2009

Sample Size Calculation for Finding Unseen Species

Hongmei Zhang
University of South Carolina - Columbia, hzhang@mailbox.sc.edu

Hal Stern

Publication Info

Published in *Bayesian Analysis*, Volume 4, Issue 4, 2009, pages 763-792.
Zhang, H., & Stern, H. (2009). Sample size calculations for finding unseen species. *Bayesian Analysis, 4*(4), 763-792.
DOI: 10.1214/09-BA429

Bayesian Analysis (2009) 4, Number 4, pp. 763–792

Sample Size Calculation for Finding Unseen Species

Hongmei Zhang* and Hal Stern†

Abstract. Estimation of the number of species extant in a geographic region has been discussed in the statistical literature for more than sixty years. The focus of this work is on the use of pilot data to design future studies in this context. A Dirichlet-multinomial probability model for species frequency data is used to obtain a posterior distribution on the number of species and to learn about the distribution of species frequencies. A geometric distribution is proposed as the prior distribution for the number of species. Simulations demonstrate that this prior distribution can handle a wide range of species frequency distributions including the problematic case with many rare species and a few exceptionally abundant species. Monte Carlo methods are used along with the Dirichlet-multinomial model to perform sample size calculations from pilot data, e.g., to determine the number of additional samples required to collect a certain proportion of all the species with a pre-specified coverage probability. Simulations and real data applications are discussed.

Keywords: Generalized multinomial model, Bayesian hierarchical model, Markov Chain Monte Carlo (MCMC), Dirichlet distribution, geometric distribution.

1 Introduction

The "species problem" is a term used to refer to studies in which objects are sampled and categorized with interest on the number of categories represented. Research related to the species problem dates back to the 1940's. Corbet (1942) proposed that a mathematical relation exists between the number of sampled individuals and the total number of observed species in a random sample of insects or other animals. Fisher et al. (1943) developed an expression for the relationship using a negative binomial model. Their proposed relationship works well over the whole range of observed abundance, and gives a very good fit to practical situations.

The focus of most statistical research on the species problem has been to estimate the number of unseen species. Bunge and Fitzpatrick (1993) give a review of numerous statistical methods to estimate the number of unseen species. Some notable references are mentioned briefly here. Good and Toulmin (1956) address the estimation of the expected number of unseen species based on a Poisson model. Efron and Thisted (1976) use two different empirical Bayes approaches, both based on a similar Poisson model, to estimate the number of unseen words in Shakespeare's vocabulary. Pitman (1996) proposes species sampling models utilizing a Dirichlet random measure. The negative

*Department of Epidemiology and Biostatistics, University of South Carolina, Columbia, SC, mailto:hzhang@sc.edu

†Department of Statistics, University of California, Irvine, CA, mailto:sternh@uci.edu

 DOI:10.1214/09-BA429

binomial model proposed by Fisher et al. (1943) is also discussed there. Boender and Rinnooy Kan (1987) suggest a Bayesian analysis of a multinomial model that can be used to estimate the number of species. Their model is the starting point for the work presented in this paper.

As seen from the references above, previous applications of the species problem have included animal ecology where individual animals are sampled and categorized into species (Fisher, Corbet, and Williams (1943)), and word usage where individual words are sampled and each word defines its own category (Efron and Thisted (1976)). More recent studies extend the species problem to applications in bioinformatics (Morris, Baggerly, and Coombes (2003)), where the sample items might be DNA fragments and each sequenced DNA segment represents a unique sequence. Our work is motivated by a bioinformatics problem of this type. Some other studies focus on drawing inferences for abundant species or rare species, e.g. Cao et al. (2001) and Zhang (2007). For consistency with the earlier literature, we use the familiar terminology of animals and species.

In this paper, we use the model of Boender and Rinnooy Kan (1987), a generalized multinomial model, as our starting point. The major contribution of this paper is to address sample size calculation for future data collection based on a pilot study. The goal is to determine the sample size in order to achieve a specified degree of population coverage. Non-parametric Bayesian methods have been developed for a related problem, inferring from a given data set the probability of discovering new species, e.g. Tiwari and Tripathi (1989) and Lijoi et al. (2007). In these studies the total number of species is either assumed to be known or to be infinite. The method proposed in this paper, on the other hand, is a two-phase design with the first phase used to infer the number of species and the second phase to estimate the required sample size. The sample size required to achieve a specified degree of population coverage is obtained by Monte Carlo simulations.

The first step is a fully Bayesian approach to drawing inferences regarding the parameters for a generalized Dirichlet-multinomial model for species frequency data. The posterior distribution of the model parameters is used in our Monte Carlo simulation method for sample size determination. For parametric Bayesian analysis of species frequency data selecting an appropriate prior distribution for the number of species in the population is very important (see, for example, Zhang and Stern (2005)). The prior distributions proposed by previous studies (Zhang and Stern (2005); Boender and Rinnooy Kan (1987)) perform poorly in situations in which a population has many rare species (each with very small number of representatives) and a few abundant species. In this case, as indicated by Sethuraman (1994) and discussed in Zhang and Stern (2005), the proportions of each species in the population are crowded at the vertexes of a multi-dimensional simplex such that most proportions are close to zero. For this type of population, inferences for the number of species in the population are often unrealistic. In this paper, we propose to use a geometric distribution as the prior distribution for the number of species. Geometric distributions have been used in many studies, but we have not seen any applications to the species problem. The geometric prior distribution can be used to reflect prior beliefs about the minimum number of species

in the population and prior belief about the range within which the number of species is believed to lie. The flexibility provided in this manner seems to allow the geometric prior distribution to adapt well to different species frequency distributions.

The rest of the paper is organized as follows. In Section 2, we review the hierarchical Bayesian model for species data, describe our choice of prior distributions, and state the conditions required to guarantee a proper posterior distribution for our model. Section 3 focuses on posterior inferences for the model's parameters. Issues related to the implementation of MCMC are also discussed. In Section 4, we develop a Monte Carlo simulation approach for designing future data collection. Section 5 provides results for a simulated data set where the proposed approach works reasonably well. Sensitivity of results to the choice of prior distribution is also discussed. We apply our method to a bioinformatics data set in Section 6. Finally we summarize our results in Section 7.

2 A Dirichlet-multinomial model

2.1 The likelihood function

Let y_i denote the number of observed animals of species i in a sample of size N. Suppose s_o is the number of different species observed and S is the number of species in a population. Then $\mathbf{y} = \{y_1, y_2, ..., y_{s_o}\}$ is one way to represent the observed sample. An alternative description for data of this type based on frequency counts has often been used in the literature. Let $x_o \leq N$ be the maximum frequency over all observed species and n_x be the number of species captured x times, $x = 1, 2, ..., x_o$. Then $\mathbf{n} = (n_1, n_2, \cdots, n_{x_o})$ is another way to represent the sample with

$$N = \sum_{x=1}^{x_o} x n_x = \sum_{i=1}^{s_o} y_i \ .$$

Here we motivate and describe the generalized multinomial probability model for $\mathbf{y}$ of Boender and Rinnooy Kan (1987). To start we introduce notation $\mathbf{y}_{complete}$ for the S-dimensional vector of species counts. The basic sampling model for the counts $\mathbf{y}_{complete}$ is multinomial with the probability for species i to be captured as θ_i, $i = 1, \cdots, S$. There are several possible interpretations for the θ_i's. If we assume all animals are equally likely to be caught, then θ_i represents the relative proportion of species i among the animal population. If not, then θ_i combines the likelihood of being caught and the abundance. If the number of species S is known, the population size of animals is large, and each species has a reasonably large number of representatives in the population, then a plausible model for $\mathbf{y}_{complete}$ is the multinomial distribution with parameters N and $\boldsymbol{\theta} = \{\theta_1, \cdots, \theta_S\}$, i.e. $\mathbf{y}_{complete}|\boldsymbol{\theta}, S \sim Mult(N, \boldsymbol{\theta})$, where $\mathbf{y}_{complete} = \{y_1, y_2, ..., y_S\}$. When S is not known, however, we don't know the dimension of $\mathbf{y}_{complete}$. Then it makes sense to consider the observed data $\mathbf{y} = \{y_1, y_2, ..., y_{s_o}\}$ which only indicates counts for the s_o species that have been observed. There is a complication in that $\mathbf{y}$ provides counts but does not indicate which elements of $\boldsymbol{\theta}$ correspond to the observed species. Though subtle, this point is important in that it invalidates the usual multinomial likelihood.

Since the correspondence between the y_i's and θ_i's can not be determined, the data $\mathbf{y}$ represent a generalized multinomial distribution where we sum over all possible choices for the s_o observed species. Let $W(s_o)$ denote all subsets $\{i_1, \ldots, i_{s_o}\}$ of s_o distinct species labels from $\{1, \ldots, S\}$, then the conditional distribution of $\mathbf{y}$ given $\boldsymbol{\theta}$ and S can be expressed as

$$Pr(\mathbf{y}|\boldsymbol{\theta}, S) = \frac{1}{\prod_{x=1}^{N} n_x!} \frac{N!}{y_1! \ldots y_{s_o}!} \sum_{\{i_1, \cdots, i_{s_o}\} \in W(s_o)} \theta_{i_1}{}^{y_1} \ldots \theta_{i_{s_o}}{}^{y_{s_o}} \quad (1)$$

which is the same result as the one given by Boender and Rinnooy Kan (1987).

