Pakistan Journal of Statistics and Operation Research

A new three-parameter discrete distribution to model over-dispersed count data

Anupama Nandi[1], Partha Jyoti Hazarika[2*], Aniket Biswas[3], G. G. Hamedani[4]

*Corresponding author

1. Department of Statistics, Dibrugarh University, Assam, India, anupamanandi0702@gmail.com
2. Department of Statistics, Dibrugarh University, Assam, India, parthajhazarika@gmail.com
3. Department of Statistics, Dibrugarh University, Assam, India, biswasaniket44@gmail.com
4. Department of Mathematical and Statistical Sciences, Marquette University, Wisconsin, U.S.A., gholamhoss.hamedani@marquette.edu

Abstract

A novel discrete distribution with three parameters, referred to as the PoiNB distribution, is formulated through the convolution of a Poisson variable and an independently distributed negative binomial random variable. This distribution generalizes some well known count distributions and can be used for modelling over-dispersed as well as equi-dispersed count data. Numerous essential statistical properties of this proposed count model are thoroughly examined. Characterizations of this distribution in terms of conditional expectation are studied in details. The estimation of the unknown parameters of this proposed distribution is carried out using the maximum likelihood estimation approach. Additionally, we introduce a count regression model based on the PoiNB distribution through the generalized linear model approach. Two real life modelling applications demonstrate that the proposed distribution may prove to be useful for modelling over-dispersed count data compared to its closest competitors.

Key Words: Negative-binomial distribution; Poisson distribution; Conway-Maxwell Poisson distribution; BerG distribution; Confluent hypergeometric function; Incomplete beta function.

Mathematical Subject Classification: 60E05, 62E15.

1. Introduction

The phenomenon of over-dispersion, where the variance of count data exceeds its mean, is a commonly discussed topic in literature. Over-dispersion is frequently encountered in various modelling applications and it is encountered more often compared to the phenomena of under-dispersion and equi-dispersion. Numerous count models have been proposed for over-dispersed data, reflecting the ongoing research interest in this field (Wongrin and Bodhisuwan (2017), Moghimbeigi et al. (2008), Tapak et al. (2020), Wang et al. (2017), Sarvi et al. (2019), Rodrigues-Motta et al. (2013), Campbell et al. (1999), Hassanzadeh and Kazemi (2016), Moqaddasi Amiri et al. (2019), Tüzen et al. (2020), Bar-Lev and Ridder (2021), Altun (2020), and Wang et al. (2001)). The simplest and most common count data model is the Poisson distribution, known for its equi-dispersion characteristic. However, this simplicity poses a limitation, leading to the development of several alternative distributions that offer advantages over the classical Poisson model. Notable among these alternatives are the hyper-Poisson (HP) (Bardwell and Crow, 1964), generalized Poisson distribution (Jain and Consul, 1971), double-Poisson (Efron, 1986), weighted Poisson (Del Castillo and Pérez-Casany, 1998), weighted generalized Poisson distribution (Chakraborty, 2010), Mittag-Leffler function distribution (Chakraborty and Ong, 2017), and the COM-Poisson distribution (Sellers and Shmueli, 2010), which generalizes the Poisson distribution to accommodate over-dispersion.

The classical geometric and negative binomial models are also employed for over-dispersed count datasets. The gamma mixture of the Poisson distribution generates the negative binomial distribution (Fisher et al., 1943). Several extensions of the geometric distribution have been introduced for over-dispersed count data modelling (Chakraborty and Bhati (2016), Chakraborty and Gupta (2015), Gómez-Déniz (2010), Jain and Consul (1971), Makcutek (2008), Nekoukhou et al. (2012), Philippou et al. (1983), and Tripathi et al. (1987)). Despite the prevalence of the negative binomial and COM-Poisson distributions, there remains ample opportunity to develop new discrete distributions with simple structures and explicit interpretations suitable for over-dispersed data.

Bourguignon and Weiß (2017) introduced the BerG distribution by combining a Bernoulli random variable with a geometric random variable through convolution. Similarly, (Bourguignon et al., 2022) introduced the BerPoi distribution using a comparable method, convolving a Bernoulli random variable with a Poisson random variable. The BerG distribution can effectively handle over-dispersed, under-dispersed, and equi-dispersed data by blending the characteristics of Bernoulli and geometric distributions. On the other hand, the BerPoi distribution is suitable for modelling under-dispersed and equi-dispersed data by merging Bernoulli and Poisson distribution features.

