Objective Priors for Discrete Parameter Spaces

James O. BERGER, Jose M. BERNARDO, and Dongchu SUN

This article considers the development of objective prior distributions for discrete parameter spaces. 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, and the binomial distribution. The recommended objective priors for the first, third, and fourth problems are new.

KEY WORDS: Binomial; Discrete parameter space; Hypergeometric; Jeffreys-rule priors; Parameter-based asymptotics; 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 and 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 nonconstant 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 probability density function (PDF) $\text{Bi}(x \mid n, p)$, 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 $\mathcal{N} = \{1, 2, \ldots\}$ 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 the 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 the 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 $\{0, 1, \ldots, N\}$. The data are 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 nonuniform 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 suggested using $1/\theta$ for any unbounded positive parameter θ, including discrete. This choice has the appeal of having the largest polynomial decay that retains impropriety. Rissanen (1983) proposed a prior for positive integers that in some sense is the vaguest proper prior. Barger and Bunge (2008) proposed using the formal Jeffreys-rule method or the reference prior method, based on an extension of the concept of an information matrix to discrete parameter problems that was developed by Lindsay and 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 that 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 that will maximize the missing information is the constant prior. And we have been unable to find any modification of the approach—for example, 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

James O. Berger is the Arts and Sciences Professor in the Department of Statistical Science, Duke University, Durham, NC 27708 (E-mail: *berger@stat.duke.edu*). Jose M. Bernardo is Professor, Departament de Estadística i Investigació Operativa, Universidad de Valencia, Valencia, Spain (E-mail: *jose.m.bernardo@uv.es*). Dongchu Sun is a Professor of statistics, School of Finance and Statistics, East China Normal University, Shanghai 200241, and a Professer or Statistics, Department of Statistics, University of Missouri–Columbia, Columbia, MO 65211 (E-mail: *sund@missouri.edu*). Berger's work was supported by NSF grants AST-0507481, DMS-0757549-001, and DMS-1007773. Sun's work was supported by NSF grants DMS-1007874 and SES-1024080. The authors gratefully acknowledge the very constructive comments of the editor, an associate editor, and two referees. *AMS 2000 subject classification:* Primary 62F15; secondary 62A01, 62B10, 62E20.

Journal of the American Statistical Association
June 2012, Vol. 107, No. 498, Theory and Methods
DOI: 10.1080/01621459.2012.682538

ordinary reference prior theory (Bernardo 1979, 2005; Berger and Bernardo 1992; Berger, Bernardo, and 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 $p(x \mid \theta)$. Typically, however, $C(\theta) = \int p(x \mid \theta)\, dx$ will not equal 1 for $\theta \notin \Theta$, so that the analysis would have to be with the density $C(\theta)^{-1} p(x \mid \theta)$, and it can be very difficult to work with a new normalizing factor. Sometimes, however, one can also treat the available data x as continuous and analyze the resulting completely continuous problem without an additional normalizing factor being introduced. For example, suppose that x is uniform on the integers $\{0, 1, \ldots, \theta\}$. Treating both θ and x as continuous variables, we have that x is uniform on $(0, \theta)$; no additional normalizing factor is introduced 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, Jeffreys (1946, 1961) postulated that the unknown R arises as a binomial random variable $\mathrm{Bi}(R \mid N, p)$ 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 $\pi^*(R) = \int \mathrm{Bi}(R \mid N, p)\pi^R(p)dp$. (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 $\boldsymbol{x}^{(k)} = (x_1, \ldots, x_k)$, where the x_i are k independent replications of the data x from the model under consideration. (As usual in the reference prior theory, x will, itself, typically be a vector of observations; the replication being considered is an imaginary replication of the entire original experiment.) In the reference prior approach, one considers the asymptotic behavior of an information-based criterion as $k \to \infty$ (cf. Berger and Bernardo 1992).

Approach 3 proceeds by

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

Note 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 $k \to \infty$ 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 desired objective prior is (cf. theorem 9 of Bernardo 2005) $\pi^*(\theta) \propto \sigma(\theta)^{-1}$.

That this approach requires the use of inefficient estimators is both philosophically problematical and practically ambiguous, because the 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 and Roeder 1987 for definition), it is possible to define an analog 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 and Bunge (2008), who analyzed several examples. The only overlap with the examples in this article (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 and Bunge (2008) thus provided 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 and Roeder (1987) and Sweeting (1992),

this approach uses a formal limiting operation in θ to make the problem continuous. For instance, in the uniform problem mentioned in Section 1.2.1, x/θ has a uniform distribution on the discrete set $\{0, \frac{1}{\theta}, \ldots, \frac{\theta-1}{\theta}, 1\}$. 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.

More formally, the steps of this approach are as follows:

- Let $\theta \to \infty$ (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 an objective prior.

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 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, $\pi^*(\theta)$, is infinite at the endpoints of the discrete parameter space Θ. 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 Θ^*: write $\Theta^* = \cup_j \Theta_j^*$, where Θ_j^* is a set containing the discrete parameter value θ_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 article; for the hypergeometric problem, we can use Approach 2, which directly yields a finite (indeed proper) discrete objective prior.

1.3 Overview

In the remainder of the article, we consider a variety of discrete parameter problems and explore the application of the above four approaches to the 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 the asymptotic missing information in the original problem.

