

**Bayesian estimation : Number of species from Poisson mixed
exponential-gamma distribution using objective priors**

Sandeep Kumar

*Department of Statistics
Central University of Haryana
Mahendergarh 123031
Haryana
India*

Anurag Pathak

*Department of Community Medicine
Dr. Ram Manohar Lohia Institute of Medical Sciences
Lucknow 226010
Uttar Pradesh
India*

Manoj Kumar *

*Department of Statistics
Central University of Haryana
Mahendergarh 123031
Haryana
India*

and

*Department of Statistics
University of Delhi
New Delhi 110007
India*

Sanjay Kumar Singh

*Department of Statistics
Banaras Hindu University
Varanasi 221005
Uttar Pradesh
India*

* E-mail: manustats@gmail.com

Abstract

In this paper, we have considered the Poisson-exponential-gamma mixture model as a species problem. The classical estimation of the parameter has been proposed through profile and conditional likelihood methods. While in the Bayesian paradigm, Bernardo's reference and Jeffery's priors have been used to estimate the number of species. Although a Markov Chain Monte Carlo (MCMC) technique has been used for obtaining the Bayes estimates of the number of species parameter. The performances of the proposed estimators are compared in terms of their risks through the simulation study. The applicability, and suitability of the proposed model have been performed using Mount Kenya insect species data.

Subject Classification: 62F10, 62F15, 62F25, 62P10.

Keywords: Likelihood, Jeffrey's prior, Reference prior, MCMC, Posterior estimates.

1. Introduction

Nowadays the expansion of statistics in ecology is increasing day by day. The estimation of the number of unseen species in a single abundance is an interesting problem in ecology. But, it is a tedious task to get the number of unseen species in a habitat or parks, etc. Fisher et al. (1943) discussed the relation between the number of species and the number of individuals in a random sample of animal populations. After that many authors worked on such dimensions for estimating the number of species (see Wilson and Collins (1992), Bunge and Fitzpatrick (1993), and Colwell and Coddington (1994)). Furthermore, Fisher et al. (1943) developed the logarithmic series distribution for species frequency; Bulmer (1974) used a Poisson log-normal distribution of species-abundance data, Sichel (1986) and Ord and Whitmore (1986) work on the Poisson-inverse Gaussian distribution for their models of species abundance. Bunge and Fitzpatrick (1993) carried out a review study focused on estimating the number of species, and Leite et al. (2000) focused on the Poisson mixed as well as in-truncated distributions as per the demand of the problem.

Recently, Shanker (2016) introduced a Poisson-exponential-gamma mixed distribution and they have discussed the moments and survival properties. While Sankaran (1970) has also used the Poisson mixed F-distribution for count-based data. Which is very useful for estimating the number of species in the ecosystem. After extensive study, authors have observed that till now there is no attempt has been made to use the Poisson-exponential-gamma distribution as a model for the number of species estimation in classical as well as Bayesian paradigms. Hence, the authors wish to express a Poisson exponential-gamma mixed distribution as an estimation of the number of species in habitats/parks, etc.

A few years back many authors such as Hong et al. (2006), Behnke et al. (2006), Barger, K. and Bunge, J. (2008), and Barger and Bunge (2010) have studied the Bayesian estimation for the number of species population parameters

based on mixed Poisson models. Now, in this sequence, authors have extended another poisson based mixed model ie. Poisson exponential-gamma model. The Jeffreys (1946) and Bernardo (1979) priors have been used to construct the Bayesian estimation for the number of species. Also, the comparison of proposed estimators of the parameters under classical and Bayesian paradigms are based on the square root of average risk (expected loss over sample space) are obtained (see Kaur et al. (2021) and Rabie (2021)).

The organization of the article is as follows: In section 2, the species model is described, In sections 3 & 4, we use a classical and Bayesian estimation with Jeffrey's and Bernardo's reference priors. In section 5, a simulation study was performed. In section 6, insect species estimation data analysis is done, and In section 7, conclusions are discussed.

2. The Species Model

Let X_i be the abundance of species i where $X_i|\lambda \sim \text{Poisson}(\lambda_i), i = 1, 2, \dots, N$ and X_i 's are independent. It means each species contributes a Poisson random number of individuals to the sample. Here, the difference in the abundance of each species by describing the mean number of individuals contributing to the sample to the random, where $\lambda_i|\theta \sim F(\lambda|\theta)$ describes heterogeneity in the population. Taking $F(\lambda|\theta)$ as the exponential-gamma distribution, the marginal distribution of X_i is $p_\theta(x) = \int \frac{e^{-\lambda} \lambda^x}{x!} f(\lambda|\theta) d\lambda$ representing the F-mixed Poisson distribution, where $f(\lambda|\theta) = \frac{\partial F(\lambda|\theta)}{\partial \lambda}$. Here, we consider the exponential-gamma distribution as an abundance distribution. The probability density function (pdf) of the exponential-gamma distribution takes the form

$$f(\lambda|\theta) = \frac{\theta}{\theta + 2} (1 + \theta + \theta\lambda) e^{-\lambda\theta}; \lambda > 0; \theta > 0$$

where $\theta > 0$ is the shape parameter. As seen from the pdf of the exponential-gamma distribution, the exponential-gamma is a mixture of two distributions: the exponential distribution with the parameter θ and the gamma distribution, i.e. $\text{Gamma}(2, \theta)$. Its cumulative distribution function (cdf) is

$$F(\lambda|\theta) = 1 - \left(1 + \frac{\theta\lambda}{\theta + 2}\right) e^{-\lambda\theta}; \lambda > 0; \theta > 0$$

Now, the compounding of Poisson with exponential-gamma distribution discussed by Shanker (2017) as derived the F-mixed Poisson distribution is given below

$$p_{\theta}(x) = \left(\frac{\theta}{\theta+2}\right) \left(\frac{(\theta x + (\theta^2 + 3\theta + 1))}{(\theta+1)^{x+2}}\right); \theta > 0; x = 0, 1, 2, 3, \dots, \quad (1)$$

when $x = 0$ then equation (1) become

$$p_{\theta}(0) = \left(\frac{\theta}{\theta+2}\right) \left(\frac{(\theta^2 + 3\theta + 1)}{(\theta+1)^2}\right); \theta > 0. \quad (2)$$

We observe the number of individuals contributed to the sample by each species, but only if the contribution is greater than 0. The species that contribute zero individuals to the sample are unobserved. Since we do not observe if $X_i = 0$, let us write the observed data as $N_j = \sum_{i=1}^N I(X_i = j)$ for $j \geq 1$. Thus, N_j represents the number of species that contribute j individuals to the sample. The observed number of species is $w = \sum_{j \geq 1} n_j$, and the observed number of individuals is $n = \sum_{j \geq 1} j n_j$, where n_j are realized values of N_j .