The above model assumes infinite population sizes, which can produce limitations in practice. A hypergeometric formulation, which recognizes the finiteness of the populations, seems more appropriate in this context. However, due to computational inefficiency of using hypergeometric models, we use model (1) to describe the distribution of counts $\mathbf{y}$, implicitly assuming that population sizes are large enough to validate the assumptions of multinomial distributions.

2.2 Prior distribution for $\boldsymbol{\theta}$

We model $\boldsymbol{\theta}$ given S as a random draw from a symmetric Dirichlet distribution with parameter α, which is a conjugate prior distribution for the multinomial distribution. We write

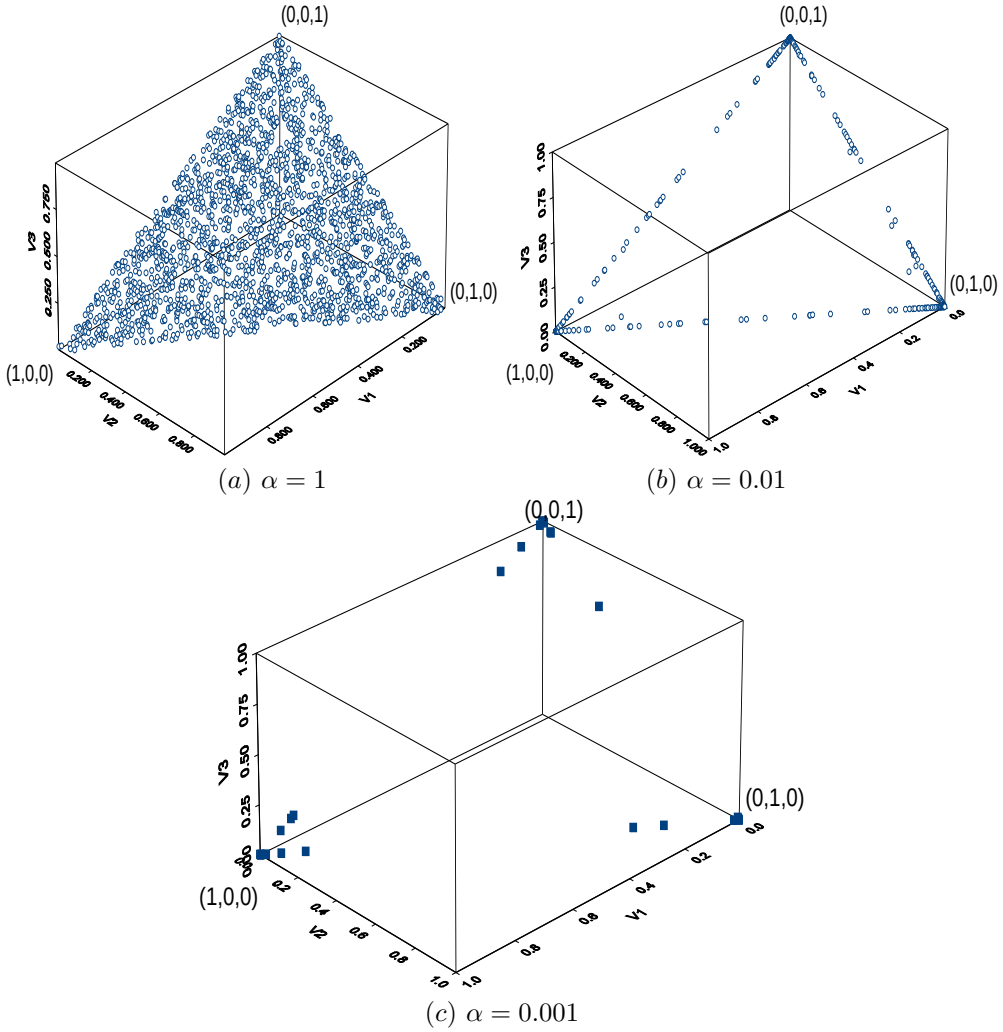
$$\boldsymbol{\theta}|S, \alpha \sim Dirichlet(\mathbf{1}_S \alpha) \quad (2)$$

with

$$p(\boldsymbol{\theta}|S, \alpha) = \frac{\Gamma(S\alpha)}{\Gamma(\alpha)^S} (\prod_{i=1}^{S} \theta_i)^{\alpha - 1}$$

where $\mathbf{1}_S$ is a vector length S with all entries equal to one, and $\boldsymbol{\theta} = \{\theta_1, \cdots, \theta_S\}$ with $\sum_{i=1}^{S} \theta_i = 1$. For a symmetric Dirichlet distribution, $E(\theta_i) = 1/S$, so the prior distribution of $\boldsymbol{\theta}$ assumes all species are a priori equally likely (in expectation) to be captured. The prior variance for each θ_i is $Var(\theta_i) = (1/S)(1 - 1/S)(1/(S\alpha + 1))$. The variance of θ_i becomes smaller as α grows, and tends to 0 as α approaches to infinity. In the limiting case of α being infinity, $\theta_i = 1/S$ and animals from each species are equally likely to be captured. Smaller values of α correspond to greater variation among the elements of $\boldsymbol{\theta}$. Small α can yield many small elements in the vector $\boldsymbol{\theta}$, which corresponds to the case in which the population has many rare species. The reason for this is that as α gets smaller, the vector of θ_i's generated from $Dirichlet(\mathbf{1}_S \alpha)$ is more concentrated on the vertices of the S-dimensional simplex containing vectors $\boldsymbol{\theta}$ that sum to one (Sethuraman (1994); Zhang and Stern (2005)). For instance, with $S = 2$ the Dirichlet distribution reduces to a Beta distribution. When α is small, the density function is U-shaped, with density concentrated near 0 and 1 for both of the proportions. We obtain further insight by considering the distribution of $\boldsymbol{\theta}$ in three dimensions. Figure 1

shows the distribution of $\boldsymbol{\theta}$ in three dimensions for $\alpha = 1, 0.01, 0.001$. When α is larger (e.g. $\alpha = 1$), the probability values are distributed evenly on the simplex. As α gets smaller, θ_i's tend to move toward the vertices of the simplex which have value 1 or 0, which implies more smaller elements in the vector $\boldsymbol{\theta}$ will be generated from the Dirichlet distribution.



(a) $\alpha = 1$ (b) $\alpha = 0.01$

(c) $\alpha = 0.001$

Figure 1: Distribution of $\boldsymbol{\theta}$ for different α's

One might expect this prior to be a bit unrealistic in that we likely know that some species are a priori more likely to be observed than others. The reason for choosing a symmetric Dirichlet distribution is that we do not know S, so we have no information to distinguish any of the θ_i's, $i = 1, \cdots, S$, from any of the others. In this case, the prior

distribution for $\boldsymbol{\theta}$ has to be exchangeable. A possible solution that can address known heterogeneity but retain exchangeability is to consider a mixture of two symmetric Dirichlet distributions corresponding to two different subpopulations, abundant species and scarce species. This approach is used by Morris et al. (2003) in the case when S is known.

2.3 Prior distribution for S and α

We apply a fully Bayesian approach to analyzing species frequency data and conclude model specification by giving prior distributions for S and α.

We specify independent prior distributions for S and α. The two parameters become dependent in their joint posterior distribution as is shown below in Section 2.4. For S we would like to use a relatively flat prior distribution without specifying a strict upper bound on the number of species. We find it useful to have the prior probability density be a decreasing function of S so that there is a slight preference for smaller number of species (this is discussed further in Section 5.5). A prior distribution for S with these characteristics is the geometric distribution with probability mass function

$$Pr(S) = f(1-f)^{S-S_{min}}, \quad S \geq S_{min}, \tag{3}$$

where S_{min} is a specified minimum number of species and f is the geometric probability parameter. Because of Theorem 1 (below) we generally take $S_{min} = 2$, the smallest value for the number of observed species that yields a proper posterior distribution for our model. One interpretation for the parameter f is as the prior probability that there are exactly S_{min} species but this would not ordinarily be a quantity that scientists are able to specify. Instead we propose to obtain a suitable value of f by specifying a plausible maximum value S_{max} for the number of unique species and a measure of prior certainty that S lies between S_{min} and S_{max}. The value of S_{max} will usually be suggested by scientific collaborators as in our application. Under the geometric distribution $Pr(S_{min} \leq S \leq S_{max}) = 1-(1-f)^{S_{max}-S_{min}}$ which can be inverted to find f for specified values of S_{min}, S_{max} and the prior certainty. For instance, if we would like to express high confidence, say probability .999 that S is between $S_{min} = 2$ and $S_{max} = 10000$, then we find $f = .0007$. On the other hand if we are less confident, say 95% certain that the true number of species is in this interval, then $f = .0003$. Note that despite the name we have assigned, we do not assume that S_{max} is an actual upper bound. S_{max} is a device used for elicitation of the geometric probability parameter f. Alternative prior distributions for S (and α) are described below. The sensitivity of posterior inferences to different choices of f and to different prior distributions is considered in Section 5.

The parameter α is important in characterizing the distribution of frequencies. As discussed in an earlier section, large α values lead to uniform distributions and small α leads to a skewed distribution with a few popular species and many rare species. As we don't have much information on α we follow an approach that appears to provide a relatively noninformative hyperprior distribution. We note that the prior standard deviation for each element of $\boldsymbol{\theta}$ is roughly proportional to $1/\sqrt{\alpha}$. By setting a noninfor-

mative prior distribution on this quantity, $p(1/\sqrt{\alpha}) \propto 1$, and doing a change of variable, we obtain a hyperprior distribution for α, i.e.

$$p(\alpha) = \alpha^{-\frac{3}{2}}, \quad \alpha > 0. \tag{4}$$

This is not a proper hyperprior distribution but the following theorem indicates that the posterior distribution is a proper distribution under fairly weak conditions.

