k+1}}}{\binom{N}{n}}, \quad \mathbf{r}_k \in \mathcal{R}_{k,n}, \quad (12)$$

where the unknown parameters  $\mathbf{R}_k = (R_1, \dots, R_k)$  are in the parameter space  $\mathcal{R}_{k,N} = \{\mathbf{R}_k = (R_1, \dots, R_k) : R_j \in \mathcal{N}_+, R_1 + \dots + R_k \leq N\}$ . Here and in the following,  $R_{k+1} = N - (R_1 + \dots + R_k)$ . Note the hypergeometric distribution is the special case when  $k = 1$ .

We again consider Approach 2 here. A natural hierarchical model for the unknown  $\mathbf{R}_k$  is to assume that it is multinomial  $\text{Mu}_k(N, \mathbf{p}_k)$ , with unknown parameters  $\mathbf{p}_k \in \mathcal{P}_k \equiv \{\mathbf{p}_k = (p_1, \dots, p_k) : 0 \leq p_j \leq 1, p_1 + \dots + p_k \leq 1\}$ . The probability mass function of  $\mathbf{R}_k$  is then

$$\text{Mu}_k(\mathbf{R}_k | N, \mathbf{p}_k) = \frac{N!}{\prod_{j=1}^{k+1} R_j!} \prod_{j=1}^{k+1} p_j^{R_j}. \quad (13)$$

The problem then reduces to finding the objective priors for the continuous hyperparameter  $\mathbf{p}_k$ . In contrast to the situation with the hypergeometric distribution, there are at least two natural choices for this objective prior.

The most commonly used objective prior for  $\mathbf{p}_k$  is the Jeffreys prior  $\pi_J$ , given by

$$\pi_J(\mathbf{p}_k) = \frac{\Gamma(\frac{k+1}{2})}{\Gamma(\frac{1}{2})^{k+1}} \prod_{j=1}^{k+1} p_j^{\frac{1}{2}-1}. \quad (14)$$

This is the objective prior when the whole vector  $\mathbf{R}_k$  is of interest and all the corresponding hyper-parameters  $\mathbf{p}_k$  are of interest.

Another reasonable objective prior arises when a particular component, say  $R_1$  is of interest. Noting that  $E(R_j | \mathbf{p}_k) = Np_j$ , for all  $j = 1, \dots, k$ , this would imply that  $p_1$  is the hyperparameter of interest in the hierarchical model. For this situation, Berger & Bernardo (1992) introduce the one-at-a-time reference prior, corresponding to the importance ordering  $\{p_1, \dots, p_k, p_{k+1}\}$  for parameters, as

$$\pi_R(\mathbf{p}_k) = \frac{1}{\pi^k} \prod_{j=1}^k \frac{1}{\sqrt{p_j(1 - \delta_j)}}, \quad \delta_j = \sum_{i=1}^j p_i. \quad (15)$$

In the following we consider analysis under both of these hyperpriors, without further comment as to the choice between them. See Berger & Bernardo (1992) for extensive discussion of the considerations involved in making this choice.

*The Marginal Prior of  $\mathbf{R}_k$  given  $N$ .* We first give the marginal priors for  $\mathbf{R}_k$  based on the two objective priors for  $\mathbf{p}_k$ . The proof and some general results are given in Appendix B.

**Theorem 4.1** *Define the function,*

$$f(y) = \frac{\Gamma(y + 1/2)}{\sqrt{\pi} \Gamma(y + 1)}, \quad y \geq 0. \quad (16)$$

(a) *Under the Jeffreys prior  $\pi_J$  for  $\mathbf{p}_k$ , the marginal prior of  $\mathbf{R}_k$  given  $N$  is*

$$p_J(\mathbf{R}_k | N) = \left\{ \prod_{i=1}^{k+1} f(R_i) \right\} \frac{N! \Gamma(\frac{k+1}{2})}{\Gamma(N + \frac{k+1}{2})}, \quad \mathbf{R}_k \in \mathcal{R}_{k,N}. \quad (17)$$