More recently, Nandi et al. (2024) proposed the PoiG distribution, which combines a Poisson random variable with a geometric random variable. The geometric distribution is known for over-dispersion, while the Poisson distribution is equi-dispersed. Consequently, the PoiG distribution also exhibits over-dispersion due to the convolution process. For Y following the $PoiG(\lambda, \theta)$ distribution, $0 < \lambda$ and $0 < \theta < 1$, the probability mass function (pmf) of Y is given by

$$p_Y(y) = \frac{\theta(1-\theta)^y}{\Gamma(y+1)} \exp\left(\frac{\lambda\theta}{1-\theta}\right) \Gamma\left(y+1, \frac{\lambda}{1-\theta}\right), \qquad y = 0, 1, 2, \dots \,. \tag{1}$$

Although the negative binomial distribution is commonly used for modelling over-dispersed count data, there is a need for better alternative models in statistical literature. This motivation led us to develop a new distribution. Utilizing the simple and effective convolution approach mentioned above, we introduce a novel over-dispersed count model called PoiNB. This new model is derived from the convolution of two independent count random variables: Poisson and negative binomial.

The Poisson distribution, exhibiting equi-dispersion, and the negative binomial distribution, demonstrating over-dispersion, result in the over-dispersion characteristic of the PoiNB model. This three-parameter distribution offers several advantages, including structural simplicity, easy comprehensibility compared to the COM-Poisson distribution, and closed-form expressions for mean and variance. Unlike the COM-Poisson distribution, the proposed distribution extends the Poisson, the geometric and the negative binomial distributions.

Rest of the article is organized as follows. In Section 2, we present the PoiNB distribution. In Section 3, we describe its important statistical properties such as recurrence relation, generating functions, moments, dispersion index, reliability properties. In Section 4, we present the characterizations of the proposed distribution. In Section 5, we present the maximum likelihood methods of parameter estimation. In Section 6, we introduce the count regression model based on the PoiNB distribution. In Section 7, two real datasets are analyzed to illustrate the practical utility of the PoiNB distribution. We conclude the article with a few limitations and future scopes of the current study.

2. The PoiNB distribution

In this section, we introduce a novel discrete distribution that arises from the analysis of two separate, independent discrete random variables Y_1 and Y_2. Let us denote the set of non-negative integers, $\{0, 1, 2, ...\}$ by N_0. Also let, Y_1 follow the Poisson distribution $P(\lambda)$ where $\lambda > 0$ and Y_2 follow the negative binomial distribution $NB(r, \theta)$ where $r \in \mathbb{N}$ and $0 < \theta < 1$, respectively. Both the variables are restricted to non-negative integer values. Consider, $Y = Y_1 + Y_2$. Then, the probability mass function (pmf) of Y is

$$\begin{aligned} p_Y(y) &= \sum_{i=0}^{y} \Pr(Y_1 = i) \Pr(Y_2 = y - i) \\ &= \sum_{i=0}^{y} \frac{e^{-\lambda}\lambda^i}{i!} \binom{y-i+r-1}{y-i} (1-\theta)^{y-i} \end{aligned}$$

$$\begin{aligned} &= e^{-\lambda}\theta^r(1-\theta)^y \binom{y+r-1}{y} {}_1F_1\left(-y; 1-r-y; \frac{\lambda}{1-\theta}\right) \\ &= e^{-\lambda}\theta^r(1-\theta)^y \frac{(r+y-1)!}{(r-1)!y!} {}_1F_1\left(-y; 1-r-y; \frac{\lambda}{1-\theta}\right), \quad y = 0, 1, 2, \ldots . \end{aligned} \tag{2}$$

