Hacettepe Journal of Mathematics & Statistics

Hacet. J. Math. Stat.
Volume 51 (3) (2022), 834 – 856
DOI : 10.15672/hujms.959817

Research Article

Bayesian and frequentist approaches on estimation and testing for a zero-inflated binomial distribution

Seungji Nam[1], Seong W. Kim*[2], Hon Keung Tony Ng[3]

[1] *Department of Statistics and Data Science, Yonsei University, Seoul, 03722, South Korea*
[2] *Department of Applied Mathematics, Hanyang University, Ansan, 15588, South Korea*
[3] *Department of Statistical Science, Southern Methodist University, Dallas, TX 75275, USA*

Abstract

To analyze discrete count data with excessive zeros, different zero-inflated statistical models that allow for frequent zero-valued observations have been developed. When the underlying data generation process of non-zero values is based on the number of successes in a sequence of independent Bernoulli trials, the zero-inflated binomial distribution is perhaps adequate for modeling purposes. In this paper, we discuss statistical inference for a zero-inflated binomial distribution using the objective Bayesian and frequentist approaches. Point and interval estimation of the model parameters and hypothesis testing for excessive zeros in a zero-inflated binomial distribution are developed. A Monte Carlo simulation study is used to assess the performance of estimation and hypothesis testing procedures. A comparative study of the objective Bayesian approach and the frequentist approach is provided. The proposed statistical inferential methods are applied to analyze an earthquake dataset and a baseball dataset for illustration.

Mathematics Subject Classification (2020). 62F03, 62F15

Keywords. Bayes factor, binomial distribution, EM algorithm, Jeffreys prior, maximum likelihood estimate, zero-inflated models

1. Introduction

Countable discrete data are widely encountered in diverse fields, including social science, natural science, engineering, and sport science. In many instances, these count data tend to possess zero-inflated patterns in the sense that there are more zero frequencies than usual. For instance, many power hitters or sluggers in baseball seldom produce triple-base hits rather than home runs. Events that rarely occur, such as triple-base hits in the baseball example, usually result in zero counts and the data are referred to as zero-inflated data. Analyzing discrete data with excessive zeros under traditional methodologies and models can result in biased estimators or loss of information. Therefore, to analyze discrete count data with excessive zeros, different zero-inflated statistical models that allow for frequent

*Corresponding Author.
Email addresses: seungji07@yonsei.ac.kr (S. Nam), seong125@gmail.com (S.W. Kim), ngh@mail.smu.edu (H.K.T. Ng)
Received: 30.06.2021; Accepted: 20.01.2022

zero-valued observations have been developed. Cohen [8] initially proposed zero-inflated Poisson models, which were later developed by [15]. Several methodologies related to zero-inflated models have been proposed and applied in various areas since then.

Most studies on zero-inflated models have focused on the Poisson and negative binomial distributions [21,29]. Both frequentist and Bayesian approaches have been used to conduct statistical inference for zero-inflated models. Dong et al. [12] described a multivariate random-parameter zero-inflated negative binomial regression model for modeling crash counts. A full Bayesian methodology was utilized to estimate model parameters. Ridout et al. [23] proposed a score test for conducting hypothesis testing of zero-inflated Poisson (ZIP) regression models against zero-inflated negative binomial alternatives. When the underlying data generation process of the non-zero values is based on the number of successes in a sequence of independent Bernoulli trials, a zero-inflated binomial distribution is appropriate for modeling purposes. Amek et al. [2] developed zero-inflated binomial (ZIB) geostatistical models and compared them with standard binomial distributions, where a Bayesian approach was used to analyze a dataset from epidemiology (the sporozoite rates) and the corresponding highest posterior density (HPD) intervals were obtained. Recently Zhang et al. [19] integrated a zero-inflated negative binomial with a Gaussian process to analyze spatial transcriptomics data in which analysis was conducted under a Bayesian framework. Jiang et al. [14] also employed a zero-inflated negative binomial regression model to perform an integrative analysis on microbiome data. See [20] and [28] for detailed descriptions and related references to account for excess zeros in sequential count data emerging in biological fields.

In the past decade, the analysis of count data using ZIB distributions has been discussed [6,11,26]. As pointed out by [4], it is important to conduct hypothesis testing of the zero-inflated model versus the regular (non-zero-inflated) model when zero-inflated models are adapted. For this reason, we consider hypothesis testing for excessive zeros in a ZIB distribution along with the point and interval estimation of the model parameters. In this paper, we focus on the ZIB distribution described as follows. Let X_i $(i = 1, 2, \ldots, N)$ be a zero-inflated binomial random variable, denoted by $X_i \sim \text{ZIB}(n_i, \theta, \omega)$ having the following probability mass function (pmf):

$$X_i \sim \begin{cases} 0, & \text{with probability } \omega, \\ \text{Bin}(n_i, \theta), & \text{with probability } 1-\omega, \end{cases}$$

for $0 \leq \omega \leq 1$ and $0 < \theta \leq 1$. That is, $\text{ZIB}(n_i, \theta, \omega)$ can be characterized as

$$f_1(x_i|\theta, \omega) = \omega I(x_i = 0) + (1-\omega) f_0(x_i|\theta), \; x_i = 0, 1, \ldots, n_i, \tag{1.1}$$

where

$$f_0(x_i|\theta) = \binom{n_i}{x_i} \theta_i^x (1-\theta)^{n_i - x_i}, \; x_i = 0, 1, \ldots, n_i, \tag{1.2}$$

and $I(\cdot)$ is an indicator function. That is, $I(A) = 1$ if A is true and $I(A) = 0$ otherwise. The parameter ω is often called the *zero-inflation* parameter.

Based on the ZIB model in Eq. (1.1), an objective Bayesian approach and a frequentist approach for statistical inference are discussed and compared in this paper. The rest of this paper is organized as follows. In Section 2, the objective Bayesian approach for estimation and hypothesis testing of the ZIB is developed. We first review the use of the Bayes factor for model selection. Since prior elicitation is one of the important issues in the Bayesian framework, we present elicitation procedures to specify default objective priors in testing the zero-inflation parameter of the ZIB. Bayesian model selection and testing procedures using the Bayes factor, which is an integrated likelihood ratio of two contending models, are proposed. In Section 3, the frequentist approach for estimation and

hypothesis testing of the ZIB is discussed. We review the direct maximization method and the expectation-maximization (EM) algorithm to obtain the maximum likelihood estimates of the model parameters. Moreover, we propose a likelihood ratio test and a bootstrap testing procedure for hypothesis testing of excessive zeros in the ZIB model. A Monte Carlo simulation study is used in Section 4 to evaluate the performance of the estimation and hypothesis testing procedures. The proposed Bayesian and frequentist approaches are compared and discussions are provided. The main results are illustrated by analyzing two real datasets in Section 5. Finally, brief concluding remarks are provided in Section 6.

2. Objective Bayesian approach

2.1. Review of the Bayes factor for model selection

Suppose that there are L different models, denoted by $M_1, M_2, \ldots M_L$, being considered as candidate models, and these models contend with each other in determining the most plausible model. If model M_i holds, then the data $\boldsymbol{X}$ follow a parametric distribution with probability density function (pdf) or probability mass function (pmf) $f_i(\boldsymbol{x}|\theta_i)$ depending upon types of random variables, where θ_i is an unknown parameter (possibly a vector). Let Θ_i be the parameter space for θ_i in which Θ_i may or may not be nested. Bayesian model selection proceeds by choosing a prior distribution $\pi_i(\theta_i)$ for θ_i under M_i, and the prior model probability $p(M_i)$ of model M_i being the true model (before the data are observed) for $i = 1, 2, \ldots, L$. The posterior probability that M_i is the true model can be expressed as

$$\Pr(M_i; \boldsymbol{x}) = \frac{p(M_i) m_i(\boldsymbol{x})}{\sum_{j=1}^{L} p(M_j) m_j(\boldsymbol{x})}, \tag{2.1}$$

where $m_i(\boldsymbol{x}) = \int_{\Theta_i} f_i(\boldsymbol{x}|\theta_i)\pi_i(\theta_i)d\theta_i$ is called the marginal or predictive density of $\boldsymbol{X}$ under model M_i, $i = 1, 2, \ldots, L$. Subsequently, for given data $\boldsymbol{x}$, the model with the largest posterior probability in (2.1) can be regarded as the most plausible model. Further, the Bayes factor of model M_j to model M_i is defined as

$$B_{ji}(\boldsymbol{x}) = \frac{m_j(\boldsymbol{x})}{m_i(\boldsymbol{x})} = \frac{\int_{\Theta_j} f_j(\boldsymbol{x}|\theta_j)\pi_j(\theta_j)d\theta_j}{\int_{\Theta_i} f_i(\boldsymbol{x}|\theta_i)\pi_i(\theta_i)d\theta_i}. \tag{2.2}$$

As a special case, when $L = 2$ with an equal prior model probability of $1/2$, we denote the two models as M_0 and M_1, and we have

$$\Pr(M_0; \boldsymbol{x}) = \frac{1}{1 + B_{10}} \text{ and } \Pr(M_1; \boldsymbol{x}) = \frac{B_{10}}{1 + B_{10}}.$$

In other words, for the case with two candidate models M_0 and M_1, based on the Bayes factor B_{10}, we would select model M_0 as the true model if

$$\Pr(M_0; \boldsymbol{x}) = \frac{1}{1 + B_{10}} > \frac{1}{2} \Longrightarrow B_{10} < 1$$

and select model M_1 as the true model if $B_{10} > 1$. Kass and Raftery [16] suggested the scale for interpretation of Bayes factors supporting evidence against M_0. If B_{10} is between 1 and 3.2, then we say "Not worth more than a bare mention". If it is between 3.2 to 10, "Substantial", 10 and 100, "Strong", and greater than 100, we dare to say "Decisive".

2.2. Objective priors in the ZIB model

Based on the ZIB distribution presented in Eq. (1.1), we develop Bayesian testing for excess zeros by considering the following two candidate models:

$$M_0 \ : \ X_i \overset{i.i.d.}{\sim} f_0(\cdot|\theta) \ , i = 1, 2, \ldots, N, \tag{2.3}$$

versus

$$M_1 \ : \ X_i \overset{i.i.d.}{\sim} f_1(\cdot|\theta, \omega) \ , i = 1, 2, \ldots, N, \tag{2.4}$$

where f_0 and f_1 are the probability functions of the binomial and the ZIB as given in Eqs. (1.1) and (1.2), respectively. Alternately, we can formulate the problem as a hypothesis test within the ZIB model as

$$H_0 : \omega = 0 \text{ vs } H_1 : \omega > 0. \tag{2.5}$$

Following similar arguments in [4], the prior specifications are the choice of $\pi_0(\theta)$ and $\pi_1(\theta)\pi_1(\omega|\theta)$. As mentioned in [17], if the common parameters are *orthogonal* to the remaining parameters in each model, i.e., the Fisher information matrix is diagonal, then the resulting Jeffreys priors have the same prior distributions. Since the parameters ω and θ in the ZIB model (M_1) are not orthogonal, we need to reparameterize the original model as follows. Let $n_i = n$ and $x_i = x$ for all $i = 1, 2, \ldots, N$ for simplicity. First, $f_1(x|\theta, \omega)$ is written as

$$f_1^*(x|\theta, \omega^*) \quad = \quad \omega^* I(x = 0) + (1 - \omega^*) f_T(x|\theta), \ x = 0, 1, \ldots, n,$$

where $\omega^* = \omega + (1 - \omega)(1 - \theta)^n$ and $f_T(x|\theta)$ is the zero-truncated version of the standard binomial distribution with parameters n and θ. Note that $(1 - \theta)^n \leq \omega^* \leq 1$, and $f_T(x|\theta)$ can be expressed as

$$f_T(x|\theta) \quad = \quad \frac{f(x)}{1 - f(0)} = \frac{\binom{n}{x}\theta^x(1 - \theta)^{n-x}}{1 - (1 - \theta)^n}, \ x = 1, 2, \ldots, n.$$

Subsequently, model M_0 can be expressed as

$$f_0^*(x|\theta) = (1 - \theta)^n I(x = 0) + [1 - (1 - \theta)^n] f_T(x|\theta), \ x = 0, 1, \ldots, n.$$

After the reparameterization, the Fisher information matrix for ω^* and θ can be shown to be diagonal.

