Brazilian Journal of Biometrics, **44**, 1–12 (2026)
doi:10.28951/bjb.v44i1.889

BRAZILIAN JOURNAL OF
BIOMΣTRICS
ISSN:2764-5290

ARTICLE

Estimation of the number of species using Poisson-Mixed model: a Bayesian approach

Sandeep Kumar,[1] Manoj Kumar,[*,2] Anurag Pathak,[3] Satya Prakash Mishra,[4] and Sanjay Kumar Singh[5]

[1]Department of Community Medicine, Autonomous State Medical College, Kaushambi, India.
[2]Department of Statistics, University of Delhi, Delhi, India
[3]Assistant Research officer, Community Health Centre, Payagpur, India
[4]Department of Statistics, Central University of Haryana, Mahendergarh, India
[5]Department of Statistics, Banaras Hindu University, Varanasi, India
[*]Corresponding author. Email: manustats@gmail.com

(Received: May 07, 2025; Revised: June 25, 2025; Accepted: August 18, 2025; Published: March 16, 2026)

Section Editor: Camilla Marques Barroso

Abstract

This study presents an innovative method for estimating the number of species by using Poisson-Xgamma distribution. Classical and Bayesian estimation methods are applied to determine the number of species. The Jeffrey's and Reference prior has been proposed for estimating the number of species under Bayesian framework. The proposed Bayes estimators via Jeffrey's and reference prior have been compared through simulated risks (RMSE). The applicability of the proposed work have been validated through Mount Kenya's insect species dataset.

Keywords: Poisson-Xgamma distribution; Jeffrey's Prior; Reference prior; Profile likelihood; Conditional likelihood.

1. Introduction

In ecological field estimating the number of unseen species with in a population is a challenging task. It offers valuable insights into ecosystem processes and supports effective decision-making for conservation and resource management. Statistical modeling is instrumental in this endeavor, particularly for analyzing count data in ecological and biological research. As we know in statistics the Poisson distribution is a popular choice for analyzing count data due to its simplicity and adaptability. But, their application is limited for the cases of over-dispersion, i.e. variance > mean. To overcome this drawback, authors have taken alternative models, as Poisson-Xgamma distribution. Hence, author has consider Poisson-Xgamma distribution for further study.

Mathematically, we assume that, there are $\mathbb{S}$ species or classes in the ecological population, with their sizes denoted as $\$_1, \$_2, ..., \$_{\mathbb{S}}$. If $\chi_\imath$ represent the number of individuals of the $\imath^{th}$ species observed during the time period $[0, t]$, where $\chi_\imath | \Upsilon \sim$ Poisson $(\Upsilon_\imath)$, $\imath = 1, 2, ..., \mathbb{S}$ and $\chi_\imath' s$ are independent. This means that each species assigns a Poisson-distributed random number of individuals to the sample. We account for variations in species abundance by modeling the mean number of individuals contributing to the sample as a random variable, with $\Upsilon_\imath | \phi \sim \mathcal{F}(\Upsilon | \phi)$ which captures heterogeneity in the population. While, we observe the number of individuals contributed to the sample by each species, when their contribution is greater than 0. Species that contribute zero individuals to the sample are unobserved. Since, we do not observe when $\chi_\imath = 0$, we define the observed data as $\mathbb{S}_\jmath = \sum_{\imath=1}^{\mathbb{S}} I(\chi_\imath = \jmath)$ for $\jmath \geqslant 1$. Thus, we say that $\mathbb{S}_\jmath$ represents the number of species that contribute $\jmath$ individuals to the sample. The total observed number of species is $\omega = \sum_{J \geqslant 1} \eta_J$, and the total observed number of individuals $\eta = \sum_{J \geqslant 1} J\eta_J$, where η_J are the observed values of $\mathbb{S}_J$.

Let Υ be an abundance parameter with an Xgamma distribution (Sen *et al.,* 2016) as the abundance model. The probability density function (pdf) of the Xgamma distribution is given as:

$$f(\Upsilon|\phi) = \frac{\phi^2}{\phi + 1}(1 + \frac{\phi}{2}\Upsilon^2)e^{-\Upsilon\phi}; \qquad \Upsilon > 0; \phi > 0, \tag{1.1}$$

where $\phi > 0$ is the shape parameter. Now by compounding the Poisson distribution with the Xgamma distribution, we get $\mathcal{F}$-mixed Poisson distribution or the Poisson-Xgamma (PXG) distribution is given as:

$$\gamma_\phi(\chi) = \left(\frac{\phi}{\phi + 1}\right)^2 \left(\frac{2(\phi + 1)^2 + \phi(\chi + 1)(\chi + 2)}{2(\phi + 1)^{\chi + 2}}\right); \qquad \phi > 0; \chi = 0, 1, 2, 3, ..., \tag{1.2}$$

when, $\chi = 0$ then equation (1.2) become

$$\gamma_\phi(0) = \left(\frac{\phi}{1 + \phi}\right)^2 \left(\frac{(\phi + 1)^2 + \phi}{(\phi + 1)^2}\right); \qquad \phi > 0. \tag{1.3}$$

For estimation methods, classical and Bayesian inference are the mainstream approaches to estimate the unknown parameter from PXG distribution. Classical parameter estimation is addressed through profile and conditional likelihood methods, while in Bayesian estimation authors introduced Bernardo's reference prior Bernardo, 1979 and Jeffreys' prior Jeffreys, 1946 to estimate the number of species. The Markov Chain Monte Carlo (MCMC) technique has been used for Bayesian estimation. The effectiveness of the proposed estimators have been assessed through simulated risks (over the sample space). The application of the model have been shown through Mount Kenya's insect species data set.

The estimation of species numbers has an emergent field of research in biology and ecology. In nineteenth century, a prominent research was proposed by Greenwood & Yule, 1920, who applied parametric methods for estimating species. Within this parametric framework, various researchers have further explored species estimation under the classical paradigm, with notable contributions from Wilson & Collins, 1992, Colwell & Coddington, 1994, and Bunge & Fitzpatrick, 1993. Fisher *et al.,* 1943 introduced the logarithmic series distribution to model species frequency and evaluated the suitability of different models. They suggested that, for a given ecological population in a well-defined area, the number of individuals of a species would follow Poisson's probability law over equal time intervals. Building on this idea, Bulmer, 1974 utilized the Poisson log-normal distribution for modeling species abundance, while Sichel, 1986 and Ord & Whitmore, 1986 applied the Poisson-inverse Gaussian distribution for the same purpose. Later, Bunge & Fitzpatrick, 1993 provided a review of the literature on methods for estimating species numbers. Several years ago, the estimation of species numbers using the Poisson mixed model within the Bayesian framework was explored by multiple authors, including Leite *et al.,* 2000, Hong *et al.,* 2006, Behnke *et al.,* 2006, Barger & Bunge, 2008, and Barger, Bunge, *et al.,* 2010, among others. Meanwhile, Rodrigues *et al.,* 2001 introduced fully hierarchical and empirical Bayesian estimation methods for species numbers, specifically focusing on the Poisson-gamma mixed model. Further, count data-based model was introduced by Para *et al.,* 2020. Recently, Kumar *et al.,* 2023 discusses the predator-prey interaction using non-informative prior, and Pathak *et al.,* 2024 studied the estimation of number of species using Poisson-Lindley abundance model.