3. Classical Estimation

The present event follows the assumption that the Poisson sample is contributed independently. We consider the likelihood to be a product of the total number of species (N) and Poisson random variables with means λ_i . Then sum across the different sets of the full data, $x_1, x_2, \dots$, that correspond to the observed data, $n_1, n_2, \dots$, using the law of total probability. The likelihood function is,

$$L(data|\theta, N) = \sum_{x \in A} \prod_{i=1}^N p_{\theta}(x_i).$$

Where A is the set of $(x_1, \dots, x_N)$ which correspond to the observed frequencies $(n_1, n_2, \dots)$. Sanathanan (1972) has demonstrated that the likelihood can be written as

$$\begin{aligned} L(data|\theta, N) &= \binom{N}{w} (1 - p_{\theta}(0))^w (p_{\theta}(0))^{N-w} \frac{w!}{\prod_{j \geq 1} n_j!} \prod_{j \geq 1} \left(\frac{p_{\theta}(j)}{1 - p_{\theta}(0)}\right)^{n_j}, \\ L(data|\theta, N) &= \binom{N}{w} \left(-\left(\frac{\theta}{\theta+2}\right) \left(\frac{\theta^2 + 3\theta + 1}{(\theta+1)^2}\right)\right)^w \left(\left(\frac{\theta}{\theta+2}\right) \left(\frac{\theta^2 + 3\theta + 1}{(\theta+1)^2}\right)\right)^{N-w} \\ &\quad \frac{w!}{\prod_{j \geq 1} n_j!} \prod_{j \geq 1} \left(\frac{\left(\left(\frac{\theta}{\theta+2}\right) \left(\frac{\theta j + (\theta^2 + 3\theta + 1)}{(\theta+1)^{j+2}}\right)\right)}{\left(1 - \left(\frac{\theta}{\theta+2}\right) \left(\frac{\theta^2 + 3\theta + 1}{(\theta+1)^2}\right)\right)}\right)^{n_j}, \\ &= A(N, \theta) B(\theta). \end{aligned} \quad (3)$$

Where $N \geq w$, the likelihood is a function of parameters N and θ in the species model, where $\theta = (\theta_1, \theta_2, \dots, \theta_m)$. Here, θ is a nuisance parameter and our interest is in estimating the discrete parameter (N). This shows the likelihood can be factored into a binomial likelihood for w that corresponds $A(N, \theta)$ and it gives the profile likelihood estimates, and a multinomial likelihood for the

observed frequencies corresponds to the $B(\theta)$ gives the conditional likelihood estimate of the parameter. This factorization of the integrated likelihood has an important role. It was first formulated by Sanathanan (1972) who derived the asymptotic theory for the maximum likelihood estimation for N and θ . The likelihood function is not differentiable with respect to the discrete parameter (N). Lindsay and Roeder (1987) discussed the estimation of discrete parameters using the linear difference score (LDS) defined as

$$U(N) = \frac{L(N) - L(N-1)}{L(N)},$$

where, $L(N)$ is the likelihood for an integer parameter N .

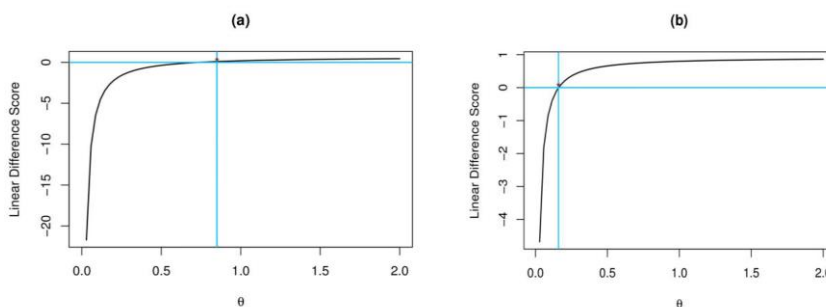


Figure 1
Plot of LDS of N with respect to different θ for fixed in
(a) $w=40$, $N=60$, and (b) $w=55$, $N=60$

Figure (1), shows that $U(N)=0$ and then gives the maxima of N with respect to the different choices of θ . Using the method described by Lindsay and Roeder (1987) to calculate the information for N and θ , we obtain from Fisher information matrix given below,

$$F(N, \theta) = \begin{pmatrix} \frac{1}{N} \frac{1-p_{\theta}(0)}{p_{\theta}(0)} & \left(-\frac{\partial}{\partial \theta} \log p_{\theta}(0)\right)^T \\ -\frac{\partial}{\partial \theta} \log p_{\theta}(0) & N(-\rho(\theta)) \end{pmatrix}.$$

Where, $\frac{\partial}{\partial \theta} p_{\theta}(0)$ is the column vector of partial derivatives. The $\rho(\theta) = \left[E_x \frac{\partial^2}{\partial \eta^2} \log p_{\theta}(x)\right]$, has taken expectation with respect to p_{θ} . Here, we notice that the diagonal elements of this partitioned matrix contain elements that factor into the function of N times with function of θ . Thus we have,

$$F(N, \theta) = \begin{pmatrix} \frac{1}{N} \frac{\theta^2 + 4\theta + 2}{\theta(\theta^2 + 3\theta + 1)} & -\left(\frac{\theta^3 + 7\theta^2 + 10\theta + 2}{\theta(\theta + 1)(\theta + 2)(\theta^2 + 3\theta + 1)}\right)^T \\ -\left(\frac{\theta^3 + 7\theta^2 + 10\theta + 2}{\theta(\theta + 1)(\theta + 2)(\theta^2 + 3\theta + 1)}\right) & N\left(\frac{4(\theta + 1)}{\theta^2(\theta + 2)^2} - \frac{2\theta^2 + 5\theta + 3}{\theta(\theta + 2)(\theta + 1)^2} + \psi(j, \theta)\right) \end{pmatrix} \quad (4)$$

$$\text{where, } \psi(j, \theta) = \sum_{j=0}^{\infty} \left(\frac{\theta}{\theta + 2}\right) \frac{1}{(\theta + 1)^{j+2}} \left(\frac{(j + 2\theta + 3)^2 - (2)(\theta j + \theta^2 + 3\theta + 1)}{\theta j + \theta^2 + 3\theta + 1}\right)$$

4. Bayesian estimation

In classical estimation theory, parameters are fixed in nature, but when the nature of the parameter changes, it shows up randomly, so we opt for the Bayesian estimation theory.

