Appl. Math. J. Chinese Univ.
2022, 37(2): 187-198

Zero-inflated non-central negative binomial distribution

TIAN Wei-zhong[1,*] LIU Ting-ting[2] YANG Yao-ting[2]

Abstract. In this article, the zero-inflated non-central negative binomial (ZINNB) distribution is introduced. Some of its basic properties are obtained. In addition, we use the maximum likelihood estimation method to estimate the parameters of the ZINNB distribution, and illustrate its application by fitting the actual data sets.

§1 Introduction

In statistics, counting data is a type of statistical data that is usually used to record the results of the occurrence and frequency of events. If the count data contains more zeros than expected, we generally called them zero-inflated data. Zero-inflated data are common in many disciplines, like engineering, manufacturing, economics, public health, epidemiology, psychology, sociology, political science, agriculture, road safety, species abundance, and criminology. One approach to analysis such data is to use zero-inflated Poisson (ZIP) distribution, which were introduced by Lambert [6], and several authors have utilized the ZIP distribution for modeling count data with an excessive number of zeros, see [1, 2, 5]. Later on, a fair amount of statistical methodology has been required in order to account for the feature of excess zeros. Greene [3] introduced the zero-inflated negative binomial (ZINB) to test zero inflation and overdispersion, Hall [4] studied the zero-inflated binomial (ZIB) distribution on an upper bounded count situation, and Sim et al. [12] discussed a zero-inflated Conway-Maxwell Poisson (ZICMP) distribution and developed the score and likelihood ratio tests. Recently, Sellers and Young [11] considered a zero-inflated sum-of-Conway-Maxwell-Poissons (ZISCMP) regression as a flexible analysis tool to model count data that express significant data dispersion and contain excess zeros, which contains zero-inflated Poisson (ZIP), zero-inflated negative binomial (ZINB), zero-inflated binomial (ZIB), and the zero-inflated Conway-Maxwell-Poisson (ZICMP).

The non-central negative binomial (NNB) distribution was found in neural counting mechanisms and photon counting, which was introduced by Ong and Lee [9]. The random variable X is said to have a NNB distribution, with the parameters $v > 0, \lambda > 0$ and $0 < p < 1$, denoted by $X \sim \text{NNB}(p, v, \lambda)$, if the probability mass function (pmf) is of the form,

$$P(X = k) = P(k; p, v, \lambda) = e^{-\lambda p} p^k q^v L_k^{v-1}(-\lambda q),$$

where $q = 1 - p$, and $L_n^\alpha(x)$ is the generalised Laguerre polynomials defined as follows,

Received: 2020-03-19. Revised: 2021-09-14.
MR Subject Classification: 62D05, 62E10, 62F03, 62F10, 62P10.
Keywords: zero-inflated non-central negative binomial distribution, maximum likelihood estimation, goodness of fit.
Digital Object Identifier(DOI): https://doi.org/10.1007/s11766-022-4070-0.

[*] Corresponding author.

$$L_n^\alpha(x) = \frac{(\alpha+1)_n}{n!} {}_1F_1(-n, \alpha+1; x), \tag{1}$$

with $\alpha \in \mathbb{Z}$, $n = 0, 1, 2, \cdots$ and ${}_1F_1(a, b; z)$ is the confluent hypergeometric function.

In fact, if $\lambda = 0$, the NNB distribution is reduced to the negative binomial (NB) distribution, $NB(p, v)$. If $p \to 0$, $v \to \infty$ such that vp is constant, the NNB distribution is reduced to the Poisson distribution, $P(vp)$, see Ong and Lee [9].

The NNB distribution has many important properties and applications, Lee and Ong [7] discussed the higher-order and non-stationary properties of the stochastic reversible counter based on NNB distribution, Ong and Lee [8, 9] studied a bivariate generalization of the NNB distribution. The NNB distribution is derived by mixing the Poisson distribution with a certain Bessel function distribution or the negative binomial distribution with the Poisson distribution, and Ong et al. [10] proposed various important probabilistic properties of the NNB distribution in practical applications.

The aim of this paper is to develop a zero-inflated non-central negative binomial (ZINNB) distribution, which is the generalization of the ZINB distribution and ZIP distribution. The rest of the article is organized as follows. The definition of the ZINNB distribution and some of its basic properties, such as probability generating function (pgf), moments, mean and variance are studied in Section 2. The test method for data existence of zero inflation is given in section 3. Maximum Likelihood Estimation of the parameters and the simulation for the proposed method are investigated in Section 4. Two real data applications are discussed in Section 5. A conclusion is provided in Section 6.

§2 ZINNB distribution and its properties

In this section, we present the definition of the ZINNB distribution and some useful properties.

Definition 2.1. Let Y be a discrete random variable which follows a ZINNB distribution with $w \in [0, 1]$, $v > 0, \lambda > 0$ and $0 < p = 1 - q < 1$. The pmf of Y is

$$f(y; w, p, v, \lambda) = P(Y = y) = \begin{cases} w + (1-w)e^{-\lambda p} q^v, & if \quad y = 0, \\ (1-w)e^{-\lambda p} p^y q^v L_y^{v-1}(-\lambda q), & if \quad y = 1, 2, 3, \cdots, \end{cases} \tag{2}$$

and we denote it as $Y \sim \text{ZINNB}(w, p, v, \lambda)$.

Remark 2.1. When $w = 0$, the ZINNB distribution is reduced to the NNB distribution, $NNB(p, v, \lambda)$. When $\lambda = 0$, the ZINNB distribution is reduced to the ZINB distribution, $ZINB(w, p, v)$ [16]. If $p \to 0$, $v \to \infty$ such that vp is constant, then the ZINNB distribution is reduced to the ZIP distribution, $ZIP(w, vp)$ [17].