(b) *Under the reference prior  $\pi_R$  for  $\mathbf{p}_k$ , the marginal prior of  $\mathbf{R}_k$  given  $N$  is*

$$p_R(\mathbf{R}_k | N) = \left\{ \prod_{i=1}^{k+1} f(R_i) \right\} \left\{ \prod_{j=2}^k f(R_j + \dots + R_k + R_{k+1}) \right\}, \quad \mathbf{R}_k \in \mathcal{R}_{k,N}. \quad (18)$$

**Remark 4.1** *It is interesting that, under the reference prior  $\pi_R$  for  $\mathbf{p}_k$ ,*

$$p_R(\mathbf{R}_{k-1} | N) = \left\{ \prod_{i=1}^{k-1} f(R_i) \right\} f(\tilde{R}_k) \left\{ \prod_{j=2}^{k-1} f(R_j + \dots + R_{k-1} + \tilde{R}_k) \right\}, \quad \mathbf{R}_{k-1} \in \mathcal{R}_{k-1,N},$$

where  $\tilde{R}_k = N - (R_1 + \dots + R_{k-1})$ . By induction, it follows that

$$p_R(R_1 | N) = f(R_1) f(N - R_1), \quad R_1 = 0, \dots, N, \quad (19)$$

which reduces to the prior for the Hypergeometric problem.

The Marginal Likelihood of  $\mathbf{r}_k$ .

**Theorem 4.2** (a) The marginal likelihood of  $\mathbf{r}_k$  given  $(\mathbf{p}_k, n, N)$  depends only on  $(n, \mathbf{p}_k)$ ,

$$\begin{aligned} p(\mathbf{r}_k | \mathbf{p}_k, n, N) &= \sum_{\mathbf{R}_k \in \mathcal{N}_{k,N}} Hy_k(\mathbf{r}_k | n, \mathbf{R}_k, N) Mu_k(\mathbf{R}_k | N, \mathbf{p}_k) \\ &= Mu_k(\mathbf{r}_k | n, \mathbf{p}_k), \mathbf{r}_k \in \mathcal{R}_{k,n}. \end{aligned} \quad (20)$$

(b) For any prior  $\pi(\mathbf{p}_k)$ , the marginal likelihood of  $\mathbf{r}_k$  given  $(n, N)$  depends only on  $n$ , and is of the same form as the marginal prior for  $\mathbf{R}_k$ , except for replacing  $N$  by  $n$ .

Proof. The proof of Part (a) is similar to (7). The rest is from algebra.  $\square$

**Remark 4.2** Under the two objective priors for  $\mathbf{p}_k$ , we have the following marginal likelihood of  $\mathbf{r}_k$ .

(a) Under the Jeffreys prior  $\pi_J$  for  $\mathbf{p}_k$ , the marginal mass function of  $\mathbf{r}_k$  given  $(n, N)$  is

$$p_J(\mathbf{r}_k | n, N) = \left\{ \prod_{i=1}^{k+1} f(r_i) \right\} \frac{n! \Gamma(\frac{k+1}{2})}{\Gamma(n + \frac{k+1}{2})}, \mathbf{r}_k \in \mathcal{R}_{k,n}. \quad (21)$$

(b) Under the reference prior  $\pi_R$  for  $\mathbf{p}_k$ , the marginal mass function of  $\mathbf{r}_k$  given  $(n, N)$  is

$$p_R(\mathbf{r}_k | n, N) = \left\{ \prod_{i=1}^{k+1} f(r_i) \right\} \left\{ \prod_{j=2}^k f(r_j + \dots + r_k + r_{k+1}) \right\}, \mathbf{r}_k \in \mathcal{R}_{k,n}, \quad (22)$$

where  $f$  is given by (16).

