Zero-Inflated Negative Binomial-Sushila Distribution and Its Application

K.M.Sakthivel¹ , C.S. Rajitha² and K.B.Alshad³

¹,²,³ Department of Statistics, Bharathiar University, Coimbatore-641046, Tamilnadu, India.
¹ sakthithebest@gmail.com

Abstract

In statistics literature, there is significant study of mixtures and compound probability distributions used for count model especially for the data contains excess zeros. In this paper, we introduce a new probability distribution which is obtained as a compound of zero-inflated negative binomial (ZINB) distribution and Sushila distribution and it is named as zero-inflated negative binomial-Sushila (ZINB-S) distribution. It can be used as an alternative and effective way of modeling over dispersed count data. The probability mass function (PMF) and some vital characteristics of ZINB-S distribution are derived. MLE method is employed for estimating the model parameters. Further the example is given to show that ZINB-S provides better fit compare to traditional models for over dispersed count data.

AMS Subject Classification: 2010: 62F10; 62P05

Key Words and Phrases: Negative binomial-Sushila distribution, Method of maximum likelihood, Zero-inflated distribution, Poisson distribution, Negative binomial distribution

1 Introduction

Modeling of count data plays a vital role in many research areas, such as public health, epidemiology and insurance etc. And one of the commonly used models for modeling the count data is Poisson model and it assumes that mean and variance of the model are equal. But in many applications this restriction is violated due to the over dispersed count data. Hence researchers utilized negative binomial model and Generalized Poisson model (Consul and Jain, 1973) for analyzing this type of data. In addition to the mean parameter this models incorporates one more parameter known as dispersion parameter, which helps to detect the over dispersion.
or under dispersion in the population. Moreover, the causes of over dispersion is
the presence of excess number of zero counts in the data known as zero inflation.
And the concept of zero inflation was first introduced by Neyman (1939) and Feller
(1943). While considering the actuarial applications, the count data often shows
excess number of zero counts. As a result, numerous research have been produced
for developing the flexible probability models such as mixed distribution models
and zero-inflated models for modeling this kind of data. And this new models pro-
vides more suitable fitting of over dispersed count data than the conventional count
distribution models such as Poisson and negative binomial models. Zero-inflated
models (Lambert, 1992, Greene, 1994) hold some desirable properties for modeling
the zero-inflated data. These models were formulated as a mixture of two com-
ponents, the zero counts and the positive counts coming from a truncated count
distribution. Some standard zero inflated distributions considered in the actuarial
literature are zero-inflated Poisson (ZIP) model and zero-inflated negative binomial
(ZINB) model (Neelon et al., 2010). Zero inflation can be observed in many other
situations such as manufacturing, road safety, epidemiology etc. Some work can be
found in health insurance (Mouatassim and Ezzahid, 2012) and dental epidemiology
(Bohning et al., 1999). Recent studies shows that negative binomial mixture distribu-
tion provides better fit than its base line distributions such as Poisson and negative
binomial distribution. (Zamani and Ismail, 2010; Gomez-Deniz et al., 2008; Wang,
2011). Recently, a negative binomial two parameter Lindley distribution (Denthet
et al., 2016) and negative binomial generalized exponential distribution (Aryuyuyen
and Bodhisuwan, 2013) was developed to model the count data.

Further Yip and Yao (2005) sufficiently discussed about some of the zero-inflated
count distribution models such as ZIP, ZINB, zero-inflated generalized Poisson
(ZIGP), zero inflated double Poisson (ZIDP) distributions etc. Subsequently Aryuyuyen
et al., (2014) developed a zero inflated negative binomial generalized exponential
(ZINB-GE) distribution for modeling the count data with excess number of zero
counts and showed that it provides better fit compared to the ZIP and ZINB distri-
butions. Saengthong et al., (2015) formulated a zero inflated negative binomial-Crack
distribution (ZINB-CR) distribution and showed that it provides better fit compared
to ZIP, ZINB and negative binomial crack (NB-CR) distribution.