We are motivated from the above work, to investigate the number of species by using PXG distribution because it provides a more flexible approach for analyzing species abundance data. This distribution can accommodate various types of dispersion in count data, including those with a univariate model and right-skewed patterns. To the best of our knowledge, after a long gap this is the first investigation of the number of species using PXG distribution through Jeffrey and reference prior. Hence, we have proposed the Bayes estimator of the unknown parameters of the number of species via different priors for proposed PXG distribution.

The remaining section of this work are arranged as follows. In Sections 2 and 3, we apply Classical and Bayesian approach to estimate the parameters. Subsection 3.1 focus on the development of objective priors, which represent Jeffrey's and Bernardo's reference priors mathematically. Section 4 presents a detailed simulation study. Section 5 demonstrates the applicability of the model using insect species dataset. Finally, Section 6 provides a summary of the findings and conclusions.

2. The likelihood

Let us suppose that species abundance model parameters involves to specifying the underlying distribution of species counts in a population and estimating its parameters through statistical techniques. The likelihood function can be written as follows:

$$L(\mathbb{S}, \phi|data) = \sum_{\chi \in \Omega} \prod_{\imath=1}^{\mathbb{S}} \gamma_\theta(\chi_\imath),$$

where, $\mathbb{S}$ and $\phi = (\phi_1, \phi_2 \ldots, \phi_m)$ represents the number of species and nuisance parameter respectively. While, Ω denotes the set of $(\chi_1, ..., \chi_{\mathbb{S}})$ corresponding to the observed frequencies $(\eta_1, \eta_2, ..., \eta_{\mathbb{S}})$. Using Sanathanan, 1972, the likelihood becomes:

$$L(\mathbb{S}, \phi|\chi) = \binom{\mathbb{S}}{\omega}(1-\gamma_\phi(0))^{\omega}(\gamma_\phi(0))^{\mathbb{S}-\omega}\frac{\omega!}{\prod_{j\geqslant 1}\eta_j!}\prod_{j\geqslant 1}\left(\frac{\gamma_\phi(j)}{1-\gamma_\phi(0)}\right)^{\eta_j}. \tag{2.1}$$

Now substituting the values of $\gamma_\phi(0)$ and $\gamma_\phi(j)$ from above equation (1.3) into equation (2.1) we have

$$\begin{aligned} L(\mathbb{S}, \phi|\chi) = \binom{\mathbb{S}}{\omega}\left(1-\left(\frac{\phi}{1+\phi}\right)^2\left(\frac{\phi+(\phi+1)^2}{(\phi+1)^2}\right)\right)^{\omega}\left(\left(\frac{\phi}{1+\phi}\right)^2\left(\frac{\phi+(\phi+1)^2}{\phi+1}\right)\right)^{\mathbb{S}-\omega} \\ \frac{\omega!}{\prod_{j\geqslant 1} n_j!}\prod_{j\geqslant 1}\left(\frac{\left(\frac{\phi}{1+\phi}\right)^2\left(\frac{2(\phi+1)^2+\phi(j+1)(j+2)}{2(\phi+1)^{j+2}}\right)}{1-\left(\frac{\phi}{\phi+1}\right)^2\left(\frac{\phi+(\phi+1)^2}{(\phi+1)^2}\right)}\right)^{\eta_j} \end{aligned} \tag{2.2}$$

$$= \Omega(\mathbb{S}, \phi)\Psi(\phi), \tag{2.3}$$

Where $\mathbb{S} \geq \omega$, meaning $\mathbb{S} - \omega = \eta_0$ represents the number of unobserved species. Now, the likelihood function for the parameters $\mathbb{S}$ and ϕ is given in the equation (2.3).

2.1 The ML estimation

Hence, the above likelihood function can be decomposed into a binomial likelihood for ω that corresponding to $\Omega(\mathbb{S}, \phi)$, and a multinomial likelihood for the observed frequencies corresponding to $\Psi(\phi)$. Now, using the techniques of linear difference score given by Sanathanan, 1972, to estimate the discrete parameter $\mathbb{S}$, which has defined as follows:

$$\Phi(\mathbb{S}) = \frac{L(\mathbb{S}) - L(\mathbb{S}-1)}{L(\mathbb{S})},$$

where $L(\mathbb{S})$ represents the likelihood for the discrete parameter $\mathbb{S}$. When $\Phi(\mathbb{S})$ satisfies the form $\Phi(\mathbb{S}) = (\chi_i - \delta_{\mathbb{S}})/\nu_{\mathbb{S}}$, where $\delta_{\mathbb{S}}$ and $\nu_{\mathbb{S}}$ are function of $\mathbb{S}$ and χ_i which shows random data. However, $\mathrm{var}(\Phi(\mathbb{S}))$ quantifies the information about $\mathbb{S}$. Thus, we have obtained the information for $\mathbb{S}$ and ϕ (Lindsay & Roeder, 1987) as given below

$$\Upsilon(\mathbb{S}, \phi) = \begin{pmatrix} \frac{1}{\mathbb{S}}\frac{1-\gamma_\phi(0)}{\gamma_\phi(0)} & \left(-\frac{\partial}{\partial\phi} log\gamma_\phi(0)\right)^T \\ -\frac{\partial}{\partial\phi} log\gamma_\phi(0) & \mathbb{S}(-\rho(\phi)) \end{pmatrix}.$$

Here, $\frac{\partial}{\partial\phi}\gamma_\phi(0)$ represents the column vector of partial derivatives, and $\rho(\phi) = \left[E_\chi \frac{\partial^2}{\partial\eta^2} log\gamma_\phi(\chi)\right]$ where the expectation is taken with respect to γ_ϕ. The diagonal elements of this partitioned matrix factor into a product of a function of $\mathbb{S}$ and a function of the nuisance parameter ϕ. Consequently, the Fisher information matrix may be expressed as follows:

$$\Upsilon(\mathbb{S}, \phi) = \begin{pmatrix} \frac{1}{\mathbb{S}}\frac{(1+\phi)^4-\phi^2(\phi+(\phi+1)^2)}{\phi^2(\phi+(\phi+1)^2)} & -\left(\frac{5\phi^2+3\phi+2}{\phi(\phi+1)(\phi+(\phi+1)^2)}\right)^T \\ -\left(\frac{5\phi^2+3\phi+2}{\phi(\phi+1)(\phi+(\phi+1)^2)}\right) & \mathbb{S}\left(\frac{2}{\phi^2} - \frac{5\phi^2+4\phi+3}{\phi(\phi+1)^3} + \psi(j, \phi)\right) \end{pmatrix}, \tag{2.4}$$

where, $\psi(j, \phi) = \sum_{j=0}^{\infty}\left(\frac{\phi}{\phi+1}\right)^2 \frac{1}{2(1+\phi)^{j+2}}\left(\frac{(4(1+\phi)+(j+1)(j+2))^2-4\left(2(1+\phi)^2+\phi(j+1)(j+2)\right)}{2(1+\phi)^2+\phi(j+1)(j+2)}\right)$.