When Approach 1 (without renormalization) or Approach 2 (when there is a single natural hierarchical model) is applicable, we view the resulting prior as being the natural objective prior, and look no further. For problems where neither approach applies, we use 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 θ. This is probably unavoidable; unless there is enough structure to use Approaches 1 and 2, unequivocal determination of an objective prior for small θ seems not to be possible.

Experience has demonstrated that exact reference priors tend to have excellent performance when evaluated, for example, for frequentist coverage of resulting credible intervals. This performance presumably also exists for the reference priors in the embedding problems considered through the approaches above (especially Approaches 1 and 2), but there is no guarantee that this performance will extend to the original discrete problems.

The particular discrete problems that are considered in the article (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), and the *binomial distribution* (in Section 6). The objective priors derived in Sections 2, 4, and 5 are new. The analyses in Section 6 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, sec. 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 $t_1 \le \cdots \le t_R$ denote the resulting failure times. Here, $N \ge R$ and λ are both unknown, while R is prespecified.

The problem of estimating N has many interesting applications. For example, Starr (1974) and Kramer and 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 and 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 and Ebrahimi (2001).

The density of $(t_1, \ldots, t_R)$ is

$$p(t_1, \ldots, t_R \mid N, \lambda) = \frac{N!}{(N-R)!}\lambda^R \exp\{-\lambda[t_1 + t_2 + \cdots + t_R + (N-R)t_R]\}.$$

The pair $(t_1 + \cdots + t_R, t_R)$ is minimal sufficient for (N, λ) and, hence, so is the pair $(V, W) = (t_1 + \cdots + t_R)/t_R, t_R)$. For any given $c > 0$, the transformation $(t_1, \ldots, t_R)$ to $(ct_1, \ldots, ct_R)$ 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 and Goldie (1981) [formula (4)] derived the joint density of (V, W) as follows:

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

for $1 < V < R$, $W > 0$, where, letting $[V]$ denote the integer part of V,

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

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

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

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 the 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 an 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 and 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 an objective Bayesian analysis, note that the likelihood $p(V \mid N)$ in (3) tends to one as $N \to \infty$, 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) was used. Similar problems were encountered by 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, $\{R, R+1, \ldots\}$ into the continuous space $(R-0.5, \infty)$, since, for each $N \geq R - 0.5$, $p(v \mid N)$ is still a probability density for v (with the same normalization, with the factorials replaced by the corresponding gamma functions). The reference prior for N in this continuous problem is simply the Jeffreys-rule prior

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

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 (but evaluated at integer values of N) is

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

where

$$J_{R,N} = \sum_{i=1}^{R-1} \frac{(-1)^{i-1}}{(N-R+i)^3} \left(\frac{R-i}{N}\right)^{R-1} \binom{R-1}{i-1} \times \left[\frac{1}{R-1} - \frac{2(R-i)}{RN} + \frac{(R-i)^2}{(R+1)N^2}\right]. \quad (5)$$

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 θ is then the positive integers $\mathcal{N} = \{1, 2, \ldots\}$. The objective prior for θ is then

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

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

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

Note that these are proper priors, and so their normalization constants are included. It appears that the tail of $\pi^*(\theta \mid R)$ is always of the 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 $N \to \infty$; 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. Although 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.)

3. THE HYPERGEOMETRIC DISTRIBUTION

Consider the hypergeometric distribution,

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

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

As mentioned in Section 1, Approach 2 can be implemented by assuming that the unknown R has the hierarchical model $\text{Bi}(R \mid N, p)$, 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

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

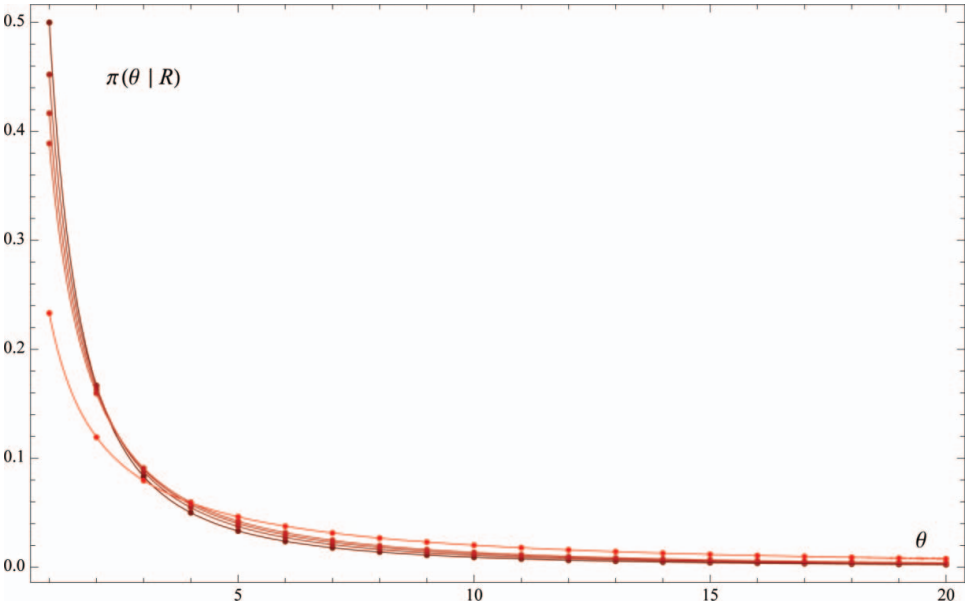

Figure 1. Objective priors for θ for $R = 2, 3, 4, 5$, and 20 (top to bottom at $\theta = 1$). The online version of this figure is in color.