As suggested by [4], we use the same Jeffreys prior for the common parameter θ and a proper prior for the remaining zero-inflation parameter ω. It is well known that the Jeffreys prior for θ in the (regular) binomial model is

$$\pi_0^J(\theta) \propto 1/\sqrt{\theta(1 - \theta)}.$$

Note that the Jeffreys prior for the orthogonalized ZIB model is the same for the truncated distribution $f_T(x|\theta)$. Thus, we have

$$\pi_1^J(\theta) \propto \frac{1}{\sqrt{\theta(1 - \theta)}} \ c(\theta), \tag{2.6}$$

where

$$c(\theta) = \frac{\sqrt{1 - (1 - \theta)^n - n\theta(1 - \theta)^{n-1}}}{1 - (1 - \theta)^n}, \text{ for } 0 < \theta < 1.$$

The derivation of Eq. (2.6) is presented in Appendix A. It is a little unusual that the Jeffreys prior for the common θ is different for each model. So, we need to justify which prior for θ is more plausible. To resolve this issue for choice, we specify a proper prior for ω^* given θ. We assume that it takes a uniform distribution over the interval $((1 - \theta)^n, 1)$, i.e.,

$$\pi_1(\omega^*|\theta) = \frac{I[(1 - \theta)^n < \omega^* \leq 1]}{1 - (1 - \theta)^n}.$$

Thus, we can impose the prior distributions for the two models $f_0^*(x|\theta)$ and $f_1^*(x|\theta,\omega^*)$, which are given respectively by

$$\pi_0^l(\theta) = \frac{c(\theta)^l}{\sqrt{\theta(1-\theta)}} \quad \text{and} \quad \pi_1^l(\theta,\omega^*) = \frac{c(\theta)^l}{\sqrt{\theta(1-\theta)}} \frac{I[(1-\theta)^n < \omega^* \leq 1]}{1-(1-\theta)^n}, \tag{2.7}$$

where l is 0 or 1 so that we utilize one of the two Jeffreys priors for θ. However, it is computationally more durable to work with the original parameterization for (θ,ω). After applying the change of variable technique with the Jacobian transformation, we have the Jefferys priors for the original model as

$$\begin{aligned} \pi_1^l(\theta,\omega) &= \pi_1^l(\theta,\omega^*)|J| \\ &= \frac{c(\theta)^l}{\sqrt{\theta(1-\theta)}} I(0<\omega<1). \end{aligned}$$

In the preliminary study, we have investigated the relationship between the two sets of prior distributions when $l=0$ and $l=1$ and found that there are not many differences in the resulting Bayes factors.

It can be shown that $c(\theta)$ is a strictly increasing function of θ regardless of n. To calculate the infimum and supremum, we apply the L'Hospital's rule with n being fixed. Then, we have

$$\lim_{\theta\to 0} c(\theta) = \sqrt{\frac{n-1}{2n}}.$$

As a result, $c(\theta) \approx 1/\sqrt{2}$ as $n \to \infty$. Finally, we notice that the value of $c(\theta)$ equals one when $\theta = 1$, i.e., $c(1) = 1$. Thus, it follows that

$$\inf c(\theta) = \frac{1}{\sqrt{2}} \quad \text{and} \quad \sup c(\theta) = 1.$$

Therefore, the corresponding Bayes factors B_{10}^l, $(l = 0,\ 1)$ have the following relationship:

$$B_{10}^0/\sqrt{2} \leq B_{10}^1 \leq \sqrt{2}B_{10}^0,$$

which is congruent with the results of [4] and supports the well-known fact that the Poisson distribution can be used as an approximation to the binomial distribution for a large number of trials. Furthermore, it is not necessary to consider 'training sample' computation for the intrinsic Bayes factors (IBF) of [5] since arbitrary constants would be cancelled out in the computation of the IBF. Since the prior with $l=0$ is simpler than the prior with $l=1$, we use the prior with $l=0$ for subsequent analyses and studies. Thus, the joint prior for the ZIB model can be expressed as

$$\pi_1(\theta,\omega) \propto \theta^{-1/2}(1-\theta)^{-1/2},\ 0<\theta<1,\ 0<\omega<1. \tag{2.8}$$

Let $\alpha = \sum_{i=1}^N I(X_i = 0)$ be the number of zero observations, and $t = \sum_{i=1}^N X_i$ be the total number of successes. Then, the likelihood function under M_1 can be expressed as

$$L_1(\boldsymbol{x}|\theta,\omega) \quad \propto \quad \Big[\omega + (1-\omega)(1-\theta)^n\Big]^\alpha (1-\omega)^{N-\alpha}\theta^t(1-\theta)^{n(N-\alpha)-t}. \tag{2.9}$$

Applying the binomial expansion on $[\omega + (1-\omega)(1-\theta)^n]^\alpha$ in Eq. (2.9) yields

$$\Big[\omega + (1-\omega)(1-\theta)^n\Big]^\alpha \quad = \quad \sum_{j=0}^{\alpha} \binom{\alpha}{j} \omega^j (1-\omega)^{\alpha-j}(1-\theta)^{n(\alpha-j)},$$

and the Bayes factor $B_{10} = m_1(\boldsymbol{x})/m_0(\boldsymbol{x})$ can be obtained as

$$B_{10} = \frac{\Gamma(Nn+1)}{\Gamma(Nn-t+1/2)} \frac{\alpha!}{(N+1)!} \sum_{j=0}^{\alpha} \frac{(N-j)!}{(\alpha-j)!} \frac{\Gamma(Nn-nj-t+1/2)}{\Gamma(Nn-nj+1)}. \tag{2.10}$$

2.3. The posterior distributions of the parameters for the ZIB

In this section, we present closed forms of the marginal posterior distributions of θ and ω, respectively, based on the Jeffreys priors presented in Section 2.2. The joint posterior distribution of θ and ω can be expressed as

$$p(\theta,\omega|\boldsymbol{x}) \propto \sum_{j=0}^{\alpha}\binom{\alpha}{j}\omega^j(1-\omega)^{N-j}\theta^{t-1/2}(1-\theta)^{Nn-nj-t-1/2}. \tag{2.11}$$

The following proposition provides the marginal posterior distributions of ω and θ with mixture distributions.

Proposition 2.1. (1) *The marginal posterior of* ω *is*

$$p(\omega|\boldsymbol{x}) = c_0 f_0(\omega) + c_1 f_1(\omega) + \cdots + c_\alpha f_\alpha(\omega),$$

where $f_\ell \sim Beta(\ell+1, N-\ell+1)$ *for* $\ell = 0,1,\ldots,\alpha$.

(2) *The marginal posterior of* θ *is*

$$p(\theta|\boldsymbol{x}) = c_0 g_0(\theta) + c_1 g_1(\theta) + \cdots + c_\alpha g_\alpha(\theta),$$

where $g_\ell \sim Beta(t+1/2, Nn-n\ell-t+1/2)$ *for* $\ell = 0,1,\ldots,\alpha$.

Here, the weights of the mixture distributions are

$$c_0 = k_0^{(\alpha)}/K^{(\alpha)},\ c_1 = k_1^{(\alpha)}/K^{(\alpha)},\cdots,\ c_\alpha = k_\alpha^{(\alpha)}/K^{(\alpha)},$$

where

$$K^{(\alpha)} = \sum_{j=0}^{\alpha}\frac{\alpha!}{(N-\alpha)!}A_j D_j. \tag{2.12}$$

Here,

$$A_j = \frac{(N-j)!}{(\alpha-j)!},\quad B_j = Nn-t-nj+\frac{1}{2},\quad C_j = Nn-nj+1, D_j = \Gamma(B_j)/\Gamma(C_j) \tag{2.13}$$

Finally, we have $K^{(\alpha)} = k_0^{(\alpha)} + k_1^{(\alpha)} + \cdots + k_\alpha^{(\alpha)}$ for which $k_\ell^{(\alpha)}$ is the ℓ-th term of $K^{(\alpha)}$ in Eq. (2.12) for $\ell = 0,1,\ldots,\alpha$.

Remark 2.2. When $\alpha = 1$, the marginal posterior distribution of ω is

$$p(\omega|\boldsymbol{x}) = N(N+1)\frac{D_0(1-\omega)^N + D_1\omega(1-\omega)^{N-1}}{ND_0+D_1},$$

where the corresponding weights are

$$c_0 = \frac{ND_0}{ND_0+D_1} \text{ and } c_1 = \frac{D_1}{ND_0+D_1}.$$

Based on the posterior distributions of ω and θ, we can obtain the posterior means, denoted by $\tilde{\omega}$ and $\tilde{\theta}$, respectively. We can also construct the credible HPD intervals for the model parameters.

Note that the interpretations of confidence intervals and credible intervals are different. A $100(1-\delta)\%$ credible interval can be interpreted as a probabilistic statement about the parameter, as there is $100(1-\delta)\%$ probability that the true parameter would lie within the interval given the observed data. For instance, a 95% credible interval (which is known as Bayesian 95% confidence interval) can be interpreted as there being a 95% probability that the true (unknown) parameter would lie within the interval, given the evidence provided by the observed data. Meanwhile, a 95% confidence interval in a frequentist setting can be interpreted as 95% of all samples giving an interval that contains the true parameter in repeated sampling.

2.4. Bayesian predictive distributions

In this subsection, we discuss the predictive distribution for a future observation that follows the ZIB distribution. Recall that the joint posterior distribution of θ and ω based on a random sample of size N is given by (2.11). Let $Z \equiv X_{N+1}$ be a future observation having the ZIB distribution, denoted by $f(z|\theta,\omega)$, i.e., $Z \sim \text{ZIB}(n_z,\theta,\omega)$, where n_z is the number of trials for the random variable Z. Then, the predictive distribution of the future observation Z given the past data $\boldsymbol{X} = (X_1, X_2, \ldots, X_N)$ based on sample sizes $n_1 = n_2 = \ldots = n_N = n$ is

$$Z|\boldsymbol{X} \sim g(z|\boldsymbol{x}) = \int_0^1 \int_0^1 f(z|\theta,\omega)p(\theta,\omega|\boldsymbol{x})d\theta d\omega. \tag{2.14}$$

Since the ZIB distribution has two sub-distributions; one degenerated at zero and the other following a conventional binomial distribution, the predictive distribution can be obtained with two separate cases. That is,

- If $Z = 0$, then $f(z|\theta,\omega) = \omega + (1-\omega)(1-\theta)^{n_z}$;
- If $Z \neq 0$, then $f(z|\theta,\omega) = (1-\omega)\binom{n_z}{z}\theta^z(1-\theta)^{n_z-z}$.

Thus, we have a closed form with a two-fold predictive distribution of Z. When $Z = 0$, the predictive distribution of Z is

$$g(z=0|\boldsymbol{x}) = \frac{1}{N+2}\frac{\sum_{j=0}^{\alpha} A_j\Big[(j+1)\Gamma(B_j)/\Gamma(C_j) + (N-j+1)\Gamma(B_j+n_z)/\Gamma(C_j+n_z)\Big]}{\sum_{j=0}^{\alpha} A_j\Gamma(B_j)/\Gamma(C_j)},$$

and the predictive distribution of Z for $z \neq 0$ is

$$g(z\neq 0|\boldsymbol{x}) = \binom{n_z}{z}\frac{1}{N+2}\frac{\Gamma(t+z+1/2)}{\Gamma(t+1/2)}\frac{\sum_{j=0}^{\alpha} A_j(N-j+1)\Gamma(B_j+n_z-z)/\Gamma(C_j+n)}{\sum_{j=0}^{\alpha} A_j\Gamma(B_j)/\Gamma(C_j)},$$

where A_j, B_j, and C_j are given in Eq. (2.13).