Here, ${}_1F_1(a, b, z) = \sum_{k=0}^{\infty} \frac{(a)_k z^k}{(b)_k k!}$ is Kummer's confluent hypergeometric function Zarzo et al. (1995). It is to be mentioned that $(a)_0 = 1$ and $(a)_k = a(a+1)\ldots(a+k-1)$. The pmf given in (2) is well defined as $a = -y$ and $b - a = 1 - r$ are non-positive integers while $z = \lambda/(1-\theta)$ is a real number. Being the convolution of Poisson and negative binomial, this distribution is named the PoiNB distribution and we write $Y \sim PoiNB(\lambda, r, \theta)$, where $\lambda > 0, \theta \in (0, 1), r \in \mathbb{N}$. From the formulation of the $PoiNB(\lambda, r, \theta)$ model, it is very convenient to obtain its mean and variance explicitly. Note that,

$$\begin{aligned} \mu = E(Y) = E(Y_1) + E(Y_2) = \lambda + \frac{r(1-\theta)}{\theta}, \\ \sigma^2 = V(Y) = V(Y_1) + V(Y_2) = \lambda + \frac{r(1-\theta)}{\theta^2}. \end{aligned} \tag{3}$$

Remark 1. The $PoiNB(\lambda, r, \theta)$ distribution behaves like the Poisson distribution with parameter λ, negative binomial distribution with parameter r and p and geometric distribution with parameter θ as $r \to 0$ or $\theta \to 1$, $\lambda \to 0$, and $\lambda \to 0$ and $r \to 1$, respectively.

Remark 2. Let $C = e^{-\lambda}\theta^r$ and $G(y) = \frac{(r+y-1)!}{(r-1)!y!} {}_1F_1\left(-y; 1-r-y; \frac{\lambda}{1-\theta}\right)$, then the pmf given in (2) can be expressed as

$$p_Y(y) = C\,(1-\theta)^y\,G(y), \quad y = 0, 1, 2, \ldots . \tag{4}$$

The cumulative distribution function (cdf) of PoiNB distribution is

$$\begin{aligned} F_Y(y) &= \Pr(Y_1 + Y_2 \le y) \\ &= \sum_{y_1=0}^{y} F_{NB}(y - y_1) p_Y(y_1) \\ &= \sum_{y_1=0}^{y} I_\theta(r, y - y_1 + 1) \frac{e^{-\lambda}\lambda^{y_1}}{y_1!}, \end{aligned} \tag{5}$$

where, $I_x(a,b) = B(x,a,b)/B(a,b)$ is the regularized incomplete beta function. The incomplete beta function in the above expressions is given by

$$B(x,a,b) = \int_0^X t^{a-1}(1-t)^{b-1}dt.$$

From Remark 2, the cdf given in (5) can also be expressed as

$$F_Y(y) = \sum_{a=0}^{y} C\,(1-\theta)^a\,G(a) = C\,Q(y), \qquad y = 0,1,2,\ldots \tag{6}$$

where $Q(y) = \sum_{a=0}^{y}(1-\theta)^a\,G(a)$.

3. Properties of the PoiNB distribution

In this section, we explore several important statistical properties of the proposed $PoiNB(\lambda, r, \theta)$ distribution. Some of the distributional properties studied here are the recurrence relation, generating functions, moments related concepts, index of dispersion and coefficient of variation.

3.1. Recurrence relation

The recurrence relation of the probability mass function's assists in determining the probability mass at a subsequent point based on the probability mass at a previous point. From the representation of the pmf in (2),

$$p_Y(y+1) = e^{-\lambda}\theta^r(1-\theta)^{y+1}\frac{(r+y)!}{(r-1)!(y+1)!}\;{}_1F_1(-(y+1);-r-y;\frac{\lambda}{1-\theta})$$

The ratio between two successive point is

$$\frac{p_Y(y+1)}{p_Y(y)} = (1-\theta)\,\frac{(r+y)}{(y+1)}\;\frac{{}_1F_1(-(y+1);-r-y;\dfrac{\lambda}{1-\theta})}{{}_1F_1(-y;1-r-y;\dfrac{\lambda}{1-\theta})}.$$

3.2. Generating functions

If Y is the sum of independent random variables Y_1 and Y_2, then the probability generating function (pgf) of Y, denoted as $G_Y(s)$ with argument s, can be easily determined using the product of the pgfs of Y_1 and Y_2, denoted as $H_{Y_1}(s)$ and $H_{Y_2}(s)$, respectively.