Theorem 1: For the model defined by (1) through (4), the posterior distribution $p(S, \alpha|\mathbf{y})$ is proper if at least two species are captured, i.e. $s_o \geq 2$.

Proof: The proof is included in the Appendix.

Naturally other prior distributions are possible and several have appeared in the literature. For example, Zhang and Stern (2005) use a noninformative prior for S, which is discrete uniform distribution on an interval of plausible values, and they use the same prior distribution of α that we use here. However, as discussed by Zhang and Stern (2005), this set of prior distributions can provide misleading posterior inferences when data are consistent with a small value of α. Another set of prior distributions is given by Boender and Rinnooy Kan (1987), where independent proper prior distributions on S and α are proposed:

$$Pr(S) \quad \propto \quad \begin{cases} 1, & S < S_{cut} \\ \frac{1}{(S - S_{cut}+1)^2}, & S \geq S_{cut} \end{cases}, \tag{5}$$

where S_{cut} is a positive number to be set, and

$$p(\alpha) = \begin{cases} 1/2, & \alpha \leq 1 \\ 1/2\alpha^{-2}, & \alpha > 1 \end{cases}, \tag{6}$$

which was earlier proposed by Good (1965). When using this set of prior distributions, as indicated by Boender and Rinnooy Kan (1987) and also by our later simulation results, the posterior inferences can be very sensitive to the choice of S_{cut}, especially when data suggest small values of α. We comment on the sensitivity of results to the prior distributions $P(S, \alpha)$ further in the data analyses of Section 5.

2.4 The posterior distribution

The joint posterior distribution of $\boldsymbol{\theta}, S$, and α for the probability model specified in (1) through (4) is, up to a normalizing constant,

$$\begin{aligned} p(\boldsymbol{\theta}, S, \alpha|\mathbf{y}) \quad &\propto \quad Pr(\mathbf{y}|\boldsymbol{\theta}, S)p(\boldsymbol{\theta}|S, \alpha)Pr(S)p(\alpha) \\ &\propto \quad [\frac{1}{\prod_{x=1}^{N} n_x!}\frac{N!}{y_1!...y_{s_o}!} \sum_{\{i_1,\cdots,i_{s_o}\}\in W(s_o)} {\theta_{i_1}}^{y_1}...{\theta_{i_{s_o}}}^{y_{s_o}}]\frac{\Gamma(S\alpha)}{\Gamma(\alpha)^S}\prod_{i=1}^{s}\theta_i^{\alpha-1} \\ &\quad (1-f)^{S-S_{min}}\alpha^{-\frac{3}{2}} \end{aligned}$$

where, $S \in \{\max(s_o, S_{min}), \max(s_o, S_{min}) + 1, \cdots\}, \alpha > 0$. It should be noted that the posterior distribution is defined over both continuous (for α and $\boldsymbol{\theta}$) and discrete (for S) sample spaces.

The joint posterior distribution can be factored as

$$p(\boldsymbol{\theta}, S, \alpha|\mathbf{y}) = p(\boldsymbol{\theta}|\mathbf{y}, S, \alpha)p(S, \alpha|\mathbf{y}), \tag{7}$$

where $p(\boldsymbol{\theta}|\mathbf{y}, S, \alpha)$ is the conditional posterior distribution of $\boldsymbol{\theta}$ given S and α,

$$\begin{aligned} p(\boldsymbol{\theta}|\mathbf{y}, S, \alpha) &\propto \Big[\sum_{\{i_1,\cdots,i_{s_o}\}\in W(s_o)} {\theta_{i_1}}^{y_1} \cdots {\theta_{i_{s_o}}}^{y_{s_o}}\Big] \prod_{i=1}^{S} \theta_i^{\alpha-1} \\ &= \sum_{\{i_1,\cdots,i_{s_o}\}\in W(s_o)} \left[{\theta_{i_1}}^{y_1+\alpha-1} \cdots {\theta_{i_{s_o}}}^{y_{s_o}+\alpha-1} \prod_{\substack{j=1 \\ j\notin\{i_1,\ldots,i_{s_o}\}}}^{S} \theta_j^{\alpha-1} \right]. \end{aligned} \tag{8}$$

Note that the conditional posterior distribution of $\boldsymbol{\theta}$ is proportional to the sum of $S!/(S-s_o)!$ Dirichlet densities. Also note that every Dirichlet distribution in the summation is identical up to permutation of the species indices.

The other factor in (7), $p(S, \alpha|\mathbf{y})$, is

$$\begin{aligned} p(S, \alpha|\mathbf{y}) &\propto \frac{S!}{(S-s_o)!} \frac{\Gamma(S\alpha)}{\Gamma(N+S\alpha)} \frac{\Gamma(y_1+\alpha)\cdots\Gamma(y_{s_o}+\alpha)}{(\Gamma(\alpha))^{s_o}} (1-f)^S \alpha^{-\frac{3}{2}}, \\ & S \in \{\max(s_o, S_{min}), \max(s_o, S_{min}) + 1, \cdots\}, \alpha > 0. \end{aligned} \tag{9}$$

This can be obtained in either of two ways, as the quotient $p(\boldsymbol{\theta}, S, \alpha|\mathbf{y})/p(\boldsymbol{\theta}|S, \alpha, \mathbf{y})$, or by integrating out $\boldsymbol{\theta}$ from the joint distribution $p(\mathbf{y}, \boldsymbol{\theta}|S, \alpha)$ and working with the reduced likelihood $p(\mathbf{y}|S, \alpha)$.

3 Posterior inferences

3.1 Posterior inferences for S and α

The posterior distribution of S and α as given by (9) is difficult to study analytically. Instead we use MCMC, specifically a Gibbs sampling algorithm with Metropolis-Hastings steps for each parameter, to generate draws from the joint posterior distribution. In applications we run multiple chains from dispersed starting values. Convergence of the sampled sequences is evaluated using the methods developed by Gelman and Rubin (1992a,b) and described for example by Gelman et al. (2003).

The conditional posterior distribution of S given $\mathbf{y}$ and α and the conditional posterior distribution of α given $\mathbf{y}$ and S, up to a normalizing constant, are

$$
\begin{aligned}
Pr(S|\mathbf{y},\alpha) &\propto \frac{S!}{(S-s_o)!}\frac{\Gamma(S\alpha)}{\Gamma(N+S\alpha)}(1-f)^S, \\
&\quad S \in \{\max(s_o, S_{min}), \max(s_o, S_{min})+1, \cdots\}, \\
p(\alpha|\mathbf{y},S) &\propto \frac{\Gamma(S\alpha)}{\Gamma(N+S\alpha)}\frac{\Gamma(y_1+\alpha)\cdots\Gamma(y_{s_o}+\alpha)}{(\Gamma(\alpha))^{s_o}}\alpha^{-\frac{3}{2}}, \quad \alpha>0,
\end{aligned} \tag{10}
$$

respectively. For Metropolis-Hastings steps for these parameters we used jumping or transition distributions that are essentially random walks. Specifically, the jumping function for iteration t for S is a discrete uniform distribution centered at the $(t-1)^{th}$ sampled point; and the jumping function for α is selected as a log-normal distribution with location parameter being the logarithm of the $(t-1)^{th}$ draw. The jumping distributions are discussed more fully in the Appendix.

3.2 Posterior inference for $\boldsymbol{\theta}$

In this paper, posterior inference of $\boldsymbol{\theta}$ is not of interest, but as it may be relevant for other applications we discuss it briefly. The conditional posterior distribution $p(\boldsymbol{\theta}|S,\alpha,\mathbf{y})$ given by (8) is a mixture of $S!/(S-s_o)!$ Dirichlet distributions, one for each choice of the s_o observed species from among the S total species. The component Dirichlet distributions are identical up to permutation of the category indices. Because of this feature of the mixture, each θ_i actually has the same marginal posterior distribution. This makes interpretation of θ_i difficult. We can however talk about posterior inference for a θ corresponding to a particular value of $y_i > 0$. For example, if we define θ_{y_i} as the θ corresponding to an observed species with frequency y_i then $p(\theta_{y_i}|\mathbf{y},S,\alpha)$ is Beta$(y_i+\alpha, N-y_i+(S-1)\alpha)$. The marginal posterior distribution, $p(\theta_{y_i}|\mathbf{y})$, is obtained by averaging this beta distribution over the posterior distribution of S and α that is obtained in Section 3.1.

4 Planning for future data collection

The previous section describes an approach to obtaining posterior inferences for $S, \alpha, \boldsymbol{\theta}$. This section considers an additional inference question. Suppose it is possible to collect additional data beyond the initial N observations. Then one might be interested in questions related to the design of future data collection efforts, such as, "What is the probability of observing at least 90% of all species if the current data are augmented by an additional M animals?", or the closely related question "How large an additional sample is required in order to observe at least 90% of all species with a specified confidence level?". This section addresses the answer to these types of questions.