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

$$\pi^*(R \mid N) = \int_0^1 \text{Bi}(R \mid N, p)\,\text{Be}\left(p \,\middle|\, \frac{1}{2}, \frac{1}{2}\right) dp = \frac{1}{\pi}\,\frac{\Gamma\left(R + \frac{1}{2}\right)\Gamma\left(N - R + \frac{1}{2}\right)}{\Gamma(R+1)\,\Gamma(N - R + 1)}, \tag{8}$$

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

The upper panel of Figure 2 graphs the objective priors $\pi^*(\theta \mid N)$ for the proportion $\theta = R/N, \theta \in \{0, 1/N, \ldots, 1\}$, 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 $\pi^R(p) = \text{Be}(p \mid \frac{1}{2}, \frac{1}{2})$ for the binomial probability model. Indeed, using Stirling's approximation for the gamma functions in (8), one obtains, in terms of $\theta = R/N$,

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

which is basically proportional to $\text{Be}(\theta \mid \frac{1}{2}, \frac{1}{2})$. As illustrated in the lower panel of Figure 2, where the solid line represents (9) as a continuous function of θ, the approximation is very good, even if N is moderate.

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 $N \to \infty$ results in using 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 Section 1, we view Approach 2

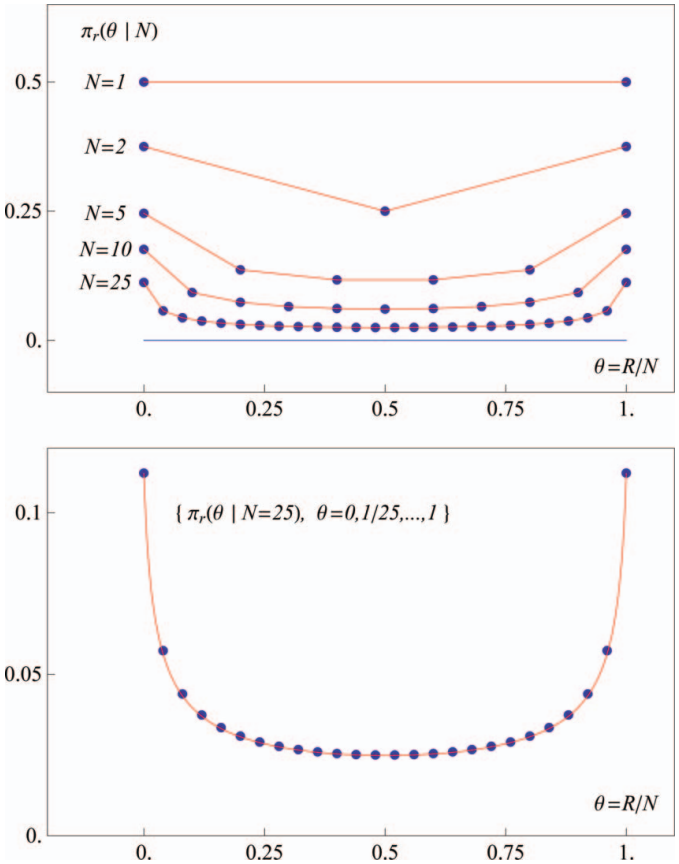

Figure 2. Objective priors $\pi^*(\theta \mid N)$ for the proportion $\theta = R/N$ of conforming items in a population of size N, for several N values (upper panel), and its continuous approximation (lower panel). The online version of this figure is in color.

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 $1/\sqrt{R(N-R)}$.

Remark. A related 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. The classical answer, known as Laplace's rule of succession (Laplace 1774), is

$$\Pr(+ \mid \text{previous n +'s}) = \frac{n+1}{n+2}\ . \tag{10}$$

When the population is known to be finite of size N, the same rule was derived by Broad (1918). If the analysis is carried out using the reference prior $\pi^R(R \mid N)$ in (8), the resulting "reference prior" law of succession is (see Berger, Bernardo, and Sun 2009b)

$$\Pr^R(+ \mid N, \text{previous n +'s}) = \frac{n+1/2}{n+1}\ . \tag{11}$$

4. MULTIVARIATE HYPERGEOMETRIC DISTRIBUTION

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

$$\mathrm{Hy}_k(\boldsymbol{r}_k \mid n, \boldsymbol{R}_k, N) = \frac{\binom{R_1}{r_1}\cdots\binom{R_k}{r_k}\binom{R_{k+1}}{r_{k+1}}}{\binom{N}{n}}, \boldsymbol{r}_k \in \mathcal{R}_{k,n}, \tag{12}$$

where the unknown parameters $\boldsymbol{R}_k = (R_1, \ldots, R_k)$ are in the parameter space $\mathcal{R}_{k,N} = \{\boldsymbol{R}_k = (R_1, \ldots, R_k) : R_j \in \mathcal{N}_+, R_1 + \cdots + R_k \le N\}$. Here and in the following, $R_{k+1} = N - (R_1 + \cdots + R_k)$.

Note that the hypergeometric distribution is the special case when $k = 1$.