The Marginal Posterior Distribution of  $\mathbf{R}_k$  given  $(\mathbf{r}_k, n, N)$ . The marginal posterior mass function of  $\mathbf{R}_k$  given  $\mathbf{r}_k$  is then

$$p(\mathbf{R}_k | n, N, \mathbf{r}_k) = \frac{Hy_k(\mathbf{r}_k | n, \mathbf{R}_k, N) Pr(\mathbf{R}_k | N)}{p(\mathbf{r}_k | n, N)}, \mathbf{R}_k \in \mathcal{R}_{k,N}. \quad (23)$$

The following results can be proved using Lemma B.2 in Appendix B.

**Theorem 4.3** (a) Under the Jeffreys prior  $\pi_J$  for  $\mathbf{p}_k$ , the marginal posterior of  $\mathbf{R}_k$  given  $(\mathbf{r}_k, n, N)$  is

$$\pi_J(\mathbf{R}_k | \mathbf{r}_k, n, N) = \frac{(N-n)!}{\prod_{j=1}^{k+1} (R_j - r_j)!} \prod_{j=1}^k \frac{Beta(\frac{1}{2} + R_j, \frac{k-j+1}{2} + R_{j+1} + \dots + R_{k+1})}{Beta(\frac{1}{2} + r_j, \frac{k-j+1}{2} + r_{j+1} + \dots + r_{k+1})}. \quad (24)$$

(b) Under the reference prior  $\pi_R$  for  $\mathbf{p}_k$ , the marginal posterior of  $\mathbf{R}_k$  given  $(\mathbf{r}_k, n, N)$  is

$$\pi_R(\mathbf{R}_k | \mathbf{r}_k, n, N) = \frac{(N-n)!}{\prod_{j=1}^{k+1} (R_j - r_j)!} \prod_{j=1}^k \frac{Beta(\frac{1}{2} + R_j, \frac{1}{2} + R_{j+1} + \dots + R_{k+1})}{Beta(\frac{1}{2} + r_j, \frac{1}{2} + r_{j+1} + \dots + r_{k+1})}. \quad (25)$$

## 5 The Binomial-Beta Distribution

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

$$x \mid (n, p) \sim \text{Bi}(x \mid n, p) \text{ and } p \sim \text{Be}(p \mid a, b), \quad (26)$$

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

$$\begin{aligned} p(x \mid n) &= \frac{n!}{x!(n-x)!} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \int_0^1 p^{x+a-1} (1-p)^{n-x+b-1} dp \\ &= \frac{n!}{x!(n-x)!} \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} \frac{\Gamma(x+a)\Gamma(n-x+b)}{\Gamma(n+a+b)}, \end{aligned} \quad (27)$$

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

$$p(x \mid n) \propto \frac{n!}{(n-x)!} \frac{\Gamma(n-x+b)}{\Gamma(n+a+b)} \approx \frac{1}{n^{a+b-1}(n-x)^{1-b}} \approx \frac{1}{n^a}, \quad (28)$$

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 objective 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.

### 5.1 Approach 3

Going to the asymptotics of reference prior theory, let  $\{x_1, \dots, x_k\}$  be  $k$  independent replications from (27). The mean and variance of the  $x_i$  are easily found to be

$$\text{E}(x_i \mid n) = \text{E}(np \mid n) = n \frac{a}{a+b}, \quad (29)$$

$$\text{Var}(x_i \mid n) = \frac{n(n+a+b)}{(a+b)^2(a+b+1)}. \quad (30)$$

A simple estimate of  $n$  is the linear estimate

$$\hat{n} = \frac{a+b}{a} \bar{x} = \frac{a+b}{a k} \sum_{j=1}^k x_j. \quad (31)$$

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

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

It follows from the central limit theorem that

$$p(\hat{n} | n) \approx N\left(\hat{n} \mid n, \frac{n(n+a+b)}{a^2(a+b+1)k}\right). \quad (32)$$

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

$$\pi_1(n) \propto \left(\frac{n(n+a+b)}{a^2(a+b+1)k}\right)^{-\frac{1}{2}} \propto \frac{1}{\sqrt{n(n+a+b)}}. \quad (33)$$

## 5.2 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  $n \rightarrow \infty$ ,

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

so that the limiting density of  $y = x/n$  is  $\text{Be}(y | a, b)$ . The parameter-based asymptotic distribution of  $x$  thus has density  $n^{-1}\text{Be}(x/n | a, b, n)$ . 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 6.)

