A New Mixture Distribution for Extreme Excess Zeros:
Negative Binomial-Generalized Exponential (NB-GE) Distribution

Junifsa Afly Prameswari, Ida Fithriani, Siti Nurrohmah
{ichajunifsa@sci.ui.ac.id, ida.f@sci.ui.ac.id, snurrohmah@sci.ui.ac.id}

Department of Mathematics, Universitas Indonesia, Department of Mathematics, Universitas Indonesia, Department of Mathematics, Universitas Indonesia

Abstract. Negative Binomial-Generalized Exponential (NB-GE) distribution is a distribution that capable for modeling overdispersion data with extreme excess zeros, which is more than 80% zeros in a data. The distribution is a mixture distribution that obtained by mixing the Negative Binomial (NB) distribution with the Generalized Exponential (GE) distribution. The formation of the Negative Binomial-Generalized Exponential (NB-GE) distribution and the characteristics of the Negative Binomial-Generalized Exponential (NB-GE) distribution such as the probability density function, kth moment, mean, variance, skewness and kurtosis are discussed in this paper. Estimation of the parameters from the Negative Binomial-Generalized Exponential (NB-GE) distribution using the maximum likelihood method. As an illustration, Negative Binomial-Generalized Exponential (NB-GE) distribution used to model the data of fatal crash that has more than 80% zeros.

Keywords: Extreme Excess Zeros, Mixing Method, Overdispersion.

1 Introduction

In real-world problems, there are rare occasions that make the data obtained has a large number of zeros. Excess zeros, besides resulting skewed data, also the value of the variance data to be greater than its mean, known as overdispersion [1]. Some distributions that can handle overdispersion problems are the Negative Binomial distribution and the Zero-Inflated Poisson distribution. In the distribution Binomial Negative, there are parameters that explain the size of the dispersion of the data [2]. However, the Negative Binomial distribution is less effective when the data distribution is heavy tail and contains excess zeros [3]. In the Zero-Inflated Poisson (ZIP) distribution which usually handles the problem of overdispersion with excess zeros, the number of zeros that can be overcome with the ZIP distribution is only about 50-80% zeros in the data [4]. Therefore, another alternative distribution is needed to overcome the problem of overdispersion with extreme excess zeros, because rare cases sometimes have zeros of more than 80% in the data. In 2013, Sirinapa Aryuyuen and Winai Bodhisuwan introduced a new distribution, namely the Negative Binomial-Generalized Exponential (NB-GE) distribution which a mixture distribution from Negative Binomial (NB) distribution and Generalized Exponential (GE) distribution, aims to deal with the problem of overdispersion caused by extreme excess zeros (more than 80% zeros) [5].
In this paper, we will discuss the formation of the Negative Binomial-Generalized Exponential (NB-GE) distribution and its characteristics. Subsequently, the parameters of NB-GE distribution will be estimated using the Maximum Likelihood Estimation (MLE) method.

2 Methods

2.1 Negative Binomial (NB) Distribution

The context of the Negative Binomial distribution is an observation of the number of failures before ther-success in a series of Bernoulli experiments [2]. If the random variable \( X \) states the number of failures before \( r \)-success occurs, \( X \sim NB(r, p) \), the PMF of \( X \) is:

\[
\Pr(X = x) = f(x) = \binom{x + r - 1}{x} p^r (1 - p)^x, \quad x = 0, 1, 2, ...
\]

(1)

2.2 Generalized Exponential (GE) Distribution

The Generalized Exponential distribution also known as Exponentiated Exponential is the generalization of distribution Exponential [6]. The random variable \( X \) of the GE distribution has a PDF given by:

\[
f(x) = \begin{cases} \alpha \beta \left(1 - e^{-\beta x}\right)^{\alpha - 1} e^{-\beta x}, & x, \alpha, \beta > 0 \\ 0, & \text{otherwise} \end{cases}
\]

(2)

The following graph is a PDF of the GE distribution with several \( \alpha \) parameter values when \( \beta = 1 \) and some \( \beta \) parameter values when \( \alpha = 1 \). The figure shows that the GE distribution is flexible for small \( x \) values.