For modeling count data, various methods for estimation of parameters are used
in the literature. Famoye (1997) used the method of maximum likelihood estimation
(MLE), minimum chi-square (MC), first two moments and proportion of zeros etc
for estimating the parameters of generalized negative binomial distribution (GNBD)
and compares all the estimators through relative efficiencies. For estimating the pa-
rameters of the generalized Poisson distribution (GPD), Wagh and Kamalija (2017)
used both moment estimation (ME) and MLE and showed that ME performs rela-
tively better compared to MLE when sample size is small.

In this paper we introduced a new zero-inflated negative binomial-Sushila (ZINB-
S) distribution as an alternative distribution to the count data with excess number of
zero counts, which is obtained by considering the excess zero counts and zeros coming
from the negative binomial- Sushila (NB-S) distribution in one part of the model
and the positive counts are coming from a truncated NB-S distribution. The NB-S
distribution has recently been developed by Yamrubboon et al., (2017). Negative
binomial Lindley (NB-L) distribution (Zamani and Ismail, 2010) is a special case of this distribution. The rest of the paper has been organized as follows. In section 2, the PMF and graphical representation of the probability mass function of the zero-inflated negative binomial- Sushila distribution is given. Some characteristics such as the factorial moments, mean and variance of the ZINB-S are provided in section 3. Section 4 discusses about the parameter estimation of the ZINB-S distribution using maximum likelihood estimation (MLE) method. Application of the ZINB-S distribution to the real data set is given in section 5. Section 6 provides the discussions and conclusions about the study.

2 Zero-inflated Negative Binomial-Sushila Distribution

2.1 Zero-inflated Count Models

The probability mass function of the zero-inflated count model can be written in the following form

\[ P(X = x) = \begin{cases} \pi + (1 - \pi)g(0), & \text{if } x = 0 \\ (1 - \pi)g(x, \Theta), & \text{if } x = 1, 2, \ldots \end{cases} \] (1)

where \( X \) is the count random variable and \( \pi \) is the proportion of the extra zero counts. \( g(x, \Theta) \) is the pmf of \( X \) with parameter space \( \theta \) and \( \pi \) represents the zero-inflation parameter and \( 0 < \pi < 1 \).

2.2 Negative Binomial-Sushila Distribution

The probability mass function of the NB-S distribution is given below:

\[ f(x, r, \alpha, \theta) = \frac{\theta^x}{\theta + 1} \left( \frac{r}{\theta} \right)^{x+r-1} \sum_{j=0}^{x} \binom{x}{j} (-1)^j \frac{(\theta + \alpha(r+j) + 1)}{(\theta + \alpha(r+j))^2} \] (2)

Where \( x = 0, 1, 2, \ldots ; r, \alpha, \theta > 0 \).

The mean, variance and factorial moments of the NB-S distribution are given below

\[ E(X) = r \left[ \frac{\theta^2(\theta - \alpha + 1)}{(\theta + 1)(\theta - \alpha)^2} - 1 \right] \] (3)

\[ V(X) = \frac{r^2\delta_3 + r\delta_3 - 2r\delta_2}{\delta_1} - \frac{r^2\delta_2^2}{\delta_1^2} + r \] (4)

where

\[ \delta_1 = \frac{\theta + 1}{\theta^2}, \delta_2 = \frac{\theta - \alpha + 1}{(\theta - \alpha)^2}, \delta_3 = \frac{\theta - 2\alpha + 1}{(\theta - 2\alpha)^2} \]
and

\[ \mu_k[X] = E[X(X-1)(X-2)\ldots(X-k+1)] = \left( \frac{\Gamma(r+k)}{\Gamma(r)} \right) \sum_{j=0}^{k} \binom{k}{j} (-1)^j \frac{\theta^2(\theta - (k-j)\alpha + 1)}{(\theta + 1)(\theta - (k-j)\alpha)^2} \] (5)

where \( k = 1, 2, \ldots \) and \( \Gamma(.) \) is the incomplete gamma function represented by

\[ \Gamma(t) = \int_{0}^{\infty} x^{t-1}e^{-x}dx, \quad t > 0 \]