Remark 2.3. When we have observations with unequal sample sizes, i.e., if there is at least one pair of (i,j) such that $n_i \neq n_j$, a closed form of the predictive distribution is not available.

3. Frequentist approach

In this section, we present the frequentist approach for statistical inference of the ZIB model based on the likelihood method. Note that there are existing R functions available for fitting the ZIB distribution by the maximum likelihood estimation method. For instance, the R functions `zibinomial` and `zibinomialff` in the R package VGAM [27] fit the ZIB distribution based on the maximum likelihood estimation method. Here, we consider the maximum likelihood estimators (MLE) for (θ,ω) under M_1 obtained by the direct maximization method and the EM algorithm. We also present a likelihood ratio test (LRT) procedure and a bootstrap procedure for testing the hypotheses in Eq. (2.5).

3.1. Maximum likelihood estimation

3.1.1. Direct maximization methods. Based on the observed values x_i from ZIB(n_i, θ, ω), $i = 1, 2, \ldots, N$, from the distribution presented in Eqs. (1.1) and (1.2), the likelihood function can be expressed as

$$L(\theta, \omega) = \prod_{i=1}^{N} \omega I(x_i = 0) + (1 - \omega) \binom{n_i}{x_i} \theta^{x_i} (1 - \theta)^{n_i - x_i}$$

and the log-likelihood function can be written as

$$\ell(\theta, \omega) = \ln L(\theta, \omega) = \sum_{i=1}^{N} \left\{ \ln \left[\omega + (1 - \omega)(1 - \theta)^{n_i}\right] I(x_i = 0) + \ln \left[(1 - \omega) \binom{n_i}{x_i} \theta^{x_i} (1 - \theta)^{n_i - x_i} \right] I(x_i > 0) \right\}. \quad (3.1)$$

The MLEs of the parameter θ and ω can be obtained by maximizing the log-likelihood function in Eq. (3.1). Numerical methods such as the Nelder-Mead method with the constraints $0 < \theta < 0$ and $0 < \omega < 1$ can be used to maximize the log-likelihood function and obtain the MLEs. Possible initial estimates of θ and ω for the iterative procedure in finding the MLEs are $\sum_{i=1}^{N} x_i / \sum_{i=1}^{N} n_i$ and $\sum_{i=1}^{N} I(x_i = 0)/N$, respectively.

Here, we denote the MLEs of θ and ω as $\hat{\theta}$ and $\hat{\omega}$, respectively. The observed Fisher information matrix can be expressed as

$$\mathcal{I}(\theta, \omega) = \begin{bmatrix} -\frac{\partial^2 \ell(\theta,\omega)}{\partial \theta^2} & -\frac{\partial^2 \ell(\theta,\omega)}{\partial \theta \partial \omega} \\ & -\frac{\partial^2 \ell(\theta,\omega)}{\partial \omega^2} \end{bmatrix}_{(\theta,\omega)=(\hat{\theta},\hat{\omega})}, \quad (3.2)$$

where the second derivatives of the log-likelihood function are presented in Appendix B. Then, the asymptotic variance-covariance matrix of the parameter estimates can be obtained by inverting the observed Fisher information matrix, i.e.,

$$V(\theta, \omega) = [\mathcal{I}(\theta, \omega)]^{-1} = \begin{bmatrix} \widehat{Var}(\hat{\theta}) & \widehat{Cov}(\hat{\theta}, \hat{\omega}) \\ & \widehat{Var}(\hat{\omega}) \end{bmatrix}. \quad (3.3)$$

The expected Fisher information matrix can also be considered here, and the elements of the expected Fisher information matrix are presented in Appendix C.

3.1.2. The EM algorithm. In the perspective of parameter estimation for zero-inflated models, it is well known that the EM-algorithm works out nicely [3,13,22]. Instead of using numerical methods for the direct maximization of the likelihood function, it is typical to treat the estimation of parameters in the ZIB distribution as a missing data problem and solve it by using the EM algorithm. Specifically, we define an indicator variable, J_i, to entitle if X_i is an observation from the zero population or an observation from the binomial population, i.e.,

$$J_i = \begin{cases} 0, & \text{if } X_i \text{ belongs to the zero population;} \\ 1, & \text{if } X_i \text{ belongs to the binomial population.} \end{cases}$$

Note that J_i $(i = 1, 2, \ldots, N)$ is a latent variable that cannot be observed when $X_i = 0$.

The conditional expectation of the latent variable given X_i can be expressed as

$$E(J_i | X_i = 0) = 1 \text{ and } E(J_i | X_i > 0) = \frac{\omega}{\omega + (1 - \omega)(1 - \theta)^{n_i}}.$$

The likelihood function of ω and θ based on complete data (denoted as $\mathcal{D}$) is

$$L_C(\omega,\theta;\mathcal{D}) = \prod_{i=1}^{N} \omega^{(1-j_i)}(1-\omega)^{j_i}\left[\binom{n_i}{x_i}\theta^{x_i}(1-\theta)^{n_i-x_i}\right]^{j_i}.$$

Hence, the log-likelihood function based on the complete data is

$$\begin{aligned}\ell_C(\omega,\theta;\mathcal{D}) &= \ln L(\omega,\theta;\mathcal{D})\\ &= \left(N-\sum_{i=1}^{N} j_i\right)\ln\omega + \ln(1-\omega)\sum_{i=1}^{N} j_i + \sum_{i=1}^{N}\ln\binom{n_i}{x_i} j_i\\ &\quad + \left(\sum_{i=1}^{N} x_i j_i\right)\ln\theta + \left[\sum_{i=1}^{N}(n_i-x_i)j_i\right]\ln(1-\theta). \qquad (3.4)\end{aligned}$$

The first derivatives of the log-likelihood function with respect to ω and θ, respectively, are

$$S_\omega(\omega,\theta;\mathcal{D}) = \frac{\partial\ell(\omega,\theta;\mathcal{D})}{\partial\omega} = \frac{1}{\omega}\left(N-\sum_{i=1}^{N} j_i\right) - \frac{1}{(1-\omega)}\sum_{i=1}^{N} j_i \qquad (3.5)$$

$$\text{and } S_\theta(\omega,\theta;\mathcal{D}) = \frac{\partial\ell(\omega,\theta;\mathcal{D})}{\partial\theta} = \frac{1}{\theta}\sum_{i=1}^{N} x_i j_i - \frac{1}{(1-\theta)}\left(\sum_{i=1}^{N} n_i j_i - \sum_{i=1}^{N} x_i j_i\right). \qquad (3.6)$$

Furthermore, the negative of the second derivatives of the log-likelihood function with respect to ω and θ can be obtained as

$$B_{\omega\omega}(\omega,\theta;\mathcal{D}) = -\frac{\partial^2\ell(\omega,\theta;\mathcal{D})}{\partial\omega^2} = \frac{1}{\omega^2}\left(N-\sum_{i=1}^{N} j_i\right) + \frac{1}{(1-\omega)^2}\sum_{i=1}^{N} j_i, \qquad (3.7)$$

$$B_{\theta\theta}(\omega,\theta;\mathcal{D}) = -\frac{\partial^2\ell(\omega,\theta;\mathcal{D})}{\partial\theta^2} = \frac{1}{\theta^2}\sum_{i=1}^{N} x_i j_i + \frac{1}{(1-\theta)^2}\left(\sum_{i=1}^{N} n_i j_i - \sum_{i=1}^{N} x_i j_i\right), \qquad (3.8)$$

$$B_{\omega\theta}(\omega,\theta;\mathcal{D}) = B_{\theta\omega}(\omega,\theta;\mathcal{D}) = -\frac{\partial^2\ell(\omega,\theta;\mathcal{D})}{\partial\omega\partial\theta} = 0. \qquad (3.9)$$

Here, we take advantage of the existence of explicit solutions to the complete likelihood equations and propose an EM algorithm as an alternative way to obtain the MLEs of ω and θ.

Suppose $\hat{\omega}^{(0)}$ and $\hat{\theta}^{(0)}$ are the initial estimates of the parameters ω and θ, the EM algorithm can be described as follows:

Step 1. Given the current estimates $\hat{\omega}^{(h)}$ and $\hat{\theta}^{(h)}$ in the h-th iteration, J_i $(i=1,2,\ldots,\alpha)$ follows a Bernoulli distribution

$$\Pr(J_i=0) = q_i^{(h)}, \text{ and } \Pr(J_i=1) = 1-q_i^{(h)}$$

for $X_i=0$ $(i=1,2,\ldots,\alpha)$, where

$$q_i^{(h)} = \frac{\hat{\omega}^{(h)}}{\hat{\omega}^{(h)} + (1-\hat{\omega}^{(h)})(1-\hat{\theta}^{(h)})^{n_i}}.$$

In the E-step, we compute

$$\hat{\gamma}_i = E(J_i|X_i>0,\hat{\theta}^{(h)},\hat{\omega}^{(h)}) = 1-q_i^{(h)}.$$

Step 2. In the M-step, based on the solutions of Eqs. (3.5) and (3.6) we have

$$\omega = 1-\frac{\sum_{i=1}^{N} j_i}{N} \text{ and } \theta = \frac{\sum_{i=1}^{N} x_i j_i}{\sum_{i=1}^{N} n_i j_i}.$$

Thus, the updated estimates of parameters ω and θ can be computed as

$$\hat{\omega}^{(h+1)} = 1 - \frac{\sum_{i=1}^{\alpha} \hat{\gamma}_i + (N-\alpha)}{N} \text{ and } \hat{\theta}^{(h+1)} = \frac{\sum_{i=\alpha+1}^{N} x_i}{\sum_{i=1}^{\alpha} n_i \hat{\gamma}_i + \sum_{i=\alpha+1}^{N} n_i}.$$

Step 3. Repeat Steps 1 and 2 until convergence occurs. For example, convergence can be defined as

$$\max(|\hat{\theta}^{(h+1)} - \hat{\theta}^{(h)}|, |\hat{\omega}^{(h+1)} - \hat{\omega}^{(h)}|) < \varepsilon,$$

for a small value of ε.

Once the MLEs of ω and θ are obtained from the EM-algorithm, the observed information matrix and the corresponding variance-covariance matrix can be computed based on Eqs. (3.2) and (3.3).

3.1.3. Confidence intervals of the parameters. Since both the parameters ω and θ are in (0, 1), we consider the normal approximated $100(1-\delta)\%$ confidence intervals of ω and θ based on a logit-transformation. That is,

$$\text{logit}^{-1}\left[\ln\left(\frac{\hat{\omega}}{1-\hat{\omega}}\right) \pm z_{1-\delta/2}\frac{\sqrt{V_{\omega\omega}}}{\hat{\omega}(1-\hat{\omega})}\right] \tag{3.10}$$

and

$$\text{logit}^{-1}\left[\ln\left(\frac{\hat{\theta}}{1-\hat{\theta}}\right) \pm z_{1-\delta/2}\frac{\sqrt{V_{\theta\theta}}}{\hat{\theta}(1-\hat{\theta})}\right], \tag{3.11}$$

where z_q is the upper q-th percentile of the standard normal distribution, $\text{logit}^{-1}(y) = 1/(1+e^{-y})$ is the inverse logit function, $V_{\theta\theta} = \widehat{Var}(\hat{\theta})$, and $V_{\omega\omega} = \widehat{Var}(\hat{\omega})$ from Eq. (3.3).

In addition to constructing confidence intervals of the parameters based on normal approximation, bootstrap procedures based on maximum likelihood estimates can also be used. The following bootstrap procedure is used to compute the bootstrap confidence intervals for parameters ω and θ:

Step A1. Based on the observed data $(x_1, x_2, \ldots, x_N)$, the MLEs of parameters θ and ω in the ZIB model are obtained by the method described in Section 3.1.1 or 3.1.2. The MLEs of θ and ω are denoted as $\hat{\theta}_{obs}$ and $\hat{\omega}_{obs}$.

