



Article

# Inference for the Parameters of a Zero-Inflated Poisson Predictive Model

Min Deng , Mostafa S. Aminzadeh \*  and Banghee So

Department of Mathematics, Towson University, Towson, MD 21252, USA; mdeng@towson.edu (M.D.); bso@towson.edu (B.S.)

\* Correspondence: maminzadeh@towson.edu

**Abstract:** In the insurance sector, Zero-Inflated models are commonly used due to the unique nature of insurance data, which often contain both genuine zeros (meaning no claims made) and potential claims. Although active developments in modeling excess zero data have occurred, the use of Bayesian techniques for parameter estimation in Zero-Inflated Poisson models has not been widely explored. This research aims to introduce a new Bayesian approach for estimating the parameters of the Zero-Inflated Poisson model. The method involves employing Gamma and Beta prior distributions to derive closed formulas for Bayes estimators and predictive density. Additionally, we propose a data-driven approach for selecting hyper-parameter values that produce highly accurate Bayes estimates. Simulation studies confirm that, for small and moderate sample sizes, the Bayesian method outperforms the maximum likelihood (ML) method in terms of accuracy. To illustrate the ML and Bayesian methods proposed in the article, a real dataset is analyzed.

**Keywords:** bayesian estimation; ML estimation; zero-inflated poisson model; gamma distribution; beta distribution; score test



**Citation:** Deng, Min, Mostafa S. Aminzadeh, and Banghee So. 2024. Inference for the Parameters of a Zero-Inflated Poisson Predictive Model. *Risks* 12: 104. <https://doi.org/10.3390/risks12070104>

Academic Editors: Silvia Dedu and Emilio Gómez Déniz

Received: 9 March 2024

Revised: 8 June 2024

Accepted: 17 June 2024

Published: 24 June 2024



**Copyright:** © 2024 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

The high occurrence of zero-valued observations in insurance claim data is a well-documented phenomenon. Traditional count models, such as Poisson or Negative Binomial, struggle to accurately represent insurance claim data due to the excessive dispersion caused by the observed frequency of zeros surpassing expectations based on these models. [Perumean-Chaney et al. \(2013\)](#) emphasized the importance of considering the excess zeros to achieve satisfactory modeling of both zero and non-zero counts. In the statistical literature, two main approaches have been developed to address datasets with a large number of zeros: Hurdle models and Zero-Inflated models. Hurdle models, initially proposed by [Mullahy \(1986\)](#), adopt a truncated-at-zero approach, as seen in truncated Poisson and Negative Binomial models. Zero-Inflated models, introduced by [Lambert \(1992\)](#), utilize a mixture model approach, separating the population into two groups—one with only zero outcomes and the other with non-zero outcomes. Notable examples include Zero-Inflated Poisson and Negative Binomial regressions.

The generic Zero-Inflated distribution is defined as:

$$f_{\text{zero-inflated}}(y|\theta, p) = \begin{cases} p + (1-p)f_0(0), & y = 0 \\ (1-p)f_0(y|\theta), & y = 1, 2, \dots \end{cases}$$

where  $p$  denotes the probability of extra zeros and  $f_0(y)$  can be any count distribution, such as Poisson or Negative Binomial. If  $f_0(y)$  follows a Poisson distribution, this model simplifies to the Zero-Inflated Poisson (ZIP) model, expressed by the density:

$$f_{\text{ZIP}}(y|\mu, p) = \begin{cases} p + (1-p)e^{-\mu}, & y = 0 \\ (1-p)\frac{e^{-\mu}\mu^y}{y!}, & y = 1, 2, \dots \end{cases}$$

In cases where covariates are linked with both the probability  $p$  of a structural zero and the mean function  $\mu$  of the Poisson model, logistic regression is used to model  $p$ , and log-linear regression is applied to model  $\mu$ . This analytical framework is referred to as ZIP regression, which, while not the primary focus of this study, serves as a foundation for our exploration.

In the insurance sector, the adoption of Zero-Inflated models is widespread, reflecting the distinctive characteristics of insurance data, which often comprise both actual zeros (indicating no claims) and potential claims. Various studies exemplify this application trend: [Mouatassim and Ezzahid \(2012\)](#) applied Zero-Inflated Poisson regression to a private health insurance dataset using the EM algorithm to maximize the log-likelihood function. [Chen et al. \(2019\)](#) introduced a penalized Poisson regression for subgroup analysis in claim frequency data, implementing an ADMM algorithm for optimization. [Zhang et al. \(2022\)](#) developed a multivariate zero-inflated hurdle model for multivariate count data with extra zeros, employing the EM algorithm for parameter estimation.

[Ghosh et al. \(2006\)](#) delved into a Bayesian analysis of Zero-Inflated power series ZIPS ( $p, \theta$ ) regression models, employing the log link function to correlate the mean  $\mu = \mu(\theta)$  of the power series with covariates and the logit function for modeling  $p$ . They proposed Beta ( $b_1, b_2$ ) and power series-specific priors for the unknown parameters  $p$  and  $\theta$ , respectively, overcoming analytical challenges through Monte Carlo simulation-based techniques. Their findings highlighted the superiority of the Bayesian approach over traditional methods.

Recent trends also show an inclination towards integrating machine learning techniques with Zero-Inflated models. [Zhou et al. \(2022\)](#) proposed modeling Tweedie's compound Poisson distribution using EMTboost. [Lee \(2021\)](#) used cyclic coordinate descent optimization for Zero-Inflated Poisson regression, addressing saddle points with Delta boosting. [Meng et al. \(2022\)](#) introduced innovative approaches using Gradient Boosted Decision Trees for training Zero-Inflated Poisson regression models.

To the best of our knowledge, the most recent Bayesian analysis was conducted by [Angers and Biswas \(2003\)](#). In their study, the authors discussed a Zero-Inflated generalized Poisson model that includes three parameters, namely  $p$ ,  $\mu$ , and  $\alpha$ . The Bayesian analysis employs a conditional uniform prior for the parameters  $p$  and  $\alpha$ , given  $\lambda$ , while Jeffreys' prior is used for  $\lambda$ . The authors concluded that analytical integration was not feasible, leading to the use of Monte-Carlo integration with importance sampling for parameter estimation. In contrast, our study utilizes beta and gamma priors to provide enhanced flexibility in the shapes of prior distributions, offering closed formulas for Bayes estimators, predictive density, and predictive expected values. This approach diverges from the regression-centric literature on the ZIP model. [Boucher et al. \(2007\)](#) presents models that consist of generalizations of count distributions with time dependency where the dependence between contracts of the same insureds can be modeled with Bayesian and frequentist models, based on a generalization of Poisson and negative binomial distributions.

The structure of the paper is organized as follows: In Section 2, we present the deviations for Maximum Likelihood Estimators (MLEs). Section 3 employs gamma and beta distributions as prior distributions for the parameters  $\mu$  and  $p$ , respectively. This section also elaborates on the derivation of the predictive density  $f_Z(z|\underline{y})$ , the conditional expectation  $E[Z|\underline{y}]$ , and an approximation for the percentile of the predictive distribution. Here,  $\underline{y}$  signifies a random sample from a Zero-Inflated Poisson ( $\mu, p$ ) distribution, and  $z$  represents an observed value of  $Z \sim \text{Zero-Inflated Poisson}(\mu, p)$ . Section 4 summarizes the outcomes of the simulation studies and introduces a data-driven approach for selecting hyper-parameter values of the prior distribution. Section 5 is devoted to the analysis of a real dataset, demonstrating the Bayesian inference methodology introduced in this work. Finally, Section 6 offers brief concluding remarks about the study. The Mathematica code utilized for the simulation studies and the computations performed on the real data are available upon request from the authors.

### 2. Maximum Likelihood Estimation

The probability mass function of the Zero-Inflated Poisson (ZIP) distribution, denoted as ZIP( $\mu, p$ ), is given by

$$f_{ZIP}(y|\mu, p) = \begin{cases} p + (1 - p)e^{-\mu}, & \text{if } y = 0, \\ (1 - p)f_0(y|\mu), & \text{if } y = 1, 2, \dots, \end{cases}$$

where  $f_0(y|\mu) = \frac{e^{-\mu}\mu^y}{y!}$ ,  $\mu > 0, y = 0, 1, 2, \dots$ , and  $p$  is the probability that reflects the degree of zero inflation. It can be shown that

$$E[Y] = (1 - p)\mu, \quad \text{Var}(Y) = (1 - p)(1 + p\mu)\mu,$$

Let  $y_1, y_2, \dots, y_n$  be a random sample from ZIP( $\mu, p$ ). Without a loss of generality, consider the corresponding sorted sample

$$y_1 = y_2 = \dots = y_m < y_{m+1} \leq y_{m+2} \leq \dots \leq y_n,$$

for some positive integers  $m = 0, 1, \dots, n - 1$ .