4.1 A relevant probability calculation

Let p denote the proportion of species we want to capture (e.g. $p = 0.9$), then the probability of capturing at least pS species conditional on the N observed animals and M additional animals, denoted as $\pi(M)$, can be written as

$$\pi(M) = Pr((s_o + S_{new}) \geq pS|M, \mathbf{y}), \tag{11}$$

where S_{new} is the number of previously unseen species observed in the M additional samples. Let $\mathbf{y}^*$ denote the additional data from the M additional observations. The probability (11) can be expressed as an integral over the unknown parameters $S, \alpha, \boldsymbol{\theta}$, and the yet-to-be-observed data $\mathbf{y}^*$,

$$\pi(M) = \int_{\mathbf{y}^*} \int_{\boldsymbol{\theta}} \int_{\alpha} \int_{S} I(s_o + S_{new} \geq pS) p(S, \alpha, \boldsymbol{\theta}, \mathbf{y}^*|M, \mathbf{y})\ dS\ d\alpha\ d\boldsymbol{\theta}\ d\mathbf{y}^*. \tag{12}$$

Here I is an indicator function which is easily determined given the counts $\mathbf{y}$ and $\mathbf{y}^*$, and the value of S. To describe a Monte Carlo approach to evaluating this integral we first observe that

$$p(S, \alpha, \boldsymbol{\theta}, \mathbf{y}^*|M, \mathbf{y}) \quad = \quad p(\mathbf{y}^*|\boldsymbol{\theta}, S, M) p(\boldsymbol{\theta}|\mathbf{y}, S, \alpha) p(S, \alpha|\mathbf{y}),$$

where $p(\mathbf{y}^*|\boldsymbol{\theta}, S, M)$ is a multinomial density function, $p(\boldsymbol{\theta}|\mathbf{y}, S, \alpha)$ is a mixture of Dirichlet distribution function, and $p(S, \alpha|\mathbf{y})$ is given above in (9). Given this factorization, the integration (summation in the case of S) in (12)can be carried out by first obtaining posterior draws of S and α and then applying the specified conditional distributions for $\boldsymbol{\theta}$ and $\mathbf{y}^*$. As is shown below in Section 4.2, sampling $\boldsymbol{\theta}$ from a mixture of Dirichlet distribution is no more difficult than sampling $\boldsymbol{\theta}$ from a Dirichlet distribution. We do not expect that the high dimension of $\boldsymbol{\theta}$ will cause any problem in the numerical integration process.

Carrying out the integration for a variety of values of M identifies a $\pi(M)$ curve and allows us to identify the smallest sample size for which $\pi(M)$ exceeds a given target. This approach provides a point estimate for the needed sample size but does not provide a great deal of information about the uncertainty in such an estimate. Instead, we find it useful to examine the function $\pi(M)$ for a variety of S, α values, i.e.

$$\begin{aligned} \pi(M|S, \alpha) &= Pr((s_o + S_{new}) \geq pS|M, \mathbf{y}, S, \alpha) \\ &= \int_{\mathbf{y}^*} \int_{\boldsymbol{\theta}} I(s_o + S_{new} \geq pS) p(\boldsymbol{\theta}, \mathbf{y}^*|M, \mathbf{y}, S, \alpha) d\boldsymbol{\theta} d\mathbf{y}^*. \end{aligned} \tag{13}$$

Examining $\pi(M)$ in this way allows us to use the posterior distribution of S, α to convey uncertainty about our estimate of M. The function $\pi(M|S, \alpha)$ is a complicated function of S and α, and an analytical form of its posterior distribution is not possible. Instead, we use a Monte Carlo approach to estimate the posterior distribution of $\pi(M|S, \alpha)$. Specifically, for each posterior draw of S and α, we estimate the quantity $\pi(M|S, \alpha)$ by averaging over $\boldsymbol{\theta}$ and $\mathbf{y}^*$. The posterior distribution of $\pi(M|S, \alpha)$ is obtained by repeating the Monte Carlo evaluation for the available draws of S and α.

4.2 Monte Carlo simulation procedure

The Monte Carlo approach to evaluating $\pi(M|S,\alpha)$ in (13) is made explicit by applying the identity $p(\boldsymbol{\theta}, \mathbf{y}^*|M, \mathbf{y}, S, \alpha) = p(\mathbf{y}^*|\boldsymbol{\theta}, S, M)p(\boldsymbol{\theta}|\mathbf{y}, S, \alpha)$, where we have assumed that $\mathbf{y}^*$ is conditionally independent of α and $\mathbf{y}$ given $M, \boldsymbol{\theta}$, and S. The assumption of conditional independence is based on the consideration that $\boldsymbol{\theta}$ and S fully define the probability vector for multinomial sampling of $\mathbf{y}^*$. The algorithm for computing $\pi(M|S,\alpha)$ for a given S, α pair is then given by the following steps. For $t = 1, \cdots, T$,

1) generate $\boldsymbol{\theta}^{(t)}$ from $p(\boldsymbol{\theta}|\mathbf{y}, S, \alpha)$ (a mixture of Dirichlet distributions)
2) generate $\mathbf{y}^{*(t)}$ from a Multinomial distribution with parameters M and $\boldsymbol{\theta}^{(t)}$
3) define $I_t = 1$, if $(s_o + S_{new}) \geq pS$, and $I_t = 0$ otherwise.

Estimate $\pi(M|S,\alpha)$ with $\frac{1}{T}\sum_{t=1}^{T} I_t$ and repeat steps 1 to 3 for as many different values of M as desired.

For each given pair of S, α, the result can be viewed as a curve giving the probability of covering a proportion p of the species as a function of M. If k posterior draws of S, α are available, then there are totally k such curves.

The Monte Carlo algorithm is conceptually straightforward but a number of implementation details are noteworthy. First, recall that the posterior distribution of $\boldsymbol{\theta}$ given $\mathbf{y}, S, \alpha$ is a mixture of Dirichlet distributions. All of the Dirichlet distributions in the mixture are identical up to permutation of the indices $(i_1, i_2, \cdots, i_S)$. Sampling from the mixture distribution requires that one pick a set of labels from $W(s_o)$ to correspond to the observed species and then simulate from the relevant component of the mixture distribution. The subsequent steps in the algorithm would then be done conditional on this choice of labels. In practice, because we are not interested in a specific θ_i or y_i, it is equally valid to arbitrarily assign labels to the observed species and proceed. A second noteworthy detail concerns efficiency. Steps 1 and 2 of the algorithm propose to use only a single draw of $\mathbf{y}^*$ for each $\boldsymbol{\theta}$. It is natural to ask whether the algorithm might be improved by selecting multiple $\mathbf{y}^*$ vectors for each $\boldsymbol{\theta}$, perhaps thereby estimating a separate curve for each $\boldsymbol{\theta}$. Our simulation results suggest however that variation among the curves corresponding to different $\boldsymbol{\theta}$'s for fixed S and α is relatively small and consequently the algorithm described above works best. Lastly, we note that step 3 of the Monte Carlo simulation procedure requires determining the number of new species by counting the number of positive y_i^*'s whose θ_i's correspond to species with $y_i = 0$ (or equivalently to Dirichlet parameter α). In practice it is possible to save a considerable amount of computing time by embedding iteration over the sample size M within the above loop (instead of running the above loop separately for each M).

4.3 The probability of species coverage and the required sample size

The Monte Carlo algorithm yields a collection of curves, each showing $\pi(M|S,\alpha)$ as a function of M. For any M these curves yield a posterior distribution of the quantity $\pi(M) = Pr((s_o + S_{new}) \geq pS|M,\mathbf{y})$. Posterior summaries, e.g., point estimates or posterior intervals can be constructed from this estimated posterior distribution.

Another practical question is how to find the minimum sample size required to observe at least a proportion p of all species with a specified probabilty q. We denote this value as M_q; it too can be viewed as a function of S and α. The posterior distribution of M_q is determined easily using the Monte Carlo approach. For each (S,α) pair we have developed an estimated curve showing $\pi(M)$ vs M. For each such curve we identify the smallest value of M such that $\pi(M|S,\alpha) \geq q$. The collection of identified sample sizes provides an estimated posterior distribution for M_q.

5 Simulations

To demonstrate our method we begin by simulating a single data set with $N = 2000$ observations for which S is known. In a later section, we consider the effect of increasing the sample size.

5.1 Data

The data are $N = 2000$ observations simulated from a multinomial distribution with $S = 2000$ species in the population and $\boldsymbol{\theta}$ a random sample from a Dirichlet distribution with $\alpha = 1$. The distribution of $\boldsymbol{\theta}$ is then uniform over all vectors with $\sum_{i=1}^{S} \theta_i = 1$. Table 1 and Figure 2 describe the data as the number of species that appear exactly x times, $x = 1, 2, \cdots, x_o$.

In this sample, the largest frequency $x_o = 11$ and the number of observed species is $s_o = 965$.

