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

Junifsa Afly Prameswari[1], Ida Fithriani[2], Siti Nurrohmah[3]
{ichajunifsa@sci.ui.ac.id[1],ida.f@sci.ui.ac.id[2], snurrohmah@sci.ui.ac.id[2]}

Department of Mathematics, Universitas Indonesia[1], Department of Mathematics, Universitas Indonesia[2], Department of Mathematics, Universitas Indonesia[3]

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 or 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].

ICSA 2019, August 02-03, Bogor, Indonesia

DOI 10.4108/eai.2-8-2019.2290342

In this paper, we will discuss the formation of the Negative Binomial-Generalized Exponential (NB-GE) distribution and its characteristics. Subsequently, theparameters 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 the$r$-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:

$$\Pr(X=x)=f(x)=\begin{cases}\binom{x+r-1}{x}p^r(1-p)^x, & x=0,1,2,\dots\\ 0, & \text{otherwise}\end{cases} \tag{1}$$

2.2Generalized 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:

$$f(x)=\begin{cases}\alpha\beta\left(1-e^{-\beta x}\right)^{\alpha-1}e^{-\beta x}, & x,\alpha,\beta>0\\ 0, & \text{otherwise}\end{cases} \tag{2}$$

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

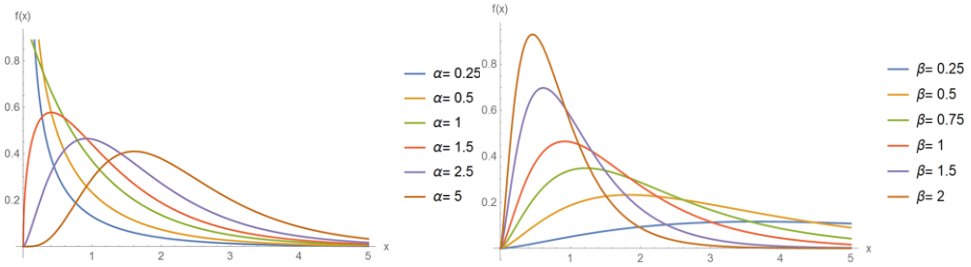

Fig. 1. The PDF of Generalized Exponential distribution.

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

$$M(t)=\frac{\Gamma(\alpha+1)\Gamma\left(1-\frac{t}{\beta}\right)}{\Gamma\left(\alpha-\frac{t}{\beta}+1\right)} \tag{3}$$

2.3Maximum Likelihood Estimation (MLE)

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

$$L(\theta) = f(x_1;\theta)f(x_2;\theta)\dots f(x_n;\theta) = \prod_{i=1}^{n} f(x_i;\theta);\ \ \theta \in \Omega$$

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

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

$$\frac{d \ln L(\theta; x_1, x_2, \dots, x_n)}{d\theta} = 0$$

$\hat{\theta}$ 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.

2.4Mixing

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

$$f_X(x) = \begin{cases} \sum_{\lambda} f_{X|\Lambda}(x|\lambda) f_\Lambda(\lambda) & \text{, for } \Lambda \text{ discrete} \\ \int_{\lambda} f_{X|\Lambda}(x|\lambda) f_\Lambda(\lambda)\, d\lambda & \text{, for } \Lambda \text{ continuous} \end{cases} \tag{4}$$

Furthermore, $f_X(x)$is called the probability density function of a mixture. The kth moment factorialof a mixture distribution can be written as follows:

$$\mu_{[k]}(X) = E_\Lambda\left[\mu_{[k]}(X|\Lambda)\right] \tag{5}$$

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

TheNegative Binomial-Generalized Exponential(NB-GE) distribution is a mixingdistribution from the NB distribution and theGE 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:

$$f_{X|\Lambda}(x|\lambda) = \binom{x+r-1}{x} e^{-\lambda r}\left(1-e^{-\lambda}\right)^{x} = \binom{x+r-1}{x} \sum_{j=0}^{x} \binom{x}{j} (-1)^{j} e^{-\lambda(r+j)} \qquad (6)$$