3. Bayesian estimation

This section of the paper focuses on finding the Bayes estimations for the unknown parameter $\mathbb{S}$. Here, $\mathbb{S}$ and ϕ be treated as random variables with their respective prior distributions. Uncertainty about prior knowledge is taken into account when considering non-informative prior i.e. Jeffrey's and Bernardo reference prior.

3.1 Prior and posterior

The prior distribution is a crucial component in Bayesian estimation, as it represents what is currently known about the unknown parameters. It is evident that the unknown parameter $\mathbb{S}$ do not have any conjugate prior. In addition we proposed Jeffery's and Bernardo reference prior as given below subsections.

Jeffrey's prior

Jeffrey's prior is defined as being proportional to the square root of the Fisher information matrix (see Jeffreys *et al.,* 1939 and Jeffreys, 1946). For multidimensional models, the determinant of the Fisher information matrix is used to maintain the invariance property. By using the determinant of the partitioned matrix in equation (2.4), we obtain Jeffrey's prior as

$$\begin{aligned} g_J(\mathbb{S},\phi) &\propto det[\Upsilon(\mathbb{S},\phi)]^{1/2}, \\ &\propto \mathbb{S}^{\frac{m-1}{2}} g(\phi), \end{aligned} \tag{3.1}$$

$$g_J(\mathbb{S},\phi) \propto \left(\left(\frac{(\phi+1)^4-\phi^2\left((\phi+1)^2+\phi\right)}{\phi^2\left((\phi+1)^2+\phi\right)}\right)\left(\frac{2}{\phi^2}-\frac{5\phi^2+4\phi+3}{\phi(\phi+1)^3}+\psi(\jmath,\phi)\right)-\left(\frac{5\phi^2+3\phi+2}{\phi(\phi+1)(\phi+(\phi+1)^2)}\right)^2\right)^{\frac{1}{2}}. \tag{3.2}$$

Here, $g(\phi)$ represents a function of ϕ, which is influenced by the dimension of the information matrix. As the dimension grows, the complexity of this function increases. The posterior distribution of the parameters based on Jeffreys is given below.

Posterior

Now, combining the likelihood in equation (2.3) with Jeffrey's prior in equation (3.2), we have the joint posterior distribution of $\pi_J(\mathbb{S},\phi|\chi)$ as follows:

$$\pi_J(\mathbb{S},\phi|\chi) \propto L(\mathbb{S},\phi|\chi) g_J(\mathbb{S},\phi)$$

$$\begin{aligned} \pi_J(\mathbb{S},\phi|\chi) \propto \binom{\mathbb{S}}{\omega}\left(1-\left(\frac{\phi}{1+\phi}\right)^2\left(\frac{\phi+(\phi+1)^2}{(\phi+1)^2}\right)\right)^{\omega}\left(\left(\frac{\phi}{1+\phi}\right)^2\left(\frac{\phi+(\phi+1)^2}{(\phi+1)^2}\right)\right)^{\mathbb{S}-\omega} \\ \frac{\omega!}{\prod_{\jmath\geqslant 1}\eta_\jmath!}\prod_{\jmath\geqslant 1}\left(\frac{\left(\frac{\phi}{1+\phi}\right)^2\left(\frac{2(\phi+1)^2+\phi(\jmath+1)(\jmath+2)}{2(\theta+1)^{\jmath+2}}\right)}{1-\left(\frac{\phi}{1+\phi}\right)^2\left(\frac{\phi+(\phi+1)^2}{(\phi+1)^2}\right)}\right)^{\eta_\jmath} \\ \left(\left(\frac{(\phi+1)^4-\phi^2(\phi+1)^2+\phi}{\phi^2(\phi+1)^2+\phi}\right)\left(\frac{2}{\phi^2}-\frac{5\phi^2+4\phi+3}{\phi(\phi+1)^3}+\psi(\jmath,\phi)\right)-\left(\frac{5\phi^2+3\phi+2}{\phi(\phi+1)((\phi+1)^2)}\right)^2\right)^{\frac{1}{2}}. \end{aligned} \tag{3.3}$$

The full conditional posterior for $\mathbb{S}$ is as follows:

$$\begin{aligned} \pi_J(\mathbb{S}|\phi,\chi) &\propto \binom{\mathbb{S}}{\omega}\left(1-\left(\frac{\phi}{1+\phi}\right)^2\left(\frac{(\phi+1)^2+\phi}{(\phi+1)^2}\right)\right)^{\omega+1}\left(\left(\frac{\phi}{\phi+1}\right)^2\left(\frac{\phi+(\phi+1)^2}{(\phi+1)^2}\right)\right)^{\mathbb{S}-\omega}, \\ &\propto B\left(\omega+1,\left(1-\left(\frac{\phi}{1+\phi}\right)^2\left(\frac{\phi+(\phi+1)^2}{(\phi+1)^2}\right)\right)\right). \end{aligned} \tag{3.4}$$

The full conditional posterior for ϕ is as follows:

$$\begin{aligned} \pi_J(\phi|\mathbb{S},\chi) \propto \frac{\omega!}{\prod_{\jmath\geqslant 1}\eta_\jmath!}\prod_{\jmath\geqslant 1}\left(\left(\frac{\phi}{1+\phi}\right)^2\left(\frac{2(\phi+1)^2+\phi(\jmath+1)(\jmath+2)}{2(\phi+1)^{\jmath+2}}\right)\right)^{\eta_\jmath}\left(\frac{1}{1-\left(\frac{\phi}{1+\phi}\right)^2\left(\frac{(\phi)^2+(\phi+1)^2}{(\phi+1)^2}\right)}\right)^{\omega+1} \\ \left(\left(\frac{(\phi+1)^4-\phi^2(\phi+(\phi+1)^2)}{\phi^2(\phi+1)^2}\right)\left(\frac{2}{\phi^2}-\frac{5\phi^2+4\phi+3}{\phi(\phi+1)^3}+\psi(\jmath,\phi)\right)-\left(\frac{5\phi^2+3\phi+2}{\phi(\phi+1)(\phi+(\phi+1)^2)}\right)^2\right)^{\frac{1}{2}}. \end{aligned} \tag{3.5}$$

The full conditional posterior for $\mathbb{S}$ can be directly sampled from a negative binomial distribution with size $(\omega+1)$ and probability $\left(1-\left(\frac{\phi}{1+\phi}\right)^2\left(\frac{\phi+(\phi+1)^2}{(\phi+1)^2}\right)\right)$. As for the full conditional posterior for ϕ, it does not have a closed-form expression, so the Metropolis-Hastings (M-H) algorithm is applied with a normal proposal distribution to obtain posterior samples.