The likelihood function is

$$\begin{aligned} L(\mu, p) &= (p + (1 - p)e^{-\mu})^m \prod_{k=m+1}^n \frac{e^{-\mu}\mu^{y_k}}{y_k!} (1 - p) \\ &= (p + (1 - p)e^{-\mu})^m \frac{e^{-(n-m)\mu} (1 - p)^{n-m} \mu^{\sum_{k=m+1}^n y_k}}{\prod_{k=m+1}^n y_k!}. \end{aligned}$$

The log of the likelihood function  $L(\mu, p)$  can be written as

$$l = \ln(L(\mu, p)) = m(\ln(p + (1 - p)e^{-\mu})) + (n - m) \ln(1 - p) - (n - m)\mu + \ln(\mu) \sum_{k=m+1}^n y_k - \ln\left(\prod_{k=m+1}^n y_k!\right). \tag{1}$$

To find the Maximum Likelihood Estimators (MLEs) of  $\mu$  and  $p$ , the following equations need to be solved simultaneously:

$$\begin{cases} \frac{\partial l}{\partial \mu} = \frac{-m(1-p)e^{-\mu}}{p+(1-p)e^{-\mu}} - (n - m) + \frac{\sum_{k=m+1}^n y_k}{\mu} = 0, \\ \frac{\partial l}{\partial p} = \frac{m}{p+(1-p)e^{-\mu}} (1 - e^{-\mu}) - \frac{n-m}{1-p} = 0. \end{cases}$$

After some algebraic manipulations, it can be demonstrated that the MLEs are solutions to the equations

$$\frac{\mu}{1 - e^{-\mu}} = \bar{y}, \quad p = \frac{n\mu - (n - m)\bar{y}}{n\mu},$$

where  $\bar{y} = \frac{\sum_{k=m+1}^n y_k}{n-m}$  represents the average of the non-zero observations. It is important to note that the equation  $\frac{\mu}{1 - e^{-\mu}} = \bar{y}$  is nonlinear in  $\mu$ . To solve for  $\mu$ , one can employ Mathematica's 'FindRoot' function, which iteratively seeks the root of the equation. The function can be invoked as follows:

$$\text{FindRoot}\left[\frac{\mu}{1 - e^{-\mu}} == \bar{y}, \{\mu, \mu_0\}\right],$$

where  $\mu_0$  is an initial "good" guess for the solution. This equation is guaranteed to have a unique solution under appropriate conditions. Once the MLE for  $\mu$  has been determined, the MLE for  $p$  can subsequently be computed.

### 3. Bayesian Inference

This section delineates a Bayesian methodology for the computation of Bayes estimates for the parameters, the formulation of the predictive density, and the approximation of percentiles for the predictive distribution.

#### 3.1. Bayes Estimates of Parameters $\mu, p$

Applying the Binomial expansion,

$$(p + (1 - p)e^{-\mu})^m = \sum_{x=0}^m \binom{m}{x} p^x ((1 - p)e^{-\mu})^{m-x} = \sum_{x=0}^m \binom{m}{x} p^x (1 - p)^{m-x} e^{-\mu(m-x)},$$

the likelihood function  $L(\mu, p)$ , as presented in Section 2, can be expressed as

$$\begin{aligned} L(\mu, p) &= \sum_{x=0}^m \binom{m}{x} p^x (1 - p)^{m-x+n-m} e^{-\mu(m-x)-(n-m)\mu} \mu^S \frac{1}{\prod_{k=m+1}^n y_k!} \\ &= c \sum_{x=0}^m \binom{m}{x} p^x (1 - p)^{n-x} e^{-\mu(n-x)} \mu^S, \end{aligned} \tag{2}$$

where  $c$  is a normalizing constant and it does not need to be found, as it will be canceled out once the posterior PDF has been formulated.  $S = \sum_{k=m+1}^n y_k$ .

For the parameters  $\mu$  and  $p$ , respectively, we consider the gamma( $\alpha, \beta$ ) and beta( $\gamma, \xi$ ) conjugate prior distributions. The choice of the gamma distribution as a prior for  $\mu$  is due to its flexibility in shaping the probability density function (pdf) for a positive random variable like  $\mu$ , given the selection of parameter values. Similarly, the beta distribution is chosen as a prior for  $p$  because its support matches the possible range of  $p$  values, and like the gamma distribution, it offers flexibility in shaping its pdf for a range of parameter values. Employing the gamma and beta prior distributions, we define

$$\begin{aligned} h(\mu|\alpha, \beta) &= \frac{\beta^\alpha}{\Gamma(\alpha)} \mu^{\alpha-1} e^{-\beta\mu}, \quad \alpha > 0, \quad \beta > 0, \\ g(p|\gamma, \xi) &= \frac{\Gamma(\gamma + \xi)}{\Gamma(\gamma)\Gamma(\xi)} p^{\gamma-1} (1 - p)^{\xi-1}, \quad \gamma > 0, \quad \xi > 0. \end{aligned} \tag{3}$$

Utilizing Equations (2) and (3), the joint posterior distribution of  $(p, \mu)$  can be expressed as follows:

$$\begin{aligned} \pi(p, \mu|\alpha, \beta, \gamma, \xi, \underline{y}) &\propto \sum_{x=0}^m \binom{m}{x} p^{x+\gamma-1} (1 - p)^{n+\xi-x-1} e^{-\mu(n+\beta-x)} \mu^{S+\alpha-1} \\ &\propto \sum_{x=0}^m \binom{m}{x} h(\mu|S + \alpha, n + \beta - x) \cdot g(p|x + \gamma, n + \xi - x). \end{aligned} \tag{4}$$

However, it is necessary that

$$\int_0^1 \int_0^\infty \pi(p, \mu|\alpha, \beta, \gamma, \xi, \underline{y}) d\mu dp = 1.$$

Therefore, utilizing

$$\sum_{x=0}^m \binom{m}{x} = 2^m,$$

the joint posterior is

$$\pi(p, \mu|\alpha, \beta, \gamma, \xi, \underline{y}) = \sum_{x=0}^m \frac{\binom{m}{x}}{2^m} h(\mu|S + \alpha, n + \beta - x) \cdot g(p|x + \gamma, n + \xi - x).$$

Since

$$\int_0^\infty h(\mu|S + \alpha, n + \beta - x)d\mu = 1 \quad \text{and} \quad \int_0^1 g(p|x + \gamma, n + \xi - x)dp = 1,$$

the marginal posteriors are

$$\begin{aligned} \pi(\mu|\xi, \gamma, x, n) &= \frac{\sum_{x=0}^m \binom{m}{x}}{2^m} h(\mu|S + \alpha, n + \beta - x), \\ \pi(p|\alpha, \beta, S, x) &= \frac{\sum_{x=0}^m \binom{m}{x}}{2^m} g(p|x + \gamma, n + \xi - x). \end{aligned}$$

Under the squared-error loss function, the Bayes estimates of  $\mu$  and  $p$  are the expected values of the corresponding priors. Therefore,

$$\begin{aligned} \hat{\mu}_{\text{Bayes}} &= \sum_{x=0}^m \binom{m}{x} \frac{(\alpha + S)}{2^m(n + \beta - x)}, \\ \hat{p}_{\text{Bayes}} &= \sum_{x=0}^m \binom{m}{x} \frac{(x + \gamma)}{2^m(n + \xi + \gamma)}. \end{aligned} \tag{5}$$

The Bayes estimators presented in Equation (5) are expressed as closed-form solutions, obviating the need for numerical integration. This represents a significant advantage of our methodology over that presented in [Angers and Biswas \(2003\)](#).

### 3.2. Predictive pdf $f(z|\underline{y})$ and Expected Value $E[z|\underline{y}]$

One of the predictive measures of interest is the expected value of a new observation  $z$  from the Zero-Inflated Poisson (ZIP) distribution, given a set of observations  $\underline{y}$ :  $y_1 = y_2 = \dots = y_m < y_{m+1} \leq y_{m+2} \leq \dots \leq y_n$ . To calculate the expected value, it is necessary to derive the predictive probability density function (pdf)  $f(z|\underline{y})$ , expressed as

$$f(z|\underline{y}) = \int_0^\infty \int_0^1 f_{\text{ZIP}}(z|\mu, p)\pi(p, \mu|\alpha, \beta, \gamma, \xi, \underline{y})dpd\mu.$$

For the  $z = 0$  case, we have,

$$\begin{aligned} f(0|\underline{y}) &= \int_0^\infty \int_0^1 (p + (1 - p)e^{-\mu})h(\mu|S + \alpha, n + \beta - x) \cdot g(p|x + \gamma, n + \xi - x)dpd\mu \\ &= \sum_{x=0}^m \frac{\binom{m}{x}}{2^m} \int_0^1 p \frac{\Gamma(n + \gamma + \xi)}{\Gamma(x + \gamma)\Gamma(n + \xi - x)} p^{x+\gamma-1} (1 - p)^{n+\xi-x-1} dp \int_0^\infty h(\mu|S + \alpha, n + \beta - x)d\mu \\ &+ \sum_{x=0}^m \frac{\binom{m}{x}}{2^m} \int_0^1 (1 - p) \frac{\Gamma(n + \gamma + \xi)}{\Gamma(x + \gamma)\Gamma(n + \xi - x)} p^{x+\gamma-1} (1 - p)^{n+\xi-x-1} dp \int_0^\infty \frac{(n + \beta - x)^{S+\alpha} e^{-\mu(n+\beta-x)} \mu^{S+\alpha-1}}{\Gamma(S + \alpha)} d\mu. \end{aligned}$$