Fig. 1. The PDF of Generalized Exponential distribution.

Moment generating function of the distribution of GE is given by:

\[
M(t) = \frac{\Gamma(\alpha + 1)\Gamma\left(1 - \frac{t}{\beta}\right)}{\Gamma\left(\alpha - \frac{t}{\beta} + 1\right)}
\]

(3)
2.3 Maximum Likelihood Estimation (MLE)

Let \( X_1, X_2, ..., X_n \) be random sample size \( n \) from a certain distribution with PDF \( f(x; \theta) \) that depends on \( \theta \in \Omega \), where \( \Omega \) is a space of parameters. Then likelihood function can be obtained as joint PDF of \( X_1, X_2, ..., X_n \), denoted by \( L(\theta; x_1, x_2, ..., x_n) \) or \( L(\theta) \), as follows:

\[
L(\theta) = f(x_1; \theta)f(x_2; \theta) ... f(x_n; \theta) = \prod_{i=1}^{n} f(x_i; \theta); \quad \theta \in \Omega
\]

Let \( u(X_1, X_2, ..., X_n) \) be a function from \( x_1, x_2, ..., x_n \) so that if \( \theta \) replaced by \( u(x_1, x_2, ..., x_n) \), the likelihood function \( L(\theta; x_1, x_2, ..., x_n) \) reach maximum value. So, \( u(X_1, X_2, ..., X_n) \) be the maximum likelihood estimator (MLE) for \( \theta \) which denoted by \( \hat{\theta} \). Therefore, \( \hat{\theta} = u(X_1, X_2, ..., X_n) \).

To find the maximum likelihood estimator, \( L(\theta) \) can be modified to log-likelihood function denoted by \( \ln L(\theta) \). Value of \( \theta \) which maximize \( L(\theta) \) equal to value of \( \theta \) which maximize \( \ln L(\theta) \), so it can be obtain by solving first derivatives of the \( \ln L(\theta) \) with respect to \( \theta \) and equating them to zero.

\[
\frac{d \ln L(\theta; x_1, x_2, ..., x_n)}{d\theta} = 0
\]

\( \hat{\theta} \) is the solution of the equation above. If the distribution has more than one parameter, let say it has \( a, b, \) and \( c \) parameters, the value of each parameter that can maximize the likelihood function can be obtain by solving first derivatives of the \( \ln L(a, b, c) \) with respect to each \( a, b, \) and \( c \), and equating them to zero.

2.4 Mixing

Let \( X \) be a random variable from a certain distribution with PDF \( f_{X|\Lambda}(x|\Lambda) \), where the distribution of random variable \( X \) depends on the parameter \( \Lambda \) and the random variable can be written \( X|\Lambda \). Suppose that \( \Lambda \) is a random variable with PDF \( f_{\Lambda}(\Lambda) \). Thus, the unconditional PDF of \( X \) is given by:

\[
f_X(x) = \begin{cases} 
\sum_{\Lambda} f_{X|\Lambda}(x|\Lambda)f_{\Lambda}(\Lambda) & \text{for } \Lambda \text{ discrete} \\
\int f_{X|\Lambda}(x|\Lambda)f_{\Lambda}(\Lambda) \, d\Lambda & \text{for } \Lambda \text{ continuous}
\end{cases}
\]

Furthermore, \( f_X(x) \) is called the probability density function of a mixture. The \( k \)th moment factorial of a mixture distribution can be written as follows:

\[
\mu_{[k]}(X) = E_A[\mu_{[k]}(X|\Lambda)]
\]

3 Negative Binomial-Generalized Exponential (NB-GE) Distribution

The Negative Binomial-Generalized Exponential (NB-GE) distribution is a mixing distribution from the NB distribution and the GE distribution. The NB distribution is the main distribution in this formation, one of the parameters in the NB distribution will be modified so that the NB distribution depends on a random variable that has GE distribution.
3.1 The PMF