Table 1: Species frequencies

x	1	2	3	4	5	6	7	8	9	10	11
# species	451	268	116	61	35	16	5	7	2	3	1

5.2 Posterior inference for S

We used the approach described in Section 3 to find the posterior distribution of S and α given the simulated data. We assume the plausible upper limit on S is $S_{max} =$

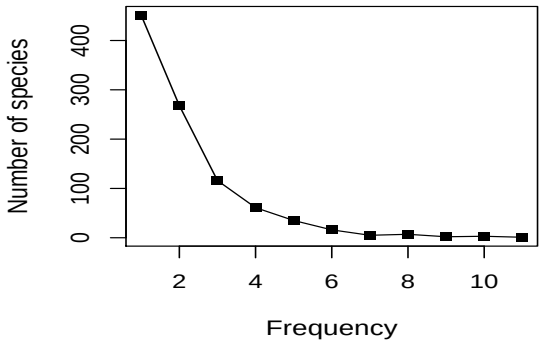


Figure 2: The distribution of frequencies for the simulated data

10000 with prior belief about 0.999 that $S_{min} \leq S \leq S_{max}$. This gives the value of f in our geometric prior distribution as $f = 0.0007$. In a later section, Section 5.5, we evaluate the effect of choosing different values of f (i.e. different geometric prior distributions). The posterior inferences in this section are based on 4000 draws from the posterior distribution after a MCMC burn-in period of 4000 iterations. Figure 3 shows a contour plot of the joint posterior distribution of S and α. The distribution has a single mode around $S = 1800$ and $\alpha = 1.0$. Figure 4 and 5 are histograms of the posterior distribution of S and α. The posterior mean of S is 1844. A 95% central posterior interval for S is (1559, 2301). The true value, $S = 2000$, is contained in the interval. Note that the method of Efron and Thisted (1976) based on the Poisson-Gamma model yields a similar estimate of S which is $\widehat{S} = 1639$ with standard error 226. A 95% central posterior interval of α is (0.64,1.69), which includes the true value $\alpha = 1$. The posterior mean of α is 1.07.

5.3 Sample size calculation for future sampling

As discussed in Section 4, we can estimate the probability of observing a proportion p of the total number of species given an additional M animals, and the sample size required to ensure that future sampling covers a specified proportion of the species with a given probability of coverage. As a first step we estimate $\pi(M|S, \alpha)$ for $M = 2000$ to 30000 in steps of size 20 for a number of (S, α) pairs. Each point of the $\pi(M|S, \alpha)$ vs M curve is based on $T = 1500$ Monte Carlo evaluations in order to make the Monte Carlo simulation error of a given probability less than 0.015.

Figure 6 is a plot giving $\pi(M|S, \alpha)$ for 100 draws of (S, α) from the posterior distribution $p(S, \alpha|\mathbf{y})$ (including more curves makes the figure more difficult to read). Each curve in the figure shows the relationship between the probability of seeing 90% of the

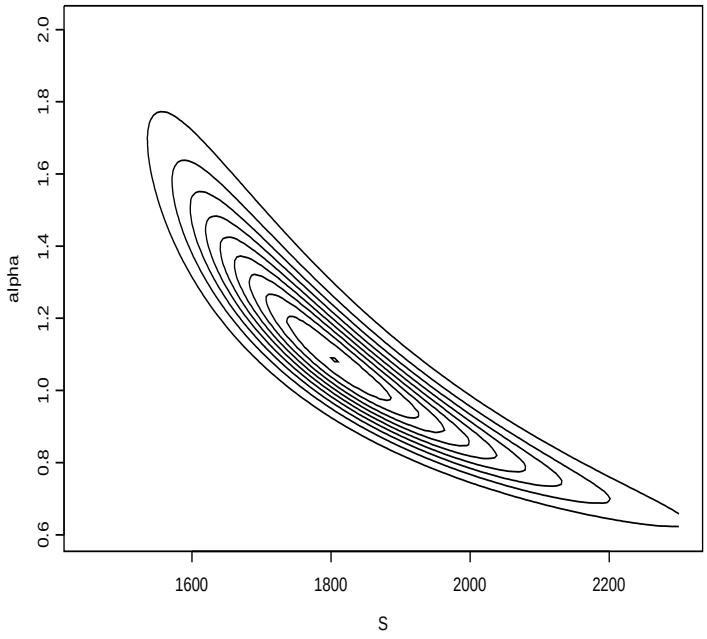


Figure 3: Contour plot for the data from $S = 2000,\ \alpha = 1$ with N=2000 ($f = 0.0007$)

species and the additional sample size M for a single posterior draw of S and α. From the figure, we can see the curves are spread out especially for larger coverage probabilities, which implies large uncertainty about the probability of seeing 90% of the species with a given M, and also large uncertainty about the minimum sample size required to see at least 90% of the species for a specified confidence level π. This reflects the uncertainty about the parameter α which has a very substantial impact on the species frequency distribution. Posterior draws with large α values will tend to have smaller S values, and hence greater likelihood of observing 90% of the species with M additional animals. These values correspond to the curves on the left in Figure 6. A small α suggests the true S is larger, so we are less likely to observe 90% of the species with M additional animals.

Probability of observing at Least pS species with M additional animals

We next use the curves in Figure 6 to draw posterior inference for the probability of observing at least $0.9S$ species with M additional observations. Table 2 gives the posterior median of $\pi(M)$ and a 90% central posterior interval for $\pi(M)$ for a range of M

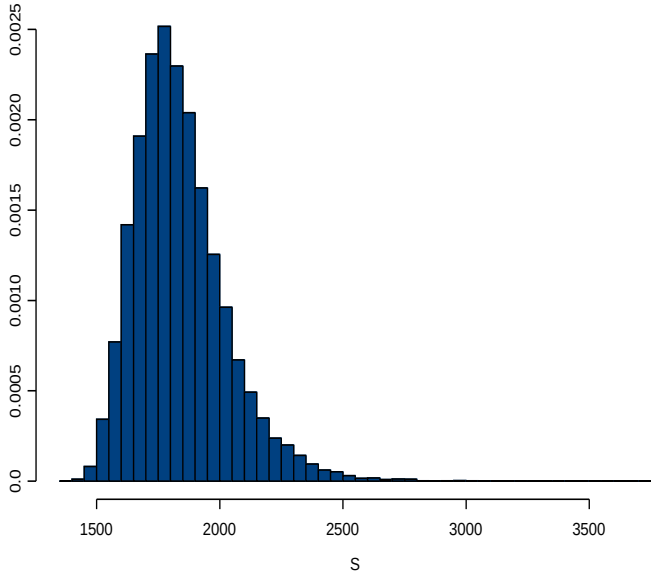


Figure 4: Histogram of the posterior of S

values. These inferences are based on $k = 100$ (S, α) pairs chosen randomly among the 4000 posterior samples. Figure 7 shows the posterior median and pointwise 90% central posterior intervals graphically for M ranging from 400 to 20000. Posterior intervals for a given value of M tend to be wide when M is relatively small (e.g. $M = 5000$), but the length of the intervals decreases quickly with the increase of M. This reflects the form of Figure 6; for a given M most curves have probability values of attaining the target number of species near zero (if that value of M is relatively small compared to the value of S on which the curve is based) or near one (if the value of M is relatively large compared to the relevant value of S).

Table 2: Estimates of $\pi(M)$ for different M values

M	5000	8000	10000	12000
$\widehat{\pi}(M)$	0.52	1	1	1
90% emp. post. int.	(0, 1)	(0.02, 1)	(0.48, 1)	(0.75,1)

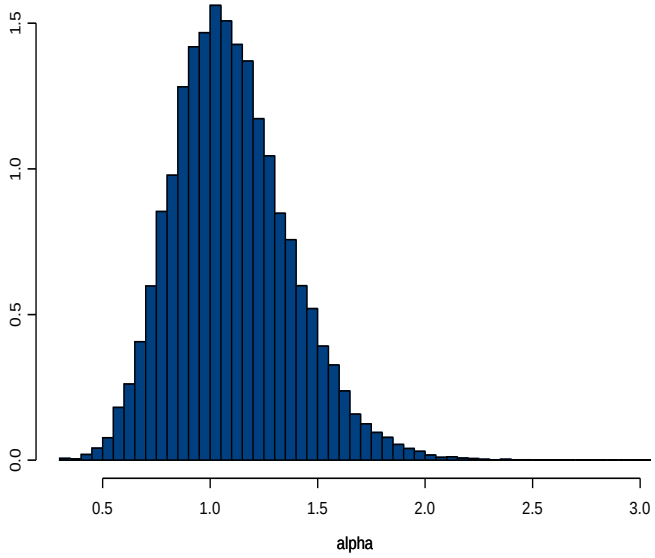


Figure 5: Histogram of the posterior distribution of α

Required sample size to capture at least pS species with coverage probability 0.9

Recall that for the simulated data the initial sample of 2000 animals captured approximately half the species. How many additional animals have to be collected if we want to see at least $0.9S$ species with probability $q = 0.9$? Once again the Monte Carlo simulations in Figure 6 can be used to answer the question. Drawing a horizontal line with coverage probability $q = 0.9$, the intersection points of the horizontal line and the curves give estimates of M_q, one from each curve. The distribution of these values provides the desired posterior inferences. The posterior median is 5330, and an empirical 90% central posterior interval is $(2980, 13080)$. Table 3 gives the posterior median of the needed sample size together with 90% central posterior intervals for various values of the target proportion of species (p) and the desired coverage probability (q). As the proportion of the species to be captured (p) increases, the number of additional samples required increases quickly, which is natural because the more common species have undoubtedly been observed. Further, as indicated in the table, for each p, the sample size required to achieve different coverage probabilities (q) changes slowly. This is once again due to the pattern observed in Figure 6, in which all curves display a similar trend: there is a steep increase in the probability of coverage q near a threshold value of M (though this threshold varies depending on S and α).