After multiplying and dividing the above integrands by appropriate terms to convert them into PDFs, and then integrating, we obtain,

$$\begin{aligned} f(0|\underline{y}) &= \sum_{x=0}^m \frac{\binom{m}{x}}{2^m} \frac{\Gamma(n + \gamma + \xi)}{\Gamma(x + \gamma)\Gamma(n + \xi - x)} \cdot \frac{\Gamma(x + \gamma + 1)\Gamma(n + \xi - x)}{\Gamma(n + \gamma + \xi + 1)} \cdot (1) \\ &+ \sum_{x=0}^m \frac{\binom{m}{x}}{2^m} \frac{\Gamma(n + \gamma + \xi)}{\Gamma(x + \gamma)\Gamma(n + \xi - x)} \cdot \frac{\Gamma(x + \gamma)\Gamma(n + \xi - x + 1)}{\Gamma(n + \gamma + \xi + 1)} \cdot \frac{(n + \beta - x)^{S+\alpha}}{(n + \beta - x + 1)^{S+\alpha}} \\ &= \sum_{x=0}^m \frac{\binom{m}{x}}{2^m} \left[ \frac{x + \gamma}{n + \gamma + \xi} + \left( \frac{n + \xi - x}{n + \gamma + \xi} \right) \left( \frac{n + \beta - x}{n + \beta - x + 1} \right)^{S+\alpha} \right]. \end{aligned} \tag{6}$$

For  $z = 1, 2, 3, \dots$ ,

$$\begin{aligned} f(z|\underline{y}) &= \int_0^1 \int_0^\infty (1-p) \frac{e^{-\mu} \mu^z}{z!} \pi(p, \mu | \alpha, \beta, \gamma, \xi, \underline{y}) d\mu dp \\ &= \sum_{x=0}^m \frac{\binom{m}{x}}{2^m} \int_0^\infty \frac{e^{-\mu(n+\beta-x+1)} \mu^{S+\alpha+z-1}}{z! \Gamma(S+\alpha)} d\mu \int_0^1 (1-p) g(p|x+\gamma, n+\xi-x) dp \\ &= \sum_{x=0}^m \frac{\binom{m}{x}}{2^m} \int_0^\infty \frac{(n+\beta-x)^{S+\alpha} e^{-\mu(n+\beta-x+1)} \mu^{S+\alpha+z-1}}{z! \Gamma(S+\alpha)} d\mu \int_0^1 (1-p) g(p|x+\gamma, n+\xi-x) dp. \end{aligned}$$

After applying the appropriate conversions to PDFs and integrating them,

$$f(z|\underline{y}) = \sum_{x=0}^m \frac{\binom{m}{x}}{2^m} \left[ \frac{(n+\beta-x)^{S+\alpha}}{z! \Gamma(S+\alpha)} \frac{\Gamma(S+\alpha+z)}{(n+\beta-x+1)^{S+\alpha+z}} \right] \left[ \frac{\Gamma(n+\gamma+\xi)}{\Gamma(x+\gamma)\Gamma(n+\xi-x)} \cdot \frac{\Gamma(x+\gamma)\Gamma(n+\xi-x+1)}{\Gamma(n+\gamma+\xi+1)} \right].$$

Note that the term inside the first square bracket on the left side of the sum corresponds to the PMF of a Negative Binomial distribution with the parameters  $(S+\alpha, q)$ . Consequently, the predictive density  $f(z|\underline{y})$  can be expressed as

$$f(z|\underline{y}) = \sum_{x=0}^m \frac{\binom{m}{x}}{2^m} \left( \frac{n+\xi-x}{n+\xi+\gamma} \right) \binom{S+\alpha+z-1}{z} q^{S+\alpha} (1-q)^z, \tag{7}$$

where  $q = \frac{n+\beta-x}{n+\beta-x+1}$ . Therefore, the predictive density is

$$f(z|\underline{y}) = \begin{cases} \sum_{x=0}^m \frac{\binom{m}{x}}{2^m} \left[ \frac{x+\gamma}{n+\gamma+\xi} + \left( \frac{n+\xi-x}{n+\gamma+\xi} \right) \left( \frac{n+\beta-x}{n+\beta-x+1} \right)^{S+\alpha} \right], & \text{for } z = 0, \\ \sum_{x=0}^m \frac{\binom{m}{x}}{2^m} \left( \frac{n+\xi-x}{n+\gamma+\xi} \right) \binom{S+\alpha+z-1}{z} q^{S+\alpha} (1-q)^z, & \text{for } z = 1, 2, 3, \dots \end{cases}$$

For example, for  $n = 50, m = 20, S = 50, \beta = 5, \alpha = 60, \xi = 350$ , and  $\gamma = 1169$ , the numerical values of  $f(z|\underline{y})$  for  $z = 0, 1, 2, 3, \dots, 9$  are given below.

{0, 0.773}, {1, 0.052}, {2, 0.063}, {3, 0.051}, {4, 0.031}, {5, 0.015}, {6, 0.006}, {7, 0.002}, {8, 0.0007}, {9, 0.0002}.

For instance, if  $y$  represents the number of claims per month, then based on the observed sample  $y = y_1, \dots, y_n$ , we would predict that the number of claims  $z$  in the upcoming month  $(\bar{n} + 1)$  would be zero with a probability of 0.773, one with a probability of 0.052, and so forth. The probabilities detailed above can be visually represented in a bar chart, as depicted in Figure 1. The x-axis is labeled as  $z$ , and the y-axis is labeled as  $f(z|\underline{y})$ .

The conditional expected value  $E[z|\underline{y}]$  is derived via the predictive density.

$$\begin{aligned} E[z|\underline{y}] &= 0 \cdot f(0|\underline{y}) + \sum_{z=1}^\infty z f(z|\underline{y}) \\ &= \sum_{z=1}^\infty z \sum_{x=0}^m \frac{\binom{m}{x}}{2^m} \left( \frac{n+\xi-x}{n+\xi+\gamma} \right) \binom{S+\alpha+z-1}{z} q^{S+\alpha} (1-q)^z \\ &= \sum_{x=0}^m \frac{\binom{m}{x}}{2^m} \left( \frac{n+\xi-x}{n+\xi+\gamma} \right) \sum_{z=1}^\infty z \binom{S+\alpha+z-1}{z} q^{S+\alpha} (1-q)^z. \end{aligned}$$

Recall that  $q = \frac{n+\beta-x}{n+\beta-x+1}$ . Also, the term  $\sum_{z=1}^\infty z \binom{S+\alpha+z-1}{z} q^{S+\alpha} (1-q)^z$  represents the expected value of a Negative Binomial distribution with parameters  $S+\alpha$  and  $q$ . Hence,

$$\sum_{z=1}^\infty z \binom{S+\alpha+z-1}{z} q^{S+\alpha} (1-q)^z = \frac{q(S+\alpha)}{1-q} = \frac{S+\alpha}{n+\beta-x}.$$

Consequently, the conditional expected value  $E[z|y]$  in the closed form is given by

$$E[z|y] = \sum_{x=0}^m \frac{\binom{m}{x}}{2^m} \left( \frac{n + \zeta - x}{n + \zeta + \gamma} \right) \left( \frac{S + \alpha}{n + \beta - x} \right). \tag{8}$$

It should be noted that the objective of formulating the predictive density and the corresponding computations is for predictive purposes. However, to accomplish this, Bayes estimates of the parameters should be computed.

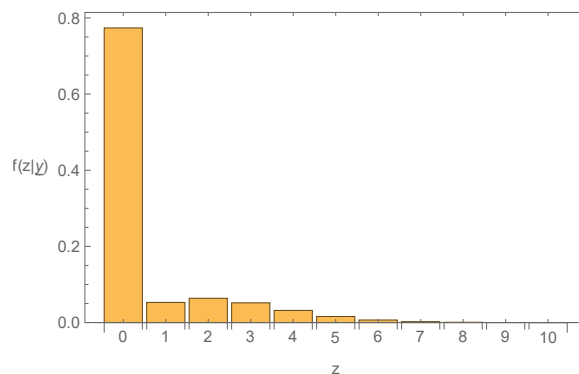


Figure 1. Graph of the predictive density.

### 3.3. Approximate Percentile for the Predictive Distribution

Recall that the value of the predictive pdf  $f(0|y)$ , derived in Section 3.2,

$$f(0|y) = \sum_{x=0}^m \frac{\binom{m}{x}}{2^m} \left[ \frac{x + \gamma}{n + \gamma + \zeta} + \left( \frac{n + \zeta - x}{n + \gamma + \zeta} \right) \left( \frac{n + \beta - x}{n + \beta - x + 1} \right)^{S+\alpha} \right],$$

is independent of  $z$ . For instance, utilizing the same parameter values for  $m, S, n, \alpha, \beta, \zeta,$  and  $\gamma$  specified in Section 3.2, we calculate  $f(0|y) = 0.77363$ . Let us define the cumulative distribution function (CDF)  $F(a|y) = P(z \leq a|y)$  as

$$F(a|y) = f(0|y) + \sum_{z=1}^a \left[ \sum_{x=0}^m \frac{\binom{m}{x}}{2^m} \left( \frac{n + \zeta - x}{n + \zeta + \gamma} \right) \binom{S + \alpha + z - 1}{z} q^{S+\alpha} (1 - q)^z \right].$$

Note, the expression within the square brackets corresponds to  $f(z|y)$  for  $z = 1, 2, \dots,$  as previously established in Section 3.2. To determine an approximate 95th percentile for the predictive distribution, we require that

$$F(a|y) \approx 0.95.$$

Mathematica can be utilized to enumerate the values of  $(a, F(a - 1|y), F(a|y))$ , starting with  $a = 1$ , and to identify the interval  $(F(a - 1|y), F(a|y))$  that encloses 0.95. Employing the same input values for  $m, n, S, \dots,$  we obtain:

- {1, 0.736, 0.826}, {2, 0.826, 0.890}, {3, 0.890, 0.941}, {4, 0.941, 0.973}, {5, 0.973, 0.989}, {6, 0.989, 0.996},
- {7, 0.996, 0.998}, {8, 0.998, 0.999}, {9, 0.999, 0.999}, {10, 0.999, 0.999}.