Step A2. Generate the random variate $x_i^{(b)}$ from $\text{ZIB}(n_i, \hat{\theta}_{obs}, \hat{\omega}_{obs})$, $i = 1, 2, \ldots, N$. Denote the bootstrap sample as $\boldsymbol{x}^{(b)} = (x_1^{(b)}, x_2^{(b)}, \ldots, x_N^{(b)})$.

Step A3. Based on the bootstrap sample $\boldsymbol{x}^{(b)} = (x_1^{(b)}, x_2^{(b)}, \ldots, x_N^{(b)})$, the MLEs of parameters θ and ω in the ZIB model are obtained by the method described in Section 3.1.1 or 3.1.2, and denotes the estimates as $\hat{\theta}^{(b)}$ and $\hat{\omega}^{(b)}$.

Step A4. Repeat Steps A2 and A3 B times to obtain $\hat{\omega}^{(1)}, \hat{\omega}^{(2)}, \ldots, \hat{\omega}^{(B)}$ and $\hat{\theta}^{(1)}, \hat{\theta}^{(2)}, \ldots, \hat{\theta}^{(B)}$.

Step A5. Arrange the sequences of the bootstrap estimates in ascending order to obtain $\hat{\omega}^{[1]} < \hat{\omega}^{[2]} < \ldots < \hat{\omega}^{[B]}$ and $\hat{\theta}^{[1]} < \hat{\theta}^{[2]} < \ldots < \hat{\theta}^{[B]}$.

Step A6. The $100(1-\delta)\%$ bootstrap confidence intervals of ω and θ are $(\hat{\omega}^{[B(\delta/2)]}, \hat{\omega}^{[B(1-\delta/2)]})$ and $(\hat{\theta}^{[B(\delta/2)]}, \hat{\theta}^{[B(1-\delta/2)]})$, respectively, where $[a]$ denotes the integer part of a.

3.2. Hypothesis testing for excess zeros in the ZIB model

To test the hypotheses in Eq. (2.5) based on the likelihood approach, we consider an LRT, a score test, and a bootstrap test in the following subsections.

3.2.1. Likelihood ratio test. Under the null hypothesis that $\omega = 0$, the MLE of θ can be readily obtained as

$$\hat{\theta}_0 = \frac{\sum_{i=1}^{N} x_i}{\sum_{i=1}^{N} n_i},$$

and the maximum log-likelihood function can be obtained as $\ell(\hat{\theta}_0, 0)$ from Eq. (3.1). Under the alternative hypothesis that $\omega \neq 0$, the MLEs of θ and ω are $\hat{\theta}$ and $\hat{\omega}$ which can be obtained by direct maximization or by the EM algorithm as described in Sections 3.1.1 and 3.1.2, respectively. Subsequently, the maximum log-likelihood function can be obtained as $\ell(\hat{\theta}, \hat{\omega})$ from Eq. (3.1). Then, the LRT statistic can be computed as

$$\lambda = -2\big[\ell(\hat{\theta}_0, 0) - \ell(\hat{\theta}, \hat{\omega})\big]. \tag{3.12}$$

It is well-known that this LRT statistic follows a chi-square distribution with one degree of freedom. Since we are dealing with the null hypothesis for which the parameter value is at the boundary of the parameter space, an adjustment on the computation of p-value is needed [7, 25]. The p-value of the LRT for testing the hypotheses in Eq. (2.5) can be calculated as

$$p_{LR} \quad = \quad 1 - \Big[0.5 + 0.5\chi_1^2(\lambda)\Big], \tag{3.13}$$

where $\chi_1^2(\lambda)$ is the value of the cumulative distribution function a chi-square random variable with one degree of freedom evaluated at an observed test statistic λ in Eq. (3.12). The null hypothesis in Eq. (2.5) is rejected at the δ level if $p_{LR} < \delta$.

3.2.2. Score test. When testing the hypotheses in Eq. (2.5), the score test statistic developed under the null hypothesis that $\omega = 0$ based on the expected Fisher information matrix can be applied. From [10], the score test statistic can be expressed as

$$S^2 = \frac{\left\{\sum_{i=1}^{N}\Big[I(x_i = 0)(1-\hat{\theta}_0)^{-n_i} - 1\Big]\right\}^2}{\sum_{i=1}^{N}\Big[(1-\hat{\theta}_0)^{-n_i} - 1 - n_i\hat{\theta}_0(1-\hat{\theta}_0)^{-1}\Big]}.$$

Under the null hypothesis, the score test statistic S^2 follows a chi-square distribution with one degree of freedom. Hence, the p-value of the score for testing the hypotheses in Eq. (2.5) can be calculated as

$$p_S \quad = \quad 1 - \chi_1^2(S^2), \tag{3.14}$$

and the null hypothesis is rejected at the δ level if $p_S < \delta$.

3.2.3. Bootstrap test based on maximum likelihood estimates. In this subsection, we propose a bootstrap test based on maximum likelihood estimates. The following bootstrap procedure is used to compute the p-value.

Step B1. Based on the observed data $(x_1, x_2, \ldots, x_N)$, the MLEs of parameters θ and ω in the ZIB model are obtained by the method described in Section 3.1.1 or 3.1.2. The MLEs of θ and ω are denoted as $\hat{\theta}_{obs}$ and $\hat{\omega}_{obs}$.

Step B2. Under the null hypothesis that $\omega = 0$, generate the random variate $x_i^{(b)}$ from the binomial model in Eq. (1.2) with parameters $\theta = \hat{\theta}_0 = \sum_{i=1}^{N} x_i \Big/ \sum_{i=1}^{N} n_i$ and n_i, $i = 1, 2, \ldots, N$. Denote the bootstrap sample as $\boldsymbol{x}^{(b)} = (x_1^{(b)}, x_2^{(b)}, \ldots, x_N^{(b)})$.

Step B3. Based on the bootstrap sample $\boldsymbol{x}^{(b)} = (x_1^{(b)}, x_2^{(b)}, \ldots, x_N^{(b)})$, the MLEs of parameters θ and ω in the ZIB model are obtained by the method described in Section 3.1.1 or 3.1.2, and denotes the estimates as $\hat{\theta}^{(b)}$ and $\hat{\omega}^{(b)}$.

Step B4. Repeat Steps B2 and B3 B times to obtain $\hat{\omega}^{(1)}, \hat{\omega}^{(2)}, \ldots, \hat{\omega}^{(B)}$.

Step B5. The bootstrap p-value is computed as

$$p_B = \frac{1}{B} \sum_{b=1}^{B} I(\hat{\omega}^{(b)} > \hat{\omega}_{obs}).$$

The null hypothesis in Eq. (2.5) is rejected at the δ level if $p_B < \delta$.

4. Monte Carlo simulation study

In this section, a Monte Carlo simulation study is used to evaluate the performance of the Bayesian and frequentist approaches for point and interval parameter estimation and testing for excess zeros in the ZIB model. The datasets are simulated from the ZIB model with parameters $\theta = 0.05, 0.10, 0.15, 0.20, 0.30, 0.40, 0.50, 0.60$ and $\omega = 0, 0.05, 0.10, 0.15$. We consider the sample sizes $N = 30, n_i = 15$ for $i = 1, 2, \ldots, 30$ and $N = 50, n_i = 20$ for $i = 1, 2, \ldots, 50$.

The simulated biases and MSEs for point estimation of the parameters ω and θ are defined by

$$Bias = \frac{1}{M} \sum_{j=1}^{M} (\hat{\delta}_j - \delta) \text{ and } MSE = \frac{1}{M} \sum_{j=1}^{M} (\hat{\delta}_j - \delta)^2,$$

respectively, where M is the number of replications in the simulation. Here, $\delta = \theta$ or ω is the true value of the parameter, and $\hat{\delta}_j$ is the estimate of δ in the j^{th} replication. On the other hand, the simulated coverage probabilities (CP) and average widths (AW) of 95% confidence/credible intervals of the parameters ω and θ are defined by

$$CP = \frac{1}{M}\sum_{j=1}^{M} I(\delta \in [\delta_{Lj}, \delta_{Uj}]) \text{ and } AW = \frac{1}{M}\sum_{j=1}^{M}(\min\{1, \hat{\delta}_{Uj}\} - \max\{0, \hat{\delta}_{Lj}\}),$$

respectively. We note that δ_{Lj} is the lower bound of the interval for parameter δ, and δ_{Uj} is the upper bound of the interval for parameter δ in the j^{th} replication. We also note that the HPD intervals were calculated using a normal approximation. In particular, when calculating HPD intervals along with CP and AW, we used the normal approximation method proposed by [24]. This method enables us to conduct a comparative analysis between the frequentist approach and the Bayesian approach. Another method that can be used is to approximate the HPD intervals based on the expected Fisher information [9]. However, based on our preliminary study, there is no significant difference between these two approaches. The simulated rejection rates (at 5% level for frequentist procedures) of the testing procedures for excessive zeros based on both Bayesian and frequentist approaches are presented in Tables 1 and 2. A total of $M = 1000$ replications is used for entire simulations.

From Tables 1 and 2, we observe that the performance of point and interval estimates of the Bayesian and frequentist approaches are comparable. There are some situations where the Bayesian approach outperforms the frequentist approach and some that the frequentist approach outperforms the Bayesian approach. We use bold faces on promising values showing the best results in terms of the rejection rate. More precisely, among the Bayesian, LRT, score test and the bootstrap method, the best rejection rate that is greater than a value of 0.9 was boldfaced.

For point estimation, the MSEs of the MLE of ω are smaller than the MSEs of the Bayesian estimate of ω in most cases, while the MSEs of the MLE of θ are similar to the MSEs of the Bayesian estimate of θ. For interval estimation, except for the cases when the true value of θ is close to zero (i.e., $\theta = 0.05$ or 0.1), the simulated CP's of the interval estimates are at or above a nominal level of 95%. For testing the hypothesis, the procedure based on the Bayes factor B_{10} is more powerful in detecting the excessive zeros (i.e., $\omega > 0$) compared to the LRT and bootstrap test in general. We observe that the performance of the testing procedures for excess zeros depends on the true value of θ. Specifically, the testing procedures considered here, especially for the LRT and bootstrap test, are less powerful when θ is close to zero. Overall, the estimation and the hypothesis testing procedures discussed here are performing reasonably well in most cases. However, one should be cautious with making conclusions based on the estimation and the hypothesis testing procedures when the value of parameter θ is close to zero. In regards to calculation efficiency, we note that the total CPU time for one set of parameter configurations using the Bayesian method was 15.15 for $N = 25$ and 19.27 seconds for $N = 50$. On the other hand, it took almost the same computing times for all frequentist approaches, which was 1.73 seconds for each configuration.

Figures 1 and 2 show some graphical displays regarding the results of Tables 1 and 2, respectively. In these figures we provide several performance measures such as bias, MSE, CP, and AW with three different values of ω: 0.05, 0.1, and 0.15. There seem to be no remarkable differences among the three approaches. However, we can see that the bootstrap method yields slightly poor results on CP values as the value of θ increases. We also notice that the rejection rates vary among different approaches. That is, no particular method shows dominating results in all cases in terms of the rejection rate. Finally, Figure 3 shows comparison of rejection rates between two noninformative priors, i.e., the Jeffreys and uniform priors. There are not much differences in the rejection rates between the two priors.