The parameters N and θ are treated as random variables and their prior distribution of parameters is non-informative prior. Here, the Jeffrey's prior and Bernardo's reference's prior are used. Jeffrey's prior (see Jeffreys (1946)) is based on invariance property under one-to-one re-parameterization. Jeffery's prior is defined to be proportional to the square root of the Fisher information matrix. For multidimensional models, the determinant of the Fisher information is used, which preserves the invariance property. By calculating the determinant of the partitioned matrix in equation (4), Jeffrey's prior is

$$g_j(N, \theta) \propto \det[F(N, \theta)]^{1/2} \quad (5)$$

$$\propto N^{\frac{m-1}{2}} g(\theta),$$

$$g_j(N, \theta) \propto$$

$$\left(\left(\frac{\theta^2 + 4\theta + 2}{\theta(\theta^2 + 3\theta + 1)} \right) \left(\frac{4(\theta + 1)}{\theta^2(\theta + 2)^2} - \frac{2\theta^2 + 5\theta + 3}{\theta(\theta + 2)(\theta + 1)^2} + \psi(j, \theta) \right) - \left(\frac{\theta^3 + 7\theta^2 + 10\theta + 2}{\theta(\theta + 1)(\theta + 2)(\theta^2 + 3\theta + 1)} \right)^2 \right)^{\frac{1}{2}}. \quad (6)$$

Where $g(\theta)$ is the function of θ which depends on the dimension of the information matrix. As the dimension increases, this becomes increasingly complex. The posterior distribution is the combination of likelihood function and Jeffery's prior that become the joint posterior distribution of $\pi_j(N, \theta | \text{data})$ is

$$\pi_j(N, \theta | \text{data}) \propto L(\text{data} | \theta, N) g_j(N, \theta),$$

$$\pi_j(N, \theta | \text{data}) \propto$$

$$\frac{w!}{\prod_{j \geq 1} n_j!} \prod_{j \geq 1} \left(\frac{\left(\frac{\theta}{\theta + 2} \right) \left(\frac{j\theta + (\theta^2 + 3\theta + 1)}{(\theta + 1)^{j+2}} \right)}{1 - \left(\frac{\theta}{\theta + 2} \right) \left(\frac{(\theta^2 + 3\theta + 1)}{(\theta + 1)^2} \right)} \right)^{n_j} \left(\left(\frac{\theta^2 + 4\theta + 2}{\theta(\theta^2 + 3\theta + 1)} \right) \left(\frac{4(\theta + 1)}{\theta^2(\theta + 2)^2} - \frac{2\theta^2 + 5\theta + 3}{\theta(\theta + 2)(\theta + 1)^2} + \right. \right.$$

$$\psi(j, \theta) - \left(\frac{\theta^3 + 7\theta^2 + 10\theta + 2}{\theta(\theta+1)(\theta+2)(\theta^2+3\theta+1)} \right)^2 \Big)^{\frac{1}{2}}. \quad (7)$$

Now full conditional posterior for N is

$$\begin{aligned} \pi_j(N|\theta, data) \\ \propto \binom{N}{w} \left(1 - \left(\frac{\theta}{\theta+2} \right) \left(\frac{\theta^2 + 3\theta + 1}{(\theta+1)^2} \right) \right)^{w+1} \left(\left(\frac{\theta}{\theta+2} \right) \left(\frac{\theta^2 + 3\theta + 1}{(\theta+1)^2} \right) \right)^{N-w}, \\ \propto NB \left(w+1, \left(1 - \left(\frac{\theta}{\theta+2} \right) \left(\frac{\theta^2+3\theta+1}{(\theta+1)^2} \right) \right) \right) \end{aligned} \quad (8)$$

and full conditional posterior for θ is

$$\begin{aligned} \pi_j(\theta|N, data) = \frac{w!}{\prod_{j \geq 1} n_j!} \prod_{j \geq 1} \left(\left(\frac{\theta}{\theta+2} \right) \left(\frac{j\theta + (\theta^2 + 3\theta + 1)}{(\theta+1)^{j+2}} \right) \right)^{n_j} \\ \left(\left(\frac{\theta^2+4\theta+2}{\theta^2(\theta^2+3\theta+1)} \right) \left(\frac{4(\theta+1)}{\theta^2(\theta+2)^2} - \frac{2\theta^2+5\theta+3}{\theta(\theta+2)(\theta+1)^2} + \psi(j, \theta) \right) - \left(\frac{\theta^3+7\theta^2+10\theta+2}{\theta(\theta+1)(\theta+2)(\theta^2+3\theta+1)} \right)^2 \right)^{\frac{1}{2}} \end{aligned} \quad (9)$$

Full conditional posterior for N , we can use direct sampling from negative binomial distribution with size $(w+1)$ and probability $\left(1 - \left(\frac{\theta}{\theta+2} \right) \left(\frac{\theta^2+3\theta+1}{(\theta+1)^2} \right) \right)$, and full conditional posterior for θ does not come in closed form then using the Metropolis-Hastings steps uses a normal proposal distribution to get posterior samples.