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

$$\begin{aligned} f(x) &= \int_0^{\infty} f_{X|\Lambda}(x|\lambda)\, f_{\Lambda}(\lambda)\, d\lambda \\ &= \int_0^{\infty} \binom{x+r-1}{x} \sum_{j=0}^{x} \binom{x}{j} (-1)^{j} e^{-\lambda(r+j)} f_2(\lambda)\, d\lambda \\ &= \binom{x+r-1}{x} \sum_{j=0}^{x} \binom{x}{j} (-1)^{j} \int_0^{\infty} e^{-\lambda(r+j)} f_2(\lambda)\, d\lambda \\ &= \binom{x+r-1}{x} \sum_{j=0}^{x} \binom{x}{j} (-1)^{j} M_{\Lambda}(-(r+j)) \end{aligned}$$

thus we obtain

$$f(x) = \begin{cases} \binom{x+r-1}{x} \sum_{j=0}^{x} \binom{x}{j} (-1)^{j} \left(\dfrac{\Gamma(\alpha+1)\Gamma\left(1+\frac{r+j}{\beta}\right)}{\Gamma\left(\alpha+\frac{r+j}{\beta}+1\right)} \right), & x = 0,1,2,\ldots \\ 0, & \text{otherwise} \end{cases} \qquad (7)$$

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

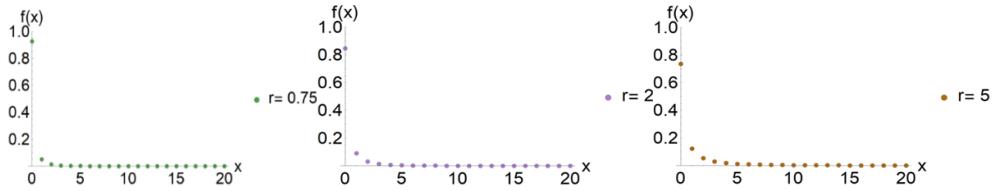

Fig. 2. The PMF of NB-GE distributionwith several values of r and $\alpha = 0.25$, $\beta = 3$.

Based on Figure 2, r is a scale parameter.

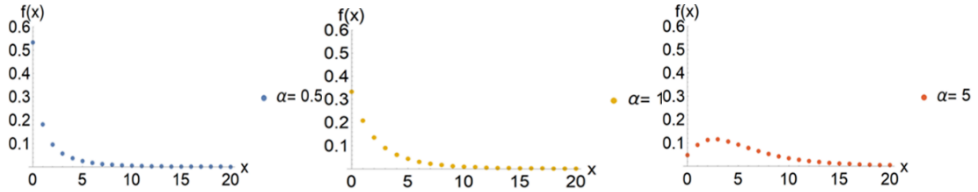

Fig. 3. The PMF of NB-GE distribution with several values of α and $r = 10$, $\beta = 5$.

Based on Figure 3, α is a shape parameter.

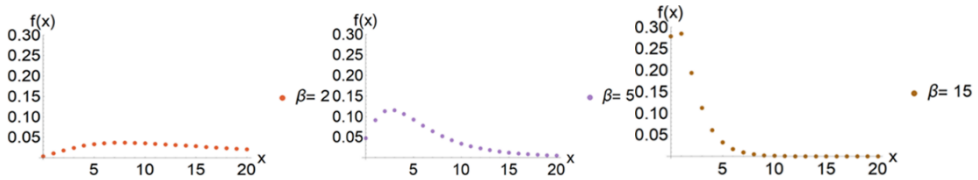

Fig. 4. The PMF of NB-GE distribution with several values of β and $r = 10, \alpha = 5$**.**

Based on Figure 4, β is a scale parameter.