Table 1. Simulated biases, MSEs for point estimation of the parameters ω and θ, simulated coverage probabilities (CP) and average widths (AW) of 95% credible intervals, asymptotic confidence intervals (ACI), and bootstrap confidence intervals (BCI) of the parameters ω and θ, and simulated rejection rates of the testing procedures for excess zeros based on both Bayesian and frequentist approaches for sample sizes $N = 25$ and $n_i = 15$ for $i = 1, 2, \ldots, 25$ with 1000 replications

		Objective Bayesian Approach									Frequentist Approach														
		Estimation of ω				Estimation of θ				Test $\omega = 0$	Estimation of ω							Estimation of θ						Test $\omega = 0$		
										B_{10}			ACI	ACI	BCI	BCI			ACI	ACI	BCI	BCI			
ω	θ	Bias	MSE	CP	AW	Bias	MSE	CP	AW	> 1	Bias	MSE	CP	AW	CP	AW	Bias	MSE	CP	AW	CP	AW	LRT	Boot	Score
0.00	0.05	0.204	0.048	0.868	0.504	0.014	0.000	0.958	0.062	0.143	0.089	0.027	0.000	0.920	0.000	0.487	0.006	0.000	0.958	0.082	0.959	0.067	0.039	0.046	0.048
0.00	0.10	0.115	0.016	0.954	0.296	0.011	0.000	0.952	0.069	0.049	0.042	0.006	0.000	0.880	0.000	0.275	0.006	0.000	0.972	0.080	0.967	0.073	0.032	0.033	0.050
0.00	0.15	0.077	0.007	0.975	0.174	0.008	0.000	0.959	0.076	0.031	0.026	0.002	0.000	0.848	0.000	0.172	0.006	0.000	0.948	0.082	0.965	0.078	0.043	0.049	0.037
0.00	0.20	0.057	0.004	0.990	0.140	0.005	0.000	0.956	0.083	0.016	0.015	0.001	0.000	0.842	0.000	0.112	0.003	0.000	0.955	0.086	0.950	0.084	0.037	0.042	0.043
0.00	0.30	0.041	0.002	1.000	0.140	0.004	0.001	0.955	0.092	0.005	0.000	0.000	0.000	0.992	0.000	0.049	0.000	0.001	0.952	0.099	0.928	0.092	0.010	0.036	0.077
0.00	0.40	0.037	0.001	1.000	0.140	0.001	0.001	0.950	0.099	0.010	0.004	0.000	0.000	0.923	0.000	0.005	0.001	0.001	0.957	0.093	0.938	0.093	0.035	0.048	0.010
0.00	0.50	0.037	0.001	1.000	0.140	-0.003	0.001	0.951	0.101	0.001	0.000	0.000	0.000	0.998	0.000	0.000	0.002	0.001	0.952	0.101	0.951	0.099	0.003	0.091	0.001
0.00	0.60	0.037	0.001	1.000	0.140	-0.001	0.001	0.954	0.099	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.001	0.950	0.099	0.953	0.105	0.000	0.064	0.000
0.05	0.05	0.174	0.038	0.928	0.377	0.014	0.000	0.980	0.057	0.172	0.065	0.027	0.845	0.903	0.989	0.514	0.006	0.000	0.963	0.085	0.978	0.069	0.054	0.058	0.042
0.05	0.10	0.089	0.012	0.958	0.314	0.010	0.000	0.968	0.074	0.132	0.026	0.009	0.887	0.817	0.987	0.310	0.003	0.000	0.979	0.083	0.971	0.075	0.082	0.102	0.049
0.05	0.15	0.056	0.006	0.978	0.267	0.006	0.000	0.970	0.084	0.133	0.007	0.004	0.933	0.719	0.993	0.219	0.001	0.000	0.962	0.085	0.958	0.081	0.122	0.149	0.110
0.05	0.20	0.042	0.004	0.987	0.235	0.004	0.001	0.965	0.089	0.191	0.004	0.003	0.939	0.600	0.991	0.169	0.001	0.000	0.964	0.089	0.936	0.087	0.238	0.251	0.225
0.05	0.30	0.035	0.003	0.991	0.194	0.003	0.001	0.956	0.097	0.407	0.001	0.002	0.959	0.452	0.772	0.140	0.000	0.001	0.955	0.096	0.949	0.096	0.513	0.571	0.657
0.05	0.40	0.035	0.003	0.994	0.194	0.003	0.001	0.958	0.102	0.729	0.000	0.002	0.971	0.440	0.696	0.122	0.000	0.001	0.955	0.102	0.938	0.100	0.743	0.752	0.742
0.05	0.50	0.035	0.003	0.994	0.194	0.000	0.001	0.954	0.103	0.734	0.003	0.002	0.950	0.437	0.702	0.125	0.000	0.001	0.952	0.104	0.948	0.103	0.747	0.768	0.741
0.05	0.60	0.035	0.003	0.994	0.194	0.000	0.001	0.960	0.102	0.734	0.002	0.002	0.966	0.454	0.715	0.123	-0.001	0.001	0.949	0.101	0.949	0.103	0.724	0.747	0.741
0.10	0.05	0.148	0.032	0.938	0.377	0.013	0.000	0.979	0.057	0.238	0.028	0.027	0.890	0.899	0.991	0.541	0.004	0.000	0.955	0.088	0.978	0.071	0.062	0.067	0.033
0.10	0.10	0.066	0.009	0.975	0.314	0.007	0.000	0.962	0.074	0.207	0.006	0.010	0.929	0.760	0.988	0.348	0.001	0.000	0.968	0.086	0.964	0.077	0.133	0.147	0.085
0.10	0.15	0.040	0.006	0.976	0.267	0.004	0.000	0.953	0.084	0.308	-0.003	0.006	0.958	0.604	0.982	0.271	0.001	0.000	0.955	0.088	0.959	0.085	0.284	0.296	0.211
0.10	0.20	0.032	0.005	0.985	0.235	0.002	0.001	0.953	0.089	0.453	-0.005	0.005	0.958	0.468	0.918	0.234	0.001	0.001	0.953	0.092	0.952	0.091	0.485	0.492	0.460
0.10	0.30	0.031	0.004	0.993	0.194	0.002	0.001	0.958	0.097	0.740	0.000	0.004	0.956	0.326	0.933	0.209	-0.001	0.001	0.956	0.099	0.946	0.099	0.804	0.830	0.300
0.10	0.40	0.032	0.004	0.995	0.194	0.002	0.001	0.951	0.102	**0.938**	-0.001	0.003	0.967	0.323	0.926	0.202	0.000	0.001	0.950	0.104	0.940	0.104	0.923	0.923	0.934
0.10	0.50	0.032	0.004	0.995	0.194	0.001	0.001	0.955	0.103	**0.942**	0.003	0.004	0.961	0.319	0.922	0.204	-0.002	0.001	0.937	0.107	0.947	0.106	0.929	0.940	0.934
0.10	0.60	0.032	0.004	0.995	0.194	0.000	0.001	0.952	0.102	**0.942**	0.000	0.004	0.963	0.326	0.902	0.201	-0.001	0.001	0.951	0.104	0.945	0.105	0.918	0.924	0.934
0.15	0.05	0.119	0.025	0.958	0.471	0.012	0.000	0.987	0.062	0.293	0.005	0.034	0.925	0.883	0.988	0.582	0.003	0.000	0.968	0.091	0.985	0.075	0.085	0.090	0.053
0.15	0.10	0.045	0.009	0.977	0.350	0.006	0.000	0.967	0.075	0.312	-0.016	0.014	0.950	0.729	0.991	0.387	0.000	0.000	0.972	0.088	0.965	0.080	0.205	0.216	0.134
0.15	0.15	0.028	0.007	0.955	0.303	0.003	0.001	0.952	0.087	0.478	-0.010	0.009	0.962	0.534	0.946	0.320	-0.001	0.001	0.951	0.091	0.947	0.089	0.430	0.439	0.355
0.15	0.20	0.024	0.006	0.949	0.271	0.002	0.001	0.945	0.091	0.683	-0.003	0.007	0.957	0.387	0.914	0.285	0.000	0.001	0.946	0.095	0.965	0.094	0.718	0.730	0.667
0.15	0.30	0.026	0.005	0.959	0.264	0.001	0.001	0.958	0.102	0.899	-0.002	0.005	0.971	0.307	0.895	0.263	0.000	0.001	0.954	0.102	0.941	0.103	0.936	0.945	**0.955**
0.15	0.40	0.027	0.005	0.961	0.263	0.002	0.001	0.949	0.107	0.982	-0.002	0.005	0.975	0.305	0.891	0.255	-0.001	0.001	0.960	0.107	0.928	0.108	0.977	0.978	**0.984**
0.15	0.50	0.027	0.005	0.960	0.263	0.001	0.001	0.952	0.108	**0.985**	0.000	0.005	0.974	0.303	0.910	0.256	0.001	0.001	0.938	0.109	0.95	0.110	0.979	0.981	0.984
0.15	0.60	0.027	0.005	0.960	0.263	-0.001	0.001	0.949	0.106	**0.985**	-0.001	0.005	0.980	0.306	0.914	0.262	-0.001	0.001	0.938	0.107	0.949	0.108	0.975	0.977	0.984

Table 2. Simulated biases, MSEs for point estimation of the parameters ω and θ, simulated coverage probabilities (CP) and average widths (AW) of 95% credible intervals, asymptotic confidence intervals (ACI), and bootstrap confidence intervals (BCI) of the parameters ω and θ, and simulated rejection rates of the testing procedures for excess zeros based on both Bayesian and frequentist approaches for sample sizes $N = 50$ and $n_i = 20$ for $i = 1, 2, \ldots, 50$ with 1000 replications