Another non-informative prior is Bernardo's reference prior (Bernardo (1979)) which is based on maximizing and expected entropy (measurement of loss of information), see Bernardo and Ramon (1998). The derivation of the reference prior takes into account the order of interest of the parameter, namely N and θ . If nuisance parameter $m = 1$ thus the fisher information matrix in equation (4) is 2×2 . Now, we can see that the elements of the information matrix would be each factor into a function of N and a function of θ . Let $H = F^{-1}$ be the covariance matrix. Notice that we have the factorization $(h_{11})^{-1/2} = a_0(N)b_0(\theta)$ and $(f_{22})^{1/2} = a_1(N)b_1(\theta)$, elements of the covariance and information matrices respectively. Assuming that the nuisance parameter space, Θ , is independent of the value of N , the reference prior is

$$\begin{aligned} g_R(N, \theta) &\propto (a_0(N))^{-1/2} (b_1(\theta))^{1/2} \\ &\propto N^{-1/2} (\rho(\theta))^{1/2} \end{aligned} \quad (10)$$

$$g_R(N, \theta) \propto N^{-1/2} \left(-\frac{4(\theta+1)}{\theta^2(\theta+2)^2} + \frac{2\theta^2+5\theta+3}{\theta(\theta+2)(\theta+1)^2} - \psi(j, \theta) \right)^{1/2}. \quad (11)$$

Here, we notice that the jointly derived prior factorizes into a function of N and a function of θ . Now, the combination of likelihood and Bernardo's reference prior that become the joint posterior distribution of $\pi_R(N, \theta | data)$ is

$$\pi_R(N, \theta | data) \propto L(data | \theta, N) g_R(N, \theta)$$

$$\begin{aligned} \pi_R(N, \theta | data) &\propto \binom{N}{w} \left(1 - \left(\frac{\theta}{\theta+2} \right) \left(\frac{\theta^2+3\theta+1}{(\theta+1)^2} \right) \right)^w \left(\left(\frac{\theta}{\theta+2} \right) \left(\frac{\theta^2+3\theta+1}{(\theta+1)^2} \right) \right)^{N-w} \\ &\frac{w!}{\prod_{j \geq 1} n_j!} \prod_{j \geq 1} \left(\frac{\left(\left(\frac{\theta}{\theta+2} \right) \left(\frac{\theta j + (\theta^2+3\theta+1)}{(\theta+1)^{j+2}} \right) \right)}{\left(1 - \left(\frac{\theta}{\theta+2} \right) \left(\frac{\theta^2+3\theta+1}{(\theta+1)^2} \right) \right)} \right)^{n_j} \times N^{-1/2} \left(-\frac{4(\theta+1)}{\theta^2(\theta+2)^2} + \frac{2\theta^2+5\theta+3}{\theta(\theta+2)(\theta+1)^2} - \right. \\ &\left. \psi(j, \theta) \right)^{1/2} \end{aligned}$$

Now full conditional posterior for N is

$$\pi_R(N | \theta, data) \propto N^{-1/2} \binom{N}{w} \left(1 - \left(\frac{\theta}{\theta+2} \right) \left(\frac{\theta^2+3\theta+1}{(\theta+1)^2} \right) \right)^w \left(\left(\frac{\theta}{\theta+2} \right) \left(\frac{\theta^2+3\theta+1}{(\theta+1)^2} \right) \right)^{N-w}, \quad (12)$$

and full conditional posterior for θ is

$$\begin{aligned} \pi_R(\theta | N, data) &\propto \frac{w!}{\prod_{j \geq 1} n_j!} \prod_{j \geq 1} \left(\frac{\left(\left(\frac{\theta}{\theta+2} \right) \left(\frac{\theta j + (\theta^2+3\theta+1)}{(\theta+1)^{j+2}} \right) \right)}{\left(1 - \left(\frac{\theta}{\theta+2} \right) \left(\frac{\theta^2+3\theta+1}{(\theta+1)^2} \right) \right)} \right)^{n_j} \\ &\left(-\frac{4(\theta+1)}{\theta^2(\theta+2)^2} + \frac{2\theta^2+5\theta+3}{\theta(\theta+2)(\theta+1)^2} - \psi(j, \theta) \right)^{1/2} \end{aligned} \quad (13)$$

Here, full conditional posterior for N and θ are not come in closed form then using the Metropolis-Hestings algorithm and proposal distribution is negative binomial for N and normal distribution for θ to get the posterior samples of parameter respectively.

5. Simulation Study

In the present study simulation carried out for the authenticity of model performance. It based on the frequentist estimates i.e. profile MLE ($\hat{N}_p$), and conditional MLE ($\hat{N}_c$). Bayesian estimates with Jeffery's prior ($\hat{N}_j$), and Bernardo's reference prior ($\hat{N}_R$). For obtaining the artificial data we have stoping time t is 0.8 and 1.2 for the stochastic abundance and w for the observed

In the frequentist approach, we excluded those samples which are yielded no coverage of MLE or the coverage fraction was less than 50%. The procedure was repeated until 1000 (say I) times were generated. In Table (5) and (6), the Maximum likelihood estimate of the parameter is the number of species (N) estimated under profile likelihood and conditional likelihood, and Bayes estimates of the parameter of the number of species (N) under Jeffery's prior and reference prior, with their respective 95% confidence interval and HPD interval reported. Here, the comparisons of proposed estimators of the parameters are based on the square root of average risk (expected loss over sample space), denoted by $R(N)$, where N is the true value of the number of species and $\hat{N}$ is the estimated value of the parameter of the number of species.

We observed that the coverage fraction and frequencies increase then the observed number of species gets closer to the estimated number of species (as most of the time these species become overestimates to the parameter). The simulation study was using statistical R software R-4.3.2 version.

In the present study, considered data set was originally suggested by Lewins and Joanes (1984), and it was carried out on Mount Kenya to collect the data on the insect population within a particular region. In this data set, the non-zero observed frequency (j) to the sample and the observed number of species (n_j) are categorized as:

[illegible]

In Table 1, we have notice that, the observed number of species (w) is 32 and the observed number of individual n is 1043. Firstly, we examined the graphical technique of the data set, which comes from the Poisson-exponential-gamma model.

In the Figure (2), depicts the observed relative frequency histogram and the theorized (or expected) relative frequency histogram on the same graph for the data set. After examining the graphical technique of this data set, we have concluded that the Poisson-exponential gamma model appears to offer a sufficient approach to the agreement between two histograms of observed and expected relative frequency.