3.2The 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 kth factorial moment of NB-GE. From equation (5), we obtain:

$$\mu_{[k]}(X) = \frac{\Gamma(r+k)}{\Gamma(r)} \sum_{j=0}^{k} \binom{k}{j} (-1)^j \left(\frac{\Gamma(\alpha+1)\Gamma\left(1-\frac{k-j}{\beta}\right)}{\Gamma\left(\alpha-\frac{k-j}{\beta}+1\right)} \right) \tag{8}$$

with the following

$\mu_{[1]}(X) = r\,\delta_1 - r$

$\mu_{[2]}(X) = (r^2+r)\delta_2 - (2r^2+2r)\delta_1 + r^2 + r$

$\mu_{[3]}(X) = (r^3+3r^2+2r)\delta_3 - (3r^3+9r^2+6r)\delta_2 + (3r^3+9r^2+6r)\delta_1$
$-r^3 - 3r^2 - 2r$

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

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

$$E(X) = r\,\delta_1 - r \tag{9}$$

$$E(X^2) = (r^2+r)\delta_2 - (2r^2+r)\delta_1 + r^2 \tag{10}$$

$$E(X^3) = (r^3+3r^2+2r)\delta_3 - (3r^3+6r^2+3r)\delta_2 + (3r^3+3r^2+r)\delta_1 - r^3 \tag{11}$$

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

$$Var(X) = (r^2+r)\delta_2 + r\delta_1(1-r\delta_1) \tag{13}$$

$$\gamma_1 = (r^3+3r^2+2r)\delta_3 - (3r^2+3r)\delta_2 + r\delta_1 + 3r^2(\delta_1)^2 - (3r^3+3r^2)\delta_1\delta_2 + 2r^3(\delta_1)^3/[(r^2+r)\delta_2 + r\delta_1(1-r\delta_1)]^{3/2} \tag{14}$$

$$\gamma_2 = (r^4 + 6r^3 + 11r^2 + 6r)\delta_4 - (6r^3 + 18r^2 + 12r)\delta_3 - 3r^4(\delta_1)^4 + (7r^2 + 7r)\delta_2$$
$$-4r\delta_1 - 6r^3(\delta_1)^3 + (12r^3 + 12r^2)\delta_1\delta_2 - (4r^4 + 12r^3 + 8r^2)\delta_1\delta_3 \quad (15)$$
$$+(6r^4 + 6r^3)(\delta_1)^2\delta_2 \,/\, [(r^2 + r)\delta_2 + r\delta_1(1 - r\delta_1)]^2$$

where$\delta_1 = \left(\frac{\Gamma(\alpha+1)\Gamma\left(1-\frac{1}{\beta}\right)}{\Gamma\left(\alpha-\frac{1}{\beta}+1\right)}\right), \delta_2 = \left(\frac{\Gamma(\alpha+1)\Gamma\left(1-\frac{2}{\beta}\right)}{\Gamma\left(\alpha-\frac{2}{\beta}+1\right)}\right), \delta_3 = \left(\frac{\Gamma(\alpha+1)\Gamma\left(1-\frac{3}{\beta}\right)}{\Gamma\left(\alpha-\frac{3}{\beta}+1\right)}\right), \delta_4 = \left(\frac{\Gamma(\alpha+1)\Gamma\left(1-\frac{4}{\beta}\right)}{\Gamma\left(\alpha-\frac{4}{\beta}+1\right)}\right).$

3.3Overdispersion

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

$$Var(X) > E(X)$$
$$(r^2 + r)\delta_2 + r\delta_1(1 - r\delta_1) > r\,\delta_1 - r$$
$$r^2\delta_2 + r\delta_2 + r\delta_1 - r^2\delta_1^2 > r\,\delta_1 - r$$
$$r^2\delta_2 + r\delta_2 - r^2\delta_1^2 > -r$$

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.

3.4Parameters Estimation

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

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

with the log-likelihood function

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

By differentiating the log-likelihood function with respect to r, α and β, the optimal values of the parameters obtained.