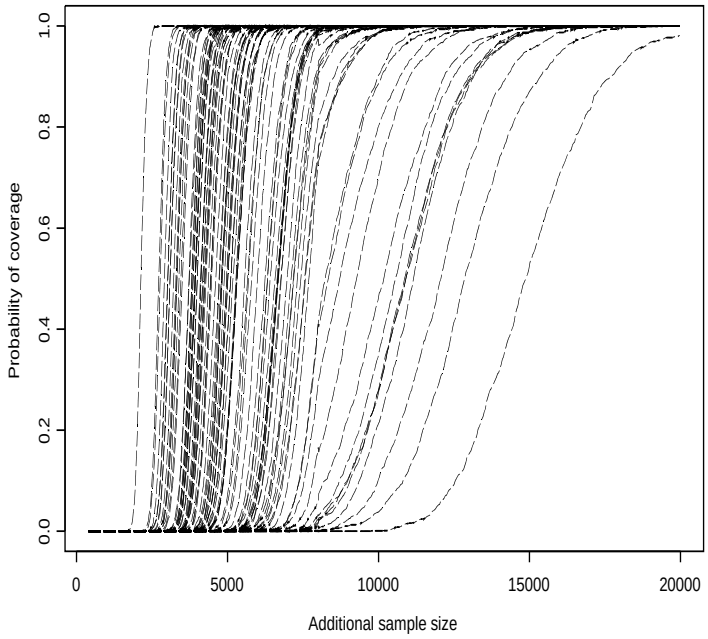


Figure 6: The relationship between probability $\pi(M|S,\alpha)$ and additional sample size M (N=2000)

Table 3: Size of additional sample required to obtain a fraction p of the total species with probability q

	Probability of covering specified fraction (q)		
Fraction of species (p)	0.5	0.7	0.9
0.7	810	830	870
	(290, 1570)	(310, 1630)	(330, 1700)
0.8	1940	2000	2090
	(1060, 3540)	(1100, 3660)	(1120, 3840)
0.9	5010	5190	5330
	(3020, 10960)	(3140, 11760)	(2980, 13080)

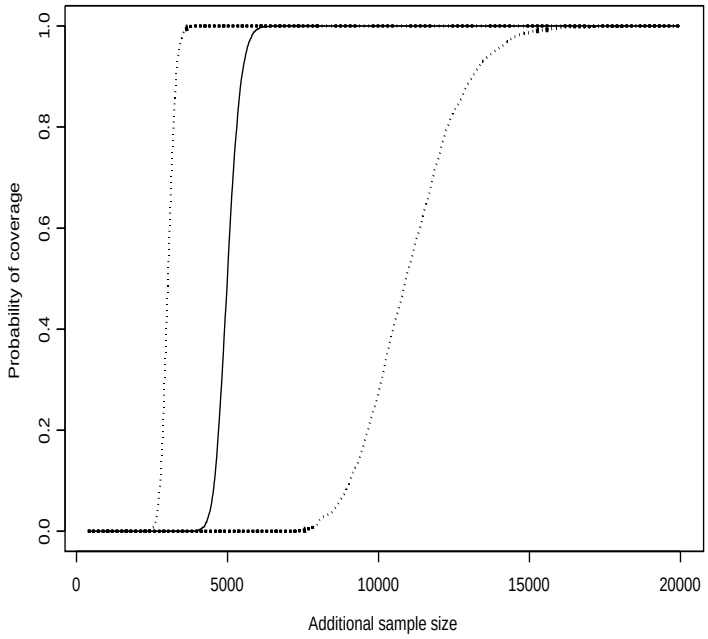


Figure 7: Trace of the estimated coverage probabilities. The middle line connects the posterior median values of $\pi(M|S,\alpha)$, and the two dotted lines besides it are the 90% pointwise central posterior intervals

5.4 Effect of sample size

Having demonstrated the approach for a sample of size 2000 from our hypothetical population in Sections 5.1 through 5.3, we now demonstrate the impact of increasing the sample size. One would expect the inferences to become more precise. We simulate a data set with size $N = 10000$ from the same population as before, i.e. $\alpha = 1, S = 2000$. For this sample, the highest frequency is $x_o = 45$, and the number of observed species is $s_o = 1663$, which is more than 80% of the total number of species.

In this example, the value of f is also selected as $f = 0.0007$. The posterior mean for S is 2030 with 95% central posterior interval (1948, 2129). The posterior mean for α is 0.93, and a 95% central posterior interval for α is (0.80, 1.08). Both intervals are much narrower than for the case with $N = 2000$. Figure 8 shows the posterior distribution of $\pi(M|S,\alpha)$ with $p = 0.9$ (i.e. capturing 90% of all species) for 100 (S,α) pairs and a number of M values. There is much less variation than is present in Figure 6.

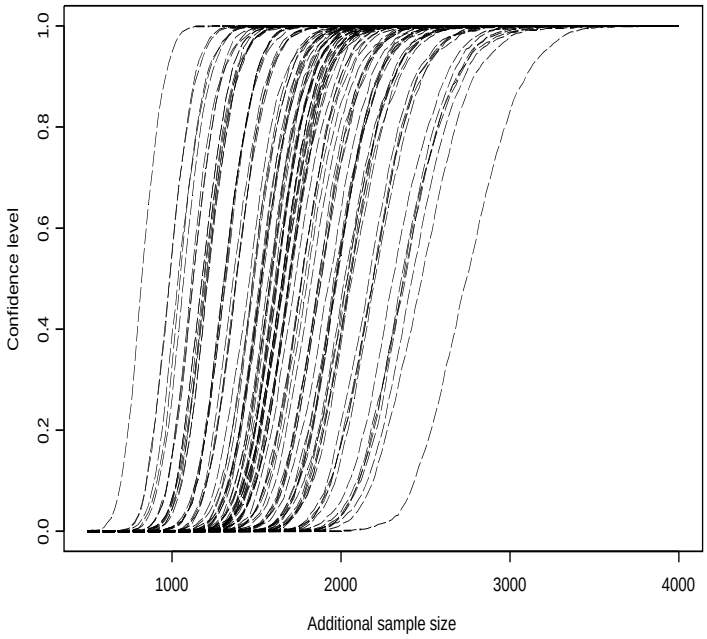


Figure 8: The relationship between probability $\pi(M|S,\alpha)$ and additional sample size M ($N = 10000$)

The posterior median of M_q for $q = 0.9$ is 1880 and a 90% central posterior interval is (1240, 2700). The required sample size is smaller because a larger number of species are observed in the pilot sample. In addition the posterior interval is much narrower than that based on $N = 2000$ observations.

5.5 Effect of prior distribution

As in any Bayesian analysis, it is critical to consider the impact of the choice of prior distribution on the inferences obtained. That is especially true here with the prior distribution for S and α. This section addresses comparisons between our prior distribution and others in the literature, as well as some practical issues associated with the use of our prior distribution.

In section 2.3, two other choices of prior distributions for S and α were mentioned. One is the proposal of Boender and Rinnooy Kan (1987) to use proper prior distribution functions for S and α, and the other suggested by Zhang and Stern (2005), where a

uniform distribution with an upper bound is assigned to S and a vague prior distribution is given to α (the same as the prior distribution for α that is used here). We applied the Dirichlet-generalized multinomial model with their prior distributions on the same simulated data discussed in Section 5.1. The posterior inferences for S and α are listed in Table 4, which indicates the results from these two alternative prior distributions are consistent with the results using the geometric prior distribution for S. The findings regarding species coverage and sample size are also similar across different methods.

We next discuss the effect of different choices of f on the posterior inferences for the simulated data. As noted earlier, different values of f imply different degrees of confidence that we might have with respect to the suggested range of S, with larger values of f corresponding to higher prior confidence of S being between S_{min} and the plausible S_{max}. Table 4 lists various choices of f for $S_{min} = 2$ and $S_{max} = 10000$, the corresponding probability of $S \in (S_{min}, S_{max})$, and the posterior inferences of S and α that result from this choice of f. The results suggest that as long as f is not too big (i.e. our prior belief in S_{max} is not too extreme), the posterior inferences for the parameters are consistent across different values of f and all agree reasonably well with the true values. We also observe that the larger the value of f, the stronger our prior information favors small values of S. This is reflected in the inferences; the posterior mean decreases as f increases. As f increases the prior distribution puts more probability mass on smaller values of S and thus the posterior mean will decrease.

The preceding discussion concerns the simulated data discussed in Section 5.1 where the population essentially does not have any rare species. The three different prior distribution choices give similar results in this case. However, as noted in Section 2.3, the methods of Boender and Rinnooy Kan (1987) and Zhang and Stern (2005) both have difficulty in inferring the number of species if the sample is consistent with small α values in the population. We use simulated data to demonstrate and compare the three methods in this context. A random sample is drawn from a population with a large number of rare or infrequent species. The same scenario given in section 5.1 is applied and a data set with $N = 2000$ observations is drawn from a population with $\alpha = 0.01, S = 2000$. In this data set, the number of observed species is $s_o = 94$ and the highest frequency is $x_o = 155$, which implies the population has some very abundant species but many more rare or hard to capture species (Zhang and Stern (2005)). The value of f is selected as before, i.e., $f = .0007$ corresponding to high confidence (.999) that the true number of species is between $S_{min} = 2$ and the suggested $S_{max} = 10000$. The posterior inferences from the three methods are listed in Table 5; posterior means are given along with 95% central posterior intervals in parentheses. The results listed in Table 5 demonstrate the poor performance using the prior distributions of Boender and Rinnooy Kan (1987) and Zhang and Stern (2005). For the new prior distribution, with $f = 0.0007$, the posterior inferences are consistent with the true values. We notice that the posterior inferences seem more sensitive to the choice of f in this context than in the high α case. We also find that the posterior interval of S using the geometric prior distribution is wide, which implies large uncertainty on the value of S due to the large number of rare species in the population. The inferences can be improved if more information is available to help construct an informative prior distribution of S.