Suppose \( X|\Lambda \sim NB(r, p = e^{-\lambda}) \) and \( \Lambda \sim GE(\alpha, \beta) \), then the PMF of \( X|\Lambda \) is given by:

\[
\tilde{f}_{X|\Lambda}(x|\lambda) = \binom{x + r - 1}{x} e^{-x \lambda} (1 - e^{-\lambda})^x = (x + r - 1) \sum_{j=0}^{x} \binom{x}{j} (-1)^j e^{-\lambda(x+j)}
\]  

(6)

PMF of \( X \) can be obtained by substituting (6) and (3) into (4) for \( \Lambda \) continuous,

\[
f(x) = \int_0^{\infty} \tilde{f}_{X|\Lambda}(x|\lambda) f_\Lambda(\lambda) d\lambda
\]

\[
= \int_0^{\infty} \left( x + r - 1 \right) \sum_{j=0}^{x} \binom{x}{j} (-1)^j e^{-x \lambda} f_2(\lambda) d\lambda
\]

\[
= \left( x + r - 1 \right) \sum_{j=0}^{x} \binom{x}{j} (-1)^j \int_0^{\infty} e^{-x \lambda} f_2(\lambda) d\lambda
\]

\[
= \left( x + r - 1 \right) \sum_{j=0}^{x} \binom{x}{j} (-1)^j M_\Lambda(-(r+j))
\]

thus we obtain

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

(7)

Figure 2, 3, and 4 shows the graph of the PMF of NB-GE distribution with several parameter values.

Fig. 2. The PMF of NB-GE distribution with several values of \( r \) and \( \alpha = 0.25, \beta = 3 \).

Based on Figure 2, \( r \) is a scale parameter.

Fig. 3. The PMF of NB-GE distribution with several values of \( \alpha \) and \( r = 10, \beta = 5 \).

Based on Figure 3, \( \alpha \) is a shape parameter.
Based on Figure 4, \( \beta \) is a scale parameter.

### 3.2 The 1st, 2nd, 3rd, and 4th Moment

Since the PMF of NB-GE distribution is a factorial form, we can obtain the first four moments by \( k \)th factorial moment of NB-GE. From equation (5), we obtain:

\[
\mu_{[k]}(X) = \frac{\Gamma(r+k)}{\Gamma(r)} \sum_{j=0}^{k} \binom{k}{j} (-1)^j \left( \frac{\Gamma(\alpha+1)\Gamma\left(1-\frac{k-j}{\beta}\right)}{\Gamma(\alpha-\frac{k-j}{\beta}+1)} \right)
\]  

with the following

- \( \mu_{[1]}(X) = r \delta_1 - r \)
- \( \mu_{[2]}(X) = (r^2 + r) \delta_2 - (2r^2 + 2r) \delta_1 + r^2 + r \)
- \( \mu_{[3]}(X) = (r^3 + 3r^2 + 2r) \delta_3 - (3r^3 + 9r^2 + 6r) \delta_2 + (3r^3 + 9r^2 + 6r) \delta_1 \)
- \( \mu_{[4]}(X) = (r^4 + 6r^3 + 11r^2 + 6r) \delta_4 - (4r^4 + 24r^3 + 44r^2 + 24r) \delta_3 \\
  + (6r^4 + 36r^3 + 66r^2 + 36r) \delta_2 - (4r^4 + 24r^3 + 44r^2 + 24r) \delta_1 \\
  + r^4 + 6r^3 + 11r^2 + 6r \)

From the factorial moment, it is straightforward to deduce the first four moments given in (9) - (12), variance (13), skewness (14) and kurtosis (15).