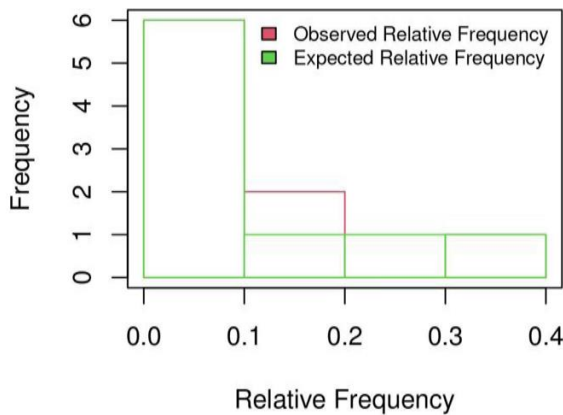


Figure 2

Histogram plot of Poisson-exponential-gamma model for observed relative frequency and expected relative frequency

However, due to some random fluctuation occurs between the two histograms and it can be assessed by the chi-square goodness of fit test. Now, chi-square goodness of fit test statistic that is $\chi^2_{cal} = 14.7850$ and $\chi^2_{3,0.95} = 15.5073$ (say χ^2_{tab}) then χ^2_{tab} value is greater than χ^2_{cal} value, so we can say that observed frequency has no significance difference between expected frequency of data set. Thus, this data set has been suggested for the Poisson-exponential-gamma model. In the Figure (3) and (4) are the diagnostic plot for parameter (N) that is trace plot and ACF plot, respectively.

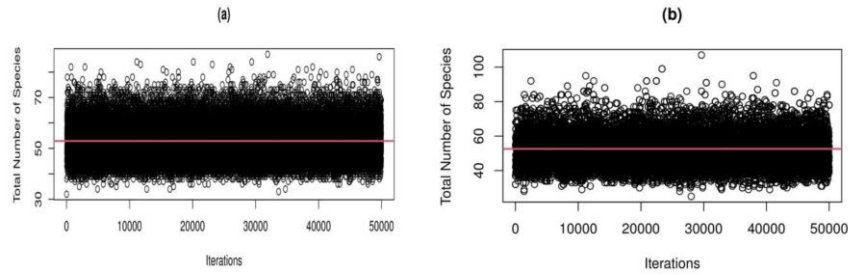


Figure 3
Trace plot for posterior mean values of Poisson-exponential-gamma model with (a) Jeffery's prior and (b) Bernardo's reference prior

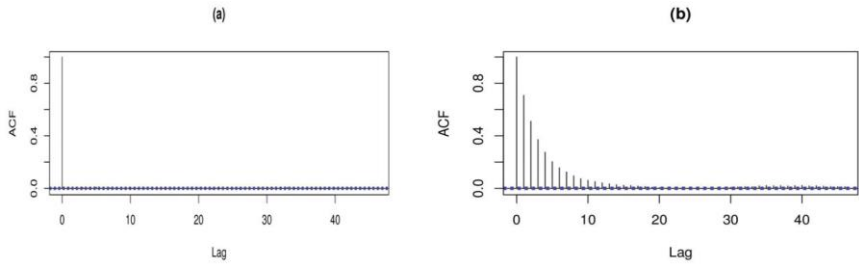


Figure 4
ACF plot for independent posterior values of Poisson-exponential-gamma with (a) Jeffery's prior and (b) Bernardo's reference prior

In the Figure (5), shows the histogram of posterior samples using Jeffery's prior and Bernardo's reference prior for parameter (N).

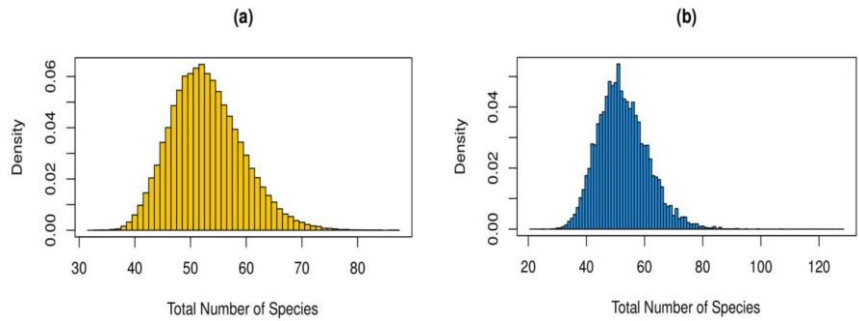


Figure 5
Posterior histogram plot of parameter N for Poisson-exponential-gamma model with (a) Jeffery's prior and (b) Bernardo's reference prior

In Table (2), the frequentist estimates i.e. Maximum Likelihood Estimate (MLE) obtained with profile likelihood N_p and conditional likelihood N_c , their respective 95% Confidence Interval (CI) of N for the Poisson-exponential-gamma. In Table (3), the Bayesian estimate of parameter (N) are summarized as posterior mean, median, and mode. Also, shown the 95% highest posterior density (HPD) intervals.

We can also see that the Bayesian estimates are always higher than the maximum likelihood estimates of parameter (N) for the Poisson-exponential-gamma model with Jeffrey's and Bernardo's reference prior. Further we found that the 95% HPD intervals have the smaller length as compared to CIs.

Table 2
MLEs of the number of species parameter under profile likelihood (N_p) and conditional likelihood (N_c) with 95% CI

Model	MLE	95% CI
Poisson-exponential-gamma model	$N_p = 37.61507$	(27.27435, 47.95578)
	$N_c = 48.95858$	(26.65971, 71.25745)

Table 3
Summary Statistics of the posterior $\pi(N|data)$ for two models. PEGM-J = Poisson-exponential-gamma with Jeffreys prior, PEGM-R = Poisson-exponential-gamma model with Bernardo's reference prior

Posterior Summary	PEGM-J	PEGM-R
Mode	51	51
Mean	52.8587	63.57278
Median	52	63
95% HPD Interval	(42,67)	(49,82)

After that, we calculate the Deviance Information Criteria (DIC) for each model is averaged over values from our posterior samples. Then DIC is defined as

$$D(data, N, \theta) = -2\log L(data|N, \theta) \quad (14)$$

We estimate the expected deviance by averaging the deviance over the posterior samples, θ^i , $i = 1, 2, \dots, I$ with I being the total number of posterior samples (say 1000). The estimates of the DIC is

$$\widehat{D}(data) = \frac{1}{I} \sum_{i=1}^I D(data, \theta^i) \quad (15)$$

Thus, DIC estimates are provided in Table (4), a smaller DIC indicates a better fits the data set.

Table 4
DIC for PEGM-J Poisson-exponential-gamma model with Jeffreys prior, and PEGM-R= Poisson-exponential-gamma model with Bernardo’s reference prior

Model	DIC
PEGM-J	290.5125
PEGM-R	129.3453

In the Figure in (6), fitted line superimposed histogram plot for the expected frequency using posterior samples of the parameters N using the Poisson-exponential-gamma model with Jeffery’s prior and Bernardo’s reference prior.