$$H_Y(s) = H_{Y_1}(s)H_{Y_2}(s).$$

This approach is applicable for obtaining the moment generating function (mgf) and the characteristic function (cf) as well. It is important to note that the probability generating functions (pgfs), moment generating functions (mgfs), and characteristic functions (cfs) for both the Poisson distribution (Y_1) and the negative binomial distribution (Y_2) are readily available. For $Y \sim PoiNB(\lambda, r, \theta)$, pgf of Y is given by

$$H_Y(s) = \frac{\theta^r e^{\lambda(s-1)}}{(1-(1-\theta)s)^r}.$$

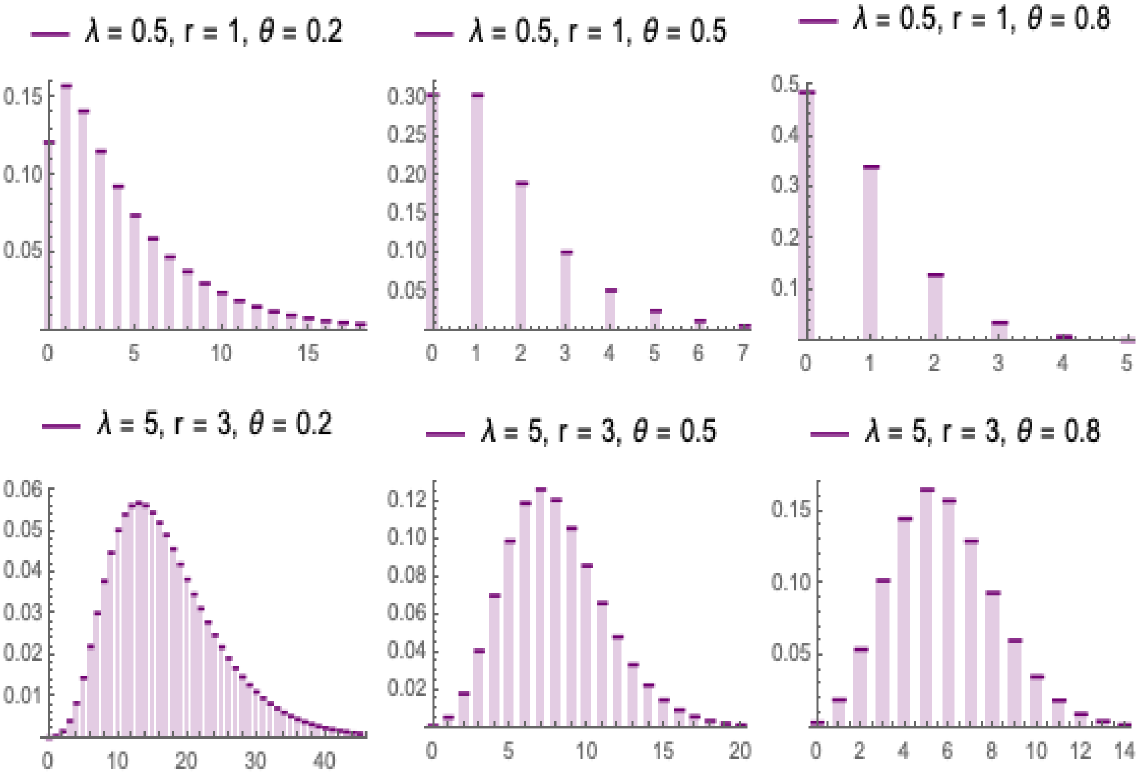

Figure 1: Probability mass function of $PoiNB(\lambda, r, \theta)$ for different choices of λ, r and θ.