### 5.3 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 objective prior. These are both clearly compatible for large  $n$ , which is reassuring, and both yield a proper posterior. The interesting question is thus whether the dependence of  $\pi_1$  on  $a$  and  $b$  results in a superior objective prior.

The following argument suggests that  $\pi_1$  is better. Suppose that  $a$  and  $b$  are very large, in which case the  $\text{Be}(p | a, b)$  distribution in (26) will be very concentrated around  $p = a/(a+b)$ . It follows that  $p(x | n)$  should behave like the  $\text{Bi}(x | n, a/(a+b))$  distribution, in this situation. In the next section, we will see that the recommended objective 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  $\pi_1$  as the objective 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 (27) for given  $n$ , and construct the one-sided credible sets that have posterior probability exceeding a desired threshold  $1 - \alpha$ , namely

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

corresponding to the two priors  $\pi_i, i = 1, 2$ . We then consider the frequentist coverage of these credible sets, namely  $\Pr(C_i(\mathbf{X}) \ni n | n)$ , 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 (27), compute  $C_i(\mathbf{x}_{(j)})$  for each sample, and approximate the

frequentist coverage and the average posterior coverage (APC) as, respectively,

$$\text{Coverage} \cong \frac{\#C_i(\mathbf{x}_{(j)}) \text{ that contain } n}{25,000}, \quad \text{APC} \cong \frac{1}{25,000} \sum_{j=1}^{25,000} \Pr(n \in C_i(\mathbf{x}_{(j)}) \mid \mathbf{x}_{(j)}). \quad (34)$$

In Tables 1 and 2, 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  $\sum_j \text{length of } C_i(\mathbf{x}_{(j)})/25,000$ .

Table 1: 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	$\pi_1$	0.549	0.579	−0.030	9.972
	$\pi_2$	0.502	0.408	0.094	8.983
20	$\pi_1$	0.559	0.608	−0.049	9.996
	$\pi_2$	0.501	0.390	0.110	8.999
50	$\pi_1$	0.562	0.618	−0.055	10.033
	$\pi_2$	0.500	0.385	0.115	9.035

It is clear that  $\pi_2$  can significantly overstate the coverage probability of the credible sets. In contrast,  $\pi_1$  is not only much more accurate in terms of its coverage statement, but also tends to err on the side of conservatism, which is desirable. And, as expected,  $\pi_2$  significantly degrades as  $a = b$  grows larger, while the performance of  $\pi_1$  remains stable as  $a = b$  grows. The performance of  $\pi_1$  is actually rather remarkable, considering that the study is only for a sample of size  $m = 1$ .

Table 2: 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	$\pi_1$	0.910	0.872	0.038	19.618
	$\pi_2$	0.908	0.872	0.037	18.328
20	$\pi_1$	0.914	0.927	-0.012	16.046
	$\pi_2$	0.916	0.927	-0.011	15.254
50	$\pi_1$	0.914	0.937	-0.022	15.202
	$\pi_2$	0.917	0.813	0.105	14.556

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

## 6 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 objective priors for  $n$ , first for the case of known  $p$  and then for the situation in which  $p$  is unknown.

### 6.1 Known $p$

To find an objective 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:  $\pi(n | p) \propto n^{-1/2}$ .