Bernardo's reference prior

Another non-informative prior is Bernardo's reference prior (Bernardo, 1979), which is based on maximizing expected entropy as discussed in Bernardo & Ramon, 1998. Thus, we derive Bernardo's reference prior for the case when the proposed model includes one nuisance parameter, i.e., $m = 1$. Therefore, the Fisher information matrix in equation (2.4) will be 2×2. Since, we observe that the elements of the information matrix each factor into a function of $\mathbb{S}$ and ϕ. Let $\xi = \Upsilon^{-1}$ represent the covariance matrix. It is important to note that we have the factorizations $(\xi_{11})^{-1/2} = \alpha_0(\mathbb{S})\beta_0(\phi)$ and $(\upsilon_{22})^{1/2} = \alpha_1(\mathbb{S})\beta_1(\phi)$, which are the elements of the covariance and information matrices, respectively. While the nuisance parameter space, Θ, is independent of the value of $\mathbb{S}$, Bernardo's reference prior is given by:

$$\begin{aligned} g_R(\mathbb{S}, \phi) &\propto (\alpha_0(\mathbb{S}))^{-1/2}(\beta_1(\phi))^{1/2} \\ &\propto \mathbb{S}^{-1/2}(\rho(\phi))^{1/2}. \end{aligned} \tag{3.6}$$

After simplification, Bernardo's reference prior is given as:

$$g_R(\mathbb{S}, \phi) \propto \mathbb{S}^{-1/2}\left(-\frac{2}{\phi^2} + \frac{5\phi^2 + 4\phi + 3\phi}{\phi(\phi + 1)^3} - \psi(j, \phi)\right)^{1/2}. \tag{3.7}$$

Posterior

The likelihood and Bernardo's reference prior together form the joint posterior distribution of $\pi_R(\mathbb{S}, \phi)$, which is expressed as:

$$\pi_R(\mathbb{S}, \phi|\chi) \propto L(\mathbb{S}, \phi|data)g_R(\mathbb{S}, \phi),$$

$$\begin{aligned} \pi_R(\mathbb{S}, \phi|\chi) \propto \binom{\mathbb{S}}{\omega}\left(1 - \left(\frac{\phi}{1+\phi}\right)^2\left(\frac{\phi + (\phi+1)^2}{(\phi+1)^2}\right)\right)^{\omega}\left(\left(\frac{\phi}{1+\phi}\right)^2\left(\frac{\phi + (\phi+1)^2}{(\phi+1)^2}\right)\right)^{\mathbb{S}-\omega} \\ \frac{\omega!}{\prod_{j\geqslant 1}\eta_j!}\prod_{j\geqslant 1}\left(\frac{\left(\frac{\phi}{1+\phi}\right)^2\left(\frac{2(\phi+1)^2+\phi(j+1)(j+2)}{2(\phi+1)^{j+2}}\right)}{1 - \left(\frac{\phi}{1+\phi}\right)^2\left(\frac{\phi^2+(\phi+1)^2}{(\phi+1)^2}\right)}\right)^{\eta_j} \\ \mathbb{S}^{-1/2}\left(-\frac{2}{\phi^2} + \frac{5\phi^2 + 4\phi + 3}{\phi(\phi + 1)^3} - \psi(j, \phi)\right)^{1/2}. \end{aligned}$$

The full conditional posterior for $\mathbb{S}$ and ϕ are given by below equations (3.8) and (3.9) respectively,

$$\pi_R(\mathbb{S}|\phi, \chi) \propto \mathbb{S}^{-1/2}\binom{\mathbb{S}}{\omega}\left(1 - \left(\frac{\phi}{1+\phi}\right)^2\left(\frac{\phi + (\phi+1)^2}{(\phi+1)^2}\right)\right)^{\omega}\left(\left(\frac{\phi}{1+\phi}\right)^2\left(\frac{\phi + (\phi+1)^2}{(\phi+1)^2}\right)\right)^{\mathbb{S}-\omega}, \tag{3.8}$$

$$\pi_R(\phi|\mathbb{S}, \chi) \propto \frac{\omega!}{\prod_{j\geqslant 1}\eta_j!}\prod_{j\geqslant 1}\left(\frac{\left(\frac{\phi}{1+\phi}\right)^2\left(\frac{2(\phi+1)^2+\phi(j+1)(j+2)}{2(\phi+1)^{j+2}}\right)}{1 - \left(\frac{\phi}{1+\phi}\right)^2\left(\frac{\phi+(\phi)^2}{(\phi+1)^2}\right)}\right)^{\eta_j}\left(-\frac{2}{\phi^2} + \frac{5\phi^2 + 4\phi + 3}{\phi(\phi + 1)^3} - \psi(j, \phi)\right)^{1/2}. \tag{3.9}$$

The full conditional posterior for $\mathbb{S}$ and ϕ given above does not have a closed-form solution. Therefore, the M-H algorithm is applied to obtain the posterior samples of $\mathbb{S}$ and ϕ. In M-H algorithm, negative binomial and normal distribution have been used as proposal distribution for $\mathbb{S}$ and ϕ respectively.

4. Simulation Study

In a simulation study, we proposed the estimators $\hat{\mathbb{S}}_p$, $\hat{\mathbb{S}}_c$, $\hat{\mathbb{S}}_J$, and $\hat{\mathbb{S}}_R$, which represent the profile MLE, conditional MLE, Bayesian estimator with Jeffreys' prior, and Bayesian estimator with Bernardo's reference prior, respectively. Here, $\mathbb{S}$ is a discrete parameter denoting the total number of species, and ϕ is a nuisance parameter representing the species abundance from the Xgamma distribution. The variables t and ω refer to the stoppage time and the observed number of species in the experiment. The proposed estimators of the parameter $\mathbb{S}$ are compared based on the square root of the average risk (expected loss over the sample space), denoted by $R(\mathbb{S})$. Additionally, in the simulation study, the number of species were fixed at $\mathbb{S} = 50$ and $\mathbb{S} = 90$, with the corresponding counts of individual species being $j = 6$, 9, and 12, as shown in tables (1) and (2), respectively. The abundance parameters were generated from the Xgamma distributions with values $\phi = 1.5$ and $\phi = 2$, as reported in tables (1) and (2). Two stoppage times, $t = 0.8$ and $t = 1.2$, were considered in this study. From the simulated data in tables (1&2), we observed that the capture fraction ($C.F = \frac{\omega}{\mathbb{S}} \times 100$) typically ranged between 87% to 95%. It is also worth noting that, in most cases, the increments in CFs were greater than 15%. For each simulated dataset, the estimators $\hat{\mathbb{S}}_p$, $\hat{\mathbb{S}}_c$, $\hat{\mathbb{S}}_J$, and $\hat{\mathbb{S}}_R$ are reported in tables (1)&2). We compare the estimators obtained under species individuals ($\jmath$) with the corresponding profile MLE, conditional MLE, Bayes estimator with Jeffrey's prior, and Bayes estimator with Bernardo's reference prior. These comparisons are made based on their root mean square errors (RMSEs). For the nuisance parameter (ϕ), the MLE has been obtained by help of numerical method. The average estimates and average risk of parameter estimates for the 1000 generated datasets are presented in tables (1&2). The number of individual species ($\jmath$) were used to calculate estimates. From tables (1&2), we noted that for various fixed values of individuals $\jmath$ and CF, as the number of species individuals increased, the CF also increased, and the risk of the estimators gradually decreased. We have also obtained the 95% confidence interval/HPD intervals for proposed estimators. Finally, we noticed that the number of individuals increases, the CF also increases, bringing the observed number of species closer to the estimated number (as these species tend to overestimate the parameter most of the times).