### 2.3 Proposed Zero-inflated Negative Binomial-Sushila Distribution

If \( X/\lambda \sim ZINB(r, p = e^{-\lambda}, \pi), \lambda \sim Sushila(\alpha, \theta) \) then the PMF of \( X \) can be obtained as

\[ g(X = x/\lambda) = \int_{0}^{\infty} p(X = x/\lambda)f(\alpha, \lambda, \theta)d\lambda \]

where

\[ p(X = x/\lambda) = \begin{cases} \pi + (1 - \pi)e^{-\lambda x}, & \text{if } x = 0 \\ (1 - \pi)(x+r-1)e^{-\lambda x}(1 - e^{-\lambda})^x, & \text{if } x \neq 0 \end{cases} \]

The PMF of the \( ZINB - S(\pi, r, \alpha, \theta) \), when \( x = 0 \) is given below

\[ g(x, r, \alpha, \theta, \pi) = \pi + (1 - \pi)\frac{\theta^2(\theta + r\alpha + 1)}{(\theta + 1)(\theta + r\alpha)^2} \]

The PMF of the ZINB-S distribution, when \( x \neq 0 \) can be obtained as

\[ g(X = x, r, \alpha, \pi) = \int_{0}^{\infty} (1 - \pi)\left(\frac{x + r - 1}{x}\right)e^{-\lambda x}(1 - e^{-\lambda})^x f(\lambda, \alpha, \theta)d\lambda \]

\[ = (1 - \pi)\left(\frac{x + r - 1}{x}\right) \sum_{j=0}^{x} (-1)^j \frac{\theta^2(\theta + (r+j)\alpha + 1)}{(\theta + 1)(\theta + (r+j)\alpha)^2} \]

Therefore the ZINB-S distribution is given by

\[ g(x, r, \pi, \alpha, \theta) = \begin{cases} \pi + (1 - \pi)\frac{\theta^2(\theta + r\alpha + 1)}{(\theta + 1)(\theta + r\alpha)^2}, & \text{when } x = 0 \\ (1 - \pi)\frac{\theta^2}{(\theta + 1)}\left(\frac{x+r-1}{x}\right) \sum_{j=0}^{x} (-1)^j \frac{\theta + \alpha(r+j) + 1}{(\theta + \alpha(r+j))^2}, & \text{when } x = 1, 2, \ldots \end{cases} \] (6)

where \( x = 0, 1, 2, \ldots; r, \alpha, \theta > 0, 0 < \pi < 1 \). When \( \pi = 0 \), ZINB-S distribution reduces to NB-S distribution and ZINB-S distribution reduces to negative binomial -Lindley(NB-L) distribution when \( \pi = 0 \) and \( \alpha = 1 \). Hence the ZINB-S is a generalized form of NB-S distribution and NB-L distribution. Figure 1 illustrates the PMF of ZINB-S distribution with different values of parameters \( \theta \) and \( \pi \).
3 Characteristics of the ZINB-S distribution

In this section, we introduce some basic characteristics of the ZINB-S distribution.

**Theorem 1.** \( X \sim ZINB - S(\pi, r, \alpha, \theta) \), the factorial moments of order \( k \) of \( X \) can be written as

\[
\mu_k[X] = (1 - \pi) \left( \frac{\Gamma(r + k)}{\Gamma(r)} \right) \sum_{j=0}^{k} \binom{k}{j} (-1)^j \frac{\theta^2(\theta - (k - j)\alpha + 1)}{(\theta + 1)(\theta - (k - j)\alpha)^2}
\] (7)

We can easily obtain the mean and variance of the ZINB-S distribution from the factorial moments.