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

$$\mathrm{Mu}_k(\boldsymbol{R}_k \mid N, \boldsymbol{p}_k) = \frac{N!}{\prod_{j=1}^{k+1} R_j!}\prod_{j=1}^{k+1} p_j^{R_j}. \tag{13}$$

The problem then reduces to finding the objective priors for the continuous hyperparameter $\boldsymbol{p}_k$. 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 $\boldsymbol{p}_k$ is the Jeffreys' prior π_J, given by

$$\pi_J(\boldsymbol{p}_k) = \frac{\Gamma\left(\frac{k+1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)^{k+1}}\prod_{j=1}^{k+1} p_j^{\frac{1}{2}-1}. \tag{14}$$

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

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

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

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

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

Theorem 4.1. Define the function,

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

(a) Under the Jeffreys' prior π_J for $\boldsymbol{p}_k$, the marginal prior of $\boldsymbol{R}_k$ given N is

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

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

$$p_R(\boldsymbol{R}_k \mid N) = \left\{\prod_{i=1}^{k+1} f(R_i)\right\} \times \left\{\prod_{j=2}^{k} f(R_j + \cdots + R_k + R_{k+1})\right\}, \boldsymbol{R}_k \in \mathcal{R}_{k,N}. \tag{18}$$

Remark 4.1. It is interesting that, under the reference prior π_R for $\boldsymbol{p}_k$,

$$p_R(\boldsymbol{R}_{k-1} \mid N) = \left\{\prod_{i=1}^{k-1} f(R_i)\right\} f(\tilde{R}_k) \times \left\{\prod_{j=2}^{k-1} f(R_j + \cdots + R_{k-1} + \tilde{R}_k)\right\}, \boldsymbol{R}_{k-1} \in \mathcal{R}_{k-1,N},$$

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

$$p_R(R_1 \mid N) = f(R_1)f(N - R_1), R_1 = 0, \ldots, N\,, \tag{19}$$

which reduces to the prior for the hypergeometric problem.

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

Theorem 4.2.

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

$$\begin{aligned} p(\boldsymbol{r}_k \mid \boldsymbol{p}_k, n, N) &= \sum_{\boldsymbol{R}_k \in \mathcal{N}_{k,N}} \mathrm{Hy}_k(\boldsymbol{r}_k \mid n, \boldsymbol{R}_k, N)\mathrm{Mu}_k(\boldsymbol{R}_k \mid N, \boldsymbol{p}_k) \\ &= \mathrm{Mu}_k(\boldsymbol{r}_k \mid n, \boldsymbol{p}_k), \boldsymbol{r}_k \in \mathcal{R}_{k,n}. \end{aligned} \tag{20}$$

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

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

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

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

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

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

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

where f is given by (16).

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

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

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

Theorem 4.3.

(a) Under the Jeffreys' prior π_J for $\boldsymbol{p}_k$, the marginal posterior of $\boldsymbol{R}_k$ given $(\boldsymbol{r}_k, n, N)$ is

$$\pi_J(\boldsymbol{R}_k \mid \boldsymbol{r}_k, n, N) = \frac{(N-n)!}{\prod_{j=1}^{k+1}(R_j - r_j)!} \prod_{j=1}^{k} \times \frac{\mathrm{Beta}\left(\frac{1}{2} + R_j, \frac{k-j+1}{2} + R_{j+1} + \cdots + R_{k+1}\right)}{\mathrm{Beta}\left(\frac{1}{2} + r_j, \frac{k-j+1}{2} + r_{j+1} + \cdots + r_{k+1}\right)}. \quad (24)$$

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

$$\pi_R(\boldsymbol{R}_k \mid \boldsymbol{r}_k, n, N) = \frac{(N-n)!}{\prod_{j=1}^{k+1}(R_j - r_j)!} \prod_{j=1}^{k} \times \frac{\mathrm{Beta}\left(\frac{1}{2} + R_j, \frac{1}{2} + R_{j+1} + \cdots + R_{k+1}\right)}{\mathrm{Beta}\left(\frac{1}{2} + r_j, \frac{1}{2} + r_{j+1} + \cdots + r_{k+1}\right)}. \quad (25)$$

5. THE BINOMIAL-BETA DISTRIBUTION

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

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

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

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

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

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

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, the application of Approach 1 would result in a nonconstant 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 $\{x_1, \ldots, x_k\}$ be k independent replications from (27). The mean and variance of the x_i are easily found to be

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

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

A simple estimate of n is the linear estimate

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

Note that this is not an efficient estimate but, as mentioned in Section 1, efficient estimates cannot be used in direct reference prior asymptotics for discrete distributions. This estimator has mean and variance

$$\mathrm{E}(\hat{n} \mid n) = n \quad \text{and} \quad \mathrm{var}(\hat{n} \mid n) = \frac{n(n+a+b)}{a^2(a+b+1)\,k}.$$

It follows from the central limit theorem that

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

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 $n \to \infty$, x/n is asymptotically normal with mean p and variance $p(1-p)/n$.

For any $y \in (0, 1)$, it follows that, as $n \to \infty$,

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

so that the limiting density of $y = x/n$ is $\mathrm{Be}(y \mid a, b)$. The parameter-based asymptotic distribution of x thus has density $n^{-1}\mathrm{Be}(x/n \mid a, b, n)$. 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 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 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 $\mathrm{Be}(p \mid a, b)$ distribution in (26) will be very concentrated around $p = a/(a+b)$. It follows that $p(x \mid n)$ should behave like the $\mathrm{Bi}(x \mid n, a/(a+b))$ 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 $1/\sqrt{n(n+a+b)}$ when $a+b$ is large compared with n. (This argument also presumes that the likelihood is such that the posterior is concentrated at moderate values of n, compared to $a+b$.) This points to choosing π_1.