### 6.1.1 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

$$\hat{n} = \frac{\bar{x}_k}{p} = \frac{1}{pk} \sum_{j=1}^k x_j. \quad (35)$$

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

$$p(\hat{n} | n, p) \approx N \left( \hat{n} \mid n, \sqrt{\frac{n(1-p)}{kp}} \right). \quad (36)$$

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

$$\pi_1(n | p) \propto \frac{1}{\sqrt{n}}. \quad (37)$$

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, (35) is not the only inefficient (but consistent) estimate of  $p$ . Another possible estimate of  $n$  here is

$$\hat{n} = \frac{1}{2}(\sqrt{1 + 16S^2} - 1), \quad (38)$$

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

$$\pi_2(n | p) \propto \frac{1}{\sqrt{n}} \times \frac{2pn + 1 - p}{\sqrt{4p^2n^2 + 2pn(3 - 5p) + 1 - 6p(1 - p)}}. \quad (39)$$

Since this differs from the prior in (37), 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  $n \rightarrow \infty$ .

### 6.1.2 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 objective prior would be the Jeffreys-rule prior for  $n$  from this distribution, which can be computed to be

$$\pi_3(n | p) \propto \frac{1}{\sqrt{n}} \times \sqrt{1 + \frac{(1 - p)}{2pn}}. \quad (40)$$

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.

### 6.1.3 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 5.3. We

Table 3: 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)	$\pi_1$	0.639	0.638	0.001	10.015
	$\pi_2$	0.639	0.638	0.001	10.011
	$\pi_3$	0.640	0.638	0.002	10.014
	$\pi_4$	0.637	0.615	0.022	9.908
	$\pi_5$	0.635	0.662	-0.027	10.103
(5, 0.1)	$\pi_1$	0.545	0.578	-0.033	10.275
	$\pi_2$	0.541	0.578	-0.037	10.218
	$\pi_3$	0.562	0.569	-0.008	10.105
	$\pi_4$	0.544	0.432	0.112	9.334
	$\pi_5$	0.539	0.598	-0.059	11.123
(50, 0.01)	$\pi_1$	0.527	0.555	-0.028	10.059
	$\pi_2$	0.531	0.555	-0.024	10.026
	$\pi_3$	0.523	0.410	0.113	9.214
	$\pi_4$	0.529	0.410	0.120	9.102
	$\pi_5$	0.523	0.573	-0.051	11.025

include in the comparison two other priors that have been considered for this problem, namely  $\pi_4(n) = n^{-1}$  and  $\pi_5(n) = 1$ . Note that it is easy to see that all five priors yield a proper posterior. Table 3 is a representative sample of the many simulation results that were examined. The following conclusions can be reached from this (and the many other simulation results):

- The prior  $\pi_4(n) = 1/n$  results in systematic overstatement of the actual coverage, and by a large amount; hence it can be eliminated from consideration.
- The prior  $\pi_5(n) = 1$  results in systematic understatement of the actual coverage by a

significant amount – up to 6% in the table – and yields too large credible sets; hence this prior can also 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 our recommended choice.

The recommended objective prior for the binomial distribution with known  $p$  is thus  $\pi^*(n) = 1/\sqrt{n}$ .

## 6.2 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

$$p(x_1, \dots, x_m | n, p) = \prod_{j=1}^m \binom{n}{x_j} p^{x_j} (1-p)^{n-x_j}, \quad (41)$$

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

$$p(x_1, \dots, x_m | n) = \frac{1}{\pi} \prod_{j=1}^m \binom{n}{x_j} \frac{\Gamma(s + 1/2) \Gamma(mn - s + 1/2)}{\Gamma(mn + 1)}. \quad (42)$$

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

### 6.2.1 Approach 3

We first compute the mean and variance of  $m^{-1} \sum_{i=1}^m x_i$ , from the distribution (42). From (29) and (30), 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  $i \neq j$ ,

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

Thus

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

Now going to  $k$  independent replications  $(x_{i1}, \dots, x_{im}), i = 1, \dots, k$ , from (42) for the reference prior asymptotics, a simple consistent linear estimate of  $n$  is

$$\hat{n} = \frac{2}{km} \sum_{i=1}^k \sum_{j=1}^m x_{ij}. \quad (43)$$

It follows from the central limit theorem that

$$p(\hat{n} \mid n, p) \approx N\left(\hat{n} \mid n, \sqrt{\frac{2n}{km} \left\{ \frac{(m+3)n}{4} + 1 \right\}}\right). \quad (44)$$

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

$$\pi_1(n) \propto \left( \frac{n}{m} \left\{ \frac{(m+3)n}{4} + 1 \right\} \right)^{-\frac{1}{2}} \propto \frac{1}{\sqrt{n(n + \frac{4}{m+3})}}. \quad (45)$$

Note that, when  $m = 1$ , this is (necessarily) the same prior as that in (33) for  $a = b = 1/2$ , since the two models are then the same.