Since 0.95 is within the interval (0.941, 0.973), corresponding to  $a = 4$ , it can be inferred that, based on the observed sample  $y = y_1, \dots, y_n$  and the specified hyper-parameter values, the forthcoming value of  $z$  would not exceed 4 with approximately 0.95 probability.

#### 4. Simulation

Simulation studies were conducted to evaluate the accuracy of Bayesian and Maximum Likelihood estimators for the parameters  $\mu$  and  $p$ . The simulation studies proceeded as follows:

Step 1: Generate  $N = 1000$  samples of size  $n = 20, 50,$  and  $100,$  respectively, from the Zero-Inflated Poisson distribution  $f_{ZIP}(y|\mu, p)$  using selected “true” values of the parameters  $\mu$  and  $p$  listed in the tables.

Step 2: For each set of “true” parameter values of  $\mu$  and  $p,$  as outlined in Section 2, the MLEs are determined using the equations

$$\frac{\mu}{1 - e^{-\mu}} = \bar{y}, \quad p = \frac{n\mu - (n - m)\bar{y}}{n\mu}.$$

Step 3: Selecting optimal hyper-parameter values is crucial for obtaining accurate Bayes estimates.

Recall that the conjugate prior distribution for  $\mu$  is a gamma( $\alpha, \beta$ ) distribution. Given that

$$E(\mu) = \frac{\alpha}{\beta}, \quad \text{Var}(\mu) = \frac{\alpha}{\beta^2} \approx \frac{\beta \times \mu}{\beta^2} = \frac{\mu}{\beta},$$

we let  $E(\mu) \approx \mu,$  or in other words,  $\alpha \approx \mu \times \beta.$  This ensures that the prior distribution of  $\mu$  is centered at the selected “true” value for  $\mu.$  By substituting this into the  $\text{Var}(\mu)$  formula, we can deduce that, to achieve high accuracy for the Bayes estimate of  $\mu,$  the hyper-parameters  $\alpha$  and  $\beta$  should be selected in such a way that they are consistent with the expected value and variance of  $\mu.$  Since  $\text{Var}(\mu) \approx \frac{\mu}{\beta},$  we therefore choose a large value for  $\beta$  to ensure that  $\text{Var}(\mu)$  is small, and let  $\alpha = \beta \times \mu$  to ensure that  $\frac{\alpha}{\beta} \approx \mu.$  It is noted that there is no unique pair of hyper-parameter values for  $\alpha$  and  $\beta;$  any pair that meets the above criteria should suffice. However, simulation studies confirm that a larger value of  $\beta$  provides a more accurate Bayes estimate.

Recall that  $p$  follows a beta( $\gamma, \zeta$ ) distribution which is also considered a conjugate prior. A similar rationale is used for selecting the hyper-parameters  $\gamma$  and  $\zeta.$  We have

$$E(p) = \frac{\gamma}{\gamma + \zeta}, \quad \text{Var}(p) = \frac{\gamma \times \zeta}{(\gamma + \zeta)^2(\gamma + \zeta + 1)}.$$

By substituting  $p \approx E(p) = \frac{\gamma}{\gamma + \zeta}$  into  $\text{Var}(p)$  and after a few algebraic manipulations, we obtain

$$\text{Var}(p) \approx \frac{p(1 - p)^2}{(\zeta + 1 - p)}.$$

Note that  $\text{Var}(p)$  is a decreasing function of  $\zeta.$  Therefore, for a given value of  $p,$  to make  $\text{Var}(p)$  small, we choose a large value for  $\zeta.$

The goal is to select  $\zeta$  and  $\gamma$  such that the true selected value of  $p$  in simulation studies is closely approximated by its expected value. That is,  $p \approx \frac{\gamma}{\gamma + \zeta},$  or  $\gamma = \frac{p \times \zeta}{1 - p},$  and a large value for  $\zeta$  to minimize  $\text{Var}(p).$  Although multiple pairs of hyper-parameter values can meet these conditions, larger values of  $\zeta$  have been shown to yield more accurate Bayes estimates of  $p.$

Step 4: Bayes estimates are computed as detailed in Section 3.

$$\hat{\mu}_{\text{Bayes}} = \sum_{x=0}^m \binom{m}{x} \frac{(\alpha + S)}{2^m(n + \beta - x)}, \quad \hat{p}_{\text{Bayes}} = \sum_{x=0}^m \binom{m}{x} \frac{(x + \gamma)}{2^m(n + \zeta + \gamma)}.$$



Step 5: The simulation involves calculating the average of the estimates and the square root of the Average Square Error (ASE) based on  $N = 1000$  simulated samples.

$$\bar{\hat{\mu}}_{\text{est}} = \frac{\sum_{i=1}^N \hat{\mu}_i}{N}, \quad \bar{\hat{p}}_{\text{est}} = \frac{\sum_{i=1}^N \hat{p}_i}{N},$$

$$\sqrt{\text{ASE}}_{\mu} = \epsilon_{\mu} = \sqrt{\frac{\sum_{i=1}^N (\hat{\mu}_i - \mu)^2}{N}}, \quad \sqrt{\text{ASE}}_p = \epsilon_p = \sqrt{\frac{\sum_{i=1}^N (\hat{p}_i - p)^2}{N}}.$$

The smaller the ASE, the more accurate the estimate. Note that ASE of each parameter is computed separately based on 1000 simulated samples.

In simulation studies, it is important to note that  $n$  and the “true” values of  $\mu$  and  $p$  should be selected so that the nonlinear Equation (9) has a solution. Recall from Section 2, the equations

$$\frac{\mu}{1 - e^{-\mu}} = \bar{y}, \quad p = \frac{n\mu - (n - m)\bar{y}}{n\mu}$$

for finding MLEs of  $\mu$  and  $p$ . The first equation is nonlinear in  $\mu$  and can be written as

$$\frac{x}{1 - e^{-x}} = \frac{S}{n - m}, \tag{9}$$

where  $\bar{y} = \frac{S}{n-m}$  is the sample mean of non-zero values in the data, and  $S = \sum_{i=m+1}^n y_i$ . Mathematica can be used to find its unique solution.

Recall that the expected value of  $Y$  is given by

$$E(Y) = (1 - p)\mu.$$

By equating the first population moment of  $\text{Zip}(\mu, p)$  to the first sample moment, we obtain  $(1 - p)\mu = \frac{S}{n}$ , which reduces to

$$p = 1 - \frac{S}{n\mu}.$$

The above relationship between  $n$ ,  $\mu$ , and  $p$  provides guidance for selecting the “true” values for  $\mu$ ,  $p$  in simulation studies. For small values of  $n$  and  $\mu$ , a large value for  $p$  must be selected. For example, if  $n = 10$ ,  $\mu = 0.5$ , we expect to have substantial zeros in the data, as  $\mu$  is very small. This means  $p$ , the percentage of zeros in the data, must be large. The value of  $p$  depends on  $S$ , the sum of non-zero values in the data. Using the same values for  $n, \mu$ , if  $S = 1.4$  (a reasonable value, as  $\frac{1.4}{n} < \mu$ ), then  $p = 0.72$ , but if  $S = 6.1$  (not expected, as  $\frac{6.1}{n} > \mu$ ), we obtain  $p = -0.22$ , which is not an accepted value.

Simulation studies have confirmed that when  $n$  and  $\mu$  are small, but the selected value of  $p$  is also small, Equation (9) fails to have a solution.

Furthermore, the mean and square root of the Average Square Error ( $\sqrt{\text{ASE}}$ ) for each parameter are detailed in Tables 1–4. For instance,  $\sqrt{\text{ASE}}$  for the MLE of  $\mu$  is defined as

$$\epsilon_{\mu ML} = \sqrt{\frac{\sum_{i=1}^N (\hat{\mu}_{ML} - \mu)^2}{N}}.$$

As previously discussed, there are multiple options for selecting hyper-parameter values. In Tables 1–4, two sets of hyper-parameter values are utilized, following the proposed method for their selection. These tables demonstrate that as the sample size increases, the Average Squared Error (ASE) for the Maximum Likelihood Estimator (MLE) decreases, as anticipated. Notably, for smaller sample sizes, the Bayes estimator exhibits superior accuracy compared to the MLE. Furthermore, for larger values of  $\zeta$  and  $\beta$ , the accuracies of Bayes estimators for  $p$  and  $\mu$  are enhanced. For instance, in Table 1, with true parameters  $\mu = 2$  and  $p = 0.2$ , two sets of hyper-parameters are selected. Set 1:  $\beta = 275$ ,

$\alpha = \beta \times \mu = 550, \zeta = 215, \gamma = \frac{\zeta \times p}{1-p} = 53.75$ . The  $\sqrt{ASE}$  for  $\mu$  is 0.02014 and for  $p$  is 0.00475 when  $n = 20$ . Set 2:  $\beta = 1150, \alpha = \beta \times \mu = 2300, \zeta = 800, \gamma = \frac{\zeta \times p}{1-p} = 200$  results in  $\sqrt{ASE}$  for  $\mu$  of 0.00504 and for  $p$  of 0.00134 when  $n = 20$ . It is important to note that the Bayes estimators' formulas (8) include  $n$  in their denominators. Thus, for a larger sample size  $n$ , the Average Squared Error for both Bayes estimators somewhat increases, yet they still significantly surpass the MLE in terms of accuracy, as measured by the Average Squared Error (ASE).