		Objective Bayesian Approach									Frequentist Approach														
		Estimation of ω				Estimation of θ				Test $\omega = 0$	Estimation of ω						Estimation of θ						Test $\omega = 0$		
										B_{10}			ACI	ACI	BCI	BCI			ACI	ACI	BCI	BCI			
ω	θ	Bias	MSE	CP	AW	Bias	MSE	CP	AW	> 1	Bias	MSE	CP	AW	CP	AW	Bias	MSE	CP	AW	CP	AW	LRT	Boot	Score
0.00	0.05	0.124	0.019	0.913	0.295	0.007	0.000	0.961	0.021	0.076	0.051	0.009	0.000	0.880	0.000	0.297	0.003	0.000	0.964	0.042	0.966	0.036	0.041	0.050	0.053
0.00	0.10	0.063	0.005	0.955	0.128	0.006	0.000	0.915	0.040	0.036	0.022	0.002	0.000	0.823	0.000	0.145	0.002	0.000	0.960	0.044	0.952	0.041	0.044	0.054	0.047
0.00	0.15	0.038	0.002	0.985	0.074	0.003	0.000	0.896	0.082	0.017	0.012	0.000	0.000	0.791	0.000	0.080	0.002	0.000	0.961	0.047	0.949	0.046	0.038	0.046	0.047
0.00	0.20	0.027	0.001	0.997	0.074	0.002	0.000	0.911	0.071	0.008	0.006	0.000	0.000	0.755	0.000	0.047	0.002	0.000	0.942	0.051	0.933	0.049	0.042	0.052	0.044
0.00	0.30	0.020	0.000	1.000	0.074	0.000	0.000	0.911	0.059	0.003	0.001	0.000	0.000	0.968	0.000	0.015	0.001	0.000	0.957	0.057	0.949	0.057	0.037	0.058	0.035
0.00	0.40	0.019	0.000	1.000	0.074	0.000	0.000	0.951	0.059	0.004	0.000	0.000	0.000	0.997	0.000	0.000	0.000	0.000	0.950	0.061	0.949	0.062	0.003	0.068	0.002
0.00	0.50	0.019	0.000	1.000	0.074	0.000	0.000	0.947	0.062	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.927	0.062	0.931	0.062	0.000	0.059	0.000
0.00	0.60	0.019	0.000	1.000	0.074	0.000	0.000	0.952	0.059	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.945	0.061	0.948	0.060	0.000	0.053	0.000
0.05	0.05	0.098	0.014	0.946	0.191	0.006	0.000	0.973	0.017	0.125	0.031	0.010	0.861	0.825	0.977	0.328	0.002	0.000	0.965	0.044	0.963	0.037	0.083	0.087	0.050
0.05	0.10	0.042	0.004	0.975	0.150	0.004	0.000	0.943	0.019	0.169	0.007	0.003	0.909	0.631	0.985	0.181	0.001	0.000	0.964	0.046	0.956	0.043	0.176	0.190	0.108
0.05	0.15	0.023	0.002	0.982	0.117	0.001	0.000	0.922	0.067	0.257	0.002	0.002	0.943	0.404	0.985	0.136	0.000	0.000	0.953	0.049	0.944	0.048	0.377	0.380	0.354
0.05	0.20	0.019	0.001	0.986	0.102	0.001	0.000	0.902	0.031	0.498	0.000	0.001	0.970	0.267	0.913	0.114	0.001	0.000	0.948	0.052	0.945	0.052	0.621	0.661	0.640
0.05	0.30	0.019	0.001	0.983	0.104	0.001	0.000	0.932	0.059	0.760	0.000	0.001	0.958	0.212	0.912	0.106	0.000	0.000	0.942	0.058	0.948	0.058	0.922	0.921	0.921
0.05	0.40	0.019	0.001	0.984	0.104	0.003	0.003	0.961	0.064	**0.935**	-0.001	0.001	0.960	0.220	0.910	0.103	0.000	0.000	0.955	0.062	0.944	0.063	0.913	0.920	0.920
0.05	0.50	0.019	0.001	0.984	0.104	0.001	0.000	0.964	0.063	0.935	0.000	0.001	0.965	0.202	0.903	0.104	0.001	0.000	0.955	0.064	0.956	0.064	0.935	**0.940**	0.920
0.05	0.60	0.019	0.001	0.984	0.104	0.003	0.008	0.961	0.064	**0.935**	0.000	0.001	0.952	0.221	0.910	0.106	0.000	0.000	0.962	0.062	0.934	0.062	0.914	0.922	0.920
0.10	0.05	0.075	0.011	0.967	0.274	0.006	0.000	0.962	0.020	0.224	0.005	0.012	0.919	0.794	0.986	0.359	0.001	0.000	0.968	0.045	0.968	0.039	0.110	0.120	0.067
0.10	0.10	0.027	0.004	0.965	0.191	0.003	0.000	0.944	0.031	0.375	-0.003	0.005	0.941	0.487	0.978	0.227	0.000	0.000	0.960	0.047	0.934	0.045	0.352	0.363	0.285
0.10	0.15	0.016	0.003	0.956	0.178	0.001	0.000	0.916	0.058	0.639	-0.002	0.003	0.966	0.275	0.915	0.183	-0.001	0.000	0.943	0.050	0.946	0.050	0.719	0.736	0.697
0.10	0.20	0.017	0.003	0.960	0.161	0.001	0.001	0.882	0.051	0.867	0.001	0.002	0.966	0.202	0.920	0.167	0.000	0.000	0.940	0.054	0.943	0.054	0.930	0.936	**0.940**
0.10	0.30	0.017	0.002	0.943	0.159	0.002	0.001	0.939	0.070	0.972	0.001	0.002	0.972	0.179	0.888	0.158	0.000	0.000	0.960	0.060	0.955	0.060	0.997	0.997	**0.999**
0.10	0.40	0.017	0.002	0.943	0.159	0.000	0.000	0.944	0.066	**0.995**	-0.001	0.002	0.981	0.179	0.900	0.159	0.001	0.000	0.950	0.064	0.946	0.064	0.994	0.994	**0.999**
0.10	0.50	0.017	0.002	0.943	0.159	0.000	0.000	0.955	0.065	0.995	-0.001	0.002	0.982	0.181	0.923	0.156	0.000	0.000	0.960	0.065	0.947	0.065	0.992	0.993	**0.999**
0.10	0.60	0.017	0.002	0.943	0.159	-0.001	0.001	0.944	0.066	0.995	0.002	0.002	0.967	0.181	0.920	0.155	0.000	0.000	0.946	0.064	0.945	0.064	0.994	0.994	**0.999**
0.15	0.05	0.053	0.010	0.960	0.292	0.005	0.000	0.964	0.022	0.300	-0.008	0.015	0.933	0.741	0.984	0.397	0.000	0.000	0.971	0.046	0.977	0.041	0.179	0.199	0.104
0.15	0.10	0.017	0.005	0.941	0.231	0.002	0.000	0.927	0.047	0.599	-0.005	0.006	0.950	0.394	0.936	0.266	0.000	0.000	0.955	0.049	0.939	0.047	0.586	0.613	0.488
0.15	0.15	0.012	0.003	0.939	0.192	0.000	0.000	0.906	0.038	0.876	0.001	0.003	0.968	0.243	0.927	0.218	0.000	0.000	0.950	0.052	0.955	0.052	0.929	0.935	0.898
0.15	0.20	0.014	0.003	0.949	0.187	0.001	0.001	0.906	0.066	0.979	0.002	0.003	0.968	0.211	0.930	0.199	0.000	0.000	0.945	0.056	0.950	0.055	0.989	0.989	**0.991**
0.15	0.30	0.015	0.003	0.945	0.184	0.001	0.001	0.959	0.073	0.999	-0.001	0.003	0.941	0.201	0.932	0.192	-0.001	0.000	0.948	0.062	0.946	0.062	0.998	0.998	**1.000**
0.15	0.40	0.015	0.003	0.944	0.184	0.005	0.004	0.954	0.067	**1.000**	0.001	0.002	0.965	0.200	0.930	0.191	0.000	0.000	0.940	0.066	0.939	0.066	1.000	**1.000**	**1.000**
0.15	0.50	0.015	0.003	0.944	0.184	0.000	0.000	0.955	0.066	**1.000**	-0.002	0.002	0.956	0.067	0.921	0.190	0.000	0.000	0.956	0.067	0.949	0.067	0.999	0.999	**1.000**
0.15	0.60	0.015	0.003	0.944	0.184	0.005	0.009	0.954	0.067	**1.000**	-0.001	0.003	0.954	0.199	0.930	0.190	0.000	0.000	0.953	0.066	0.932	0.066	1.000	**1.000**	**1.000**

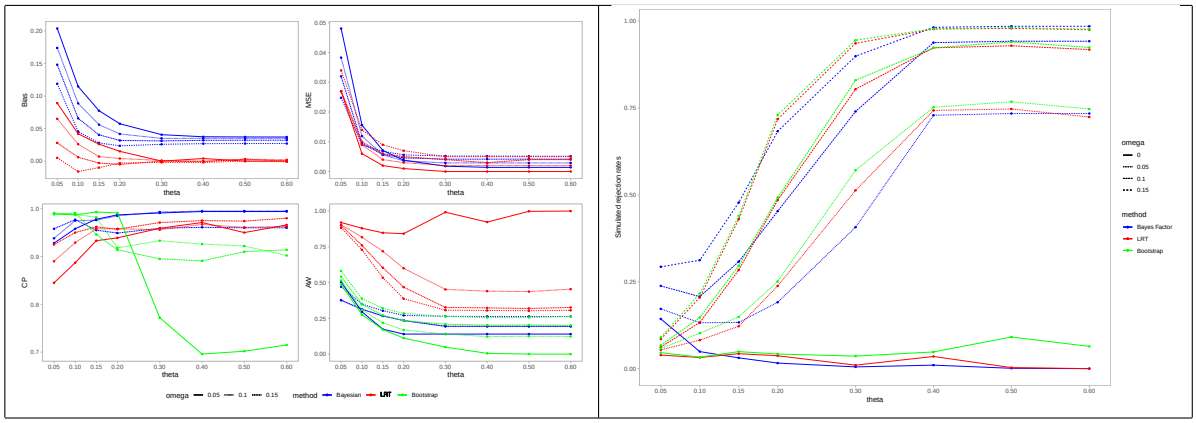

Figure 1. Bias, MSE, CP, and AW of the Bayesian, LRT, and bootstrap methods with different values of ω and θ (left panel), and the rejection rate of the three methods (right panel) based on the results of Table 1

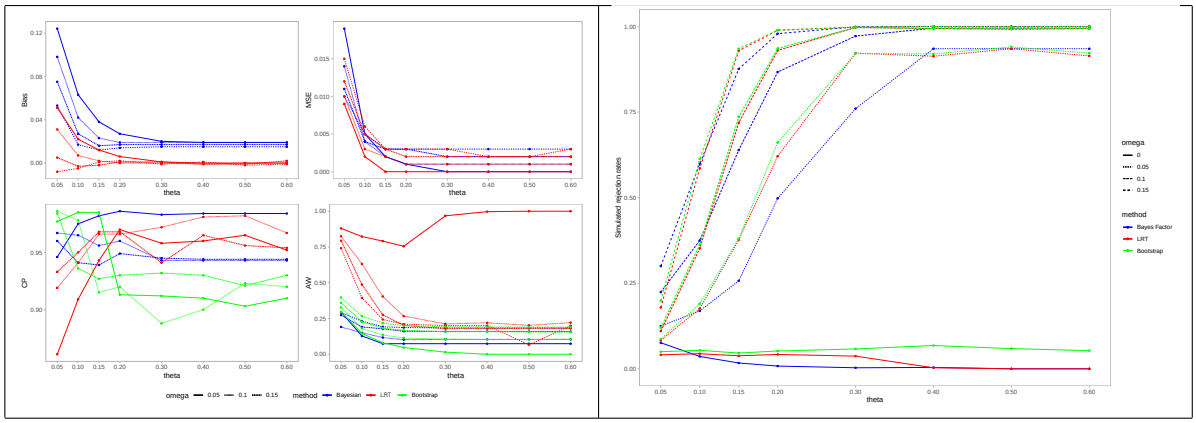

Figure 2. Bias, MSE, CP, and AW of the Bayesian, LRT, and bootstrap methods with different values of ω and θ (left panel), and the rejection rate of the three methods (right panel) based on the results of Table 2

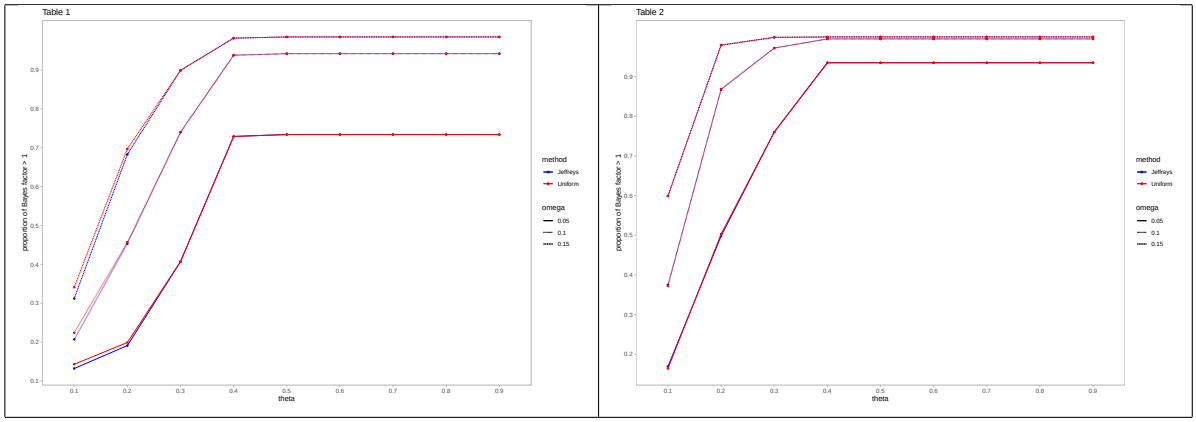

Figure 3. The rejection rate calculated with the uniform and Jeffreys priors based on the results of Table 1 (left panel) and Table 2 (right panel)

5. Practical data analyses

In this section, we illustrate the proposed methodologies by using two real datasets that may have excessive zeros.