\[
E(X) = r \delta_1 - r
\]  

\[
E(X^2) = (r^2 + r) \delta_2 - (2r^2 + r) \delta_1 + r^2
\]  

\[
E(X^3) = (r^3 + 3r^2 + 2r) \delta_3 - (3r^3 + 6r^2 + 3r) \delta_2 + (3r^3 + 3r^2 + r) \delta_1 - r^3
\]  

\[
E(X^4) = (r^4 + 6r^3 + 11r^2 + 6r) \delta_4 - (4r^4 + 18r^3 + 26r^2 + 12r) \delta_3 \\
  + (6r^4 + 18r^3 + 19r^2 + 7r) \delta_2 - (4r^4 + 6r^3 + 4r^2 + r) \delta_1 + r^4
\]  

\[
Var(X) = (r^2 + r) \delta_2 + r \delta_1(1 - r \delta_1)
\]  

\[
\gamma_1 = (r^3 + 3r^2 + 2r) \delta_3 - (3r^3 + 3r^2) \delta_2 + r \delta_1 + 3r^2(\delta_1)^2 - (3r^3 + 3r^2) \delta_1 \delta_2 \\
  + 2r^3(\delta_1)^3/[r(\delta_2 + r \delta_1(1 - r \delta_1))]^{3/2}
\]
\[ y_2 = (r^4 + 6r^3 + 11r^2 + 6r)\delta_4 - (6r^3 + 18r^2 + 12r)\delta_3 - 3r^2(\delta_1)^4 + (7r^2 + 7r)\delta_2 \]

\[-4r\delta_1 - 6r^3(\delta_1)^3 + (12r^3 + 12r^2)\delta_1\delta_2 - (4r^4 + 12r^3 + 8r^2)\delta_1\delta_3 \]

\[+ (6r^4 + 6r^3)(\delta_1)^2 \delta_2 / [(r^2 + r)\delta_2 + r\delta_1(1 - r\delta_1)]^2 \]

where \( \delta_1 = \left( \frac{\Gamma(a+1)\Gamma(\frac{1}{a})}{\Gamma(a-\frac{1}{a}+1)} \right), \delta_2 = \left( \frac{\Gamma(a+1)\Gamma(\frac{1-2}{a})}{\Gamma(a-\frac{2}{a}+1)} \right), \delta_3 = \left( \frac{\Gamma(a+1)\Gamma(\frac{1-2}{a})}{\Gamma(a-\frac{2}{a}+1)} \right), \delta_4 = \left( \frac{\Gamma(a+1)\Gamma(\frac{1-4}{a})}{\Gamma(a-\frac{4}{a}+1)} \right). \]

### 3.3 Overdispersion

As an alternative distribution for overdispersion, the characteristic can be seen by the variance and mean:

\[ \text{Var}(X) > E(X) \]

\[ (r^2 + r)\delta_2 + r\delta_1(1 - r\delta_1) > r\delta_1 - r \]

\[ r^2\delta_2 + r\delta_2 + r\delta_1 - r^2\delta_1^2 > r\delta_1 - r \]

\[ r^2\delta_2 + r\delta_2 - r^2\delta_1^2 > -r \]

For \( r, \alpha, \beta > 0 \), the equation above shows that the variance of NB-GE is always greater than its mean, so it is fulfilled.

### 3.4 Parameters Estimation

The likelihood function of the NB-GE \((r, \alpha, \beta)\) is written as follows:

\[ L(r, \alpha, \beta; x) = \prod_{i=1}^{n} \frac{\Gamma(r+x_i)}{\Gamma(x_i-1) \Gamma(r)} \sum_{j=0}^{x_i} \binom{x_i}{j} (-1)^j \left( \frac{\Gamma(a+1)\Gamma(1+\frac{r+j}{\beta})}{\Gamma(a+\frac{r+j}{\beta}+1)} \right) \]

with the log-likelihood function

\[ L(r, \alpha, \beta) = \ln L(r, \alpha, \beta; x) \]