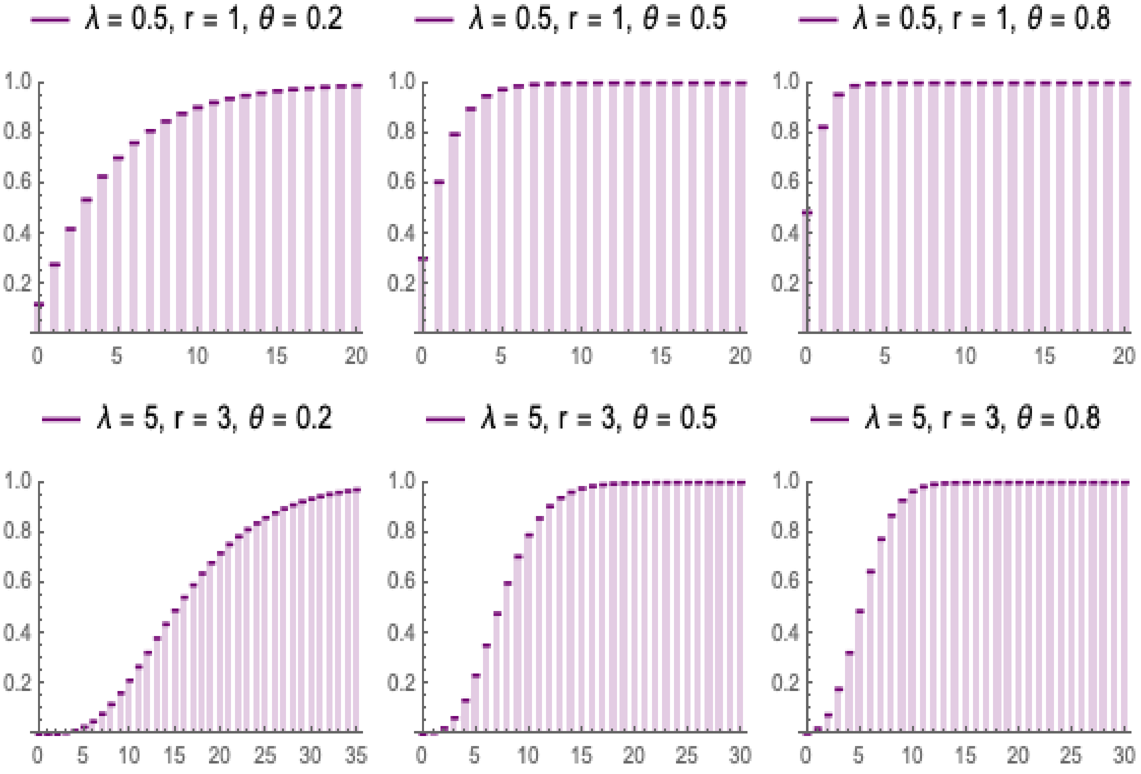

Figure 2: Cumulative distribution function of $PoiNB(\lambda, r, \theta)$ for different choices of λ, r and θ.

The mgf of Y is obtained by replacing s by e^t

$$M_Y(t) = \frac{\theta^r e^{\lambda(e^t-1)}}{(1-(1-\theta)e^t)^r}.$$

The cf of Y is

$$\phi_Y(t) = \frac{\theta^r e^{\lambda(e^{it}-1)}}{(1-(1-\theta)e^{it})^r}.$$

3.3. Moments and related concepts

The k^{th} order raw moment of $Y \sim PoiNB(\lambda, r, \theta)$ can be obtained using the general expressions of the raw moments of $Y_1 \sim P(\lambda)$ and $Y_2 \sim NB(r, \theta)$ as follows.

$$\begin{aligned} E(Y^k) &= E\left[\sum_{j=0}^{k} \binom{k}{j} {Y_1}^j {Y_2}^{k-j}\right] \\ &= \sum_{j=0}^{k} \binom{k}{j} E({Y_1}^j) E({Y_2}^{k-j}) \end{aligned}$$

$$E({Y_1}^j) = \sum_{y_1=0}^{\infty} {y_1}^j \ \frac{e^{-\lambda}\lambda^{y_1}}{y_1!} = B_j(\lambda).$$

Here, $B_n(\alpha)$ denotes the ordinary Bell polynomial defined by

$$e^{\alpha(e^t-1)} = \sum_{n=0}^{\infty} B_n(\alpha) \frac{t^n}{n!}.$$

For the definition of $B_n(\alpha)$ and the derivation of the j^{th} order raw moment of the Poisson random variable in terms of Bell polynomials, one may refer to the work of Kim et al. (2021).