According to the definition of the confluent hypergeometric function, we have

$${}_1F_1(a, b; z) = \sum_{i=0}^{\infty} \frac{(a)_i}{(b)_i} \frac{z^i}{i!}, \tag{3}$$

and for $\alpha > 0$,

$$(\alpha)_i = \frac{\Gamma(\alpha+i)}{\Gamma(\alpha)} = \frac{(\alpha+i-1)!}{(\alpha-1)!},$$

$$(-\alpha)_i = (-1)^i \frac{\alpha!}{(\alpha-i)!}. \tag{4}$$

Therefore, for $y \neq 0$,

$$f(y; w, p, v, \lambda) = (1-w)e^{-\lambda p} p^y q^v (v+y-1)! \sum_{i=0}^{y} \frac{(\lambda q)^i}{i!(y-i)!(v+i-1)!}.$$

To emphasise the usefulness of parameters w, p, v and λ, the plots of pmf of ZINNB distributions with different values of parameters are given in Figure 1.

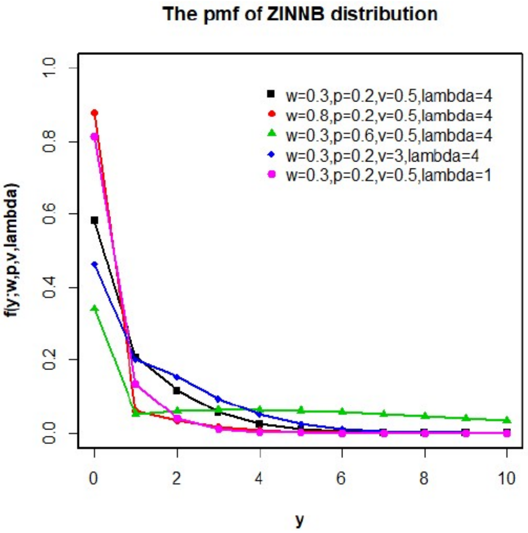

Figure 1. pmf for ZINNB (0.3,0.2,0.5,4), ZINNB (0.8,0.2,0.5,4), ZINNB (0.3,0.6,0.5,4), ZINNB (0.3,0.2,3,4) and ZINNB (0.3,0.2,0.5,1).

Next, we study some basic properties of $Y \sim \text{ZINNB}(w, p, v, \lambda)$.

Proposition 2.1. *If $Y \sim$ ZINNB(w, p, v, λ), then the pgf of Y is given*

$$G(t) = w + (1-w)\left(\frac{q}{1-pt}\right)^v e^{\lambda\left(\frac{q}{1-pt}-1\right)}. \tag{5}$$

Proof.

According to Ong and Lee [9], we know the pgf of NNB distribution is

$$g(t) = \sum_{k=0}^{\infty} P(k; p, v, \lambda)t^k = \left(\frac{q}{1-pt}\right)^v e^{\lambda\left(\frac{q}{1-pt}-1\right)}.$$

According to equation (2), we have

$$\begin{aligned} G(t) &= \sum_{y=0}^{\infty} f(y; w, p, v, \lambda)t^y \\ &= w + (1-w)e^{-\lambda p} q^v + \sum_{y=1}^{\infty} (1-w)e^{-\lambda p} p^y q^v L_y^{v-1}(-\lambda q)t^y \\ &= w + (1-w)P(0; p, v, \lambda) + (1-w)\sum_{y=1}^{\infty} P(y; p, v, \lambda)t^y \\ &= w + (1-w)g(t). \end{aligned}$$

Proposition 2.2. *If $Y \sim$ ZINNB(w, p, v, λ), the recursion formula of $f(y; w, p, v, \lambda)$ with $y \neq 0$*

is

$$(y+1)f(y+1;w,p,v,\lambda) = (1-w)\big[(2y+v+\lambda q)pf(y;w,p,v,\lambda)-p^2(y+v-1)f(y-1;w,p,v,\lambda)\big]. \tag{6}$$

Proof. For $y \neq 0$, we have

$$(y+1)f(y+1) = (y+1)(1-w)e^{-\lambda p}p^{y+1}q^v L_{y+1}^{v-1}(-\lambda q),$$

and according to the recursion formula of $L_n^\alpha(x)$, see Ong and Lee [9]. The recursion formula is

$$(y+1)L_{y+1}^\alpha(x) = (2y+\alpha+1-x)L_y^\alpha(x) - (y+\alpha)L_{y-1}^\alpha(x), \tag{7}$$

thus, we get equation (6).

Corollary 2.1. *If $Y \sim$ ZINNB(w,p,v,λ), for $j \geq 1$, the $j-th$ factorial moments of Y, $E[(Y)_j] = E[Y(Y-1)(Y-2)\cdots(Y-j+1)]$, are*

$$\mu_{[j]} = (1-w)j!\left(\frac{p}{q}\right)^j L_j^{v-1}(-\lambda). \tag{8}$$

Proof. From equation (5), we have

$$G(t) = w + (1-w)\left(\frac{q}{1-pt}\right)^v e^{\lambda\left(\frac{q}{1-pt}-1\right)}.$$

On differentiating the above equation j times with respect to t and putting $t=1$, we get equation (8).