\[ = \ln \left( \prod_{i=1}^{n} \frac{\Gamma(r+x_i)}{\Gamma(x_i-1) \Gamma(r)} \sum_{j=0}^{x_i} \binom{x_i}{j} (-1)^j \left( \frac{\Gamma(a+1)\Gamma(1+\frac{r+j}{\beta})}{\Gamma(a+\frac{r+j}{\beta}+1)} \right) \right) \]

\[ = \sum_{i=1}^{n} \ln(\Gamma(r+x_i) - \Gamma(r) - \Gamma(x_i+1)) + \sum_{i=1}^{n} \ln \left( \sum_{j=0}^{x_i} \binom{x_i}{j} (-1)^j \left( \frac{\Gamma(a+1)\Gamma(1+\frac{r+j}{\beta})}{\Gamma(a+\frac{r+j}{\beta}+1)} \right) \right) \]

By differentiating the log-likelihood function with respect to \( r, \alpha \) and \( \beta \), the optimal values of the parameters obtained.

\[ \frac{\partial}{\partial r} L(r, \alpha, \beta) = \sum_{i=1}^{n} \psi(r+x_i) - n\psi(r) + \sum_{i=1}^{n} \left( \sum_{j=0}^{x_i} \binom{x_i}{j} (-1)^j \frac{\Gamma(1+r+\frac{j}{\beta})}{\Gamma(a+\frac{r+j}{\beta}+1)} \right) \]

\[ = 0 \]
\[
\frac{\partial}{\partial \alpha} 
\mathcal{L}(r, \alpha, \beta) = \sum_{i=1}^{n} \left( \frac{\sum_{j=0}^{\chi_i} (\chi_i^j) (-1)^j \frac{\Gamma(x+1)\Gamma(1+\frac{j}{r})}{\Gamma(\alpha+ sol x+\frac{j}{r}+1)} \right) = 0
\]

\[
\frac{\partial}{\partial \beta} 
\mathcal{L}(r, \alpha, \beta) = \sum_{i=1}^{n} \left( \frac{\sum_{j=0}^{\chi_i} (\chi_i^j) (-1)^j \frac{\Gamma(1+\frac{j}{r})}{\Gamma(\beta+sol x+\frac{j}{r}+1)} \right) = 0
\]

where \( \psi(x) = \frac{\Gamma'(x)}{\Gamma(x)} \) is the digamma function. Solving the equations above are difficult and complicated. So, the solution of the nonlinear equations above can be obtained by numerical techniques.

4 Result and Discussion

To see how well the NB-GE performs, we used a real data set which the number of fatal crashes in Michigan on 2006 from Michigan Department of State Police [7] for the applications of NB-GE distribution. There are 32,672 roads in this data which are divided into 11 classes.

<table>
<thead>
<tr>
<th>Fatal Crash</th>
<th>0</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Roads</td>
<td>29087</td>
<td>2952</td>
<td>464</td>
<td>108</td>
<td>40</td>
<td>9</td>
<td>5</td>
<td>2</td>
<td>3</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

We can see the descriptive of the data above by Table 2. It shows that the variance of the Fatal Crash Data is larger than the mean.
Table 2. Descriptive Statistics of Fatal Crash Data.

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Valid</td>
<td>32672</td>
</tr>
<tr>
<td>Missing</td>
<td>0</td>
</tr>
<tr>
<td>Mean</td>
<td>0.14</td>
</tr>
<tr>
<td>Median</td>
<td>0.00</td>
</tr>
<tr>
<td>Mode</td>
<td>0</td>
</tr>
<tr>
<td>Std. Deviation</td>
<td>0.452</td>
</tr>
<tr>
<td>Variance</td>
<td>0.204</td>
</tr>
<tr>
<td>Skewness</td>
<td>5.120</td>
</tr>
<tr>
<td>Std. Error of Skewness</td>
<td>0.014</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>45.255</td>
</tr>
<tr>
<td>Std. Error of Kurtosis</td>
<td>0.027</td>
</tr>
<tr>
<td>Minimum</td>
<td>0</td>
</tr>
<tr>
<td>Maximum</td>
<td>10</td>
</tr>
<tr>
<td>25 Percentiles</td>
<td>.00</td>
</tr>
<tr>
<td>50 Percentiles</td>
<td>.00</td>
</tr>
<tr>
<td>75 Percentiles</td>
<td>.00</td>
</tr>
</tbody>
</table>