Table 4: Posterior inferences on S and α from different methods (95% posterior intervals are in the parentheses)

		$\widehat{S}$	$\widehat{\alpha}$
Boender and Rinnooy Kan (1987)			
	$S_{cut} = 500$	1833	1.08
		(1597, 2160)	(0.72, 1.55)
	$S_{cut} = 1000$	1817	1.1
		(1573, 2122)	(0.74, 1.62)
Zhang and Stern (2005)			
		1866	1.09
		(1576, 2337)	(0.64,1.75)
Geometric prior			
$p(S_{min} \le S \le S_{max})$	f	$\widehat{S}$	$\widehat{\alpha}$
0.63	0.0001	1870	1.03
		(1556, 2320)	(0.61, 1.69)
0.90	0.00023	1865	1.04
		(1574, 2335)	(0.61, 1.69)
0.999	0.0007	1844	1.07
		(1559, 2301)	(0.64, 1.69)
0.9999	0.001	1842	1.07
		(1537, 2252)	(0.64, 1.70)
≈ 1	0.002	1803	1.12
		(1537, 2252)	(0.70, 1.32)
≈ 1	0.003	1782	1.16
		(1530, 2153)	(0.72, 1.81)
≈ 1	0.006	1721	1.26
		(1501, 2022)	(0.82, 1.93)
≈ 1	0.01	1660	1.39
		(1467, 1909)	(0.93, 2.09)

Table 5: Posterior inferences on S and α from different methods when population α is small ($\alpha = 0.01$)

	$\widehat{S}$	$\widehat{\alpha}$
Boender and Rinnooy Kan (1987)		
$S_{cut} = 50$	297	0.12
	(129, 1021)	(0.021, 0.31)
$S_{cut} = 500$	318	0.087
	(151, 490)	(0.045, 0.23)
$S_{cut} = 5000$	1506	0.023
	(174, 4611)	(0.0044, 0.18)
Zhang and Stern (2005)		
	5159	0.0054
	(451, 9822)	(0.0019, 0.052)
Geometric prior		
f=0.0002	2548	0.013
	(247, 9100)	(0.0022, 0.11)
f=0.0005	2307	0.014
	(270, 7323)	(0.0027, 0.096)
f=0.0007	2180	0.014
	(267, 6734)	(0.0030, 0.098)
f=0.001	1650	0.017
	(255, 5280)	(0.0038, 0.11)

6 Application to sequence data

6.1 Description of the data

This work was motivated by a bioinformatics problem arising during a genome sequencing project. Details of the technological approach are not particularly crucial here – for one thing, the approach is no longer used by the company. A key issue that came up during the project was the desire to identify the unique elements in a set of DNA fragments. The unique elements could be easily determined by sequencing all of the fragments but this is not necessarily cost effective if there is a lot of duplication. One strategy under consideration proposed sequencing a small sample of fragments and recording the frequency with which each unique sequence was found. Framed in this way the problem is directly analogous to our species problem. The hope is that based on the small sample it will be possible to determine how large of a sequencing effort to mount.

A prototype data set was provided with sample size $N = 1677$ and $s_o = 644$, in which there were 440 species each observed once and 1 species observed 76 times. Figure 9 shows the pattern of the data in terms of frequencies. The figure shows a very sharp decreasing pattern in the distribution of frequencies, which is different from that of our simulated data in Section 5.1 and more like the small α case discussed at the end of Section 5.5. A few "species" occur with high frequencies, and a very high proportion of the observed species only occur once. This is the type of data that typically indicates a small value of α that can cause difficulties for the generalized multinomial model.

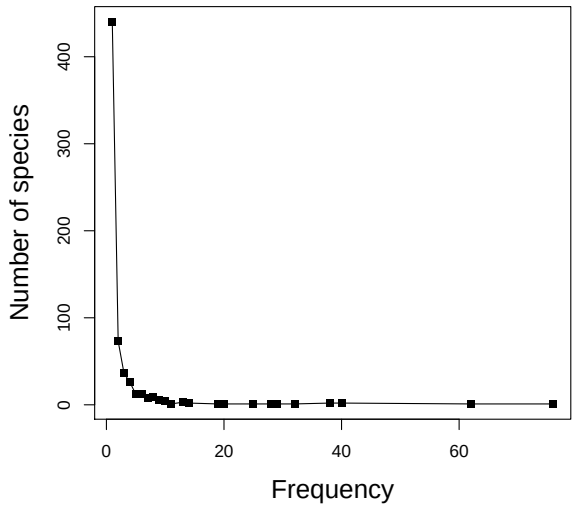


Figure 9: The distribution of the frequencies for the DNA sequence data

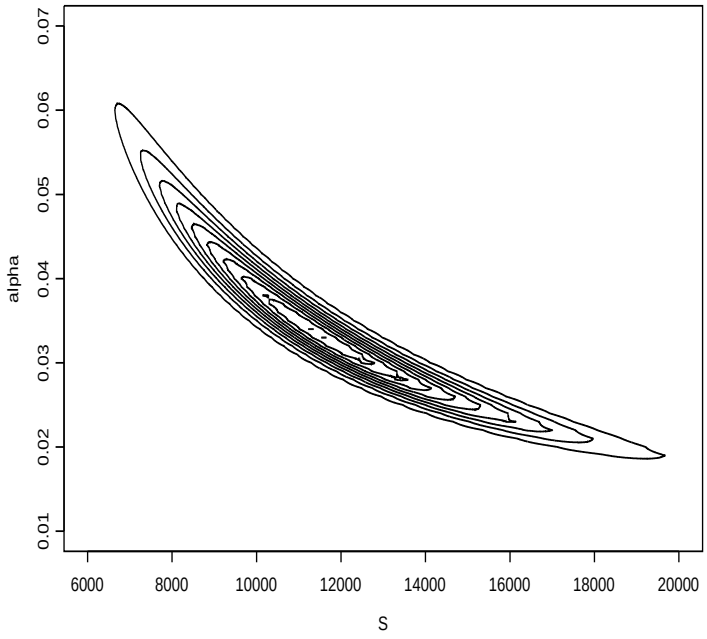


Figure 10: Contour plot for the true data

6.2 Applying the model

We apply the method proposed in previous sections to the DNA segments data. Our collaborator suggested the maximum value of S be $S = 10000$. With the value of f selected as $f = 0.0007$, as discussed earlier, our prior confidence of S between 2 and 10000 is 0.999. Figure 10 is a contour plot for the posterior distribution of S and α, which clearly shows one mode around $S = 12000$. The posterior mean for S is 12111. A 95% central posterior interval of S is (7246, 19637). The posterior mean for α is 0.033 and its 95% central posterior interval is (0.020, 0.056).

Note that although the prior confidence of S in the interval (2,10000) is 0.999, the posterior distribution is concentrated above 10000. This suggests that the data provide overwhelming evidence of a large number of rare species. As seen in the next section, this inference results in our determination that extremely large sample sizes are required to collect even a small fraction of the total number of species.

6.3 Sample size calculation for future sampling

We use the Monte Carlo simulation approach discussed in Section 4.2 to carry out the sample size calculation. The posterior inferences for S suggests a large number of distinct DNA sequences in the population. The posterior inference of α implies that the population has a large number of rare species. We thus expect a large sample is needed even for modest species coverage. Table 6 lists the estimated sample sizes in order to see 10% or 15% of all the distinct DNA sequences with different probabilities of coverage. Similar patterns are observed as in the simulations: the change in the required sample sizes across different coverage probabilities (q) is small for a given target fraction (p). On the other hand, the required sample sizes and the uncertainty both increase quickly with small increase in the target fraction of species (p). Due to the large number of rare species in the population the inferences obtained here are of limited value commercially; an extremely large sample size is required to see a substantial fraction of the species.

Table 6: Additional sample sizes needed to collect 10% or 15% of all distinct DNA sequences

	Probability of covering specified fraction (q)	
Fraction of species (p)	0.5	0.9
0.10	1900	2200
	(400, 6425)	(450, 7488)
0.15	7000	7500
	(2600, 23500)	(2800, 27000)

7 Summary

A multinomial-Dirichlet model is proposed for the analysis of data in which individual objects belong to different categories. The prior distribution for the number of categories is selected to be a geometric distribution with probability parameter set to reflect our confidence that the number of categories lies in a predetermined range. The multinomial-Dirichlet model with this prior distribution seems to work well over a range of scenarios. A new Monte Carlo simulation algorithm is introduced for determining the minimum size of an additional sample required to capture a certain proportion of categories in the population with specified coverage probability. Simulation results show that sample size calculation in this way is feasible. An application to a DNA segments data set indicates the applicability of the proposed method but also suggests continued difficulty with the problematic case with many rare species. Future study is needed to extend the model to address situations where the distribution of species is not well approximated by our model, e.g. where the relative proportion of rare species is high.