**Theorem 2.** \( X \sim ZINB - S(\pi, r, \alpha, \theta) \), the mean and variance of \( X \) can be written as

\[
E(X) = (1 - \pi)r \left[ \frac{\theta^2(\theta - \alpha + 1)}{(\theta + 1)(\theta - \alpha)^2} - 1 \right]
\] (8)

\[
V(X) = r(1 - \pi)\{(r + 1)\delta_2 - (2r + 3)\delta_1 - [(1 - \pi)r(\delta_1 - 1)^2]\}
\] (9)

where

\[
\delta_1 = \frac{\theta^2}{(\theta + 1)} \frac{(\theta - \alpha + 1)}{(\theta - \alpha)^2}, \quad \delta_2 = \frac{\theta^2}{(\theta + 1)} \frac{(\theta - 2\alpha + 1)}{(\theta - 2\alpha)^2}, \quad \delta_3 = \frac{\theta^2}{(\theta + 1)} \frac{(\theta - 3\alpha + 1)}{(\theta - 3\alpha)^2}
\]

4 Parameter Estimation

This section provides the parameter estimation of the ZINB-S distribution using the method of maximum likelihood estimation. We define an indicator function

\[
I(x) = \begin{cases} 
1, & \text{if } x = 0 \\
0, & \text{if } x \in \{1, 2, \ldots\} 
\end{cases}
\] (10)

Then the likelihood function of the ZINB-S distribution can be written as follows
\[ L = \prod_{i=1}^{n} \left\{ I[\pi + (1 - \pi) \frac{\theta^2(\theta + r\alpha + 1)}{(\theta + 1)(\theta + r\alpha)^2}] + (1 - I)[(1 - \pi) \frac{\theta^2}{(\theta + 1)} \sum_{j=0}^{x} \frac{(x)}{j} (-1)^j \frac{(\theta + \alpha(r + j) + 1)}{(\theta + \alpha(r + j))^2}] \right\} \] (11)

Take

\[ P = \pi + (1 - \pi) \frac{\theta^2(\theta + r\alpha + 1)}{(\theta + 1)(\theta + r\alpha)^2} \]

and

\[ Q = (1 - \pi) \frac{\theta^2}{(\theta + 1)} \left( x + r - 1 \right) \sum_{j=0}^{x} \frac{(x)}{j} (-1)^j \frac{(\theta + \alpha(r + j) + 1)}{(\theta + \alpha(r + j))^2} \]

Then the log likelihood function is in the form

\[ \log L = \sum_{i=1}^{n} \log(IP + (1 - I)Q) \] (12)

The partial derivatives of the log likelihood function are obtained by differentiating \( \log L \) with respect to the parameters \( r, \alpha, \theta, \pi \) as follows

\[ \frac{\partial \log L}{\partial \theta} = \sum_{i=1}^{n} \frac{1}{IP + (1 - I)Q} \left\{ I(1 - \pi)\theta \left[ \frac{(\theta + r\alpha)(\theta + 1)(3\theta + 2r\alpha + 2) - \theta(\theta + r\alpha + 1)(3\theta + r\alpha + 2)}{(\theta + 1)^2(\theta + r)^3} \right] \right. \\
\left. + (1 - I)(1 - \pi)\theta \left[ \left( x + r - 1 \right) \sum_{j=0}^{x} (-1)^j \frac{(\theta + \alpha(r + j) + 1)}{(\theta + \alpha(r + j))^2} \right] \right\} \] (13)

\[ \frac{\partial \log L}{\partial r} = \sum_{i=1}^{n} \frac{1}{IP + (1 - I)Q} \left\{ I[-(1 - \pi)\frac{\theta^2\alpha(\theta + r\alpha + 2)}{(\theta + 1)(\theta + r\alpha)^3}] + (1 - I)(1 - \pi)\frac{\theta^2}{\theta + 1} \frac{\Gamma(x + r)\sum_{j=0}^{x} \binom{x}{j}}{(\Gamma(r))} \right\} (-1)^{j+1} \frac{(\theta + \alpha(r + j) + 2)}{(\theta + \alpha(r + j))^3} + \sum_{j=0}^{x} \binom{x}{j} (-1)^j \frac{(\theta + \alpha(r + j) + 1)}{(\theta + \alpha(r + j))^2} \frac{[\Gamma(r)\Gamma(x + r)]'}{(\Gamma(r))^2} \right\} \] (14)