Figure 5 shows the values of PMF of NB-GE, NB, and ZIP distribution that obtained by substituting the estimated parameters using maximum likelihood, along the empirical probability of fatal crash data. As it can be seen in Figure 5, these data are overdispersed since the mean (0.14) is smaller than the variance (0.204) and have more than 80% zeros.

From the Figure 5, we can see that the NB-GE distribution is the most suitable for modelling the data, compared to NB and ZIP distribution. The comparison of expected values and the Chi-Squares of NB, ZIP, and NB-GE distribution can be seen in the following table.
Table 3. Observed, expected, and Chi-Square values of NB, ZIP, and NB-GE distribution.

<table>
<thead>
<tr>
<th>$x$</th>
<th>$n_x$</th>
<th>NB</th>
<th>ZIP</th>
<th>NB-GE</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>29087</td>
<td>29101.55</td>
<td>29087.00</td>
<td>29097.85</td>
</tr>
<tr>
<td>1</td>
<td>2952</td>
<td>2859.17</td>
<td>2806.25</td>
<td>2908.37</td>
</tr>
<tr>
<td>2</td>
<td>464</td>
<td>549.3700</td>
<td>661.3500</td>
<td>498.2800</td>
</tr>
<tr>
<td>3</td>
<td>108</td>
<td>122.7500</td>
<td>103.9100</td>
<td>115.9600</td>
</tr>
<tr>
<td>4</td>
<td>40</td>
<td>29.35000</td>
<td>12.24000</td>
<td>33.18000</td>
</tr>
<tr>
<td>5</td>
<td>9</td>
<td>7.29000</td>
<td>0.00000</td>
<td>11.04000</td>
</tr>
<tr>
<td>6</td>
<td>5</td>
<td>1.86000</td>
<td>0.00000</td>
<td>4.130000</td>
</tr>
<tr>
<td>7</td>
<td>2</td>
<td>0.00000</td>
<td>0.00000</td>
<td>1.700000</td>
</tr>
<tr>
<td>8</td>
<td>3</td>
<td>0.00000</td>
<td>0.00000</td>
<td>1.750000</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.540000</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>0.00000</td>
<td>0.00000</td>
<td>0.540000</td>
</tr>
<tr>
<td>Total</td>
<td>32672</td>
<td>32672</td>
<td>32672</td>
<td>32672</td>
</tr>
</tbody>
</table>

Estimates Parameters

|          | $\hat{r} = 2.9114$ | $\hat{\mu} = 0.4713$ | $\hat{r} = 0.9368$ | $\hat{\mu} = 0.4713$ | $\hat{\phi} = 0.8859$ | $\hat{\alpha} = 1.2801$ | $\hat{\beta} = 8.9989$ |

$\chi^2_{calculated}$

|          | 89.6609 | 441.63 | 11.4793 |

df

|          | 5 | 3 | 6 |

$\chi^2_{table}$

|          | 11.07 | 7.815 | 12.592 |

By comparing these distributions in Table 3, based on the Chi-Square calculated and Chi-Square table of this comparison, these results have shown that the NB-GE distribution provide better fit than the NB and ZIP distribution for the count data that have a large number of zeros.

5Conclusion

This paper introduced the NB-GE distribution which is obtained by mixing the NB and GE distribution. The properties of the NB-GE distribution which includes the factorial moments, mean, variance, skewness, and kurtosis have obtained. Parameters estimation are also obtained using maximum likelihood method and the usefulness of the NB-GE distribution is illustrated by real data set characterized by a large number of zeros. In conclusion, it is believed that the NB-GE distribution is very useful tool for modeling data with extreme excess zeros.

References