Corollary 2.2. *The expected value and variance of $Y \sim$ ZINNB(w,p,v,λ) are*

$$E(Y) = (1-w)\left(\frac{p}{q}\right)(v+\lambda) = V,$$

$$Var(Y) = (1-w)\left[V + \left(\frac{p}{q}\right)^2(v+2\lambda)\right].$$

Proposition 2.3. *If $Y \sim$ ZINNB(w,p,v,λ), the recursion formula for the $j-th$ factorial moments $\mu_{[j]}$ is*

$$\mu_{[j+1]} = \frac{p}{q}(2j+v+\lambda-1)\mu_{[j]} - j\left(\frac{p}{q}\right)^2(j+v-1)\mu_{[j-1]}, \quad j \geq 1. \tag{9}$$

Proof. From equation (8), we have

$$\mu_{[j+1]} = (1-w)(j+1)!\left(\frac{p}{q}\right)^{(j+1)} L_{j+1}^{v-1}(-\lambda),$$

and the result can be obtained by equation (7).

Proposition 2.4. *If $Y_1 \sim$ ZINNB(w,p,v_1,λ_1) and $Y_2 \sim$ ZINNB(w,p,v_2,λ_2) be independent.*
i) The probability of the sum of Y_1 and Y_2 is

$$P(Y_1+Y_2=n) = \begin{cases} w^2 + w(1-w)\big[e^{-\lambda_1 p}q^{v_1} + e^{-\lambda_2 p}q^{v_2}\big] + (1-w)^2 e^{-\lambda p}q^v, & if \quad n=0, \\ w(1-w)p^n\big[e^{-\lambda_1 p}q^{v_1}L_n^{v_1-1}(-\lambda_1 q) + e^{-\lambda_2 p}q^{v_2}L_n^{v_2-1}(-\lambda_2 q)\big] & \\ +(1-w)^2 e^{-\lambda p}p^n q^v L_n^{v-1}(-\lambda q), & if \quad n \neq 0. \end{cases} \tag{10}$$

ii) If $Y_1 = k \neq 0$, then the conditional probability of $P(Y_1 = k|Y_1+Y_2=n)$ is given by following,

$$\begin{aligned} &P(Y_1=k|Y_1+Y_2=n) \\ = &\begin{cases} (1-w)e^{-\lambda p}q^v L_k^{v_1-1}(-\lambda_1 q)L_{n-k}^{v_2-1}(-\lambda_2 q)h(\theta), & if \quad n \neq k, \\ \big[we^{-\lambda_1 p}q^{v_1}L_n^{v_1-1}(-\lambda_1 q) + (1-w)e^{-\lambda p}q^v L_n^{v_1-1}(-\lambda_1 q)\big]\,h(\theta), & if \quad k=n, \end{cases} \end{aligned} \tag{11}$$

where $h(\theta) = \left[w\left[e^{-\lambda_1 p}q^{v_1}L_n^{v_1-1}(-\lambda_1 q) + e^{-\lambda_2 p}q^{v_2}L_n^{v_2-1}(-\lambda_2 q)\right] + (1-w)e^{-\lambda p}q^{v}L_n^{v-1}(-\lambda q)\right]^{-1}$, $\theta = (w, p, v_1, v_2, \lambda_1, \lambda_2)$, $v = v_1 + v_2$ *and* $\lambda = \lambda_1 + \lambda_2$.

Proof. The results can be obtained after some algebraically calculations with considering equation (2) and the following equation,

$$\sum_{k=0}^{n} L_k^a(x_1)L_{n-k}^b(x_2) = L_n^b(x_2) + \sum_{k=1}^{n-1} L_k^a(x_1)L_{n-k}^b(x_2) + L_n^a(x_1) = L_n^{a+b-1}(x_1+x_2).$$

§3 Zero inflation test

Before using the zero-inflated model to fit and analyze the data, the zero inflation phenomenon of the data should be tested first. In this section, we analyze the data through dispersion index, zero inflation index and hypothesis testing statistics.

3.1 Over dispersion and zero inflation index

The NNB distribution provides a lot of conveniences for us to analyze the data sets. However, in many practical applications, we will encounter the phenomenon of the over-dispersion of data. In other words, the variance of the count variable exceeds its mean. At this time, the traditional discrete distribution (the NNB distribution, Poisson distribution, negative binomial distribution, etc.) cannot fit the data well.

Given a count variable Y, its dispersion index is usually defined as $d = V(Y)/E(Y)$. The variable is over dispersed if $d > 1$. Another measure of the departure from the NNB distribution is the zero inflation index.

Definition 3.1. Let Y to be a nonnegative integer random variable (count variable) such that its mean is μ and its proportion of 0 is p_0. The zero inflation index of Y is $zi = 1 + log(p_0)/\mu$.

Notice that $zi = 0$ if Y is NNB distribution and $zi > 0$ if Y is "zero-inflated". That is, its proportion of 0 is greater than the proportion of 0 of an NNB variate with the same mean.

3.2 The Chi-square test

The test of zero inflation is equivalent to the hypothesis test of zero-inflated parameter w:

$$H_0 : w = 0 \qquad VS \qquad H_1 : w > 0.$$

When the test results do not reject the null hypothesis H_0, it is considered that there is no zero inflation phenomenon in the counting data. The Chi-square test statistic χ^2 is used to test the deviation between the actual observed data and the expected observed data of the sample data,

$$\chi^2 = \sum_{i=1}^{c} \frac{(f_i - m_i)^2}{m_i} = S_1,$$

where c is the number of classes decided for a given data set, f_i and m_i are the observed frequencies and expected frequencies under the null hypothesis H_0 of the ith class, respectively. When the null hypothesis is valid, the chi-square statistic follows an asymptotic chi-square distribution with $c - 1$ degrees of freedom.

§4 Maximum likelihood estimation and simulations

In this section, we discuss the estimation of the parameters w, p, v and λ of the ZINNB distribution. Let $a(y)$ be the observed frequency of y events for any $y = 0, 1, 2, ...$ and z be the highest value of y observed. Then the likelihood function of the sample is

$$L(w,p,v,\lambda|y) = \prod_{y=0}^{z} f(y)^{a(y)} = f_0(y)^{a(0)} \prod_{y=1}^{z} f_1(y)^{a(y)},$$

where $f_0(y)$ and $f_1(y)$ are given in equation (2) when $y = 0$ and $y \neq 0$, respectively.