\[ \frac{\partial \log L}{\partial \pi} = \sum_{i=1}^{n} \frac{1}{IP + (1 - I)Q} \left\{ I[1 - \frac{\theta^2(\theta + r\alpha + 1)}{(\theta + 1)(\theta + r\alpha)^2}] - \frac{\theta^2}{(\theta + 1)} \left( x + r - 1 \right) \sum_{j=0}^{x} \frac{(x)}{j} (-1)^j \frac{(\theta + \alpha(r + j) + 1)}{(\theta + \alpha(r + j))^2} \right\} \] (15)
\[
\frac{\partial \log L}{\partial \alpha} = \sum_{i=1}^{n} \frac{1}{IP + (1 - I)Q} \left\{ -I(1 - \pi) \left[ \frac{r\theta^2(\theta + r\alpha + 2)}{(\theta + 1)(\theta + r\alpha)^3} \right] \\
+ (1 - I)(1 - \pi) \left[ \frac{\theta^2}{\theta + 1} \left( \frac{x + r - 1}{x} \right)(-1)^{j+1} \left( \frac{r + j}{(\theta + \alpha(r + j))^{3}} \right) \right] \right\} 
\]

The parameters of ZINB-S distribution are obtained by solving the above mentioned partial derivative equations using R software.

5 Application

Consider a real data set which provides information on 9,461 automobile insurance policies and it was taken from Zamani and Ismail (2010), it consists of the number of accidents per each policy. Table 1 provides the estimation of the observed and expected frequencies of Poisson, negative binomial, NB-S and ZINB-S distributions and comparison performance of these models in terms of the log-likelihood, p-values and chi-square statistic. Further, it is concluded that the ZINB-S distribution fits better than Poisson, negative binomial and NB-S distribution.

<table>
<thead>
<tr>
<th>Number of Accidents</th>
<th>Number of Policies</th>
<th>Poisson</th>
<th>NB</th>
<th>NB-S</th>
<th>ZINB-S</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>7840</td>
<td>7638.3</td>
<td>7843.3</td>
<td>7846.3</td>
<td>7840.0</td>
</tr>
<tr>
<td>1</td>
<td>1317</td>
<td>1634.6</td>
<td>1290.2</td>
<td>1299.2</td>
<td>1320.3</td>
</tr>
<tr>
<td>2</td>
<td>239</td>
<td>174.9</td>
<td>257.7</td>
<td>244.4</td>
<td>227.8</td>
</tr>
<tr>
<td>3</td>
<td>42</td>
<td>12.5</td>
<td>54.5</td>
<td>52.97</td>
<td>22.7</td>
</tr>
<tr>
<td>4</td>
<td>14</td>
<td>0.7</td>
<td>11.8</td>
<td>12.99</td>
<td>14.2</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>0</td>
<td>2.6</td>
<td>3.54</td>
<td>4.5</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>0</td>
<td>0.6</td>
<td>1.05</td>
<td>1.6</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>0</td>
<td>0.2</td>
<td>0.34</td>
<td>0.6</td>
</tr>
<tr>
<td>8+</td>
<td>0</td>
<td>0</td>
<td>0.1</td>
<td>0.11</td>
<td>0.3</td>
</tr>
</tbody>
</table>

Parameter Estimates

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{\lambda} )</td>
<td>0.214</td>
</tr>
<tr>
<td>( \hat{r} )</td>
<td>0.70</td>
</tr>
<tr>
<td>( \hat{\alpha} )</td>
<td>0.0100</td>
</tr>
<tr>
<td>( \hat{\theta} )</td>
<td>9.87</td>
</tr>
<tr>
<td>( \hat{\pi} )</td>
<td>0.8552</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Test</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Chi-square</td>
<td>293.8</td>
</tr>
<tr>
<td>p -Value</td>
<td>&lt;0.01</td>
</tr>
<tr>
<td>Chi-square</td>
<td>5490.78</td>
</tr>
<tr>
<td>Log Likelihood</td>
<td>-5530.87</td>
</tr>
</tbody>
</table>
6 Conclusion

In this paper, the zero-inflated negative binomial-Sushila distribution is introduced. We have obtained basic characteristics of the distribution such as factorial moments, mean and variance. Parameters are estimated by using the method of maximum likelihood estimation. And the efficiency of the ZINB-S distribution compare to other conventional probability distributions is illustrated with suitable application by using a real data set.

References