### 6.2.2 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  $\pi_2(n) \propto 1/n$ .

### 6.2.3 Comparison

It is easy to see that both objective priors yield proper posterior distributions, when the  $\text{Be}(p \mid \frac{1}{2}, \frac{1}{2})$  prior is used for  $p$ . Note that 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 objective prior for the binomial problem with both  $p$  and  $n$  unknown is thus

$$\pi^*(p, n) \propto \frac{1}{n} \times \text{Be}(p \mid \frac{1}{2}, \frac{1}{2}). \quad (46)$$

Although Jeffreys never explicitly studied this problem,  $\pi^*(p, n)$  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.

## 7 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,  $\pi(\mathcal{M}_i)$ .

### 7.1 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  $\pi(\mathcal{M}_i) = 1/3$ .

## 7.2 Variable Selection

Consider the variable selection problem where there are  $m$  variables (or graphical nodes or links) that can make up a model. Let  $\gamma$  be the vector of zeroes and ones that indicates which variables are in the model, and denote the resulting model as  $\mathcal{M}_\gamma$ . 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 objective 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 Approach 2 would now require marginalizing over  $\gamma$ , 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  $\text{Bi}(r | m, p)$  model, which is  $\text{Be}(p | \frac{1}{2}, \frac{1}{2})$ . Utilizing this yields objective model prior probabilities

$$\pi(\mathcal{M}_\gamma) = \int_0^1 p^r(1-p)^{m-r} \text{Be}(p | \frac{1}{2}, \frac{1}{2}) dp = \frac{\Gamma[r + \frac{1}{2}]\Gamma[m - r + \frac{1}{2}]}{\pi\Gamma[m + 1]}. \quad (47)$$

**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

$$\pi(\mathcal{M}_\gamma) = \frac{1}{(m + 1)\binom{m}{r}}. \quad (48)$$

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, (48) can also be derived as in (47) if, instead, the uniform prior is used for  $p$ ; the choice of Jeffreys is thus reasonable from either perspective.

## Appendix A: Proof of Lemma 2.1

*Proof.* It is easy to see that

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

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

$$J_{R,N} = \int_1^R \frac{1}{(v + N - 3)^{R+2}} g_R(v) dv. \quad (49)$$

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

$$J_{R,N} = \sum_{i=1}^{R-1} (-1)^{i-1} \binom{R-1}{i-1} J_{i,R,N}, \quad (50)$$

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

$$J_{i,R,N} = \int_i^R \frac{(v-i)^{R-2}}{(v+N-R)^{R+2}} dv. \quad (51)$$

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

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

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

$$\sum_{i=1}^{R-1} (-1)^{i-1} \binom{R-1}{i-1} J_{i,R,N,2} = \frac{2(-1)^R}{R^3 - R}. \quad (52)$$

Substituting these expressions into (50), 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. □

## Appendix B: Proof of Results in Section 4

In order to find the marginal priors and posteriors for  $\mathbf{R}_k$ , recall that the standard conjugate prior for  $\mathbf{p}_k$  is the Dirichlet distribution  $\text{Di}_k(\mathbf{a})$ , where  $\mathbf{a} = (a_1, \dots, a_k, a_{k+1})$  for positive constants  $a_j$ , whose density is given by

$$\text{Di}_k(\mathbf{p}_k | \mathbf{a}) = \frac{\Gamma(\sum_{j=1}^{k+1} a_j)}{\prod_{j=1}^{k+1} \Gamma(a_j)} \prod_{j=1}^{k+1} p_j^{a_j - 1}. \quad (53)$$

The Jeffreys prior  $\pi_J$  corresponds to (53) with all  $a_j = 1/2$ . Another frequently considered objective prior for  $\mathbf{p}_k$  is the constant prior  $\pi_U$ , corresponding to (53) with all  $a_j = 1$ . Note that the reference prior  $\pi_R$  for  $\mathbf{p}_k$  in (15) is not a special case of (53).