**Table 1.** Comparison of the ML and Bayesian estimates with different pairs of hyper-parameters.

$\mu = 2 \quad p = 0.2 \quad (\beta = 275, \alpha = \beta \times \mu; \zeta = 215, \gamma = \frac{\zeta \times p}{1-p})$								
$n$	$\bar{\hat{\mu}}_{ML}$	$\epsilon_{\mu_{ML}}$	$\bar{\hat{p}}_{ML}$	$\epsilon_{p_{ML}}$	$\bar{\hat{\mu}}_{Bayes}$	$\epsilon_{\mu_{Bayes}}$	$\bar{\hat{p}}_{Bayes}$	$\epsilon_{p_{Bayes}}$
20	1.97574	0.44914	0.18126	0.15170	1.99292	0.02014	0.19698	0.00475
50	2.00219	0.26253	0.19498	0.08370	1.98611	0.03007	0.19285	0.00881
100	2.00931	0.19610	0.19704	0.05838	1.97559	0.04103	0.18752	0.01395
$\mu = 2 \quad p = 0.2 \quad (\beta = 1150, \alpha = \beta \times \mu; \zeta = 800, \gamma = \frac{\zeta \times p}{1-p})$								
$n$	$\bar{\hat{\mu}}_{ML}$	$\epsilon_{\mu_{ML}}$	$\bar{\hat{p}}_{ML}$	$\epsilon_{p_{ML}}$	$\bar{\hat{\mu}}_{Bayes}$	$\epsilon_{\mu_{Bayes}}$	$\bar{\hat{p}}_{Bayes}$	$\epsilon_{p_{Bayes}}$
20	1.97574	0.44914	0.18126	0.15170	1.99823	0.00504	0.19915	0.00134
50	2.00219	0.26253	0.19498	0.08370	1.9963	0.00799	0.19783	0.00268
100	2.00931	0.19610	0.19704	0.05838	1.99289	0.01194	0.19582	0.00468

**Table 2.** Comparison of the ML and Bayesian estimates with different pairs of hyper-parameters.

$\mu = 2 \quad p = 0.7 \quad (\beta = 275, \alpha = \beta \times \mu; \zeta = 325, \gamma = \frac{\zeta \times p}{1-p})$								
$n$	$\bar{\hat{\mu}}_{ML}$	$\epsilon_{\mu_{ML}}$	$\bar{\hat{p}}_{ML}$	$\epsilon_{p_{ML}}$	$\bar{\hat{\mu}}_{Bayes}$	$\epsilon_{\mu_{Bayes}}$	$\bar{\hat{p}}_{Bayes}$	$\epsilon_{p_{Bayes}}$
20	1.98591	0.79190	0.67086	0.15811	1.95401	0.04787	0.69407	0.00599
50	1.98265	0.46426	0.69127	0.07691	1.89261	0.10929	0.68545	0.01462
100	1.98076	0.31300	0.69404	0.05428	1.80519	0.19674	0.67207	0.02799
$\mu = 2 \quad p = 0.7 \quad (\beta = 1150, \alpha = \beta \times \mu; \zeta = 1600, \gamma = \frac{\zeta \times p}{1-p})$								
$n$	$\bar{\hat{\mu}}_{ML}$	$\epsilon_{\mu_{ML}}$	$\bar{\hat{p}}_{ML}$	$\epsilon_{p_{ML}}$	$\bar{\hat{\mu}}_{Bayes}$	$\epsilon_{\mu_{Bayes}}$	$\bar{\hat{p}}_{Bayes}$	$\epsilon_{p_{Bayes}}$
20	1.98591	0.79190	0.67086	0.15811	1.98861	0.01185	0.69878	0.00124
50	1.98265	0.46426	0.69127	0.07691	1.97212	0.02836	0.69694	0.00308
100	1.98076	0.31300	0.69404	0.05428	1.94568	0.05483	0.69392	0.00610

**Table 3.** Comparison of the ML and Bayesian estimates with different pairs of hyper-parameters.

$\mu = 12 \quad p = 0.2 \quad (\beta = 675, \alpha = \beta \times \mu; \zeta = 215, \gamma = \frac{\zeta \times p}{1-p})$								
$n$	$\bar{\hat{\mu}}_{ML}$	$\epsilon_{\mu_{ML}}$	$\bar{\hat{p}}_{ML}$	$\epsilon_{p_{ML}}$	$\bar{\hat{\mu}}_{Bayes}$	$\epsilon_{\mu_{Bayes}}$	$\bar{\hat{p}}_{Bayes}$	$\epsilon_{p_{Bayes}}$
20	11.9721	0.88443	0.20039	0.09286	11.9646	0.04374	0.19309	0.00762
50	11.9764	0.53137	0.19949	0.05721	11.9156	0.09248	0.18427	0.01635
100	11.9929	0.38829	0.19972	0.03910	11.8426	0.16546	0.17285	0.02767
$\mu = 12 \quad p = 0.2 \quad (\beta = 3500, \alpha = \beta \times \mu; \zeta = 800, \gamma = \frac{\zeta \times p}{1-p})$								
$n$	$\bar{\hat{\mu}}_{ML}$	$\epsilon_{\mu_{ML}}$	$\bar{\hat{p}}_{ML}$	$\epsilon_{p_{ML}}$	$\bar{\hat{\mu}}_{Bayes}$	$\epsilon_{\mu_{Bayes}}$	$\bar{\hat{p}}_{Bayes}$	$\epsilon_{p_{Bayes}}$
20	11.9721	0.88443	0.20039	0.09286	11.9930	0.00861	0.19804	0.00216
50	11.9764	0.53137	0.19949	0.05721	11.9829	0.01878	0.19523	0.00496
100	11.9929	0.38829	0.19972	0.03910	11.9665	0.03525	0.19090	0.00928

**Table 4.** Comparison of the ML and Bayesian estimates with different pairs of hyper-parameters.

$\mu = 12 \quad p = 0.7 \quad (\beta = 675, \alpha = \beta \times \mu; \quad \zeta = 325, \gamma = \frac{\zeta \times p}{1-p})$								
$n$	$\hat{\mu}_{ML}$	$\epsilon_{\mu_{ML}}$	$\hat{p}_{ML}$	$\epsilon_{p_{ML}}$	$\hat{\mu}_{Bayes}$	$\epsilon_{\mu_{Bayes}}$	$\hat{p}_{Bayes}$	$\epsilon_{p_{Bayes}}$
20	11.9607	1.51174	0.69764	0.10061	11.8783	0.12373	0.69363	0.00643
50	12.0131	0.92864	0.69742	0.06468	11.7048	0.29726	0.68450	0.01556
100	11.9953	0.63406	0.69835	0.04626	11.4338	0.56810	0.67035	0.02971
$\mu = 12 \quad p = 0.7 \quad (\beta = 3500, \alpha = \beta \times \mu; \quad \zeta = 1600, \gamma = \frac{\zeta \times p}{1-p})$								
$n$	$\hat{\mu}_{ML}$	$\epsilon_{\mu_{ML}}$	$\hat{p}_{ML}$	$\epsilon_{p_{ML}}$	$\hat{\mu}_{Bayes}$	$\epsilon_{\mu_{Bayes}}$	$\hat{p}_{Bayes}$	$\epsilon_{p_{Bayes}}$
20	11.9607	1.51174	0.69764	0.10061	11.9761	0.02424	0.69869	0.00133
50	12.0131	0.92864	0.69742	0.06468	11.9408	0.05955	0.69674	0.00328
100	11.9953	0.63406	0.69835	0.04626	11.8824	0.11795	0.69354	0.00647