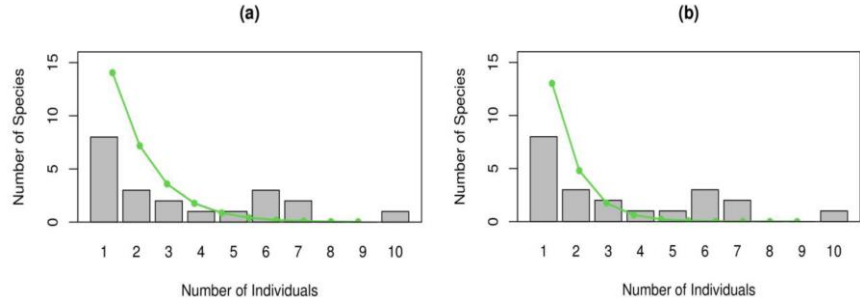


Figure 6
Fitted line superimposed histogram plot for expected frequency of Poisson-exponential-gamma with (a) Jeffreys prior and (b) Bernardo’s reference prior

7. Conclusions

In this study, model comparison based on deviance information criteria, PEGM-R has the smaller deviance from PEGM-J, then we can say that PEGM-R gives a more optimum estimate of the number of species. When j increases then CF (coverage fraction) increases but the estimate of N and $R(N)$ decreases and the estimate of the number of species $\hat{N}$ is the case of overestimation, i.e. $\hat{N} > w$. We observed that $R(N)$ under Bayesian estimate of Bernardo’s reference prior has a minimum than $R(N)$ under Bayesian estimate of Jeffery’s prior. Bayesian estimation of the number of species with a non-informative prior and a Poisson-exponential-gamma mixture model distribution is a strong tool for understanding and characterizing insect diversity. Its ability to account for uncertainty and adapt to new datasets makes it a useful asset in insect application research, leading to advances in our understanding and supporting sustainable development in a variety of fields.

Table 5
ML estimate and Bayes estimates of number of species, N and square root of average risk $R(N)$ for Poisson-exponential-gamma Model with fixed, $\theta = 0.5$.

		t=0.8				t=1.6				
N	j	CF		$\tilde{N}$	R(N)	CI/HPD	CF	$\tilde{N}$	R(N)	CI/HPD
50	5	71.20	$\tilde{N}_p$	48.5307	4.3263	(48.182*, 48.879*)	91.44	57.9474	8.1876	(57.623*, 58.271*)
			$\tilde{N}_c$	77.2711	29.0383	(77.013*, 77.529*)		72.4763	22.6623	(72.200*, 72.752*)
			$\tilde{N}_f$	78.4293	30.1866	(59.420, 96.380)		73.0790	23.2613	(60.120, 85.360)
			$\tilde{N}_R$	77.997	29.7608	(59.500, 95.220)		72.8411	23.0236	(60.080, 84.860)
	6	72.50	$\tilde{N}_p$	49.1783	4.4865	(48.831*, 49.524*)	93.08	58.3135	8.6538	(57.989*, 58.637*)
			$\tilde{N}_c$	76.7113	28.8255	(76.451*, 76.971*)		71.1924	21.6068	(70.911*, 71.473*)
			$\tilde{N}_f$	77.8047	29.9125	(59.480, 95.360)		71.7766	22.1860	(59.500, 83.3000)
			$\tilde{N}_R$	77.4133	29.5221	(59.620, 94.300)		71.5424	21.9570	(59.400, 82.9000)
	7	73.00	$\tilde{N}_p$	48.9037	4.2487	(48.555*, 49.252*)	93.12	57.4829	7.8882	(57.156*, 57.809*)
			$\tilde{N}_c$	72.4530	23.9818	(72.184*, 72.721*)		69.5791	19.9937	(69.294*, 69.863*)
			$\tilde{N}_f$	73.4474	24.9589	(56.660, 89.420)		70.1507	20.5615	(58.300, 81.400)
			$\tilde{N}_R$	73.0607	24.5774	(56.840, 88.520)		69.9083	20.3190	(58.340, 80.9600)
	8	74.00	$\tilde{N}_p$	49.4588	4.0234	(49.112*, 49.805*)	93.48	57.0825	7.4839	(56.753*, 57.411*)
			$\tilde{N}_c$	72.7945	24.2327	(72.526*, 73.063*)		68.2562	18.6640	(67.968*, 68.544*)
			$\tilde{N}_f$	73.7495	25.1759	(57.180, 89.700)		68.8059	19.2124	(57.240, 79.540)
			$\tilde{N}_R$	73.4173	24.8403	(57.220, 88.760)		68.5833	18.9882	(57.220, 79.120)
9	78.30	$\tilde{N}_p$	49.0584	4.7159	(48.710*, 49.406*)	94.52	56.6038	7.0891	(56.273*, 56.934*)	
		$\tilde{N}_c$	72.5715	24.9624	(72.301*, 72.841*)		67.3332	17.6850	(67.042*, 67.623*)	
		$\tilde{N}_f$	73.5434	25.9141	(56.780, 89.380)		67.8587	18.2067	(56.4400, 78.3800)	
		$\tilde{N}_R$	73.1882	25.5633	(56.980, 88.600)		67.6386	17.9878	(56.6600, 78.1000)	
70	5	72.70	$\tilde{N}_p$	69.5272	5.2188	(69.236*, 69.817*)	91.71	81.3961	11.8558	(81.122*, 81.669*)
			$\tilde{N}_c$	110.6870	42.6244	(110.471*, 110.902*)		101.8968	32.3721	(101.663*, 102.12*)
			$\tilde{N}_f$	111.8557	43.7846	(89.220, 133.500)		102.4896	32.9592	(86.980, 116.960)
			$\tilde{N}_R$	111.3817	43.3172	(89.6400, 132.520)		102.2451	32.7175	(87.020, 116.5800)
	6	74.0	$\tilde{N}_p$	70.3075	5.6014	(70.018*, 70.597*)	92.14	81.1071	11.7036	(80.832*, 81.381*)
			$\tilde{N}_c$	108.8285	40.3936	(108.611*, 109.04*)		99.7154	30.3217	(99.478*, 99.952*)
			$\tilde{N}_f$	109.9302	41.4760	(88.000, 130.600)		100.3082	30.9083	(85.700, 114.2400)
			$\tilde{N}_R$	109.4999	41.0504	(88.440, 129.680)		100.0696	30.6702	(85.720, 113.800)
	7	75.60	$\tilde{N}_p$	68.6799	5.7057	(68.386*, 68.973*)	92.26	80.3846	10.8279	(80.108*, 80.661*)
			$\tilde{N}_c$	104.7081	36.561	(104.485*, 104.930*)		96.7460	27.1259	(96.504*, 96.987*)

(Contd...)