Table 1. The estimators $\hat{\mathbb{S}}$ and their RMSE obtained from simulated samples for the parameter $\phi = 1.5$ with $t = 0.8$ and $t = 1.2$.

				$\phi = 1.5$, $t = 0.8$				$\phi = 1.5$, $t = 1.2$		
$\mathbb{S}$	J	CF		$\hat{\mathbb{S}}$	R($\mathbb{S}$)	CI/HPD	CF	$\hat{\mathbb{S}}$	R($\mathbb{S}$)	CI/HPD
50	6	87.73	$\hat{\mathbb{S}}_p$	55.211	6.087	(54.690*, 55.733*)	94.18	57.164	7.460	(56.602* , 57.727*)
			$\hat{\mathbb{S}}_c$	55.156	6.037	(54.634*, 55.678*)		57.413	7.697	(56.859*, 57.967*)
			$\hat{\mathbb{S}}_J$	55.491	6.327	(48.160, 62.420)		57.690	7.964	(50.800, 64.185)
			$\hat{\mathbb{S}}_R$	55.710	6.520	(48.370, 62.690)		57.874	8.143	(50.915, 64.380)
	9	90.93	$\hat{\mathbb{S}}_p$	55.735	6.394	(55.180*, 56.290*)	96.62	56.602	6.856	(55.970*, 57.235*)
			$\hat{\mathbb{S}}_c$	55.909	6.552	(55.360*, 56.458*)		57.016	7.255	(56.402*, 57.630*)
			$\hat{\mathbb{S}}_J$	56.203	6.818	(49.245, 62.760)		57.231	7.464	(51.055, 63.065)
			$\hat{\mathbb{S}}_R$	56.391	6.989	(49.390, 62.940)		57.402	7.629	(51.095, 63.235)
	12	91.17	$\hat{\mathbb{S}}_p$	55.493	6.028	(54.926*, 56.060*)	96.86	55.952	6.170	(55.283*, 56.621*)
			$\hat{\mathbb{S}}_c$	55.715	6.232	(55.156*, 56.274*)		56.394	6.598	(55.746*, 57.041*)
			$\hat{\mathbb{S}}_J$	55.998	6.492	(49.130, 62.415)		56.590	6.791	(50.760, 62.135)
			$\hat{\mathbb{S}}_R$	56.190	6.671	(49.230, 62.620)		56.739	6.937	(50.895, 62.295)
90	6	87.51	$\hat{\mathbb{S}}_p$	99.440	10.500	(99.056*,99.825*)	93.68	102.561	12.952	(102.144*, 102.978*)
			$\hat{\mathbb{S}}_c$	99.312	10.386	(98.926*, 99.697*)		103.001	13.381	(102.59*, 103.412*)
			$\hat{\mathbb{S}}_J$	99.654	10.693	(89.575, 109.040)		103.281	13.654	(93.845, 112.060)
			$\hat{\mathbb{S}}_R$	99.872	10.893	(89.780, 109.255)		103.484	13.851	(94.005, 112.260)
	9	89.21	$\hat{\mathbb{S}}_p$	99.649	10.554	(99.248*, 100.050*)	96.26	101.938	12.149	(101.475*, 102.401*)
			$\hat{\mathbb{S}}_c$	99.799	10.705	(99.401*, 100.198*)		102.673	12.872	(102.222*, 103.123*)
			$\hat{\mathbb{S}}_J$	100.111	10.988	(90.275,109.140)		102.898	13.094	(94.235,110.800)
			$\hat{\mathbb{S}}_R$	100.321	11.180	(90.540,109.390)		103.078	13.275	(94.550,111.110)
	12	90.09	$\hat{\mathbb{S}}_p$	100.061	10.841	(99.655*, 100.468*)	97.04	101.641	11.822	(101.159*, 102.124*)
			$\hat{\mathbb{S}}_c$	100.289	11.064	(99.886*, 100.692*)		102.429	12.598	(101.961*, 102.896*)
			$\hat{\mathbb{S}}_J$	100.594	11.347	(90.840, 109.510)		102.635	12.802	(94.325, 110.270)
			$\hat{\mathbb{S}}_R$	100.803	11.540	(91.150, 109.775)		102.796	12.962	(94.580, 110.530)

*CI- 95% Confidence Interval (CI) under classical approach.

Table 2. The estimators $\hat{\mathbb{S}}$ and their RMSE obtained from simulated samples for the parameter $\phi = 2$ with $t = 0.8$ and $t = 1.2$.

				$\phi = 2$, $t = 0.8$				$\phi = 2$, $t = 1.2$		
$\mathbb{S}$	j	CF		$\hat{\mathbb{S}}$	R($\mathbb{S}$)	CI/HPD	CF	$\hat{\mathbb{S}}$	R($\mathbb{S}$)	CI/HPD
50	6	91.74	$\hat{\mathbb{S}}_p$	56.536	7.114	(55.993*, 57.079*)	95.39	57.165	7.408	(56.581*, 57.749*)
			$\hat{\mathbb{S}}_c$	56.675	7.248	(56.137*, 57.213*)		57.493	7.726	(56.921*, 58.065*)
			$\hat{\mathbb{S}}_J$	56.977	7.526	(49.840, 63.665)		57.748	7.974	(51.095, 64.055)
			$\hat{\mathbb{S}}_R$	57.174	7.710	(50.030, 63.900)		57.934	8.155	(51.205, 64.225)
	9	94.48	$\hat{\mathbb{S}}_p$	56.352	6.706	(55.755*, 56.949*)	97.95	56.250	6.407	(55.568*, 56.933*)
			$\hat{\mathbb{S}}_c$	56.695	7.033	(56.111*, 57.279*)		56.701	6.848	(56.042*, 57.360*)
			$\hat{\mathbb{S}}_J$	56.941	7.268	(50.405, 63.075)		56.888	7.033	(51.165, 62.305)
			$\hat{\mathbb{S}}_R$	57.118	7.439	(50.590, 63.320)		57.042	7.185	(51.270, 62.480)
	12	95.45	$\hat{\mathbb{S}}_p$	56.306	6.639	(55.687*, 56.926*)	98.75	55.764	5.868	(55.029*, 56.498*)
			$\hat{\mathbb{S}}_c$	56.695	7.011	(56.091*, 57.298*)		56.218	6.317	(55.511*, 56.924*)
			$\hat{\mathbb{S}}_J$	56.926	7.234	(50.565, 62.855)		56.378	6.477	(51.080, 61.435)
			$\hat{\mathbb{S}}_R$	57.093	7.392	(50.730, 63.050)		56.514	6.613	(51.165, 61.575)
90	6	91.68	$\hat{\mathbb{S}}_p$	102.008	12.555	(101.608*, 102.409*)	95.62	103.361	13.594	(102.930*, 103.792*)
			$\hat{\mathbb{S}}_c$	102.222	12.761	(101.825*, 102.620*)		103.943	14.165	(103.521*, 104.366*)
			$\hat{\mathbb{S}}_J$	102.526	13.052	(92.670, 111.565)		104.201	14.419	(94.920, 112.670)
			$\hat{\mathbb{S}}_R$	102.734	13.252	(93.005, 111.885)		104.394	14.609	(95.235, 112.920)
	9	93.93	$\hat{\mathbb{S}}_p$	101.870	12.180	(101.440*, 102.301*)	97.85	101.510	11.637	(101.010*, 102.010*)
			$\hat{\mathbb{S}}_c$	102.400	12.695	(101.978*, 102.823*)		102.326	12.444	(101.843*, 102.809*)
			$\hat{\mathbb{S}}_J$	102.665	12.954	(93.440, 111.165)		102.516	12.634	(94.530, 109.900)
			$\hat{\mathbb{S}}_R$	102.855	13.139	(93.710, 111.390)		102.669	12.788	(94.715, 110.125)
	12	94.60	$\hat{\mathbb{S}}_p$	101.600	11.958	(101.156*, 102.044*)	98.41	100.790	10.899	(100.262*, 101.318*)
			$\hat{\mathbb{S}}_c$	102.224	12.562	(101.790*, 102.658*)		101.619	11.722	(101.111*, 102.128*)
			$\hat{\mathbb{S}}_J$	102.474	12.806	(93.445, 110.710)		101.792	11.895	(94.110, 108.730)
			$\hat{\mathbb{S}}_R$	102.653	12.981	(93.675, 110.955)		101.944	12.046	(94.400, 109.045)