To further study this, we use 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 $\boldsymbol{x} = (x_1, \ldots, x_m)$ from (27) for given n, and construct the one-sided credible sets that have posterior probability exceeding a desired threshold $1 - \alpha$, namely

$$C_i(\boldsymbol{x}) = \{1, \ldots, n_i^*\}, \quad n_i^* = \inf_{n^* \in \mathcal{N}} \Pr_i[n \le n^* \mid \boldsymbol{x}] \ge 1 - \alpha\,,$$

corresponding to the two priors π_i, $i = 1, 2$. We then consider the frequentist coverage of these credible sets, namely $\Pr(C_i(\boldsymbol{X}) \ni n \mid n)$, with the goal being to have frequentist coverage similar to the stated posterior probability. Because of problems caused by discreteness, we choose the target not to 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 the 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, $\boldsymbol{x}_{(j)}$, with each coordinate being drawn independently from (27), compute $C_i(\boldsymbol{x}_{(j)})$ for each sample, and approximate the frequentist coverage and the average posterior coverage (APC) as, respectively,

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

Table 1. APC, frequentist coverage, and AS of one-sided 50% (upper panel) and 90% (lower panel) 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
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

In Table 1, 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 APC and Coverage, as well as Difference = APC − Coverage, which is best if close to zero, with negative differences being better than positive differences (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 AS = E[length of $C_i(\boldsymbol{x}_{(j)}) \mid n$], and approximated by $\sum_j$ length of $C_i(\boldsymbol{x}_{(j)})/25{,}000$.

It is clear that π_2 can significantly overstate the coverage probability of the credible sets. In contrast, π_1 not only is 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, 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 $\pi^*(n) = 1/\sqrt{n(n+a+b)}$ as the objective prior for the binomial-beta distribution.

6. REFERENCE PRIOR FOR THE BINOMIAL SAMPLE SIZE

Assume that $x \sim \mathrm{Bi}(x \mid n, p)$. 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, Liseo, and Wolpert (1999) and DasGupta and Rubin (2005). Basu and Ebrahimi (2001) developed 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 when p is unknown.

6.1 Known p

To find an objective prior for n from this discrete distribution, the application of Approach 1 would again result in a nonconstant normalization factor. There is no single hierarchical extension that is natural for Approach 2, but a reasonable extension—first used by Raftery (1988a) for the case of unknown p (see also Moreno and Giron 1998)—is to assume a Poisson $\mathrm{Ps}(n-1 \mid \lambda)$ distribution for $n-1$. One would then derive the reference prior $\pi(\lambda)$ for the corresponding integrated model $p(x \mid \lambda, p) = \sum_{n=1}^{\infty} \mathrm{Bi}(x \mid n, p)\, \mathrm{Ps}(n-1 \mid \lambda)$, and use this to obtain the corresponding reference prior $\pi(n \mid p) \propto \int_0^{\infty} \mathrm{Ps}(n-1 \mid \lambda)\, \pi(\lambda \mid p)\, d\lambda$. 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: $\pi(n \mid p) \propto n^{-1/2}$. Note also that, while Raftery (1988a) dealt with the unknown p situation, the approach taken in the article would result in the prior $1/n$ for the known p case; this prior is discussed in the following sections.

6.1.1 Approach 3. Let $x_1, \ldots, x_k$ be k independent replications from $\mathrm{Bi}(x \mid n, p)$. A linear estimate of n is

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

Note that

$$\mathrm{E}(\hat{n} \mid n, p) = n, \quad \text{and} \quad \mathrm{var}(\hat{n} \mid n) = \frac{n(1-p)}{k\,p}.$$

It follows from the central limit theorem that

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

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

$$\pi_1(n \mid p) \propto \frac{1}{\sqrt{n}}. \tag{37}$$

As mentioned in Section 1.2.3, this is (necessarily) the same result derived by Barger and Bunge (2008), since we use 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

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

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

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

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 $\pi_2(n \mid p)/\pi_1(n \mid p)$ is maximized at $n = 1$, with the value $1 + p$, and decreases monotonically to 1 as $n \to \infty$.

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 $n \to \infty$ is (treating x as continuous) $\mathrm{N}(x \mid np, np(1-p))$. Treating n as also being continuous, the appropriate objective prior would be the Jeffreys-rule prior for n from this distribution, which is

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

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 attractive, 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 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 2 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 a systematic overstatement of the actual coverage, and by a large amount; hence, it can be eliminated from consideration.