The log-likelihood function of the sample is

$$lnL(w,p,v,\lambda|y) = a(0)ln[w + (1-w)e^{-\lambda p}q^v] + \sum_{y=1}^{z} a(y)ln[(1-w)e^{-\lambda p}p^y q^v L_y^{v-1}(-\lambda q)]. \quad (12)$$

On differentiating the log-likelihood function equation (12) with respect to the parameters w, p, v and λ and setting up to zero, we obtain the following equations,

$$\begin{aligned}
\frac{\partial lnL(w,p,v,\lambda|y)}{\partial w} &= \frac{a(0)(1-e^{-\lambda p}q^v)}{w+(1-w)e^{-\lambda p}q^v} + \sum_{y=1}^{z}\frac{a(y)}{1-w} = 0,\\
\frac{\partial lnL(w,p,v,\lambda|y)}{\partial p} &= -\frac{a(0)(1-w)e^{-\lambda p}q^v(\lambda+\frac{v}{q})}{w+(1-w)e^{-\lambda p}q^v} + \sum_{y=1}^{z} a(y)\left[\frac{y}{p} - \lambda - \frac{v}{q} + \lambda\frac{L_{y-1}^{v-2}(-\lambda q)}{L_y^{v-1}(-\lambda q)}\right] = 0,\\
\frac{\partial lnL(w,p,v,\lambda|y)}{\partial v} &= \frac{a(0)(1-w)e^{-\lambda p}vq^{v-1}}{w+(1-w)e^{-\lambda p}q^v} + \sum_{y=1}^{z} a(y)\left[\frac{v}{q} + \lambda\frac{L_{y-1}^{v-2}(-\lambda q)}{L_y^{v-1}(-\lambda q)}\right] = 0,
\end{aligned}$$

and

$$\frac{\partial lnL(w,p,v,\lambda|y)}{\partial \lambda} = -\frac{a(0)(1-w)e^{-\lambda p}q^v p}{w+(1-w)e^{-\lambda p}q^v} + \sum_{y=1}^{z} a(y)\left[\frac{qL_{y-1}^{v-2}(-\lambda q)}{L_y^{v-1}(-\lambda q)} - p\right] = 0.$$

Maximum likelihood estimators of w, p, v, and λ are obtained by solving the above equations simultaneously. The estimates can be obtained through numerical procedures and R programming, refering Wickham and Grolemund [13].

In this following, a simulation is conducted to illustrate the behavior of the maximum likelihood estimations. Simulations are carried out by taking samples of sizes $n = 50$, 100, 300, and 500 from $ZINNB(w,p,v,\lambda)$ distribution for all combinations of $w = 0.3, 0.8, p = 0.2, 0.6, v = 0.5, 3$, and $\lambda = 1, 4$. For each configuration of the experiments, 1000 data sets were generated. The maximum likelihood estimators for 16 groups of parameters with their standard deviations and bias are computed by simulated annealing (SANN) method with R software. The results are shown in Table 1, 2, 3 and 4.

As can be seen from Table 1, 2, 3 and 4, the sample size affects the estimator of parameters. When the sample size is getting bigger, the estimators are getting better. And the standard deviation values have decreasing trend, as the sample size increased. The bias values gets closer and closer to zero as the sample size increases.

§5 Applications

In this section, we consider to apply our proposed distribution into two real data sets. The first data set comes from labor mobility in the German Labor market, which measured how often a person changed employers over a ten-year period from 1974 to 1984. The number of

Table 1. Simulation of the maximum likelihood estimators for parameters with standard deviation (sd) and bias of the ZINNB distribution. (n=50)

w	p	v	λ	$\hat{w}$ (sd, bias)	$\hat{p}$ (sd, bias)	$\hat{v}$ (sd, bias)	$\hat{\lambda}$ (sd, bias)
0.3	0.2	0.5	4	0.212 (0.276, -0.095)	0.210 (0.071, 0.010)	0.433 (0.896, -0.083)	3.865 (0.946, -0.184)
0.3	0.2	0.5	1	0.197 (0.709, -0.131)	0.183 (0.551, -0.066)	0.542 (0.994, 0.043)	0.813 (1.325, -0.195)
0.3	0.2	3	4	0.286 (0.103, -0.014)	0.203 (0.124, 0.009)	2.928 (0.613, -0.098)	3.947 (0.777, -0.069)
0.3	0.2	3	1	0.260 (0.248, -0.060)	0.237 (0.127, 0.037)	2.819 (0.955, -0.183)	0.883 (1.019, -0.117)
0.3	0.6	0.5	4	0.321 (0.087, 0.010)	0.596 (0.046, -0.004)	0.502 (0.608, 0.044)	4.026 (0.659, 0.027)
0.3	0.6	0.5	1	0.227 (0.301, -0.063)	0.587 (0.115, -0.013)	0.595 (0.701, 0.096)	1.089 (0.826, 0.089)
0.3	0.6	3	4	0.309 (0.065, 0.032)	0.597 (0.034, -0.039)	3.001 (0.539, 0.017)	3.990 (0.555, -0.070)
0.3	0.6	3	1	0.304 (0.073, 0.031)	0.599 (0.056, -0.104)	2.998 (0.572, -0.018)	1.013 (0.633, 0.012)
0.8	0.2	0.5	4	0.725 (0.389, -0.075)	0.184 (0.208, -0.016)	0.456 (1.146, -0143)	3.839 (1.144, -0.164)
0.8	0.2	0.5	1	0.0.831 (1.223, 0.104)	0.138 (0.990, -0.313)	0.439 (0.559, -0.225)	0.859 (0.671, -0.316)
0.8	0.2	3	4	0.777 (0.095, -0.024)	0.202 (0.080, 0.063)	2.981 (0.913, -0.104)	3.931 (0.977, -0.086)
0.8	0.2	3	1	0.715 (0.338, -0.086)	0.188 (0.295, -0.094)	2.781 (1.159, -0.221)	0.949 (1.152, -0.251)
0.8	0.6	0.5	4	0.802 (0.075, 0.011)	0.587 (0.079, -0.013)	0.522 (0.111, 0.023)	4.006 (0.090, 0.025)
0.8	0.6	0.5	1	0.787 (0.358, -0.092)	0.564 (0.230, -0.037)	0.502 (0.269, 0.075)	1.016 (0.231, 0.064)
0.8	0.6	3	4	0.786 (0.061, -0.109)	0.593 (0.062, -0.085)	3.015 (0.754, 0.070)	4.022 (0.667, 0.048)
0.8	0.6	3	1	0.795 (0.068, -0.067)	0.593 (0.089, -0.081)	3.027 (0.307, 0.025)	1.059 (0.527, 0.059)