To allow for simultaneous analysis of the conjugate class and the reference prior, we consider the reparameterization of  $\mathbf{p}_k$  in terms of discrete hazard rates, as given in He (2009):

$$h_j = \frac{p_j}{p_j + \dots + p_{k+1}}, \text{ for } j = 1, \dots, k, \quad (54)$$

and  $h_{k+1} = 1$ . The ranges for the hazard rates are  $0 < h_i < 1$ . Note that the transformation from  $\mathbf{p}_k = (p_1, \dots, p_k)$  to  $\mathbf{h}_k \equiv (h_1, \dots, h_k)$  is one-to-one, and

$$p_j = \begin{cases} h_j \prod_{i=1}^{j-1} (1 - h_i), & \text{if } j = 1, 2, \dots, k. \\ \prod_{i=1}^k (1 - h_i), & \text{if } j = k + 1. \end{cases} \quad (55)$$

Because the ranges of  $h_j$  are independent, it is natural to consider independent Beta priors

$$\pi_B(\mathbf{h}) = \prod_{j=1}^k \frac{1}{\text{Beta}(c_j, d_j)} h_j^{c_j-1} (1-h_j)^{d_j-1}, \quad 0 < h_1, \dots, h_k < 1. \quad (56)$$

The following results show that the three objective priors,  $\pi_U$ ,  $\pi_J$ ,  $\pi_R$  are all special cases of (56). Their proofs can be found in He (2009).

**Lemma B.1.**

(a) The conditional prior of  $\mathbf{R}_k$  given  $\mathbf{h}_k$  is

$$p(\mathbf{R}_k | N, \mathbf{h}_k) = \frac{N!}{\prod_{j=1}^{k+1} R_j!} \prod_{j=1}^k h_j^{R_j} (1-h_j)^{R_{j+1} + \dots + R_k + R_{k+1}}.$$

(b) If  $\mathbf{p}_k$  has the Dirichlet distribution (53), then  $h_1, \dots, h_k$  are independent, and  $h_j$  has a Beta  $(a_j, a_{j+1} + \dots + a_{k+1})$  distribution.

(c) The one-at-a-time reference prior  $\pi_R$  is equivalent to  $h_1, \dots, h_k$  being iid Beta  $(1/2, 1/2)$ .

(d) If  $h_j$  has the independent beta  $(c_j, d_j)$  prior, the marginal mass function of  $\mathbf{R}_k$  is

$$p(\mathbf{R}_k | N) = \frac{N!}{\prod_{j=1}^{k+1} R_j!} \left\{ \prod_{j=1}^k \frac{1}{\text{Beta}(c_j, d_j)} \right\} \prod_{j=1}^k \text{Beta}(c_j + R_j, d_j + R_{j+1} + \dots + R_{k+1}).$$