*CI- 95% Confidence Interval (CI) under classical approach.

5. Illustrative Example

For illustration, we have considered insect population data from Mount Kenya, observed in a specific region, as reported by Lewins & Joanes, 1984. This dataset is assumed to be samples from an F-mixed Poisson model and is well-suited for the Poisson-Xgamma distribution.

This dataset includes only species with non-zero observed frequencies, with the observed species counts categorized as follows: $(j, \eta_j) : (1, 8), (2, 3), (3, 2), (4, 1), (5, 1), (6, 3), (7, 2), (10, 1), (12, 1), (18, 1), (21, 1), (25, 1), (46, 1), (56, 1), (95, 1), (98, 1), (109, 1), (157, 1), (335, 1)$. In the complete dataset, the total number of observed species (ω) is 32, and the total number of individual organisms (η) is 1043.

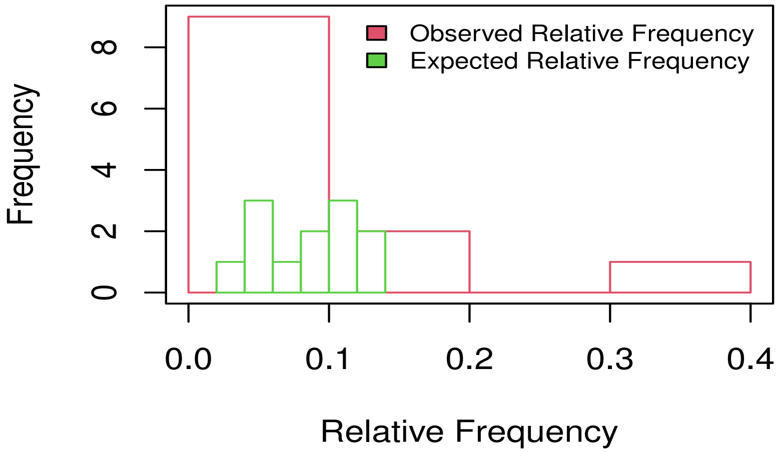

Figure 1. Histogram of fitted observed and expected relative frequency of PXG for the second data

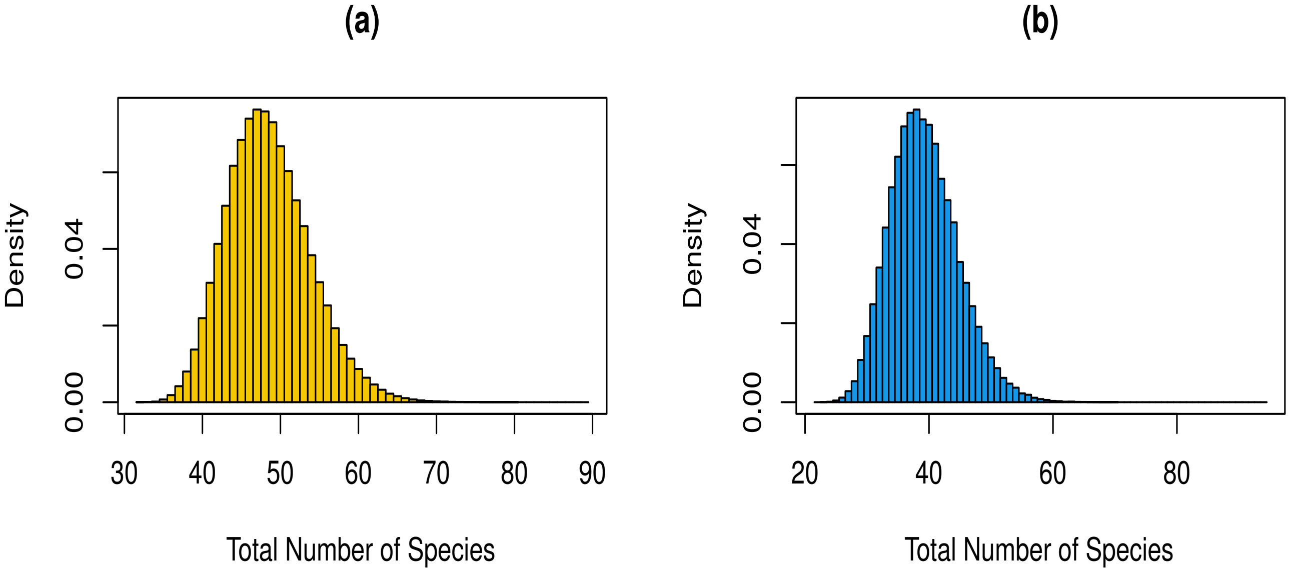

Figure 2. Posterior histogram plot of parameter $\mathbb{S}$ for PXG with (a) Jeffrey's prior and (b) Bernardo's reference prior for second dataset

To analyze the Poisson-Xgamma model, we begin by plotting graph for the dataset. The observed and predicted relative frequency histograms is displayed together in Fig. (1). Upon examining the graphical representation, the Poisson-Xgamma model appears to provide a fit, demonstrating agreement between the two histograms. The chi-square goodness-of-fit test has been performed, and we have $\chi^2_{cal} = 8.54$ and $\chi^2_{(0.01,3)} = 11.34$, ($\chi^2_{cal} < \chi^2_{tab}$). Therefore, there is no significant difference between the observed and expected (hypothesized) frequencies.