### Sensitivity Test

A sensitivity analysis, <https://www.investopedia.com/terms/s/sensitivityanalysis.asp> (accessed on 15 April 2024), tests how independent variables, under a set of assumptions, influence the outcome of a dependent variable. This section theoretically demonstrates that the accuracy of Bayesian estimates, as measured by  $\sqrt{ASE}$ , improves with larger values of  $\beta$  and  $\zeta$ . The simulation results, presented in Tables 1–4, align with our theoretical assertions, utilizing two distinct sets of hyper-parameters. To corroborate these findings further, a sensitivity test was executed for a specific set of “true” parameter values:  $\mu = 2$ ,  $p = 0.2$ , and  $n = 100$ . Under these stipulations, we considered small hyper-parameter values

$$(\beta, \zeta) = (271, 211), (273, 213), (275, 215), (277, 217), (279, 219)$$

and large hyper-parameter values

$$(\beta, \zeta) = (1146, 796), (1148, 798), (1150, 800), (1152, 802), (1154, 804),$$

yielding  $N = 1000$  samples of size  $n = 100$  from the ZIP model with  $\mu = 2$  and  $p = 0.2$ . The outcomes of the sensitivity test are summarized in Table 5, revealing a substantial decrease in both  $\sqrt{ASE}$  for estimates of  $\mu$  and  $p$ , when transitioning from smaller to larger values of hyper-parameters. For example, from a small values  $\beta = 271, \zeta = 211$  to a large values  $\beta = 1146, \zeta = 796$ ,  $\sqrt{ASE}$  changes from 0.04149 to 0.01198 for  $\mu$  and from 0.01414 to 0.00470 for  $p$ . This trend is also evident within each group. For the smaller hyper-parameter group, the  $\sqrt{ASE}$  for  $\mu$  diminishes from 0.04149 to 0.04058, and the  $\sqrt{ASE}$  for  $p$  from 0.01414 to 0.01377 corresponding to hyper-parameters  $\beta = 271, \zeta = 211$  and  $\beta = 279, \zeta = 219$ , as the hyper-parameter values incrementally increase by approximately 0.7% for  $\beta$  on each step (from 271 to 273) and 0.9% for  $\zeta$  on each step (from 211 to 213). A similar pattern is observed for the larger group, even though the increments for  $\beta$  and  $\zeta$  are about 0.17% on each step ( $\beta$  from 1146 to 1148) and 0.25% on each step ( $\zeta$  from 796 to 798), respectively. This sensitivity analysis validates that the precision of Bayesian estimates is contingent upon the choice of hyper-parameters, with larger  $\beta$  and  $\zeta$  values enhancing estimate accuracy in terms of  $\sqrt{ASE}$ . As shown in Table 5, there is a gradual increase in hyper-parameters and a corresponding steady decrease in  $\sqrt{ASE}$ . For the practical application of the proposed method, it is advisable to establish a stopping rule for the increment in the hyper-parameters. Specifically, the increase in hyper-parameters should continue until no significant reduction in  $\sqrt{ASE}$  is observed.

**Table 5.** Sensitivity test for accuracy of Bayesian estimates by selecting hyper-parameters.

$\mu = 2 \quad p = 0.2 \quad n = 100$							
$\beta$	$\xi$	$\alpha = \beta \times \mu$	$\gamma = \frac{\xi \times p}{1-p}$	$\hat{\mu}_{Bayes}$	$\epsilon_{\mu_{Bayes}}$	$\hat{p}_{Bayes}$	$\epsilon_{p_{Bayes}}$
271	211	542	52.75	1.97532	0.04149	0.18735	0.01414
273	213	546	53.25	1.97545	0.04126	0.18743	0.01405
275	215	550	53.75	1.97559	0.04103	0.18752	0.01395
277	217	554	54.25	1.97572	0.04080	0.18760	0.01386
279	219	558	54.75	1.97586	0.04058	0.18768	0.01377
1146	796	2292	199	1.99287	0.01198	0.19580	0.00470
1148	798	2296	199.5	1.99288	0.01196	0.19581	0.00469
1150	800	2300	200	1.99289	0.01194	0.19582	0.00468
1152	802	2304	200.5	1.99290	0.01192	0.195825	0.004669
1154	804	2308	201	1.99291	0.01190	0.195834	0.004656

### 5. Numerical Example

#### 5.1. The Synthetic Auto Telematics Dataset

This section focuses on the practical application of the ML and Bayesian methods discussed in the article, using a real dataset.

##### 5.1.1. Data and Basic Descriptive Statistics

The dataset selected for our analysis is the synthetic auto telematics dataset, accessible at <http://www2.math.uconn.edu/~valdez/data.html> (accessed on 20 January 2024). Using real-world insurance datasets frequently presents significant challenges due to the need to maintain the confidentiality of individual customer information. The synthetic dataset developed by So et al. (2021) was designed to closely mimic genuine telematics auto data originally sourced from a Canadian-based insurer, which offered a UBI program that was launched in 2013 to its automobile insurance policyholders. It was created to be openly accessible to researchers for their studies, enabling them to work with data that closely resembles real-world scenarios while ensuring privacy protection. Therefore, to evaluate the practical relevance of the proposed models, we decided to use this synthetic dataset instead of real data.

The synthetic dataset consists of 100,000 policies with a total of 52 variables, which includes the NB\_Claim variable representing the number of claims. For further information on the 52 variables, please refer to So et al. (2021). Figure 2 displays a histogram that visualizes the descriptive statistics for the number of claims in the synthetic auto telematics dataset. The average number of claims is 0.04494, with the minimum number of claims being zero, with a frequency of 95,728, and a maximum number of claims reaching three, with a frequency of 11. The respective frequencies for one and two claims are 4061 and 200. Furthermore, the standard deviation of the number of claims stands at 0.21813, indicating a wide dispersion across the dataset. The skewness of the number of claims, valued at 5.1138, demonstrates that the data are right-skewed. Moreover, a kurtosis of the number of claims of 31.538 signifies the presence of a heavy tail in the data distribution. Notably, the histogram reveals an exceptionally high frequency of zero claims, characteristic of Zero-Inflated model data, underscoring the limitation of regular distributions such as the Poisson or Negative Binomial in effectively modeling the number of claims.

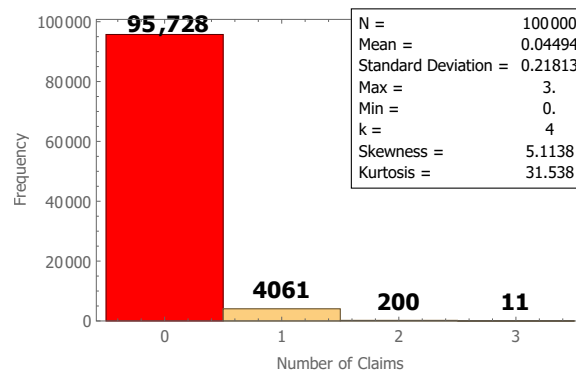


Figure 2. Histogram of the synthetic auto telematics dataset.

### 5.1.2. Goodness-of-Fit of the Synthetic Auto Telematics Dataset Chi-Square Goodness-Fit-Test

To assess the suitability of the Zero Inflated Poisson model for the synthetic auto telematics dataset, we consider the hypotheses:

$H_0$ : The data are generated from a Zero-Inflated Poisson distribution.

$H_a$ : The data are not generated from a Zero-Inflated Poisson distribution.

Klugman et al. (2012) have suggested that the Kolmogorov–Smirnov and Anderson–Darling tests apply solely to individual datasets. Hence, for a grouped dataset like the synthetic auto telematics dataset, the Chi-Square goodness-of-fit test is utilized to ascertain if a categorical variable conforms to a hypothesized frequency distribution. The degrees of freedom for the test statistic are given by  $k - \text{number of estimated parameters} - 1 = k - 3$ , where  $k$  is the number of categories.

Given that two parameters,  $\mu$  and  $p$ , are estimated, the Chi-square test formula is represented as:

$$\chi^2 = \sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i},$$

where  $O_i$  denotes the observed frequencies, and  $E_i$  refers to the expected frequencies based on the model.

Applying this approach to the specific dataset, we have:

$$\vec{O} = \begin{bmatrix} 95,728 \\ 4061 \\ 200 \\ 11 \end{bmatrix}$$

$$\vec{E} = \begin{bmatrix} 100,000 \times f_{ZIP}(0, \hat{\mu}, \hat{p}) \\ 100,000 \times f_{ZIP}(1, \hat{\mu}, \hat{p}) \\ 100,000 \times f_{ZIP}(2, \hat{\mu}, \hat{p}) \\ 100,000 \times (1 - f_{ZIP}(0, \hat{\mu}, \hat{p}) - f_{ZIP}(1, \hat{\mu}, \hat{p}) - f_{ZIP}(2, \hat{\mu}, \hat{p})) \end{bmatrix}$$

Given that we have four categories (zero, one, two, and greater than or equal to three), we determine the degrees of freedom ( $df$ ) to be 1. At significance levels of 0.05 or 0.10, the null hypothesis  $H_0$  is rejected if the  $\chi^2$  statistic,

$$\chi^2 = \sum_{i=1}^4 \frac{(O_i - E_i)^2}{E_i},$$

exceeds 3.841 or 2.706, respectively. Therefore, a smaller value of the  $\chi^2$  statistic indicates stronger support for  $H_0$ , implying a better fit of the data to the Zero-Inflated Poisson model.

### Score Test

The score test, [https://en.wikipedia.org/wiki/Score\\_test](https://en.wikipedia.org/wiki/Score_test) (accessed on 17 April 2024), is utilized to evaluate the constraints of parameters within a statistical model. The process of formulating the score test is detailed as follows:

The hypotheses are defined by:

$$H_0 : \theta = \theta_0 \text{ versus } H_a : \theta \neq \theta_0$$

The test statistic is given as:

$$S_1(\theta_0) = \frac{U^2(\theta_0)}{I(\theta_0)} \sim \chi^2_{(1)}$$

where

$$U(\theta_0) = \left. \frac{\partial \ln(L(\theta|x))}{\partial \theta} \right|_{\theta=\theta_0}, \quad I(\theta) = -E \left[ \left. \frac{\partial^2 \ln(f(X;\theta))}{\partial \theta^2} \right| \theta \right].$$

The critical values are  $\chi^2_{0.05,1} = 3.841$  and  $\chi^2_{0.10,1} = 2.706$ . Reject  $H_0$  if  $S_1(\theta_0) > \text{critical value}$ .