Table 2. APC, frequentist coverage, and AS of one-sided 50% credible intervals based on 20,000 simulations of sample size m from $\mathrm{Bi}(x \mid 10, p)$

(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

- The prior $\pi_5(n) = 1$ results in a 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 candidates; 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 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 $m \geq 2$ independent observations $x_1, \ldots, x_m$ from the $\text{Bi}(x \mid n, p)$ 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 $(x_1, \ldots, x_m)$, is

$$p(x_1, \ldots, x_m \mid n, p) = \prod_{j=1}^{m} \binom{n}{x_j} p^s (1-p)^{mn-s}, \quad (41)$$

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

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

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 $\text{E}(x_i \mid n) = E(x_1 \mid n) = n/2$ and $\text{var}(x_i \mid n) = n(n+1)/2$. It is also easy to verify that, for $i \neq j$,

$$\begin{aligned} \text{cov}(x_i, x_j \mid n) &= \text{E}[\text{cov}(x_i, x_j \mid n, p)] \\ &\quad + \text{cov}(\text{E}(x_i \mid n, p), \text{E}(x_j \mid n, p)) \quad (43) \\ &= 0 + \text{cov}(np, np \mid n) = n^2 \text{var}(p) = n^2/8. \quad (44) \end{aligned}$$

Thus,

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

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

$$\hat{n} = \frac{2}{km} \sum_{i=1}^{k} \sum_{j=1}^{m} x_{ij}. \quad (45)$$

It follows from the central limit theorem that

$$p(\hat{n} \mid n, p) \approx N\left(\hat{n} \mid n, \sqrt{\frac{2n}{km}\left\{\frac{(m+3)n}{4} + 1\right\}}\right). \quad (46)$$

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

$$\pi_1(n) \propto \left(\frac{n}{m}\left\{\frac{(m+3)n}{4} + 1\right\}\right)^{-\frac{1}{2}} \propto \frac{1}{\sqrt{n\left(n + \frac{4}{(m+3)}\right)}}. \quad (47)$$

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 $n \to \infty$, 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 $n \to \infty$,

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

The limiting distribution of $(x_1/n, \ldots, x_m/n)$ does not depend on n, and $n > 0$ is the scale parameter of the limiting distribution of $(x_1, \ldots, x_m)$. 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

Table 3. MSE comparison of the posterior medians for n arising from $\pi^*(p, n)$ and $\pi^R(p, n)$ for various values of p and n and sample sizes m equal 1 and 10

			p				
m	n	Prior	0.05	0.25	0.50	0.75	0.95
1	10	π^*	0.76	0.62	0.33	0.47	0.80
		π^R	0.88	0.63	0.33	0.47	0.80
	50	π^*	0.92	0.53	0.14	0.48	0.85
		π^R	0.92	0.53	0.14	0.50	0.88
	100	π^*	0.91	0.52	0.10	0.46	0.84
		π^R	0.91	0.52	0.10	0.50	0.89
10	10	π^*	0.79	1.22	1.35	0.39	0.01
		π^R	0.89	0.47	0.59	0.28	0.01
	50	π^*	0.63	1.23	1.40	0.43	0.03
		π^R	0.76	0.46	0.62	0.31	0.04
	100	π^*	0.64	1.18	1.28	0.43	0.03
		π^R	0.76	0.45	0.59	0.31	0.04

Table 4. Log-score comparison of the marginal posterior of p arising from $\pi^*(p, n)$ and $\pi^R(p, n)$ for various values of p and n, when $m = 2$

		p				
n	Prior	0.05	0.25	0.50	0.75	0.95
10	π^*	0.95	−0.21	−0.33	−0.17	0.53
	π^R	0.66	0.03	0.02	0.03	0.04
50	π^*	0.53	−0.20	−0.32	−0.15	0.55
	π^R	0.10	0.02	0.03	0.04	0.05
100	π^*	0.50	−0.18	−0.27	−0.09	0.62
	π^R	0.03	0.03	0.05	0.08	0.10

$\text{Be}(p \mid \frac{1}{2}, \frac{1}{2})$ 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 use 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

$$\pi^*(p, n) \propto \frac{1}{n} \times \text{Be}\left(p \,\middle|\, \frac{1}{2}, \frac{1}{2}\right). \tag{48}$$

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. Raftery (1988a) proposed the variant $\pi^R(p, n) = 1/n$, and it is interesting to compare this with the choice above. Some comparison is given by Berger, Liseo, and Wolpert (1999), where it is shown that

$$\frac{\pi^R(n \mid x_1, \ldots, x_m)}{\pi^*(n \mid x_1, \ldots, x_m)} \approx \frac{\pi\sqrt{n - s + 0.3}}{n + 1},$$

so that the Raftery posterior has a sharper tail as n grows.

To numerically study the difference in inferences resulting from the two priors, we first study the mean squared error (MSE) of the resulting posterior medians for n, for various values of m, n, and p. (The posterior mean is not finite for either posterior.) Table 3 actually presents $\sqrt{\text{MSE}}/n$ for easier comparison. The results are rather inconclusive: $\pi^*(p, n)$ seems better for more extreme values of p, and $\pi^R(p, n)$ performs better for nonextreme p. The large posterior tails here may make MSE comparison unreliable, so we also look at scoring rules. Indeed, since p is also unknown, we look at the expected log score of the marginal posterior of p at the true value of p for sample size $m = 2$, that is, $\text{E}[\log(\pi(p \mid X_1, X_2) \mid n, p]$; here, the expectation is over the data for given n and p. The results are given in Table 4 and are similar: $\pi^*(p, n)$ is better for more extreme values of p and $\pi^R(p, n)$ is better for nonextreme p. Of course, these results are not unexpected, given that $\pi^*(p, n)$ is based on the Jeffreys' prior for p while $\pi^R(p, n)$ is based on the uniform prior.