Table 2. Simulation of the maximum likelihood estimators for parameters with standard deviation (sd) and bias of the ZINNB distribution. (n=100)

w	p	v	λ	$\hat{w}$ (sd, bias)	$\hat{p}$ (sd, bias)	$\hat{v}$ (sd, bias)	$\hat{\lambda}$ (sd, bias)
0.3	0.2	0.5	4	0.243 (0.210, -0.057)	0.202 (0.045, 0.006)	0.439 (0.713, -0.062)	3.916 (0.750, -0.087)
0.3	0.2	0.5	1	0.249 (0.519, -0.103)	0.212 (0.300, 0.039)	0.523 (0.804, 0.023)	0.930 (1.111, -0.187)
0.3	0.2	3	4	0.294 (0.083, -0.006)	0.203 (0.036, 0.007)	2.942 (0.542, -0.061)	3.956 (0.513, -0.047)
0.3	0.2	3	1	0.285 (0.178, -0.042)	0.221 (0.096, 0.021)	2.899 (0.769, -0.103)	0.882 (0.792, -0.118)
0.3	0.6	0.5	4	0.302 (0.055, 0.009)	0.601 (0.032, -0.002)	0.501 (0.420, 0.016)	4.002 (0.413, 0.023)
0.3	0.6	0.5	1	0.353 (0.226, 0.058)	0.600 (0.091, 0.008)	0.533 (0.494, 0.032)	1.009 (0.602, 0.087)
0.3	0.6	3	4	0.299 (0.046, -0.029)	0.597 (0.024, -0.031)	2.997 (0.340, -0.021)	4.030 (0.384, 0.046)
0.3	0.6	3	1	0.302 (0.047, 0.028)	0.599 (0.037, -0.084)	3.014 (0.402, 0.012)	1.001 (0.409, -0.010)
0.8	0.2	0.5	4	0.767 (0.212, -0.033)	0.212 (0.113, 0.012)	0.476 (0.864, -0.124)	3.858 (0.888, -0.145)
0.8	0.2	0.5	1	0.806 (0.913, 0.073)	0.141 (0.699, -0.104)	0.426 (0.471, -0.223)	0.947 (0.459, -0.285)
0.8	0.2	3	4	0.789 (0.081, -0.011)	0.206 (0.053, 0.059)	2.898 (0.717, -0.095)	4.071 (0.737, 0.079)
0.8	0.2	3	1	0.749 (0.199, -0.051)	0.209 (0.225, 0.051)	2.830 (0.951, -0.172)	0.963 (0.939, -0.138)
0.8	0.6	0.5	4	0.802 (0.069, 0.007)	0.598 (0.059, -0.008)	0.497 (0.089, -0.013)	4.004 (0.078, 0.019)
0.8	0.6	0.5	1	0.745 (0.189, -0.056)	0.584 (0.132, -0.017)	0.501 (0.149, 0.066)	1.003 (0.157, 0.035)
0.8	0.6	3	4	0.799 (0.041, -0.105)	0.596 (0.038, -0.059)	32.994 (0.501, -0.069)	4.039 (0.406, 0.035)
0.8	0.6	3	1	0.797 (0.046, -0.032)	0.599 (0.059, -0.043)	2.997 (0.261, -0.019)	0.986 (0.494, -0.024)

Table 3. Simulation of the maximum likelihood estimators for parameters with standard deviation (sd) and bias of the ZINNB distribution. (n=300)