Again,

$$E({Y_2}^{k-j}) = \sum_{y_2=0}^{\infty} \binom{y_2+r-1}{y_2} {y_2}^{k-j}\, \theta^r (1-\theta)^{y_2},$$

Consider $N_{k-j}(r,\theta) = \sum_{y_2=0}^{\infty} \binom{y_2+r-1}{y_2} {y_2}^{k-j}, \theta^r (1-\theta)^{y_2}$. The general expressions for the raw moments of Y is as follows

$$E(Y^k) = \sum_{j=0}^{k} \binom{k}{j} B_j(\lambda) \ N_{k-j}(r,\theta). \tag{7}$$

Let μ_k' denote the raw moment of order k, that is $\mu_k' = E(Y^k)$. The explicit expressions of the first four moments are

$$
\begin{aligned}
\mu_1' &= \lambda + \frac{r(1-\theta)}{\theta}, \\
\mu_2' &= \frac{1}{\theta^2}\left[\theta^2\lambda(\lambda+1) + r^2(1-\theta)^2 + r(1-\theta)(2\theta\lambda+1)\right], \\
\mu_3' &= \frac{1}{\theta^3}[\lambda(1+\lambda(3+\lambda))\theta^3 + r^2(1-\theta)^2(1+\theta(3\lambda-1)) + r(1-\theta)(\theta^2(3\lambda(\lambda+1)-1)+ \\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad 3\theta\lambda+\theta+1)], \quad \text{and} \\
\mu_4' &= \frac{1}{\theta^4}[\lambda(1+\lambda(7+\lambda(6+\lambda)))\theta^4 + (1-\theta)^4 r^3 + (1-\theta)^2 r^2(\theta^2\left(6\lambda^2+2\lambda-3\right)+ \\
&\qquad 4\theta\lambda+3) + r(1-\theta)(2+\theta(4\lambda+\theta\left(4\theta\lambda^2(\lambda+3)-2\theta+2\lambda(3\lambda+5)+1\right)))].
\end{aligned}
$$

Using the above expressions of the first four raw moments, we can obtain the first four central moments of Y explicitly. Let μ_k denote the central moment of order k, that is $\mu_k = E(Y-\mu_1')^k$. The first central moment is always 0. The explicit expressions of the next three central moments are

$$
\begin{aligned}
\mu_2 &= \lambda + \frac{r(1-\theta)}{\theta^2}, \\
\mu_3 &= \frac{1}{\theta^3}[\theta^3\lambda + (1-\theta)^3 r^3 - \left(\theta^3 - 3\theta + 2\right) r^2 + (\theta-2)\theta^2 r + r], \ \text{and} \\
\mu_4 &= \frac{1}{\theta^4}[\theta^4\lambda(1+3\lambda) + 3(\theta-1)^4 r^4 - 3(\theta+1)(\theta-1)^3 r^3 + \\
&\qquad ((\theta-4)\theta-1)(\theta-1)^2 r^2 + (\theta-1)r\left(\theta^2(2\theta-6\lambda-1)-2\right)].
\end{aligned}
$$

The mean and variance of the $PoiNB(\lambda, r, \theta)$ distribution correspond to the first raw moment (μ_1') and second central moment (μ_2), respectively. The skewness and kurtosis measures, expressed as $\mu_3/\mu_2^{3/2}$ and μ_4/μ_2^2 respectively, can be directly derived from the central moments.

$$
\begin{aligned}
\beta_1 &= \frac{\mu_3^2}{\mu_2^3} = \frac{\left(\theta^3\lambda + (\theta-1)^3 r^3 - \left(\theta^3-3\theta+2\right) r^2 + (\theta-2)\theta^2 r + r\right)^2}{\left(\lambda\theta^2 + r(1-\theta)\right)^3} \\
\beta_2 &= \frac{\mu_4}{\mu_2^2} = \frac{1}{\left(\theta^2\lambda - \theta r + r\right)^2}[\theta^4\lambda(3\lambda+1) + 3(1-\theta)^4 r^4 + 3(1+\theta)(1-\theta)^3 r^3 + \\
&\qquad ((\theta-4)\theta-1)(1-\theta)^2 r^2 + (\theta-1)r\left(\theta^2(2\theta-6\lambda-1)-2\right)]
\end{aligned}
$$