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

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

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

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.

4Result and Discussion

To see how well the NB-GE performs, we useda 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 1. Fatal Crash Data.

Fatal Crash	0	1	2	3	4	5	6	7	8	9	10
Roads	29087	2952	464	108	40	9	5	2	3	1	1

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.

	Valid	32672
	Missing	0
Mean		.14
Median		.00
Mode		0
Std. Deviation		.452
Variance		.204
Skewness		5.120
Std. Error of Skewness		.014
Kurtosis		45.255
Std. Error of Kurtosis		.027
Minimum		0
Maximum		10
Percentiles	25	.00
	50	.00
	75	.00

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.

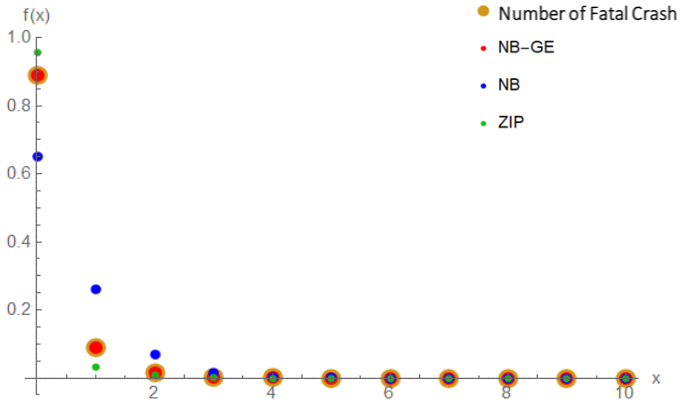

Fig. 5. Empirical probability graph data in Table 1 along with PMF of NB-GE, NB, and ZIP.

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.

x	n_x	NB	ZIP	NB-GE
0	29087	29101.55	29087.00	29097.85
1	2952	2859.170	2806.250	2908.370
2	464	549.3700	661.3500	498.2800
3	108	122.7500	103.9100	115.9600
4	40	29.35000	12.24000	33.18000
5	9	7.290000	0.25	11.04000
6	5	1.860000		4.130000
7	2	0.65		1.700000
8	3			1.750000
9	1			0.54
10	1			
Total	32672	32672	32672	32672
Estimates Parameters		$\hat{r}$ =2.9114	$\hat{\mu}$ =0.4713	$\hat{r}$ =0.9368
		$\hat{p}$ =0.1376	$\hat{\phi}$ =0.8859	$\hat{\alpha}$ =1.2801
				$\hat{\beta}$ =8.9989
$\chi^2_{calculated}$		89.6609	441.63	11.4793
df		5	3	6
χ^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

[1]Winkelmann, R. (2003). *Econometric Analysis of Count Data.* Berlin: Springer Berlin Heidelberg.

[2]Johnson, N. L., Kemp, A. W., & Kotz, S. (2005). *Univariate Discrete Distribution.* New York: Wiley-Interscience.
[3]Lord, D., & Geedipally, R. S. (2011). The negative binomial–Lindley distribution as a tool for analyzing crash data characterized by a large amount of zeros. *Accident Analysis & Prevention*, 1738-1742.
[4]Warton, D. I. (2005). Many zeros does not mean zero inflation: comparing the goodness-of-fit of parametric models to multivariate abundance data. *Environmetrics*, 275-289.
[5]Aryuyuen, S., & Bodhisuwan, W. (2013). The Negative Binomial-Generalized Exponential (NB-GE) Distribution. *Applied Mathematical Sciences*, 1093-1105.
[6]Gupta, R. D., & Kundu, D. (1999). Generalized Exponential Distribution. *Australian & New Zealand Journal of Statistics*, 173-188.
[7]Michigan Department of State Police, M. (2006). *Michigan Traffic Crash Facts for Country/Communities.* Michigan: Michigan Department of State Police.