**Proof of Theorem 4.1.**

(a) For the Jeffreys prior, we have  $c_j = 1/2$  and  $d_j = (k+1-j)/2$ . Then

$$\begin{aligned} p_J(\mathbf{R}_k | N) &= \frac{N!}{\prod_{j=1}^{k+1} R_j!} \prod_{j=1}^k \frac{\Gamma(\frac{k+1-j}{2} + \frac{1}{2})}{\sqrt{\pi} \Gamma(\frac{k+1-j}{2})} \prod_{j=1}^k \frac{\Gamma(R_j + \frac{1}{2}) \Gamma(R_{j+1} + \dots + R_{k+1} + \frac{k+1-j}{2})}{\Gamma(R_j + R_{j+1} + \dots + R_{k+1} + \frac{k+2-j}{2})} \\ &= \frac{N!}{R_{k+1}!} \left\{ \prod_{j=1}^k \frac{\Gamma(R_j + \frac{1}{2})}{\sqrt{\pi} \Gamma(R_j + 1)} \right\} \frac{\Gamma(\frac{k+1}{2}) \Gamma(R_{k+1} + \frac{1}{2})}{\Gamma(\frac{1}{2}) \Gamma(N + \frac{k+1}{2})} \\ &= \left\{ \prod_{i=1}^{k+1} f(R_i) \right\} \frac{N! \Gamma(\frac{k+1}{2})}{\Gamma(N + \frac{k+1}{2})}, \quad \mathbf{R}_k \in \mathcal{R}_{k,N}. \end{aligned}$$

(b) For the reference prior, from Lemma B.1, we have  $c_j = d_j = 1/2$ . Then

$$\begin{aligned} p_R(\mathbf{R}_k | N) &= \frac{N!}{\prod_{j=1}^{k+1} R_j!} \prod_{j=1}^k \frac{\Gamma(R_j + \frac{1}{2}) \Gamma(R_{j+1} + \dots + R_{k+1} + \frac{1}{2})}{\pi \Gamma(R_j + R_{j+1} + \dots + R_{k+1} + 1)} \\ &= \prod_{j=1}^{k+1} \frac{\Gamma(R_j + \frac{1}{2})}{\sqrt{\pi} \Gamma(R_j + 1)} \prod_{j=2}^k \frac{\Gamma(R_j + \dots + R_{k+1} + \frac{1}{2})}{\sqrt{\pi} \Gamma(R_j + \dots + R_{k+1} + 1)} \end{aligned}$$

$$\begin{aligned}
&= \left\{ \prod_{i=1}^{k+1} f(R_i) \right\} \left\{ \prod_{j=2}^k f(R_j + \cdots + R_k + R_{k+1}) \right\} \\
&= \left\{ \prod_{i=1}^{k+1} f(R_i) \right\} \left\{ \prod_{i=2}^k f(N - R_1 - \cdots - R_{i-1}) \right\}, \mathbf{R}_k \in \mathcal{R}_{k,N}.
\end{aligned}$$

Theorem 4.1 is proved. □

The following lemma gives uniform results on the class of independent beta priors for  $h_j$ . It is useful in deriving the posteriors of  $\mathbf{R}_k$  in Theorem 4.3. Its proof is omitted.

**Lemma B.2.** Assume the independent Beta prior (56).

(a) The marginal likelihood of  $\mathbf{r}_k$  has the same form as (57), replacing  $N$  by  $n$ . That is, for

$$\mathbf{r}_k \in \mathcal{R}_{k,n},$$

$$p(\mathbf{r}_k | n) = \frac{n!}{\prod_{j=1}^{k+1} r_j!} \left\{ \prod_{j=1}^k \frac{1}{\text{Beta}(c_j, d_j)} \right\} \prod_{j=1}^k \text{Beta}(c_j + r_j, d_j + r_{j+1} + \cdots + r_{k+1}). \quad (57)$$

(b) The marginal posterior mass function of  $\mathbf{R}_k \in \mathcal{R}_{k,N}$  given  $\mathbf{r}_k$  is

$$\pi(\mathbf{R}_k | \mathbf{r}_k, n, N) = \frac{(N - n)!}{\prod_{j=1}^{k+1} (R_j - r_j)!} \prod_{j=1}^k \frac{\text{Beta}(c_j + R_j, d_j + R_{j+1} + \cdots + R_{k+1})}{\text{Beta}(c_j + r_j, d_j + r_{j+1} + \cdots + r_{k+1})}. \quad (58)$$

## 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**, 905–938.

- 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**, 125–135.
- 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.
- He, C. Z. (2009), ‘Bayesian analysis for the multinomial data under noninformative priors’, p. submitted.
- 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.