50	8	77.40	$\hat{N}_J$	105.7741	37.6064	(84.900, 125.740)	94.94	97.2847	27.6610	(83.440, 110.360)
			$\hat{N}_R$	105.3445	37.1925	(85.320, 124.920)		97.0412	27.4189	(83.380, 109.880)
			$\hat{N}_p$	68.3717	5.0723	(68.077*, 68.665*)		79.8652	10.3152	(79.587*, 80.142*)
			$\hat{N}_c$	103.7168	35.2464	(103.493*, 103.940*)		95.5375	25.9729	(95.294*, 95.781*)
	9	78.00	$\hat{N}_J$	104.7624	36.2799	(84.020, 124.460)	97.63	96.0630	26.4958	(82.56, 108.900)
			$\hat{N}_R$	104.3825	35.9073	(84.220, 123.520)		95.8669	26.3032	(82.38, 108.400)
			$\hat{N}_p$	67.9035	5.5372	(67.608*, 68.198*)		80.5799	10.9761	(80.303*, 80.856*)
			$\hat{N}_c$	102.4067	34.0798	(102.181*, 102.631*)		96.6888	27.12810	(96.447*, 96.930*)
	9	78.00	$\hat{N}_J$	103.4440	35.1008	(82.880, 122.760)	97.63	97.2286	27.6667	(83.440, 110.140)
			$\hat{N}_R$	103.0287	34.6931	(83.100, 121.920)		96.9893	27.4274	(83.600, 109.860)

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

Table 6
ML estimate and Bayes estimates of number of species, N and square root of average risk $R(N)$ for Poisson-exponential gamma Model with fixed, $\theta = 0.9$.

t=0.8						t=1.6					
N	j	CF		$\hat{N}$	R(N)	CI/HPD	CF	$\hat{N}$	R(N)	CI/HPD	
50	5	64.56	$\hat{N}_p$	44.6196	7.2038	(44.257*,44.982*)	88.72	57.3398	7.9558	(57.015*, 57.663*)	
			$\hat{N}_c$	75.4195	27.9320	(75.159*,75.679*)		75.5304	26.1316	(75.262*, 75.798*)	
			$\hat{N}_j$	76.7578	29.2242	(56.840, 95.820)		76.2187	26.8151	(61.820, 89.960)	
			$\hat{N}_R$	76.3085	28.7967	(57.320, 94.680)		75.9540	26.5489	(61.740, 89.380)	
	6	66.56	$\hat{N}_p$	45.8914	5.9562	(45.533*,46.248*)	89.12	57.0774	7.5461	(56.751*, 57.403*)	
			$\hat{N}_c$	77.2023	29.9083	(76.945*,77.459*)		73.4244	23.9334	(73.151*, 73.697*)	
			$\hat{N}_j$	78.5346	31.2276	(58.620, 97.700)		74.0548	24.5611	(60.360, 87.040)	
			$\hat{N}_R$	78.0856	30.7949	(59.320, 96.840)		73.8403	24.3501	(60.300, 86.4600)	
	7	66.84	$\hat{N}_p$	45.4835	6.2164	(45.124*,45.842*)	89.60	56.4891	7.0960	(56.161*, 56.816*)	
			$\hat{N}_c$	75.1939	27.1721	(74.933*,75.454*)		71.9374	22.5507	(71.660*, 72.213*)	
$\hat{N}_j$			76.4589	28.4272	(56.920, 94.980)	72.5715		23.1803	(59.180, 85.120)		
$\hat{N}_R$			76.0103	27.9744	(57.460, 93.880)	72.3515		22.9625	(59.320, 84.700)		
8	67.08	$\hat{N}_p$	44.8215	7.4319	(44.457*, 45.185*)	89.76	56.9372	7.4359	(56.610*, 57.264*)		
		$\hat{N}_c$	70.1391	23.2153	(69.86*, 70.411*)		71.5171	21.9317	(71.239*, 71.795*)		
		$\hat{N}_j$	71.2498	24.2694	(53.740, 88.120)		72.1343	22.5385	(59.180, 84.400)		
		$\hat{N}_R$	70.8798	23.8877	(53.900, 87.040)		71.8712	22.2804	(59.180, 83.840)		
9	69.05	$\hat{N}_p$	44.4154	7.5144	(44.05*, 44.780*)	90.64	55.9868	6.6778	(55.656*, 56.316*)		
		$\hat{N}_c$	71.7028	24.3892	(71.434*,71.971*)		69.6749	20.3037	(69.392*, 69.957*)		
		$\hat{N}_j$	72.9197	25.5544	(54.380, 90.460)		70.3083	20.9247	(57.660, 82.160)		

(Contd...)