The distribution under test is defined as:

$$f_{ZIP}(y|\mu, p) = \begin{cases} p + (1-p)e^{-\mu}, & y = 0 \\ (1-p)f_0(y|\mu), & y = 1, 2, \dots \end{cases}$$

Considering the hypotheses  $H_0 : p = 0$  versus  $H_a : p \neq 0$ , which are equivalent to:

$H_0$  : The data come from the Poisson( $\mu$ )

$H_a$  : The data come from the ZIP( $\mu, p$ ).

Using  $l = \ln(L(\mu, p))$  from Section 2, for testing  $H_0 : p = 0$  versus  $H_a : p \neq 0$ , we obtain:

$$U(0) = me^\mu - n, \quad I(0) = (n - m) + \frac{m(1 - e^{-\mu})^2}{e^{-2\mu}}$$

As a result, the Score test  $S_1 = S_1(0) = \frac{U^2(0)}{I(0)}$  can be expressed as:

$$S_1 = \frac{(m - ne^{-\mu})^2}{ne^{-2\mu} + m - 2me^{-\mu}}$$

where  $n$  is the sample size, and  $m$  is the number of zeros in the sample. To compute  $S_1$ , replace  $\mu$  with its MLE under  $H_0$ , that is,  $\hat{\mu} = \bar{y}$ . The score test is asymptotically distributed as Chi-square with 1 degree of freedom. Thus, the larger the value of  $S_1$ , the stronger the evidence in support of  $H_a$ , indicating a better fit for the Zero-Inflated Poisson model.

As previously mentioned, determining the optimal hyper-parameter values poses a significant challenge. Given that the “true” values of parameters  $\mu$  and  $p$  are unknown in practice, we propose a data-driven approach to determine hyper-parameter values, utilizing their Maximum Likelihood Estimates (MLEs), as discussed in the Simulation section.

Note that in real practice, unlike in simulations, the “true” values of parameters are unknown. However, we are still required to choose the “best” values for the hyper-parameters. This can be accomplished through the MLEs of  $\mu$  and  $p$ . The method proposed here for selecting hyper-parameter values has also been employed by [Deng and Aminzadeh \(2023\)](#) and [Aminzadeh and Deng \(2022\)](#), where the authors use MLEs to choose hyper-parameter values.

The following steps are proposed for selecting hyper-parameter values for “real” data:

1. Apply a goodness-of-fit test to assess how well the data fits the Zero-Inflated Poisson distribution. Goodness-of-fit tests extensively developed in the literature for ZIP( $\mu, p$ ) include the Score and Chi-square tests considered in this article.
2. Compute the MLEs  $\hat{\mu}$  and  $\hat{p}$ , as detailed in Section 2.
3. Choose a large value for  $\beta$  and let  $\alpha = \beta \times \hat{\mu}$ .
4. Choose a large value for  $\xi$  and let  $\gamma = \frac{\hat{p} \times \xi}{1 - \hat{p}}$ , or  $\xi = \frac{\gamma \times (1 - \hat{p})}{\hat{p}}$  if selecting  $\gamma$  first.

This approach ensures the selection of hyper-parameter values that best fit the given real data under the assumption of a Zero-Inflated Poisson distribution.

The table indicates that the Zero-Inflated Poisson model provides a good fit to the data according to the Maximum Likelihood (ML) estimation method. Given the large sample size ( $n = 100,000$ ), the Bayesian method does not significantly outperform the ML method. The findings presented in Table 6 align with our expectations: larger hyper-parameter values enhance the model’s fit to the data. Utilizing the Chi-square test with a critical value of  $\chi^2_{(0.05,1)} = 3.841$ , Table 6 demonstrates that the ZIP model, employing ML estimates as well as Bayes 4 and Bayes 5 sets of hyper-parameters, adequately fits the data.

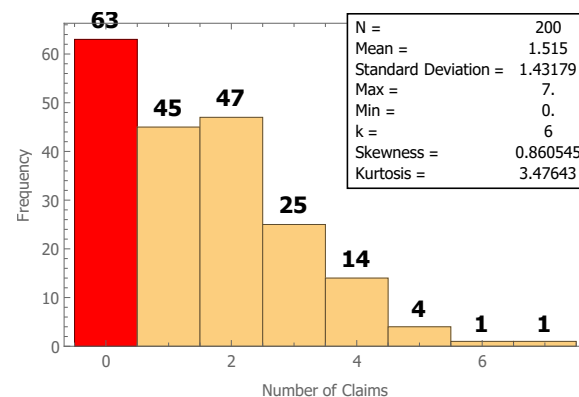
**Table 6.** Goodness-of-fit tests of the ML and Bayesian estimates with different pairs of hyper-parameters.

Hyper-Parameters		Estimates of $\mu$	Estimate of $p$	$\chi^2$ Goodness-of-Fit
MLE		0.10219	0.56024	2.20614
Bayes 1	$\beta = 675, \gamma = 325$	0.08640	0.47911	10.0378
Bayes 2	$\beta = 6500, \gamma = 4500$	0.08797	0.48471	9.24796
Bayes 3	$\beta = 146,500, \gamma = 144,500$	0.09374	0.51477	5.48988
Bayes 4	$\beta = 954,500, \gamma = 9,444,500$	0.10210	0.55976	2.20985
Bayes 5	$\beta = 39,546,500, \gamma = 39,444,500$	0.10217	0.56013	2.20692

The score test yields a value of 3.67026, which exceeds  $\chi^2_{1,0.10} = 2.706$ . Consequently, we reject the null hypothesis (that the data follows a Poisson distribution) in favor of the alternative hypothesis, affirming that the ZIP model better suits the dataset at a significance level of 0.10.

### 5.2. A Simulated Data from ZIP Model

In this section, a random sample of size 200 from a ZIP model with  $\mu = 2$  and  $p = 0.2$  is generated. The summary of the generated data is presented in Figure 3.



**Figure 3.** Histogram of the sample generated from ZIP with  $\mu = 2$  and  $p = 0.2$ .

There are six categories (zero, one, two, three, four, and greater than or equal to five). The sample mean is 1.515, and the sample standard deviation is 1.41179. The maximum value observed is seven, and the minimum is zero. The skewness = 0.86055 which indicates that the data are right-skewed, and the Kurtosis = 3.47643 which indicates that the data

have a heavy tail. The histogram confirms the high frequency of zeros. We use the same test statistics as in Section 5.1. The observed values are as follows:

$$\vec{O} = \begin{bmatrix} 63 \\ 45 \\ 47 \\ 25 \\ 14 \\ 6 \end{bmatrix}$$

The expected frequencies are

$$\vec{E} = \begin{bmatrix} 200 \times f_{ZIP}(0, \hat{\mu}, \hat{p}) \\ 200 \times f_{ZIP}(1, \hat{\mu}, \hat{p}) \\ 200 \times f_{ZIP}(2, \hat{\mu}, \hat{p}) \\ 200 \times f_{ZIP}(3, \hat{\mu}, \hat{p}) \\ 200 \times f_{ZIP}(4, \hat{\mu}, \hat{p}) \\ 200 \times (1 - \sum_{i=0}^4 f_{ZIP}(i, \hat{\mu}, \hat{p})) \end{bmatrix}$$

where  $\hat{\mu}$  and  $\hat{p}$  are estimates of  $\mu$  and  $p$ , respectively.

The table was prepared using the same method described in the Simulation Section. That is, for a selected value of  $\beta$ , we let  $\alpha = \beta \times \mu$ , and for a selected  $\gamma$ , we let  $\xi = \frac{\gamma \times (1-p)}{p}$ .

Table 7 reveals that all of the ZIP models (with different hyper-parameter values) are a good fit for the dataset, as all the test values are smaller than  $\chi^2_{3,0.05} = 7.815$ . Additionally, the  $\chi^2$  value based on ML is larger than those based on the Bayesian method, suggesting that the Bayesian method provides a better fit for the data than the ML method. Both the ML and Bayesian approaches confirm that the data fit the ZIP model well, as expected.

**Table 7.** Goodness-of-fit test of simulated data with different pairs of hyper-parameters. we confirm it is correct, There are different ways to select other hyper-parameters.

	Hyper-Parameters	Estimates of $\mu$	Estimate of $p$	$\chi^2$ Goodness-of-Fit
	MLE	1.87122	0.19037	0.71024
	Bayes 1 $\beta = 85, \gamma = 104$	1.86634	0.18819	0.70968
	Bayes 2 $\beta = 85, \gamma = 115$	1.86634	0.18903	0.70926
	Bayes 3 $\beta = 87, \gamma = 115$	1.86738	0.18903	0.70919
	Bayes 4 $\beta = 90, \gamma = 110$	1.86891	0.18867	0.70963
	Bayes 5 $\beta = 90, \gamma = 115$	1.86891	0.18903	0.70937
	Bayes 6 $\beta = 4, \gamma = 6$	1.80385	0.16304	1.10954

The calculated score test value  $S_1 = 8.06051$  is greater than the critical value of 3.841. Therefore, we reject the null hypothesis that the data follow a Poisson distribution and accept the alternative hypothesis that the Zero-Inflated Poisson (ZIP) model is a better fit for the dataset.