w	p	v	λ	$\hat{w}$ (sd, bias)	$\hat{p}$ (sd, bias)	$\hat{v}$ (sd, bias)	$\hat{\lambda}$ (sd, bias)
0.3	0.2	0.5	4	0.287 (0.102, -0.014)	0.205 (0.034, 0.005)	0.465 (0.426, -0.036)	3.957 (0.381, -0.047)
0.3	0.2	0.5	1	0.288 (0.353, -0.091)	0.209 (0.163, 0.032)	0.527 (0.531, 0.027)	0.985 (0.767, -0.109)
0.3	0.2	3	4	0.297 (0.035, -0.002)	0.201 (0.013, 0.003)	2.988 (0.276, -0.015)	3.981 (0.268, -0.023)
0.3	0.2	3	1	0.305 (0.099, 0.017)	0.208 (0.057, 0.012)	2.949 (0.471, -0.054)	0.952 (0.455, -0.049)
0.3	0.6	0.5	4	0.291 (0.026, -0.004)	0.599 (0.015, -0.001)	0.494 (0.213, -0.007)	4.001 (0.227, 0.014)
0.3	0.6	0.5	1	0.280 (0.136, -0.021)	0.587 (0.053, -0.003)	0.512 (0.294, 0.012)	1.089 (0.361, 0.016)
0.3	0.6	3	4	0.301 (0.021, 0.018)	0.601 (0.012, 0.027)	2.999 (0.163, -0.012)	4.001 (0.185, -0.013)
0.3	0.6	3	1	0.301 (0.023, 0.013)	0.597 (0.021, -0.059)	2.998 (0.194, -0.005)	1.003 (0.234, 0.009)
0.8	0.2	0.5	4	0.786 (0.099, -0.015)	0.199 (0.046, -0.009)	0.525 (0.508, 0.065)	3.965 (0.518, -0.039)
0.8	0.2	0.5	1	0.794 (0.388, -0.034)	0.202 (0.322, 0.073)	0.483 (0.213, -0.121)	0.988 (0.233, -0.167)
0.8	0.2	3	4	0.797 (0.029, -0.004)	0.201 (0.025, 0.014)	2.986 (0.350, -0.021)	3.995 (0.387, -0.072)
0.8	0.2	3	1	0.786 (0.054, -0.015)	0.205 (0.064, 0.042)	2.927 (0.561, -0.075)	0.997 (0.518, -0.042)
0.8	0.6	0.5	4	0.801 (0.031, 0.003)	0.597 (0.029, -0.005)	0.498 (0.039, -0.012)	4.002 (0.045, 0.008)
0.8	0.6	0.5	1	0.822 (0.128, 0.028)	0.601 (0.074, 0.012)	0.521 (0.087, 0.050)	1.003 (0.121, 0.028)
0.8	0.6	3	4	0.787 (0.020, -0.076)	0.601 (0.020, 0.047)	3.002 (0.237, 0.034)	3.998 (0.283, -0.018)
0.8	0.6	3	1	0.794 (0.023, -0.019)	0.589 (0.031, -0.021)	3.012 (0.169, 0.006)	1.003 (0.307, 0.017)

Table 4. Simulation of the maximum likelihood estimators for parameters with standard deviation (sd) and bias of the ZINNB distribution. (n=500)

w	p	v	λ	$\hat{w}$ (sd, bias)	$\hat{p}$ (sd, bias)	$\hat{v}$ (sd, bias)	$\hat{\lambda}$ (sd, bias)
0.3	0.2	0.5	4	0.297 (0.064, -0.010)	0.201 (0.030, 0.001)	0.497 (0.273, -0.024)	4.008 (0.283, -0.034)
0.3	0.2	0.5	1	0.302 (0.259, 0.063)	0.232 (0.139, 0.032)	0.508 (0.452, 0.014)	0.963 (0.555, -0.094)
0.3	0.2	3	4	0.299 (0.025, -0.002)	0.200 (0.010, 0.001)	2.993 (0.189, -0.010)	3.994 (0.208, -0.009)
0.3	0.2	3	1	0.292 (0.084, -0.008)	0.194 (0.035, -0.003)	2.984 (0.313, -0.018)	1.008 (0.301, 0.029)
0.3	0.6	0.5	4	0.300 (0.018, 0.002)	0.599 (0.012, -0.001)	0.492 (0.153, -0.006)	4.002 (0.166, 0.009)
0.3	0.6	0.5	1	0.287 (0.083, -0.013)	0.600 (0.034, 0.002)	0.504 (0.215, 0.004)	1.006 (0.258, 0.005)
0.3	0.6	3	4	0.301 (0.015, 0.007)	0.603 (0.008, 0.015)	3.002 (0.124, -0.007)	4.005 (0.127, 0.005)
0.3	0.6	3	1	0.305 (0.017, 0.008)	0.598 (0.015, -0.048)	2.995 (0.147, -0.003)	0.998 (0.172, -0.003)
0.8	0.2	0.5	4	0.796 (0.044, -0.005)	0.204 (0.046, 0.004)	0.513 (0.379, 0.048)	3.977 (0.402, -0.027)
0.8	0.2	0.5	1	0.778 (0.301, -0.019)	0.207 (0.195, -0.037)	0.490 (0.137, -0.086)	0.989 (0.207, -0.145)
0.8	0.2	3	4	0.797 (0.018, -0.002)	0.200 (0.017, 0.002)	2.993 (0.265, -0.017)	4.011 (0.269, 0.014)
0.8	0.2	3	1	0.795 (0.036, -0.007)	0.204 (0.048, 0.014)	2.946 (0.365, -0.055)	0.965 (0.389, -0.036)
0.8	0.6	0.5	4	0.799 (0.024, -0.002)	0.600 (0.019, -0.004)	0.498 (0.032, -0.007)	4.001 (0.039, 0.003)
0.8	0.6	0.5	1	0.804 (0.080, 0.017)	0.601 (0.058, 0.001)	0.500 (0.053, 0.021)	0.994 (0.105, -0.003)
0.8	0.6	3	4	0.799 (0.013, -0.013)	0.600 (0.012, -0.033)	2.996 (0.174, -0.011)	3.985 (0.180, -0.006)
0.8	0.6	3	1	0.799 (0.015, -0.011)	0.597 (0.023, -0.009)	3.003 (0.119, 0.004)	1.009 (0.237, 0.008)

participants was 807, and the detailed description of the data is shown in Winkelmann and Zimmermann [14]. The second data set has a sample of 5190 people from the Australian health survey, asking how many times they consulted a doctor or specialist in the two weeks prior to the interview, for details see Winkelmann [15]. ZINNB distribution was used to fit these two groups of data, and other existing distributions, such as, NNB, ZIP and ZINB distribution were selected for comparison.