APPENDIX A: PROOF OF LEMMA 2.1

Proof. Note that $p(V \mid N)$ satisfies the regularity conditions for the Fisher information to exist and be computable through the usual second derivative form; $\log(p(V \mid N))$ is twice differentiable in N for all values of V, and two derivatives of $\int h(V)p(V \mid N)dV$ with respect to N can be passed under the integral sign [because $\int_1^R h(V)p(V \mid N)dV = \frac{1}{(R-2)!}\frac{N!}{(N-R)!}\int_1^R \frac{1}{(V+N-R)^R}h(V)g_R(V)dV$, and $\frac{1}{(V+N-R)^R}$ is a twice differentiable and bounded function]. It is easy to see that

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

(Strictly speaking one must compute this with gamma functions replacing the factorials in the density, but the result is the same.) Then, the Fisher information of N is (4), where

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

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

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

It follows that

$$J_{R,N} = \sum_{i=1}^{R-1}(-1)^{i-1}\binom{R-1}{i-1}J_{i,R,N}, \tag{A.2}$$

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

$$J_{i,R,N} = \int_i^R \frac{(v-i)^{R-2}}{(v+N-R)^{R+2}}dv = \int_0^{R-i}\frac{y^{R-2}}{(y+N-R+i)^{R+2}}dy. \tag{A.3}$$

Making the transformation $s = y/(y+N-R+i)$, so $y = (N-R+i)s/(1-s)$ and $dy = (N-R+i)(1-s)^{-2}ds$, it follows that

$$\begin{aligned} J_{i,R,N} &= \frac{1}{(N-R+i)^3}\int_0^{(R-i)/N} s^{R-2}(1-s)^2 ds \\ &= \frac{1}{(N-R+i)^3}\int_0^{(R-i)/N}(s^{R-2} - 2s^{R-1} + s^R)ds \\ &= \frac{1}{(N-R+i)^3}\left(\frac{R-i}{N}\right)^{R-1}\left[\frac{1}{R-1} - \frac{2(R-i)}{RN} + \frac{(R-i)^2}{(R+1)N^2}\right]. \end{aligned}$$

Consequently,

$$\begin{aligned} J_{R,N} = \sum_{i=1}^{R-1}\frac{(-1)^{i-1}}{(N-R+i)^3}\left(\frac{R-i}{N}\right)^{R-1}\binom{R-1}{i-1} \\ \times\left[\frac{1}{R-1} - \frac{2(R-i)}{RN} + \frac{(R-i)^2}{(R+1)N^2}\right], \end{aligned}$$

thus proving the lemma. □

APPENDIX B: PROOF OF RESULTS IN SECTION 4

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

$$\text{Di}_k(\boldsymbol{p}_k \mid \boldsymbol{a}) = \frac{\Gamma\left(\sum_{j=1}^{k+1} a_j\right)}{\prod_{j=1}^{k+1}\Gamma(a_j)}\prod_{j=1}^{k+1} p_j^{a_j - 1}. \tag{B.1}$$

The Jeffreys' prior π_J corresponds to (B.1) with all $a_j = 1/2$. Another frequently considered objective prior for $\boldsymbol{p}_k$ is the constant prior π_U, corresponding to (B.1) with all $a_j = 1$.

Note that the reference prior π_R for $\boldsymbol{p}_k$ in (15) is not a special case of (B.1).

To allow for simultaneous analysis of the conjugate class and the reference prior, we consider the reparameterization of $\boldsymbol{p}_k$ in terms of discrete hazard rates, as given by He (2011):

$$h_j = \frac{p_j}{p_j + \cdots + p_{k+1}}, \text{ for } j = 1, \ldots, k, \tag{B.2}$$

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

$$p_j = \begin{cases} h_j \prod_{i=1}^{j-1}(1-h_i), & \text{if } j = 1, 2, \ldots, k, \\ \prod_{i=1}^{k}(1-h_i), & \text{if } j = k+1. \end{cases} \tag{B.3}$$

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

$$\pi_B(\boldsymbol{h}) = \prod_{j=1}^{k} \frac{1}{\text{Beta}(c_j, d_j)} h_j^{c_j-1}(1-h_j)^{d_j-1}, 0 < h_1, \ldots, h_k < 1. \tag{B.4}$$

The following results show that the three objective priors, π_U, π_J, π_R, are all special cases of (B.4). Their proofs can be found in He (2011).

Lemma B.1.

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

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

(b) If $\boldsymbol{p}_k$ has the Dirichlet distribution (B.1), then $h_1, \ldots, h_k$ are independent, and h_j has a beta $(a_j, a_{j+1} + \cdots + a_{k+1})$ distribution.

(c) The one-at-a-time reference prior π_R is equivalent to $h_1, \ldots, h_k$ 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 $\boldsymbol{R}_k$ is

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

Proof of Theorem 4.1.

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

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

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

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

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 $\boldsymbol{R}_k$ in Theorem 4.3. Its proof is omitted.

Lemma B.2. Assume the independent beta prior (B.4).