Table 8 uses a similar approach to Section 5.1 for selecting hyper-parameter values. That is, for a selected value of  $\beta$ , we let  $\alpha = \beta \times \hat{\mu}$ , and for a selected  $\gamma$ , we let  $\xi = \frac{\gamma \times (1-\hat{p})}{\hat{p}}$ , showing that all of the ZIP models fit the dataset well as the critical value is 7.815. Based on the MLE method, the test value of 0.71024 is slightly smaller than other Bayes-based test values. We noticed that, since MLE estimates of  $\mu$  and  $p$  are used to determine the hyper-parameters, the Bayesian estimates of  $\mu$  and  $p$  are somewhat underestimated. Additionally, the larger the values of the hyper-parameters, the better the model fits the data.



**Table 8.** Goodness-of-fit test of simulated data with different pairs of hyper-parameters.

	Hyper-Parameters	Estimates of $\mu$	Estimate of $p$	$\chi^2$ Goodness-of-Fit
	MLE	1.87122	0.19037	0.71024
	Bayes 1 $\beta = 85, \gamma = 104$	1.82315	0.18156	0.81937
	Bayes 2 $\beta = 85, \gamma = 115$	1.82315	0.18219	0.81904
	Bayes 3 $\beta = 87, \gamma = 115$	1.82352	0.18219	0.81720
	Bayes 4 $\beta = 90, \gamma = 110$	1.82407	0.18192	0.81464
	Bayes 5 $\beta = 90, \gamma = 115$	1.82407	0.18219	0.81450
	Bayes 6 $\beta = 210, \gamma = 315$	1.83893	0.18683	0.75708

Tables 7 and 8 demonstrate that as hyper-parameter values increase, the Chi-square test values decrease, suggesting a better fit for the data. However, a notable exception is observed with a small pair of hyper-parameters in Table 7, where the Chi-square test value is larger than those obtained from MLE. Specifically, in Table 7, the entry “Bayes 6”, which utilizes very small hyper-parameters, has a Chi-square value of 1.10954, significantly higher than other values in the table. This observation confirms that larger hyper-parameters result in a better fit.

## 6. Conclusions

Many insurance claims datasets exhibit a high frequency of no claims. To analyze such data, researchers have proposed the use of Zero-Inflated models. These models incorporate parameters with covariates using link functions and are referred to as Zero-Inflated Regression Models. Various methods, including Maximum Likelihood (ML), Bayesian, and Decision Tree, among others, have been employed to fit the data. A significant distinction of this research from prior studies is the introduction of a novel Bayesian approach for the Zero-Inflated Poisson model without covariates. This study aims to develop the statistical ZIP model by estimating the unknown parameters  $\mu$  and  $p$ . To our knowledge, similar research has not been documented in the literature. We derive analytical solutions for the Maximum Likelihood estimators of the unknown parameters  $\mu$  and  $p$ . Additionally, we present analytical closed-form solutions for Bayesian Estimators of these parameters by selecting conjugate prior distributions: Gamma for  $\mu$  and Beta for  $p$ , respectively. The comparison between the ML and Bayesian methods indicates that the Bayesian method, utilizing a data-driven approach (which employs MLEs of parameters  $p$  and  $\mu$  to select hyper-parameter values), surpasses the ML method in accuracy. We derived the predictive distribution based on the posterior distribution, predicting possible future observations from past observed values and calculating percentiles. Furthermore, we demonstrate that larger values of the hyper-parameters  $\beta$  and  $\xi$  enhance the accuracy of the Bayesian estimates. Our findings are confirmed through a sensitivity test. The real-life data from the synthetic auto telematics dataset and simulated data from a specified Zero-Inflated Poisson model, using the methods proposed in this paper, validated the goodness-of-fit (GOF) to the ZIP model based on both Chi-square and Score tests. However, the simulation has limitations, including 1. Ensuring the data adequately fit the Zero-Inflated Poisson model in real applications. 2. The sample size is sufficiently large so that the MLEs of parameters  $p$  and  $\mu$  are accurate. 3. The selection of hyper-parameter values aligns with the MLEs, as elaborated in Section 4 (Step 3). Future research could explore non-traditional discrete Time Series modeling for Zero-Inflated data to forecast the number of claims at specific future points and extend the Bayesian analysis to the Zero-Inflated Negative Binomial (ZINB) model without covariates. Furthermore, in this article the parameters  $\mu$  and  $p$  are assumed to be independent. A future line of improvement would consist of introducing some copula that allows contemplating the dependence between the parameters, as a small value of  $\mu$  corresponds to a large value of  $p$ .

**Author Contributions:** Conceptualization, M.D., M.S.A. and B.S.; Methodology, M.D. and M.S.A.; Software, M.D., M.S.A. and B.S.; Validation, M.S.A.; Formal analysis, M.D. and M.S.A.; Investigations, M.D., M.S.A. and B.S.; Resources, M.D., M.S.A. and B.S.; Data curation, B.S.; writing—original draft preparation, M.D., M.S.A. and B.S.; writing—review and editing, M.D., M.S.A. and B.S.; visualization, M.D. and M.S.A. supervision, Not applicable; project administration, Not applicable; funding acquisition, Not applicable. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research received no external funding.

**Data Availability Statement:** The data presented in this study are available upon request from the corresponding author

**Acknowledgments:** The authors are grateful for the editors' and reviewers' invaluable time and suggestions to enhance the article's presentation.

**Conflicts of Interest:** The authors declare no conflict of interest.

## References

- Aminzadeh, Mostafa S., and Min Deng. 2022. Bayesian Estimation of Renewal Function Based on Pareto-distributed Inter-arrival Times via an MCMC Algorithm. *Variance* 15: p1–p15
- Angers, Jean-François, and Atanu Biswas. 2003. A Bayesian analysis of zero-inflated generalized Poisson model. *Computational Statistics & Data Analysis* 42: 37–46. ISSN 0167-9473. [\[CrossRef\]](#)
- Boucher, Jean-Philippe, Michel Denuit, and Montserrat Guillén. 2007. Risk Classification for Claim Counts A Comparative Analysis of Various Zero-inflated Mixed Poisson and Hurdle Models. *North American Actuarial Journal* 11: 110–31. [\[CrossRef\]](#)
- Chen, Kun, Rui Huang, Ngai Hang Chan, and Chun Yip Yau. 2019. Subgroup analysis of Zero-Inflated Poisson regression model with applications to insurance data. *Insurance: Mathematics and Economics* 86: 8–18. [\[CrossRef\]](#)
- Deng, Min, and Mostafa S. Aminzadeh. 2023. Bayesian Inference for the Loss Models via Mixture Priors. *Risks* 11: 156. [\[CrossRef\]](#)
- Ghosh, Sujit K., Pabak Mukhopadhyay, and Jye-Chyi Lu. 2006. Bayesian analysis of zero-inflated regression models. *Journal of Statal Planning & Inference* 136: 1360–75.
- Klugman, Stuart A., Harry H. Panjer, and Gordon E. Willmot. 2012. *Loss Models from Data to Decisions*, 3rd ed. New York: John Wiley.
- Lambert, Diane. 1992. Zero-Inflated Poisson regression, with an application to defects in manufacturing. *Technometrics* 34: 1–14. [\[CrossRef\]](#)
- Lee, Simon C. 2021. Addressing imbalanced insurance data through Zero-Inflated Poisson regression with boosting. *ASTIN Bulletin: The Journal of the IAA* 51: 27–55. [\[CrossRef\]](#)
- Meng, Shengwang, Yaqian Gao, and Yifan Huang. 2022. Actuarial intelligence in auto insurance: Claim frequency modeling with driving behavior features and improved boosted trees. *Insurance: Mathematics and Economics* 106: 115–27. [\[CrossRef\]](#)
- Mouatassim, Younès, and El Hadj Ezzahid. 2012. Poisson regression and Zero-Inflated Poisson regression: Application to private health insurance data. *European Actuarial Journal* 2: 187–204. [\[CrossRef\]](#)
- Mullahy, John. 1986. Specification and testing of some modified count data models. *Journal of Econometrics* 33: 341–65. [\[CrossRef\]](#)
- Perumean-Chaney, Suzanne E., Charity Morgan, David McDowall, and Inmaculada Aban. 2013. Zero-inflated and overdispersed: What's one to do? *Journal of Statistical Computation and Simulation* 83: 1671–83. [\[CrossRef\]](#)
- So, Banghee, Jean-Philippe Boucher, and Emiliano A. Valdez. 2021. Synthetic dataset generation of driver telematics. *Risks* 9: 58. [\[CrossRef\]](#)
- Zhang, Pengcheng, David Pitt, and Xueyuan Wu. 2022. A new multivariate zero-inflated hurdle model with applications in automobile insurance. *ASTIN Bulletin: The Journal of the IAA* 52: 393–416. [\[CrossRef\]](#)
- Zhou, He, Wei Qian, and Yi Yang. 2022. Tweedie gradient boosting for extremely unbalanced zero-inflated data. *Communications in Statistics-Simulation and Computation* 51: 5507–29. [\[CrossRef\]](#)

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.