5.1. Earthquake data in South Korea

There are two measures in characterizing the size of an earthquake: magnitude and intensity. Magnitude is the most general measure to estimate the scale of an earthquake. The magnitude is an absolute figure regardless of the location, which is an indicator of the amplitude recorded on the seismograph at each station and considers various factors such as epicentral depth and epicentral distance. On the other hand, intensity is a relative figure that varies according to the location, which is the measure that indicates the size of the tremor in a particular location. Here, we consider a dataset on the number of earthquakes in South Korea over the last 43 years in which the data of earthquakes are classified by magnitude. Specifically, we collect the number of earthquakes with a magnitude of 5.0 or larger that have occurred in South Korea. See the data from the Korea Meteorological Administration website at `https://www.weather.go.kr/w/eqk-vol/search/korea.do`. The dataset consists of the number of earthquakes in a duration of six months (January–June or July–December) from September 1978 to December 2020 and the number of earthquakes with a magnitude > 5.0 during a six-month period. Since there is no earthquake until September in 1978, there are $2 \times 43 - 1 = 85$ observations with n_i being the number of earthquakes in a six-month period and x_i being the number of earthquakes with magnitude > 5.0 out of n_i earthquakes for $i = 1, 2, \ldots, 85$. Among $N = 85$ observations, there are 78 zeros. The dataset is presented in Table 3.

Based on the data presented in Table 3, we apply the Bayesian and frequentist approaches described in Sections 2 and 3 to estimate the parameters of the ZIB model and test for excess zeros. In the EM-algorithm, we use the tolerance limit with $\varepsilon = 10^{-8}$ for convergence and it takes 537 iterations to convergence. For the observed Fisher information matrix based on the missing information principle, we use $K = 10^6$ simulations for the approximation. The bootstrap test based on the maximum likelihood estimate is conducted using $B = 10000$ bootstrap samples. The results of the data analysis are presented in Table 4. From Table 4, we observe that both the Bayesian and frequentist approaches yield similar parameter estimates. Since the Bayes factor is $B_{10} = 1.3210 > 1.0$, we select the ZIB model with excess zeros using the Bayesian approach. However, based on the LRT, the score test, and the bootstrap test, the p-values are greater than 0.05, showing there is no sufficient evidence to support the ZIB model at the 5% level.

The LRT, score test, and bootstrap test may not effectively detect the zero-inflation for this dataset. One possible reason for the low power (i.e., ability to detect zero-inflation) in the LRT and bootstrap tests is that the estimate of θ is very close to 0. Although we do not know the true value of θ for this dataset, based on the estimate of θ, we believe that the true value of θ should be close to 0. As we observed in the simulation study, this discrepancy might be due to the poor power performance of the frequentist testing procedures when the value of θ is close to zero. For instance, when $\theta = 0.05$, the power is only 0.085 (i.e., making the right decision to claim zero-inflation 8.5% of times) even when the true value of ω is 0.15.

5.2. Major league baseball data

In this subsection we analyze an offensive measure of a player in Major League Baseball (MLB) for fitting the ZIB. It is known that extra-base hits are seldom produced by non-power hitters, and thus these can be regarded as zero-inflated. Extra base hits consist of two-base hits, three-base hits, and home runs. As mentioned in Introduction, three-base hits are very rare to occur so that we intended to use these at the initial stage of this

Table 3. Earthquake data in South Korea from 1978 – 2021

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
n_i	6	11	11	4	12	9	6	8	5	10	10	14	5	18	8
x_i	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0
i	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
n_i	13	2	2	9	3	3	14	2	14	1	13	6	4	11	16
x_i	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
i	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45
n_i	6	13	11	15	14	22	17	13	8	18	14	30	7	13	16
x_i	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
i	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
n_i	23	18	21	28	24	14	24	18	21	16	28	22	19	23	19
x_i	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0
i	61	62	63	64	65	66	67	68	69	70	71	72	73	74	75
n_i	27	39	21	24	18	30	22	31	25	50	43	20	29	18	26
x_i	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0
i	76	77	78	79	80	81	82	83	84	85					
n_i	33	219	89	134	67	48	45	43	34	34					
x_i	0	3	0	1	0	0	0	0	0	0					

Table 4. Point and interval estimates of the ZIB model parameters and results for testing excess zeros based on the earthquake data presented in Table 3

Bayesian Approach		Frequentist Approach		
			Direct Maximization	EM Algorithm
$\tilde{\theta}$	0.00890	$\hat{\theta}$	0.00915	0.00915
$Var(\tilde{\theta})$	0.00002	$Var(\hat{\theta})$	0.00003	0.00003
95% HPD interval for θ	(0.00000, 0.01811)	95% ACI for θ	(0.00302, 0.02740)	(0.00296, 0.02791)
		95% BCI for θ	(0.00205, 0.02880)	
		$\hat{\theta}_0$	0.005118	0.00512
$\tilde{\omega}$	0.3995	$\hat{\omega}$	0.496039	0.49601
$Var(\tilde{\omega})$	0.04757	$Var(\hat{\omega})$	0.092432	0.09293
95% HPD interval for ω	(0.00000, 0.82691)	95% ACI for ω	(0.08321, 0.91434)	(0.08272, 0.91483)
		95% BCI for ω	(0.00000, 0.85780)	
Bayes factor B_{10}	1.3210	$\ell(\hat{\theta}, \hat{\omega})$	-29.46504	-29.46504
		$\ell(\hat{\theta}_0, 0)$	−30.01913	
		LRT Stat. λ	1.10817	
		p_{LR}	0.14624	
		Score Test Stat. S^2	1.14518	
		p_S	0.28456	
		Bootstrap Test p_B	0.13700	

research. However, there are too many zeros in three-base hits, making the fitting of data virtually impossible. We decided to use extra-base hits as zero-inflated for our analysis. In a regular season of MLB, each team plays a total of 162 games. A player should have at least three plate appearances unless he is taken out of the game. It is known that a span of five games is a reasonable group for change point analysis [1,18]. We use a span of four games for making one observation in our analysis. Let n_i be the number of times in the plate and let x_i be the total number of extra-base hits. Players sometimes take a rest or do not play all games due to injuries or some other reasons, yielding different values of n_i for $i = 1, 2, \ldots, N$.

We analyze the data of Brandon Crawford from the San Francisco Giants in the Year 2019. The data are readily available online: `https://www.baseball-reference.com`. Table 5 shows n_i and x_i values in which there is a total of $N = 37$ observations, and 17

observations of 37 are zeros. As appeared in the earthquake data, the parameter estimates with both approaches are close to each other. In particular, the estimates of θ are very close while there is a little difference between the MLE and the Bayes estimate for ω. From a testing perspective, the Bayes factor B_{10} turned out to be 2.2725, which corresponds to a fairly small p-value of 0.0324 for the LRT. This implies that both procedures decently support the ZIB model with excessive zeros. Notice that the estimates for (θ, ω) are around $(0.1, 0.28)$. Although simulation results with this configuration are not presented in Tables 1 and 2, the corresponding outcomes are congruent with the simulation studies regarding hypothesis testing.

Table 5. MLB data for Brandon Crawford

i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
n_i	12	15	15	12	15	12	14	14	13	14	15	13	13	15	10
x_i	1	1	0	0	0	0	0	0	2	1	2	1	0	4	0
i	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30
n_i	16	15	15	13	12	16	15	15	14	13	17	14	11	13	15
x_i	0	3	3	3	1	4	2	0	0	1	0	1	0	1	2
i	31	32	33	34	35	36	37								
n_i	12	13	10	12	16	13	8								
x_i	1	1	0	0	2	0	0								

Table 6. Point and interval estimates of the ZIB model parameters and results for testing of excess zeros based on the MLB data presented in Table 5

Bayesian Approach		Frequentist Approach		
			Direct Maximization	EM Algorithm
$\tilde{\theta}$	0.1025	$\hat{\theta}$	0.10202	0.10203
$Var(\tilde{\theta})$	0.00042	$Var(\hat{\theta})$	0.00046	0.00073
95% HPD interval for θ	(0.06202, 0.14299)	95% ACI for θ	(0.06701, 0.15231)	(0.05991, 0.16845)
		95% BCI for θ	(0.06156, 0.14467)	
		$\hat{\theta}_0$	0.074	0.074
$\tilde{\omega}$	0.2760	$\hat{\omega}$	0.28078	0.28080
$Var(\tilde{\omega})$	0.01414	$Var(\hat{\omega})$	0.01670	0.01967
95% HPD interval for ω	(0.04300, 0.50898)	95% ACI for ω	(0.10020, 0.57781)	(0.09099, 0.60362)
		95% BCI for ω	(0.00000, 0.50542)	
Bayes factor B_{10}	2.2725	$\ell(\hat{\theta}, \hat{\omega})$	48.79855	48.79855
		$\ell(\hat{\theta}_0, 0)$	50.50361	
		LRT Stat. λ	3.41012	
		p_{LR}	0.03240	
		Score Test Stat. S^2	3.41611	
		p_S	0.06456	
		Bootstrap Test p_B	0.0336	

6. Concluding remarks

In this paper, we aim to provide some feasible Bayesian and frequentist methods for the analysis of proportional data which may involve zero-inflation. We conducted a comparative analysis based on both frequentist and objective Bayesian approaches for a ZIB distribution. We derived well-known noninformative Jeffreys priors after the orthogonal transformation on parameters. A full justification was presented for the feasibility of priors in hypothesis testing. For comparison purposes, we employed existing R packages for fitting the model by maximum likelihood methods. In particular, we used both conventional maximization methods and the EM algorithm to come up with parameter estimates. We also presented LRT procedures along with a bootstrap procedure for testing

the zero-inflation parameter to compare with Bayes factors. One promising finding is that both LRT and bootstrap procedures yielded a result that was slightly less powerful when parameter θ is very small and close to zero.

We consider the zero inflation binomial model, in which the occurrence of zeros is higher than the conventional binomial model. For future research, one can also consider a zero-deflated model in situations where the occurrence of zeros is lower than the conventional binomial model. Finally, in the perspective of model checking, it would be interesting to consider a posterior predictive analysis based on different data structures like time series models. For instance, an autoregressive process would be a plausible model to handle non-iid samples like baseball data.

Acknowledgment. Kim's research was partially supported by Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2021R1A2C1005271). H.K.T. Ng's work was supported by a grant from the Simons Foundation (#709773).

References

[1] J. Albert and P. Williamson, *Using model/data simulations to detect streakiness*, Amer. Statist. **55** (1), 41-50, 2001.

[2] N. Amek, N. Bayoh, M. Hamel, K.A. Lindblade, J. Gimnig, K.F. Laserson, L. Slutsker, T. Smith and P. Vounatsou, *Spatio-temporal modeling of sparse geostatistical malaria sporozoite rate data using a zero inflated binomial model*, Spat Spatiotemporal Epidemiol **2** (4), 283-290, 2011.

[3] C.C. Astuti and A.D. Mulyanto, *Estimation parameters and modelling zero inflated negative binomial*, Cauchy: Jurnal Matematika Murni dan Aplikasi **4** (3), 115-119, 2016.

[4] M.J. Bayarri, J.O. Berger and G.S. Datta, *Objective Bayes testing of Poisson versus inflated Poisson models*, IMS Collections **3**, 105-121, 2008.

[5] J.O. Berger and L.R. Pericchi, *The intrinsic Bayes factor for model selection and prediction*, J. Amer. Statist. Assoc. **91** (433), 109-122, 1996.

[6] W. Bodromurti, K.A. Notodiputro and A. Kurnia, *Zero inflated binomial model for infant mortality data in Indonesia*, Int. J. Appl. Eng. Res. **13**, 3139-3143, 2018.

[7] G. Claeskens, R. Nguti and P. Janssen, *One-sided tests in shared frailty models*, Test **17** (1), 69-82, 2008.

[8] A C. Cohen, *Estimation in mixtures of discrete distributions*, Statistical Pub, 1963.

[9] F. De Santis and S. Gubbiotti, *Sample size requirements for calibrated approximate credible intervals for proportions in clincal trials*, Int. J. Environ. Res. Public Health **18** (2) 1-11, 2021.

[10] D. Deng and S.R. Paul, *Score tests for zero inflation in generalized linear models*, Canad. J. Statist. **28** (3), 563-570, 2000.