Appendix

A.1 Proof of Theorem 1

The joint posterior distribution of S and α, as derived in Section 2, is

$$p(S,\alpha|\mathbf{y}) \quad \propto \quad \frac{S!}{(S-s_o)!}\frac{\Gamma(S\alpha)}{\Gamma(N+S\alpha)}\frac{\Gamma(y_1+\alpha)\cdots\Gamma(y_{s_o}+\alpha)}{(\Gamma(\alpha))^{s_o}}\alpha^{-\frac{3}{2}}, \tag{14}$$

for $S \geq s_o$ and $0 < \alpha < \infty$. We find the conditions required to insure that

$$\sum_{S=s_o}^{\infty} \int_0^{\infty} p(S,\alpha|\mathbf{y})d\alpha < \infty,$$

by obtaining an upper bound on the integral over α for each S.

For each S, choose $\epsilon > 0$ such that $S\epsilon < 1$. We then consider the integral over two intervals $(0,\epsilon)$ and (ϵ,∞).

On the interval $(0,\epsilon)$**:**

Recall that for the gamma function we have $\Gamma(1+z) = z\Gamma(z)$. This and other properties of the gamma function yield the following results.

1. $\Gamma(\alpha) = \Gamma(1+\alpha)/\alpha$ $(\alpha > 0)$
2. If $y_i \geq 1$ and $\alpha < \epsilon < 1$, then $\Gamma(y_i+\alpha) < \max(\Gamma(y_i+1),1)$
3. Define $\gamma_{min} > 0$ as the minimum value of the gamma function on the interval (1,2), then $\Gamma(1+\alpha) \geq \gamma_{min}$ for $0 \leq \alpha < 1$.
4. $\Gamma(S\alpha) = \Gamma(1+S\alpha)/(S\alpha) < 1/(S\alpha)$ since $S\alpha < S\epsilon < 1$.
5. If $s_o \geq 2$, then we must have $N \geq 2$ so that $\Gamma(N+S\alpha) > \Gamma(N)$

Applying these equalities and inequalities gives

$$\begin{aligned}
&\sum_{S=s_o}^{\infty} \int_0^{\epsilon} p(S,\alpha|\mathbf{y})d\alpha \\
&< \sum_{S=s_o}^{S_{max}} \frac{S!}{(S-s_o)!}\int_0^{\epsilon} \frac{\prod_{i=1}^{s_o}\max(\Gamma(y_i+1),1)}{\gamma_{min}^{s_o}\Gamma(N)}\frac{1}{S\alpha}\alpha^{s_o-3/2}(1-f)^S d\alpha \\
&= \sum_{S=s_o}^{S_{max}} \frac{(S-1)!}{(S-s_o)!}(1-f)^S \int_0^{\epsilon} C_y \alpha^{s_o-5/2}d\alpha
\end{aligned}$$

where $C_y = [\prod_{i=1}^{s_o}\max(\Gamma(y_i+1),1)]/[\gamma_{min}^{s_o}\Gamma(N)]$ is a constant depending only on $\mathbf{y}$. For $s_o \geq 2$ the integral near zero is finite and thus so is the sum since the prior distribution of S is proper.

On the interval (ϵ, ∞)**:**

Repeated application of the recurrence $\Gamma(1+z) = z\Gamma(z)$ yields

$$\begin{aligned} p(S,\alpha|\mathbf{y}) &\propto \frac{S!}{(S-s_o)!}\frac{\prod_{i=1}^{s_o}\prod_{j=1}^{y_i}(y_i+\alpha-j)}{\prod_{j=1}^{N}(S\alpha+N-j)}\alpha^{-3/2}(1-f)^S \\ &= \frac{S!(1-f)^S}{(S-s_o)!}\frac{\alpha^{-3/2}}{S^N}\frac{\prod_{i=1}^{s_o}\prod_{j=1}^{y_i}(1+\frac{(y_i-j)}{\alpha})}{\prod_{j=1}^{N}(1+\frac{(N-j)}{S\alpha})} \\ &< \frac{S!(1-f)^S}{(S-s_o)!}\frac{\alpha^{-3/2}}{S^N}\prod_{i=1}^{s_o}\prod_{j=1}^{y_i}(1+\frac{(y_i-j)}{\epsilon}) \end{aligned}$$

The final product is a constant in terms of S and α and the remaining terms yield a finite integral over α and sum over S.

Combining the information from the two intervals, we conclude that the posterior distribution is proper if $s_o \geq 2$, i.e., there are at least two categories observed.

A.2 Jumping functions for S and α

Metropolis-Hastings jumping function for S

The jumping function for S is a symmetric discrete uniform distribution centered at $S^{(t-1)}$ (an asymmetric distribution is used if $S^{(t-1)}$ is near the limit of its range) with width parameter $b^{(t-1)}$. Take $S^{(*)}$ as the proposed value of S when jumping from $S^{(t-1)}$. The jumping distribution can be written as

$$S^{(*)}|S^{(t-1)} \sim \begin{cases} DUNIF(S^{(t-1)}-b^{(t-1)}, S^{(t-1)}+b^{(t-1)}), & S^{(t-1)} \geq s_o + b^{(t-1)} \\ DUNIF(s_o, S^{(t-1)}+b^{(t-1)}), & S^{(t-1)} < s_o + b^{(t-1)}, \end{cases}$$

where the second two lines represent the cases where the current draw is near the boundary of the parameter space. The width parameter $b^{(t-1)}$ is selected to be proportional to the current value $S^{(t-1)}$.

Metropolis-Hastings jumping function for α

We use a normal jumping distribution on the logarithm of α. Define $\phi = log(\alpha)$. Let $\phi^{(t-1)}$ denote the current sampled point, and $\phi^{(*)}$ be the candidate point generated from the jumping distribution. The jumping distribution for ϕ is

$$\phi^{(*)}|\phi^{(t-1)} \sim N(\phi^{(t-1)}, V^2)$$

where the standard deviation V is chosen to make the jumping function efficient. In practice, V is selected based on a pilot sample to achieve acceptance rate near 0.44, the optimal rate suggested by Gelman et al. (2003).

References

Boender, C. G. E. and Rinnooy Kan, A. H. G. (1987). "A multinomial Bayesian approach to the estimation of population and vocabulary size." *Biometrika*, 74(4): 849–856. 764, 765, 766, 769, 781, 782, 783, 784

Bunge, J. and Fitzpatrick, M. (1993). "Estimating the number of species: a review." *Journal of The American Statistical Association*, 88: 364–373. 763

Cao, Y., Larsen, D. P., and Thorne, R. S.-J. (2001). "Rare species in multivariate analysis for bioassessment: some considerations." *Journal of North American Benthological Soc.*, 21: 144–153. 764

Corbet, A. S. (1942). "The distribution of butterflies in the Malay Peninsula." *Proc. Royal Entomological Society of London (A)*, 16: 101–116. 763

Efron, B. and Thisted, R. (1976). "Estimating the number of unseen species: How many words did Shakespeare know?" *Biometrika*, 63: 435–448. 763, 764, 775

Fisher, R. A., Corbet, A. S., and Williams, C. B. (1943). "The relation between the number of species and the number of individuals in a random sample of an animal population." *Journal of Animal Ecology*, 12: 42–58. 763, 764

Gelman, A., Carlin, J. B., Stern, H. S., and B., R. D. (2003). *Bayesian Data Analysis*. Chapman & Hall/CRC. 770, 789

Gelman, A. and Rubin, B. D. (1992a). "Inference from iterative simulation using multiple sequences." *Statistical Science*, 7: 457–511. 770

Gelman, A. and Rubin, D. B. (1992b). "A single series from the Gibbs sampler provides a false sense of security." In Bernardo, J. M., Berger, J. O., Dawid, A. P., and Smith, A. F. M. (eds.), *Bayesian Statistics 4. Proceedings of the Fourth Valencia International Meeting*, 625–631. Clarendon Press [Oxford University Press]. 770

Good, I. J. (1965). *The Estimation of Probabilities; An Essay on Modern Bayesian Methods*. Cambridge, Mass.: MIT Press. 769

Good, I. J. and Toulmin, G. H. (1956). "The number of new species, and the increase in population coverage, when a sample is increased." *Biometrika*, 43: 45–63. 763

Lijoi, A., Mena, R. H., and Prünster, I. (2007). "Bayesian nonparametric estimation of the probability of discovering new species." *Biometrika*, 94: 769–786. 764

Morris, J. S., Baggerly, K. A., and Coombes, K. R. (2003). "Bayesian shrinkage estimation of the relative abundance of MRNA transcripts using SAGE." *Biometrics*, 59(3): 476–486. 764, 768

Pitman, J. (1996). "Some Developments of the Blackwell-MacQueen Urn Scheme." In Ferguson, T. S., Shapley, L. S., and MacQueen, J. B. (eds.), *Statistics, Probability and Game Theory (IMS Lecture Notes Monograph Series, Vol. 30)*, 245–267. Institute of Mathematical Statistics. 763

Sethuraman, J. (1994). "A constructive definition of Dirichlet priors." *Statistica Sinica*, 4: 639–650. 764, 766

Tiwari, R. C. and Tripathi, R. C. (1989). "Nonparametric Bayes estimation of the probability of discovering a new species." *Communications in Statistics: Theory and Methods*, 18: 877–895. 764

Zhang, H. (2007). "Inferences on the number of unseen species and the number of abundant/rare species." *Journal of Applied Statistics*, 34(6): 725–740. 764

Zhang, H. and Stern, H. (2005). "Investigation of a generalized multinomial model for species data." *Journal of Statistical Computing and Simulation*, 75: 347–362. 764, 766, 769, 781, 782, 783, 784