For comparing all models, we calculated the Akaike information criterion (AIC) and Bayesian information criterion (BIC). All these data are shown in Table 5 and Table 6. In addition, the comparison diagrams of each distribution fitting data are drawn in Figure 2 and Figure 3.

For the labor mobility data set, the value of the empirical dispersion index and the empirical zero inflation index are $\tilde{d}_1 = 2.030 > 1$ and $\tilde{zi} = 0.303$. We calculate the Chi-square test statistic $S = 131.642 > 15.086$, so we reject the null hypothesis H_0, in favor of $H_1 : w > 0$. In chi-square test, the defined classes for this example are {{0}, {1}, {2}, {3}, {4}, {5 and above}}. Consequently, these empirical values show that this set of data is over-dispersed and zero inflation exists. Based on the above results, we conducted a fitting analysis on this group of data, and the results are shown in Table 5.

Table 5. Fit the labor mobility data using ZINNB, NNB, ZIP and ZINB distribution.

Count	Observed frequency	ZINNB	NNB	ZIP	ZINB
0	465	478.13	321.40	450.99	596.07
1	183	162.13	249.36	157.59	44.80
2	89	92.06	135.33	114.74	30.71
3	39	44.01	61.51	55.70	23.11
4	17	18.87	25.01	20.28	18.16
5	5	7.49	9.40	5.91	14.63
6	1	2.81	3.33	1.43	11.99
7	6	1.00	1.13	0.30	9.94
8	0	0.35	0.37	0.05	8.31
9	1	0.12	0.12	0.01	6.99
10	1	0.04	0.04	0.001	5.91
total	807	807	807	807	807
Estimates of the parameters		$\hat{w} = 0.341$		$\hat{w} = 0.425$	$\hat{w} = 0.624$
		$\hat{p} = 0.190$	$\hat{p} = 0.179$		
		$\hat{v} = 0.983$	$\hat{v} = 1.176$		$\hat{k} = 0.548$
		$\hat{\lambda} = 3.982$	$\hat{\lambda} = 3.844$	$\hat{\lambda} = 1.456$	$\hat{\mu} = 4.257$
AIC		1998.67	2100.95	2030.18	2454.50
BIC		2017.44	2115.03	2039.57	2468.58

In this data set, we see that the AIC and BIC of each distribution are from small to large are, ZINNB, ZIP, NNB, ZINB. Therefore, it can be seen that the ZINNB distribution gives better fit to this set of data.

For the number of doctor consultations data set, the value of the empirical dispersion index and the empirical zero inflation index are $\tilde{d}_1 = 2.111 > 1$ and $\tilde{zi} = 0.252$. The Chi-square test statistic $S = 131.642 > 15.086$, so we reject the null hypothesis H_0, in favor of $H_1 : w > 0$. In chi-square test, the defined classes for this example are {{0}, {1}, {2}, {3}, {4}, {5 and above}}. Consequently, these empirical values show that this set of data is over-dispersed and zero inflation exists. Based on the above results, we conducted a fitting analysis on this group of data, and the results are shown in Table 6.

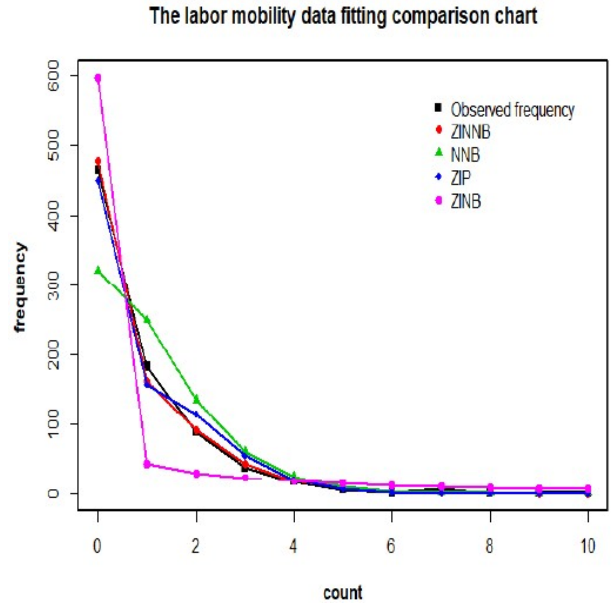

Figure 2. Using ZINNB, NNB, ZIP and ZINB distribution to fit the labor mobility data comparison chart.

Table 6. Fit the number of doctor consultations data using ZINNB, NNB, ZIP and ZINB distribution.

Count	Observed frequency	ZINNB	NNB	ZIP	ZINB
0	4141	4146.25	3032.64	4245.33	4651.26
1	782	675.48	1449.65	407.59	132.32
2	174	256.15	505.45	305.21	83.21
3	30	81.20	149.57	152.36	59.61
4	24	22.98	39.86	57.05	45.32
5	9	6.01	9.86	17.09	35.65
6	12	1.48	2.31	4.26	28.67
7	12	0.35	0.52	0.91	23.41
8	5	0.08	0.11	0.17	19.33
9	1	0.02	0.02	0.03	16.11
total	5190	5190	5190	5190	5190
Estimates of the parameters		$\hat{w} = 0.559$		$\hat{w} = 0.766$	$\hat{w} = 0.827$
		$\hat{p} = 0.121$	$\hat{p} = 0.119$		
		$\hat{v} = 0.973$	$\hat{v} = 0.627$		$\hat{k} = 0.411$
		$\hat{\lambda} = 3.975$	$\hat{\lambda} = 3.850$	$\hat{\lambda} = 1.498$	$\hat{\mu} = 3.379$
AIC		7297.99	8359.10	7669.76	8997.86
BIC		7324.21	8378.76	7682.87	9017.52