[11] A. Diallo, A. Diop and J.F. Dupuy, *Estimation in zero-inflated binomial regression with missing covariates*, Statistics **53** (5), 839-865, 2019.

[12] C. Dong, D.B. Clarke, X. Yan, A. Khattak and B. Huang, *Multivariate random-parameters zero-inflated negative binomial regression model: an application to estimate crash frequencies at intersections*, Accid Anal Prev **70**, 320-329, 2014.

[13] C. Huang, X. Liu, T. Yao and X. Wang, *An efficient EM algorithm for the mixture of negative binomial models*, J. Phys. Conf **1324** (1), 012093, 2019.

[14] S. Jiang, G. Xiao, A.Y. Koh, J. Kim, Q. Li and X. Zhan, *A Bayesian zero-inflated negative binomial regression model for the integrative analysis of microbiome data*, Biostatistics **22** (3), 522-540, 2021.

[15] N.L. Johnson and S. Kotz, *Distributions in statistics: discrete distributions*, John Wiley & Sons, 1969.
[16] R. Kass and A.E. Raftery, *Bayes Factors*, J. Amer. Statist. Assoc. **90** (430), 773-795, 1995.
[17] R. Kass and S. Vaidyanathan, *Approximate Bayes factors and orthogonal parameters with application to testing equality of two binomial proportions*, J. R. Stat. Soc. Ser. B. Stat. Methodol. **54** (1), 129-144, 1992.
[18] S.W. Kim, S. Shahin, H.K.T. Ng and J. Kim, *Binary segmentation procedures using the bivariate binomial distribution for detecting streakiness in sports data*, Comput. Statist., **36** (3), 1821-1843, 2021.
[19] Q. Li, M. Zhang, Y. Xie and G. Xiao, *Bayesian modeling of spatial molecular profiling data via Gaussian process*. Bioinformatics **37** (22), 4129-4136, 2021.
[20] Z. Li, K. Lee, M. Karagas, J. Madan, A. Hoen, A. O'Malley, and H. Li, *Conditional regression based on a multivariate zero-inflated logistic-normal model for microbiome relative abundance data*, Stat. Biosci. **10** (3), 587-608, 2018.
[21] T. Loyes, B. Moerkerke, O.D. Smet and A. Buysse, *The analysis of zero-inflated count data: beyond zero-inflated Poisson regression*, Br. J. Math. Stat. Psychol. **65** (1), 163-180, 2011.
[22] B. Quost and T. Denoeux, *Clustering and classification of fuzzy data using the fuzzy EM algorithm*, Fuzzy Sets and Systems **286**, 134-156, 2016.
[23] M. Ridout, J. Hinde and C.G.B. Demetrio, *A score test for testing a zero-inflated Poisson regression model against zero-inflated negative binomial alternatives*, Biometrics **57** (1), 219-223, 2001.
[24] L. Sahabo and S. Yi, *Normally approximated Bayesian credible interval of binomial proportion*, J Korean Stat Soc **30** (1), 233-244, 2019.
[25] S. Self and K. Liang, *Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions*, J. Amer. Statist. Assoc. **82** (398), 605-610, 1987.
[26] F. Tang and J.E. Cavanaugh, *State-space models for binomial time series with excess zeros*, J Time Ser Anal. **9**, 128-151, 2017.
[27] T. W. Yee, *VGAM: Vector generalized linear and additive models*, R package version 1.1-5, 2021.
[28] X. Zhang, H. Mallick, Z. Tang, L. Zhang, X. Cui, A. Benson and N. Yi, *Negative binomial mixed models for analyzing microbiome count data*, BMC Bioinform. **18** (4), 1-10, 2017.
[29] M. Zulkifli, I. Noriszura and A.M. Razali, *Zero-inflated Poisson versus zero-inflated negative binomial: application to theft insurance data*, The 7th IMT-GT International Conference on Mathematics, Statistics and its Applications, 2011.

Appendix

A. Derivations of Eq (8)

Note that the pmf of the zero-truncated binomial random variable X is

$$f_T(x|\theta) = \frac{\binom{n}{x}\theta^x(1-\theta)^{n-x}}{1-(1-\theta)^n},\ x = 1, 2, \ldots, n.$$

So, the expected value of X is

$$E(X) = \frac{n\theta}{1-(1-\theta)^n}.$$

Now, let us calculate the Jeffreys prior. Note that

$$\log f_T(x|\theta) = \ln \binom{n}{x} + x\ln\theta + (n-x)\ln(1-\theta) - \ln(1-(1-\theta)^n).$$

Since

$$\begin{aligned}\frac{\partial^2}{\partial\theta^2}\ln f_T(x|\theta) &= \frac{\partial}{\partial\theta}\left[\frac{x}{\theta} - \frac{n-x}{1-\theta} - \frac{n(1-\theta)^{n-1}}{1-(1-\theta)^n}\right] \\ &= -\frac{x}{\theta^2} - \frac{n-x}{(1-\theta)^2} + \frac{n\{(1-\theta)^n + n - 1\}(1-\theta)^{n-2}}{(1-(1-\theta)^n)^2},\end{aligned}$$

we have

$$\begin{aligned}I(\theta) &= -E\left[\frac{\partial^2}{\partial\theta^2}\ln f_T(x|\theta)\right] \\ &= \frac{n\theta}{\theta^2(1-(1-\theta)^n)} + \frac{n(1-(1-\theta)^n) - n\theta}{(1-\theta)^2(1-(1-\theta)^n)} - \frac{n\left[(1-\theta)^n + n - 1\right](1-\theta)^{n-2}}{(1-(1-\theta)^n)^2} \\ &= \frac{n}{1-(1-\theta)^n}\left[\frac{1}{\theta} + \frac{1-(1-\theta)^n - \theta}{(1-\theta)^2} - \frac{\left[(1-\theta)^n + n - 1\right](1-\theta)^{n-2}}{1-(1-\theta)^n}\right] \\ &= \frac{n}{1-(1-\theta)^n}\left[\frac{1}{\theta} + \frac{1-(1-\theta)^{n-1}}{1-\theta} - \frac{\left[(1-\theta)^n + n - 1\right](1-\theta)^{n-2}}{1-(1-\theta)^n}\right] \\ &= \frac{n}{1-(1-\theta)^n}\left[\frac{(1-\theta)+\theta-\theta(1-\theta)^{n-1}}{\theta(1-\theta)} - \frac{\left[(1-\theta)^n + n - 1\right](1-\theta)^{n-2}}{1-(1-\theta)^n}\right] \\ &= \frac{n\left\{1-(1-\theta)^n - \theta(1-\theta)^{n-1}\left\{1-(1-\theta)^n\right\} - \theta(1-\theta)^{2n-1} - (n-1)\theta(1-\theta)^{n-1}\right\}}{\theta(1-\theta)\left[1-(1-\theta)^n\right]^2} \\ &= \frac{1}{\theta(1-\theta)}\frac{1-(1-\theta)^n - n\theta(1-\theta)^{n-1}}{[1-(1-\theta)^n]^2}.\end{aligned}$$

Thus, the Jeffreys prior is readily available by taking a square root on $I(\theta)$.

B. First and second partial derivatives of the log-likelihood function

The first partial derivatives of the log-likelihood function in Eq. (3.1):

$$\begin{aligned}\frac{\partial\ln L(\theta,\omega)}{\partial\theta} &= \sum_{i=1}^{N}\left\{\left[\frac{-(1-\omega)n_i(1-\theta)^{n_i-1}}{\omega+(1-\omega)(1-\theta)^{n_i}}\right]I(x_i=0) + \left[\frac{x_i - n_i\theta}{\theta(1-\theta)}\right]I(x_i>0)\right\}, \\ \frac{\partial\ln L(\theta,\omega)}{\partial\omega} &= \sum_{i=1}^{N}\left\{\left[\frac{1-(1-\theta)^{n_i}}{\omega+(1-\omega)(1-\theta)^{n_i}}\right]I(x_i=0) - \left[\frac{1}{(1-\omega)}\right]I(x_i>0)\right\}.\end{aligned}$$

The second partial derivatives of the log-likelihood function in Eq. (3.1):

$$\begin{aligned}
\frac{\partial^2 \ln L(\theta,\omega)}{\partial\theta^2} &= \sum_{i=1}^{N}\Bigg\{\left[\frac{(1-\omega)n_i(n_i-1)(1-\theta)^{n_i-2}}{\omega+(1-\omega)(1-\theta)^{n_i}}-\left(\frac{(1-\omega)n_i(1-\theta)^{n_i-1}}{\omega+(1-\omega)(1-\theta)^{n_i}}\right)^2\right]I(x_i=0)\\
&\quad-\left[\frac{n_i}{(1-\theta)^2}+\frac{x_i(1-2\theta)}{\theta^2(1-\theta)^2}\right]I(x_i>0)\Bigg\},\\
\frac{\partial^2 \ln L(\theta,\omega)}{\partial\omega^2} &= \sum_{i=1}^{N}\left\{-\left[\frac{1-(1-\theta)^{n_i}}{\omega+(1-\omega)(1-\theta)^{n_i}}\right]^2 I(x_i=0)-\left[\frac{1}{(1-\omega)^2}\right]I(x_i>0)\right\},\\
\frac{\partial^2 \ln L(\theta,\omega)}{\partial\theta\partial\omega} &= \sum_{i=1}^{N}\left\{\frac{n_i(1-\theta)^{n_i-1}}{\omega+(1-\omega)(1-\theta)^{n_i}}+\frac{(1-\omega)n_i(1-\theta)^{n_i-1}[1-(1-\theta)^{n_i}]}{[\omega+(1-\omega)(1-\theta)^{n_i}]^2}\right\}I(x_i=0).
\end{aligned}$$

C. Expected Fisher information matrix

We can obtain the elements of the Fisher information matrix based on the second partial derivatives of the log-likelihood function in Eq. (3.1) as:

$$\begin{aligned}
&E\left[-\frac{\partial^2 \ln L(\theta,\omega)}{\partial\theta^2}\right]\\
=\;& -\sum_{i=1}^{N}\Bigg\{\left[\frac{(1-\omega)n_i(n_i-1)(1-\theta)^{n_i-2}}{\omega+(1-\omega)(1-\theta)^{n_i}}-\left(\frac{1-\omega)n_i(1-\theta)^{n_i-1}}{\omega+(1-\omega)(1-\theta)^{n_i}}\right)^2\right]E[I(X_i=0)]\\
&\quad-\left[\frac{n_i}{(1-\theta)^2}\right]E[I(X_i>0)]-\left[\frac{(1-2\theta)}{\theta^2(1-\theta)^2}\right]E[X_iI(X_i>0)]\Bigg\},\\
&E\left[-\frac{\partial^2 \ln L(\theta,\omega)}{\partial\omega^2}\right]\\
=\;& \sum_{i=1}^{N}\left\{\left[\frac{1-(1-\theta)^{n_i}}{\omega+(1-\omega)(1-\theta)^{n_i}}\right]^2 E[I(X_i=0)]+\left[\frac{1}{(1-\omega)^2}\right]E[I(X_i>0)]\right\},\\
&E\left[-\frac{\partial^2 \ln L(\theta,\omega)}{\partial\theta\partial\omega}\right]\\
=\;& -\sum_{i=1}^{N}\left\{\frac{n_i(1-\theta)^{n_i-1}}{\omega+(1-\omega)(1-\theta)^{n_i}}+\frac{(1-\omega)n_i(1-\theta)^{n_i-1}[1-(1-\theta)^{n_i}]}{[\omega+(1-\omega)(1-\theta)^{n_i}]^2}\right\}E[I(X_i=0)],
\end{aligned}$$

where

$$\begin{aligned}
E[I(X_i=0)] &= \Pr(X_i=0)=\omega+(1-\omega)(1-\theta)^{n_i},\\
E[I(X_i>0)] &= \Pr(X_i>0)=(1-\omega)[1-(1-\theta)^{n_i}],\\
E[X_iI(X_i>0)] &= n_i\theta.
\end{aligned}$$