The dataset was analyzed using the Poisson-Xgamma model, with the cutoff point determined based on the χ^2 goodness-of-fit measure as outlined in Behnke *et al.,* 2006. For dataset, two distinct regions can be identified depending on the cutoff point $\jmath$, when $\jmath \leq 12$, the model fit is acceptable, while for $\jmath > 12$, the χ^2_{cal} value becomes excessively high, which is unsuitable for study. As a result, we have ideally $j = 12$ was selected for dataset to utilize the maximum amount of data. The frequencies up to 12 for dataset were used, and applied the goodness-of-fit. From dataset we observed frequencies exceeding 12, only frequencies below 12 are utilized. This approach suggests that the most abundant species belong to a distinct sub-population. The model parameters are subsequently estimated using both Bayesian and frequentist methods.

For Bayesian estimation, MCMC methods with M-H algorithm are used to simulate posterior distributions, through R software. The full conditional posterior distributions of the proposed model with Jeffreys' prior are expressed as $\pi_J(\mathbb{S}|\phi,\chi)$ and $\pi_J(\phi|\mathbb{S},\chi)$ in equations (3.4&3.5), respectively. Similarly, with Bernardo's reference prior, as given by $\pi_R(\mathbb{S}|\phi,\chi)$ and $\pi_R(\phi|\mathbb{S},\chi)$ in equations (3.8&3.9), respectively.

The posterior distributions for the proposed PXG model, based on Jeffreys' and Bernardo's reference priors, are discussed in subsections (3.1). The Gaussian distribution has been considered as a proposal density in MCMC. The posterior density plot is shown in figure 2). Using MCMC techniques to get the posterior samples for Bayes estimators (see Andrieu & Thoms, 2008). The Bayes estimators and 95% credible intervals have also been obtained. The frequentist estimates for $\mathbb{S}$ are provided in table (3), while the posterior summaries of estimates and their credible intervals are also shown in table (4). We observed from table (4) that the posterior mean, mode, and median consistently exceed the maximum likelihood estimates (MLEs) for the Poisson-Xgamma model under both Jeffreys' and Bernardo's reference priors. Additionally, the profile likelihood intervals are comparable to the credible interval estimates, and the posterior mean for $\mathbb{S}$ is higher than the MLEs. The asymptotic 95% symmetric confidence interval for the MLE of the model exceeds the observed number of species, with $\omega = 32$ for dataset.

Table 3. The MLEs of $\mathbb{S}$ derived from profile likelihood (ϕ_p) and conditional likelihood (ϕ_c), along with their corresponding confidence intervals, for the PXG model

PXG Model	MLE	95% CI	MLE	95% CI
Profile likelihood	$\mathbb{S}_p = 24.21913$	$(19.47399, 27.88602)$	$\phi_p = 0.6871$	$(0.5291, 0.8846)$
Conditional likelihood	$\mathbb{S}_c = 24.20571$	$(19.47473, 27.89691)$	$\phi_c = 0.6993$	$(0.5672, 0.8863)$

Table 4. Summary statistics of the posterior distribution $\pi(\mathbb{S}|\chi)$ for two models: PXG-J with Jeffreys' prior and PXG-R with Bernardo's reference prior

Model	Mode	Median	Mean	95% Credible Interval
PXG-J	49	51	51.24578	$(41, 64)$
PXG-R	38	49	49.28501	$(40, 61)$

We have presented the relative fit between models in figure (1). The Deviation Information Criteria (DIC) has been calculated for both models PXG-J and PXG-R. DIC is defined as:

$$D(\chi, \mathbb{S}, \phi) = -2 \log L(\mathbb{S}, \phi|\chi). \tag{5.1}$$

The estimate of the expected DIC is obtained by averaging the deviance over the posterior samples, ϕ^ι, where $\iota = 1, 2, ..., N$, with N representing the total number of posterior samples. The estimates of the deviance are:

$$\hat{D}(\chi) = \frac{1}{N} \sum_{\iota=1}^{N} D(\chi, \phi^\iota). \tag{5.2}$$

We have plotted the observed dataset alongside the expected values for the frequencies, using the mean posterior values of the parameters, as shown in figure (3). From this figure, it is evident that the Poisson-Xgamma model provides an acceptable fit for the considered dataset. The expected frequencies plotted at the mean posterior value for the PXG model with Jeffreys' and Bernardo's reference priors yield approximately similar results.

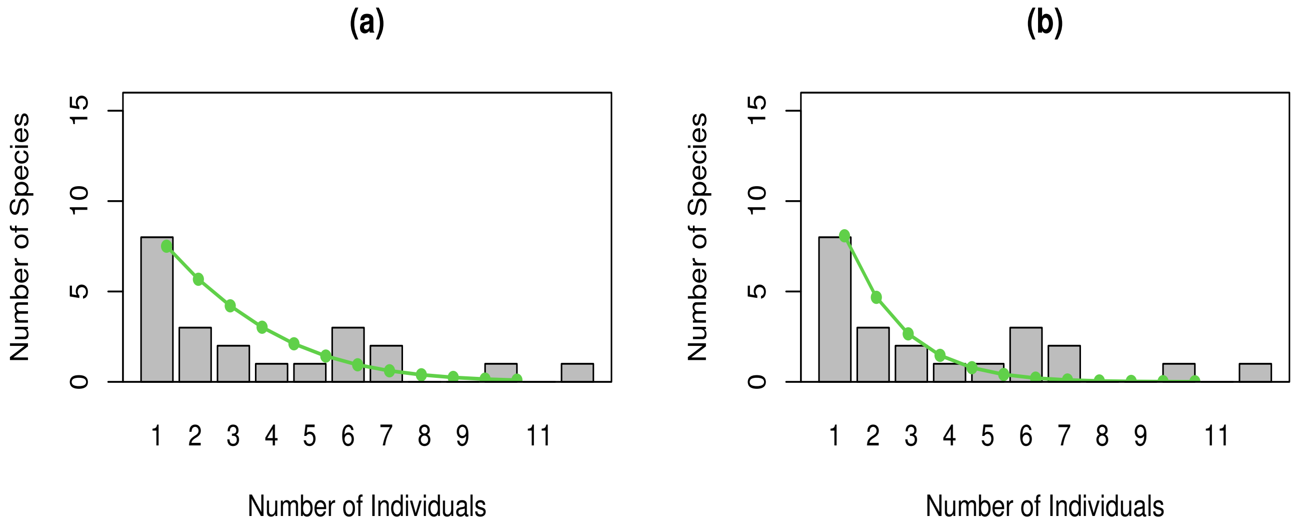

Figure 3. Expected frequency plot, evaluated at mean posterior values for PXG with (a) Jeffrey's prior and (b) Bernardo's reference prior for dataset.