			$\hat{N}_R$	72.4589	25.0937	(55.020, 89.440)		70.0360	20.6554	(57.740, 81.720)
70	5	66.03	$\hat{N}_p$	64.3233	8.2816	(64.022*, 64.624*)	88.37	79.9035	10.5699	(79.629*, 80.178*)
			$\hat{N}_c$	112.5147	45.8282	(112.302*, 112.72*)		105.0781	35.7637	(104.850*, 105.305*)
			$\hat{N}_f$	113.9688	47.2665	(88.580, 137.96)		105.7776	36.4578	(88.740, 121.940)
			$\hat{N}_R$	113.4542	46.7655	(89.140, 136.720)		105.4804	36.1667	(88.840, 121.340)
	6	66.51	$\hat{N}_p$	62.2297	10.0393	(61.922*, 62.536*)	88.83	78.8895	9.4732	(78.612*, 79.166*)
			$\hat{N}_c$	103.3718	35.8991	(103.15*, 103.59*)		101.8906	32.5003	(101.659*, 102.122*)
			$\hat{N}_f$	104.6688	37.1409	(81.840, 126.66)		102.5523	33.1620	(86.240, 117.860)
			$\hat{N}_R$	104.2111	36.7013	(82.14, 125.46)		102.2634	32.8675	(86.220, 117.400)
	7	67.89	$\hat{N}_p$	65.4929	7.1837	(65.193*, 65.792*)	89.51	78.1092	8.7538	(77.830*, 78.387*)
			$\hat{N}_c$	109.9667	42.8831	(109.751*, 110.18*)		99.2380	29.8229	(99.002*, 99.473*)
			$\hat{N}_f$	111.2956	44.2196	(87.30, 134.080)		99.8664	30.4427	(84.260, 114.620)
			$\hat{N}_R$	110.7687	43.6824	(87.940, 133.06)		99.5788	30.1626	(84.300, 114.100)
	8	68.17	$\hat{N}_p$	64.8717	8.6236	(64.56*, 65.17*)	89.63	79.2341	9.8394	(78.957*, 79.510*)
			$\hat{N}_c$	101.7009	34.9026	(101.47*, 101.92*)		99.9247	30.3979	(99.689*, 100.159*)
			$\hat{N}_f$	102.8170	35.9644	(81.440, 123.040)		100.5459	31.0118	(84.880, 115.040)
			$\hat{N}_R$	102.3668	35.5309	(81.680, 122.020)		100.2955	30.7660	(85.220, 114.720)
	9	69.43	$\hat{N}_p$	64.0604	8.2224	(63.756*, 64.364*)	90.86	78.9724	9.7146	(78.694*, 79.249*)
			$\hat{N}_c$	99.5316	31.865	(99.303*, 99.759*)		99.3684	30.0186	(99.132*, 99.604*)
			$\hat{N}_f$	100.6299	32.9182	(79.740, 120.440)		99.9873	30.6290	(84.460, 114.360)
			$\hat{N}_R$	100.1984	32.5018	(79.940, 119.660)		99.7207	30.3734	(84.780, 114.0400)

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

Acknowledgements

The authors would like to thank editors and referees for their constructive comments and suggestions which improved and enriched the presentation of the paper.

Declarations

Ethical approval: This article does not contain any studies with human participants and animal participants performed by any of the authors.

Conflict of Interest: The corresponding author declares that no conflict of interest emerged among the authors, that all co-authors have equally contributed to the research or writing of the manuscript and no funding was received to conduct this research work from any institution/university.

References

- [1] Barger, K. and Bunge, J. Bayesian estimation of the number of species using noninformative priors. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 50(6) : 1064–1076 (2008).
- [2] Barger, K. and Bunge, J. Objective Bayesian estimation for the number of species. *Bayesian Analysis*, 5(4) : 765–785 (2010).
- [3] Behnke, A., Bunge, J., Barger, K., Breiner, H.-W., Alla, V., and Stoeck, T. Microeukaryote community patterns along an o₂/h₂s gradient in a super sulfidic anoxic fjord (framvaren, norway). *Applied and environmental microbiology*, 72(5) : 3626– 3636 (2006).
- [4] Bernardo, J. M. Reference posterior distributions for Bayesian inference. *Journal of the Royal Statistical Society: Series B (Methodological)*, 41(2) : 113–128 (1979).
- [5] Bernardo, J. M. and 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(1) : 101–135 (1998).
- [6] Bulmer, M. On fitting the Poisson lognormal distribution to species-abundance data. *Biometrics*, 101–110 (1974).
- [7] Bunge, J. and Fitzpatrick, M. Estimating the number of species: a review. *Journal of the American Statistical Association*, 88(421) : 364–373 (1993).
- [8] Colwell, R. K. and Coddington, J. A., Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 345(1311) : 101–118 (1994).
- [9] Fisher, R. A., Corbet, A. S., and 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).
- [10] Hong, S.-H., Bunge, J., Jeon, S. O., and Epstein, S. S. Predicting microbial species richness. *Proceedings of the National Academy of Sciences*, 103(1) : 117–122 (2006).
- [11] 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(1007) : 453–461 (1946).
- [12] Kaur, K., Arora, S., & Mahajan, K. K. Bayesian vs semi-Bayesian estimation for income inequality measures in case of GIWD. *Journal of Statistics and Management Systems*, 24(8), 1753-1784 (2021).

- [13] Leite, J. G., Rodrigues, J., and Milan, L. A. A Bayesian Analysis for Estimating the Number of Species in a Population using Nonhomogeneous Poisson Process. *Statistics & Probability Letters*, 48(2) : 153–161 (2000).
- [14] Lewins, W. A. and Joanes, D. Bayesian estimation of the number of species. *Biometrics*, 323–328 (1984).
- [15] Lindsay, B. G. and Roeder, K. A Unified Treatment of Integer Parameter Models. *Journal of the American Statistical Association*, 82(399) : 758–764 (1987).
- [16] Ord, J. K. and Whitmore, G. A. The Poisson-inverse gaussian distribution as a model for species abundance. *Communications in Statistics-theory and Methods*, 15(3) : 853–871 (1986).
- [17] Rabie, A. E-Bayesian estimation for a constant-stress partially accelerated life test based on Burr-X Type-I hybrid censored data. *Journal of Statistics and Management Systems*, 24(8), 1649–1667 (2021).
- [18] Rodrigues, J., Milan, L. A., and Leite, J. G. Hierarchical Bayesian Estimation for the Number of Species. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 43(6) : 737–746 (2001).
- [19] Sanathanan, L. Estimating the size of a multinomial population. *The Annals of Mathematical Statistics*, 43(1) : 142–152 (1972).
- [20] Sankaran, M. 275. Note: The discrete Poisson-Lindely distribution. *Biometrics*, 145–149 (1970).
- [21] Shanker, R. Garima distribution and its application to model behavioural science data. *Biometrics and Biostatistics International Journal*, 4(7) : 1–9 (2016).
- [22] Shanker, R. The discrete poisson-garima distribution. *Biometrics and Biostatistics International Journal*, 5(2) : 1–7 (2017).
- [23] Sichel, H. Parameter estimation for a word frequency distribution based on occupancy theory. *Communications in Statistics-Theory and Methods*, 15(3) : 935–949 (1986).
- [24] Wang, X., He, C. Z., and Sun, D. Bayesian population estimation for small sample capture-recapture data using noninformative priors. *Journal of Statistical Planning and Inference*, 137(4):1099–1118 (2007).
- [25] Wilson, R. M. and Collins, M. F. Capture-recapture estimation with samples of size one using frequency data. *Biometrika*, 79(3): 543–553 (1992).

Received September, 2023