(a) The marginal likelihood of $\boldsymbol{r}_k$ has the same form as (B.5), replacing N by n. That is, for $\boldsymbol{r}_k \in \mathcal{R}_{k,n}$,

$$p(\boldsymbol{r}_k \mid n) = \frac{n!}{\prod_{j=1}^{k+1} r_j!} \left\{ \prod_{j=1}^{k} \frac{1}{\text{Beta}(c_j, d_j)} \right\} \times \prod_{j=1}^{k} \text{Beta}(c_j + r_j, d_j + r_{j+1} + \cdots + r_{k+1}). \tag{B.5}$$

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

$$\pi(\boldsymbol{R}_k \mid \boldsymbol{r}_k, n, N) = \frac{(N-n)!}{\prod_{j=1}^{k+1}(R_j - r_j)!} \prod_{j=1}^{k} \times \frac{\text{Beta}(c_j + R_j, d_j + R_{j+1} + \cdots + R_{k+1})}{\text{Beta}(c_j + r_j, d_j + r_{j+1} + \cdots + r_{k+1})}. \tag{B.6}$$

[Received October 2009. Revised September 2011.]

REFERENCES

Barger, K., and Bunge, J. (2008), "Bayesian Estimation of the Number of Species Using Noninformative Priors," *Biometrical Journal*, 50, 1064–1076. [636,637,644]

Basu, S., and Ebrahimi, N. (2001), "Bayesian Capture-Recapture Methods for Error Detection and Estimation of Population Size: Heterogeneity and Dependence," *Biometrika*, 88, 269–279. [638,643]

Berger, J. O., and Bernardo, J. M. (1992), "On the Development of Reference Priors" (with discussion), in *Bayesian Statistics 4,* eds. J. M. Bernardo, J. O. Berger, D. V. Lindley, and A. F. M. Smith, London: Oxford University Press, pp. 35–60. [636,637,641]

Berger, J. O., Bernardo, J. M., and Sun, D. (2009a), "The Formal Definition of Reference Priors," *The Annals of Statistics*, 37, 905–938. [636]

——— (2009b), "Natural Induction: An Objective Bayesian Approach," *Revista de la Real Academia de Ciencias Exactas, Físicas y Naturales. Serie A: Matemáticas*, 103, 125–135. [641]

Berger, J. O., Liseo, B., and Wolpert, R. L. (1999), "Integrated Likelihood Methods for Eliminating Nuisance Parameters" (with discussion), *Statistical Science*, 14, 1–28. [643,646]

Bernardo, J. M. (1979), "Reference Posterior Distributions for Bayesian Inference" (with discussion), *Journal of the Royal Statistical Society,* Series B, 41, 113–147; Reprinted in *Bayesian Inference* (1995, Vol. 1), eds. G. C. Tiao and N. G. Polson, Oxford: Edward Elgar, pp. 229–263. [636]

——— (2005), "Reference Analysis," in *Handbook of Statistics* (Vol. 25) , eds. D. K. Dey and C. R. Rao, Amsterdam: Elsevier, pp. 17–90. [636]

Bernardo, J. M., and Smith, A. F. M. (1994), *Bayesian Theory*, New York: Wiley. [636]

Broad, C. D. (1918), "On the Relation Between Induction and Probability," *Mind*, 27, 389–404. [641]

DasGupta, A., and Rubin, H. (2005), "Estimation of Binomial Parameters When Both n, p Are Unknown," *Journal of Statistical Planning and Inference*, 130, 391–404. [643]

Goudie, I. B. J., and Goldie, C. M. (1981), "Initial Size Estimation for the Pure Death Process," *Biometrika*, 68, 543–550. [638,639]

Haldane, J. B. S. (1941), "The Fitting of Binomial Distributions," *The Annals of Eugenics, London*, 11, 179–181. [643]

He, C. Z. (2011), "Bayesian Analysis for the Multinomial Data Under Noninformative Priors," Technical Report, University of Missouri. [647]

Jeffreys, H. (1961), *Theory of Probability*, London: Oxford University Press. [636]

——— (1946), "An Invariant from for the Prior Probability in Estimation Problems," *Proceedings of the Royal Society of London A*, 186, 453–461. [637]

Kahn, W. D. (1987), "A Cautionary Note for Bayesian Estimation of the Binomial Parameter n," *The American Statistician*, 41, 38–39. [636]

Kramer, M., and Starr, N. (1990), "Optimal Stopping in a Size Dependent Search," *Sequential Analysis*, 9, 59–80. [638]

Laplace, P. S. (1774), "Mémoire sur la Probabilité des Causes par les Événements," in *Oeuvres Complètes* (Vol. 8), Paris: Gauthier-Villars, pp. 27–68. [641]

Lawless, J. (1982), *Statistical Models and Methods for Lifetime Data*, New York: Wiley. [638]

Lindsay, B. G., and Roeder, K. (1987), "A Unified Treatment of Integer Parameter Models," *Journal of the American Statistical Association*, 82, 758–764. [636,637]

Moreno, E., and Giron, J. (1998), "Estimating With Incomplete Count Data: A Bayesian Approach," *Journal of Statistical Planning and Inference*, 66, 147–159. [644]

Raftery, A. E. (1988a), "Inference for the Binomial n Parameter: A Hierarchical Bayes Approach," *Biometrika*, 75, 223–228. [644,646]

——— (1988b), "Analysis of a Simple Debugging Model," *Applied Statistics*, 37, 12–22. [639]

Rissanen, J. (1983), "A Universal Prior for Integers and Estimation by Minimum Description Length," *The Annals of Statistics*, 11, 416–431. [636]

Starr, N. (1974), "Optimal and Adaptive Stopping Based on Capture Times," *Journal of Applied Probability*, 11, 294–301. [638]

Sweeting, T. J. (1992), "Parameter-Based Asymptotics," *Biometrika*, 79, 219–232. [637]