The value of DIC for PXG-J and PXG-R is 276.54 and 129.88 respectively. The PXG model with Bernardo's reference prior has minimum DIC as compared to the PXG model with Jeffrey's prior. A smaller deviance indicates a better fit to the models for dataset. Thus, the choice of prior significantly influences the resulting estimates, the model selection under Bernardo's reference prior performs slightly better than Jeffreys' prior.

6. Conclusion

The purpose of the study is to proposed a suitable prior for the Bayesian estimation of the number of species using PXG model. A Jeffrey and Bernardo's reference prior has been proposed for this purpose. The profile and conditional likelihood are used for the Bayesian inference. The performance of the proposed different priors has been investigated under a simulation study. From the findings of the study it can be assessed that the proposed estimator ($\hat{\mathbb{S}}_R$) with reference prior is more suitable prior for the analysis of the parameter of the PXG model. In addition, a Mount Kenya species dataset is analysed for the purpose of illustration. It indicates that these proposed estimators are practical and feasible. This study is useful for the analysis from various ecological fields dealing with the analysis of the abundance models. In future this work can be extended for various class of priors of the different F-mixed distribution.

Acknowledgement

The authors acknowledged the efforts of the reviewers for their constructive suggestions, which improved the quality of the paper. The corresponding author (Dr. Manoj Kumar) acknowledges the Institute of Eminence, university of Delhi, 110007 for providing the grant under the faculty research programme (FRP-2024-25).

Conflicts of Interest

The corresponding author declares that no competing/conflict of interest emerged among the authors.

Author Contributions

Conceptualization: Kumar, M.; Kumar, S. **Data curation**: Kumar, M.; Kumar, S.; Pathak, A; Mishra, S.P. **Formal analysis**: Pathak,A; Kumar, S; Singh, S.K.; **Funding acquisition**: Kumar, M. **Methodology**:Kumar, M.; Kumar, S.; Pathak, A; Mishra, S.P. **Software**:Pathak,A; Kumar, S **Resources**:Kumar, M.; Mishra, S.P. **Supervision**: Kumar, M **Validation**: Singh, S.K. **Writing - original draft**: Kumar, S. **Writing - review and editing**: Kumar, M.; Kumar, S.; Pathak, A; Mishra, S.P.; Singh, S.K.

References

1. Andrieu, C. & Thoms, J. A tutorial on adaptive MCMC. *Statistics and computing* **18,** 343–373 (2008).

2. Barger, K. & Bunge, J. Bayesian estimation of the number of species using noninformative priors. *Biometrical Journal: Journal of Mathematical Methods in Biosciences* **50,** 1064–1076 (2008).
3. Barger, K., Bunge, J., *et al.* Objective Bayesian estimation for the number of species. *Bayesian Analysis* **5,** 765–785 (2010).
4. Behnke, A., Bunge, J., Barger, K., Breiner, H.-W., Alla, V. & Stoeck, T. Microeukaryote community patterns along an O2/H2S gradient in a supersulfidic anoxic fjord (Framvaren, Norway). *Applied and environmental microbiology* **72,** 3626–3636 (2006).
5. Bernardo, J. M. Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)* **41,** 113–128 (1979).
6. Bernardo, J. M. & Ramon, J. M. An introduction to Bayesian reference analysis: inference on the ratio of multinomial parameters. *Journal of the Royal Statistical Society: Series D (The Statistician)* **47,** 101–135 (1998).
7. Bulmer, M. On fitting the Poisson lognormal distribution to species-abundance data. *Biometrics,* 101–110 (1974).
8. Bunge, J. & Fitzpatrick, M. Estimating the number of species: a review. *Journal of the American Statistical Association* **88,** 364–373 (1993).
9. Colwell, R. K. & Coddington, J. A. Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences* **345,** 101–118 (1994).
10. Fisher, R. A., Corbet, A. S. & Williams, C. B. The Relation Between the Number of Species and the Number of Individuals in a Random Sample of an Animal Population. *The Journal of Animal Ecology,* 42–58 (1943).
11. Greenwood, M. & Yule, G. U. An inquiry into the nature of frequency distributions representative of multiple happenings with particular reference to the occurrence of multiple attacks of disease or of repeated accidents. *Journal of the Royal statistical society* **83,** 255–279 (1920).
12. Hong, S.-H., Bunge, J., Jeon, S.-O. & Epstein, S. S. Predicting microbial species richness. *Proceedings of the National Academy of Sciences* **103,** 117–122 (2006).
13. Jeffreys, H. An Invariant Form for the Prior Probability in Estimation Problems. *Proceedings of the Royal Society of London. Series A. Mathematical and Physical Sciences* **186,** 453–461 (1946).
14. Jeffreys, H. *et al.* Theory of Probability (1939).
15. Kumar, S., Pathak, A., Kumar, M., Singh, S. K. & Gupta, R. Bayesian inference for functional response of stochastic predator-prey model using non-informative prior. *Int. J. Agricult. Stat. Sci. Vol* **19,** 891–898 (2023).
16. Leite, J. G., Rodrigues, J. & Milan, L. A. A Bayesian Analysis for Estimating the Number of Species in a Population using Nonhomogeneous Poisson Process. *Statistics & Probability Letters* **48,** 153–161 (2000).
17. Lewins, W. A. & Joanes, D. Bayesian estimation of the number of species. *Biometrics,* 323–328 (1984).
18. Lindsay, B. G. & Roeder, K. A Unified Treatment of Integer Parameter Models. *Journal of the American Statistical Association* **82,** 758–764 (1987).
19. Ord, J. K. & Whitmore, G. A. The poisson-inverse gaussian disiribuiion as a model for species abundance. *Communications in Statistics-theory and Methods* **15,** 853–871 (1986).
20. Para, B. A., Jan, T. R. & Bakouch, H. S. Poisson Xgamma distribution: A discrete model for count data analysis. *Model Assisted Statistics and Applications* **15,** 139–151 (2020).
21. Pathak, A., Kumar, M., Singh, S. K., Singh, U. & Kumar, S. Bayesian estimation of the number of species from Poisson-Lindley stochastic abundance model using non-informative priors. *Computational Statistics* **1–26** (2024).
22. Rodrigues, J., Milan, L. A. & Leite, J. G. Hierarchical Bayesian Estimation for the Number of Species. *Biometrical Journal: Journal of Mathematical Methods in Biosciences* **43,** 737–746 (2001).
23. Sanathanan, L. Estimating the size of a multinomial population. *The Annals of Mathematical Statistics* **43,** 142–152 (1972).
24. Sen, S., Maiti, S. S. & Chandra, N. The xgamma distribution: statistical properties and application. *Journal of Modern Applied Statistical Methods* **15,** 38 (2016).
25. Sichel, H. Parameter estimation for a word frequency distribution based on occupancy theory. *Communications in Statistics-Theory and Methods* **15,** 935–949 (1986).
26. Wilson, R. M. & Collins, M. F. Capture-recapture estimation with samples of size one using frequency data. *Biometrika* **79,** 543–553 (1992).