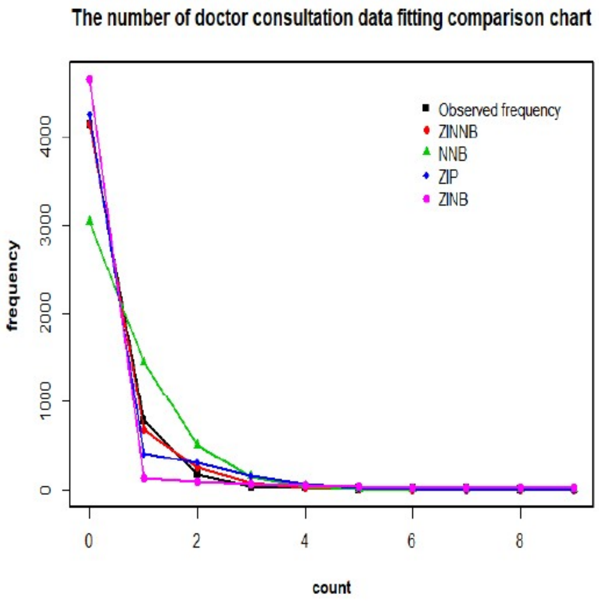

Figure 3. Using ZINNB, NNB, ZIP and ZINB distribution to fit the number of doctor consultation data comparison chart.

According to the data in Table 6, the AIC and BIC for ZINNB distribution is relatively small, which can be concluded that the ZINNB distribution gives better fit to this data.

From the above fitting results, it can be seen that the goodness of fit of different distributions in different data sets may be different. On the whole, we can get the goodness of fit of ZINNB distribution is relatively good from Figure 2 and Figure 3.

By looking at these two sets of data, we find that the data are overdispersed. In other words, the variance of the count variable exceeds its mean. The reason why ZINNB distribution can better fit these two sets of data may be that ZINNB is composed of other distributions and can degenerate into other distributions under different cases, including extra parameters. Moreover, the data sets are overdispersed, and the traditional discrete distributions cannot fit the data well. Therefore, when the data set is over-dispersed, we can try to select ZINNB distribution to fit the data.

§6 Conclusion

In this paper, the ZINNB distribution is introduced. Several important statistical properties of the distribution are studied. The maximum likelihood estimation for parameters of the ZINNB distribution is discussed. In order to illustrate the usefulness of this model, two real data application are investigated. After comparing with the existing distributions, the results showed that the ZINNB distribution had a good goodness of fit. Continued in-depth study of the nature and the related regression models of ZINNB distribution will be the subject of subsequent work.

References

[1] D Böhning. *Zero-inflated Poisson models and CA MAN: A tutorial collection of evidence*, Biometrical Journal: Journal of Mathematical Methods in Biosciences, 1998, 40(7): 833-843.

[2] F Famoye, K Singh. *Zero-inflated generalized Poisson regression model with an application to domestic violence data*, Journal of Data Science, 2006, 4(1): 117-130.

[3] W H Greene. *Accounting for excess zeros and sample selection in Poisson and negative binomial regression models*, 1994.

[4] D B Hall. *Zero-inflated Poisson and binomial regression with random effects: a case study*, Biometrics, 2000, 56(4): 1030-1039.

[5] M T Hasan, G Sneddon. *Zero-inflated Poisson regression for longitudinal data*, Communications in Statistics-Simulation and Computation, 2009, 38(3): 638-653.

[6] D Lambert. *Zero-inflated Poisson regression, with an application to defects in manufacturing*, Technometrics, 1992, 34(1): 1-14.

[7] P A Lee, S H Ong. *Higher-order and non-stationary properties of Lampard's stochastic reversible counter system*, Statistics: A Journal of Theoretical and Applied Statistics, 1986, 17(2): 261-278.

[8] Ong S H, Lee P A. *Bivariate non-central negative binomial distribution: Another generalisation*, Metrika, 1986, 33(1): 29-46.

[9] S H Ong, P A Lee. *The non-central negative binomial distribution*, Biometrical Journal, 1979, 21(7): 611-627.

[10] S H Ong, K K Toh, Y C Low. *The non-central negative binomial distribution: Further properties and applications*, Communications in Statistics-Theory and Methods, 2019, 1-16.

[11] K F Sellers, D S Young. *Zero-inflated sum of Conway-Maxwell-Poissons (ZISCMP) regression*, Journal of Statistical Computation and Simulation, 2019, 89(9): 1649-1673.

[12] S Z Sim, R C Gupta, S H Ong. *Zero-inflated Conway-Maxwell Poisson distribution to analyze discrete data*, The international journal of biostatistics, 2018, 14(1).

[13] H Wickham, G Grolemund. *R for data science: import, tidy, transform, visualize, and model data*, O'Reilly Media, 2016.

[14] R Winkelmann, K F Zimmermann. *Recent developments in count data modelling: theory and application*, Journal of economic surveys, 1995, 9(1): 1-24.

[15] R Winkelmann. *Duration dependence and dispersion in count-data models*, Journal of business economic statistics, 1995, 13(4): 467-474.

[16] K K W Yau, K Wang, A H Lee. *Zero-inflated negative binomial mixed regression modeling of over-dispersed count data with extra zeros*, Biometrical Journal: Journal of Mathematical Methods in Biosciences, 2003, 45(4): 437-452.

[17] K K W Yau, A H Lee. *Zero-inflated Poisson regression with random effects to evaluate an occupational injury prevention programme*, Statistics in medicine, 2001, 20(19): 2907-2920.

[1] College of Big Data and Internet, Shenzhen Technology University, Shenzhen 518118, China.

[2] School of Science, Xi'an University of Technology, Xi'an 710054, China.

Email: tianweizhong@sztu.edu.cn