The 3-D surface plots of the measures of coefficient of skewness $\gamma_1 = \sqrt{\beta_1}$ and coefficient of kurtosis $\gamma_2 = \beta_2 - 3$ are presented in Figure (3) and Figure (4) , respectively. It is simple to verify that the $PoiNB(\lambda, r, \theta)$ is positively skewed from Figure 3. When r increases and λ and θ get closer to zero, the distribution gets more skewed. As λ and θ grow greater, the distribution gets closer to the symmetry. This can be easily verified from the surface plots in the first and third column of Figure 3. From the second column, it appears that the parameter r seems to have an opposite effect; for lower r, it approaches symmetry, while for larger r, it exhibits more skewness. From Figure 4, it is clear that the $PoiNB(\lambda, r, \theta)$ distribution is leptokurtic. Similar to the case of skewness, for larger values of λ and θ, and smaller values of r the $PoiNB(\lambda, r, \theta)$ distribution approaches the mesokurtic characteristic. For smaller values of λ and θ, and greater values of r, the $PoiNB(\lambda, r, \theta)$ distribution is highly leptokurtic in nature.

3.4. Index of dispersion and coefficient of variation

The index of dispersion (Hoel, 1943) is a statistical measure that describes the degree of variability or dispersion in a dataset. It is used to assess the capability of a given distribution in modelling over-dispersed, under-dispersed and equi-dispersed datasets. If the index of dispersion of a distribution exceeds unity, it

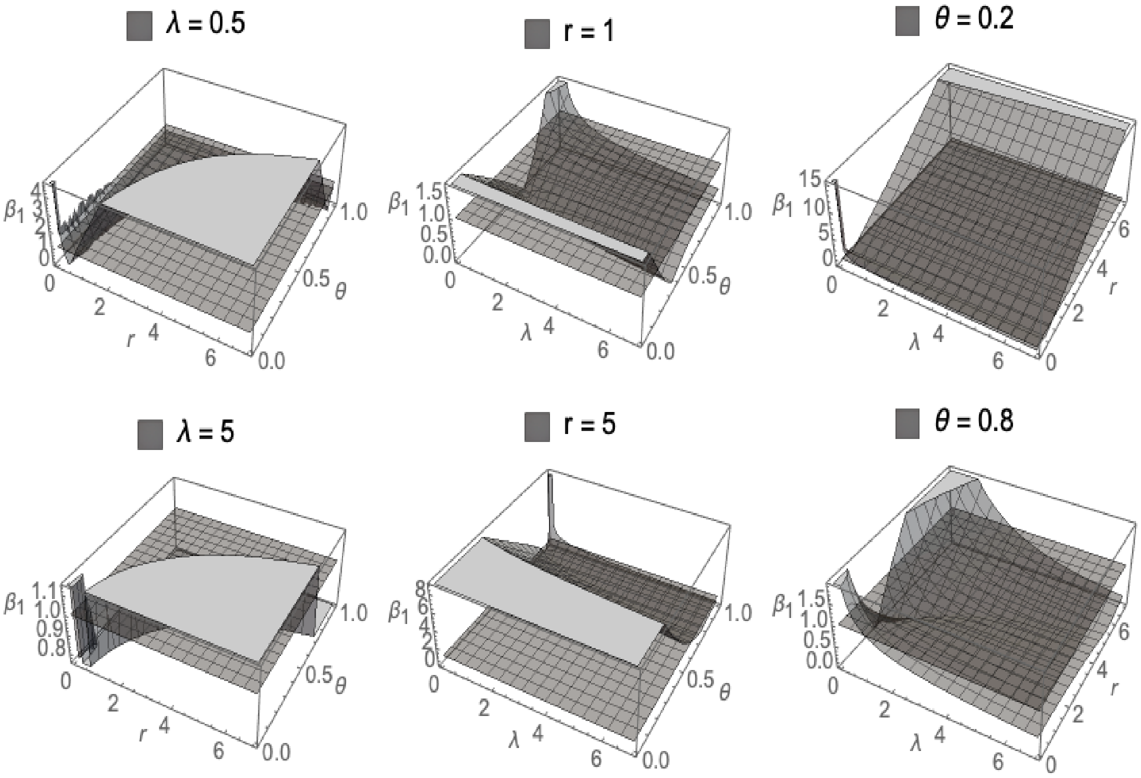

Figure 3: Skewness of $PoiNB(\lambda, r, \theta)$ for different choices of λ, r and θ.

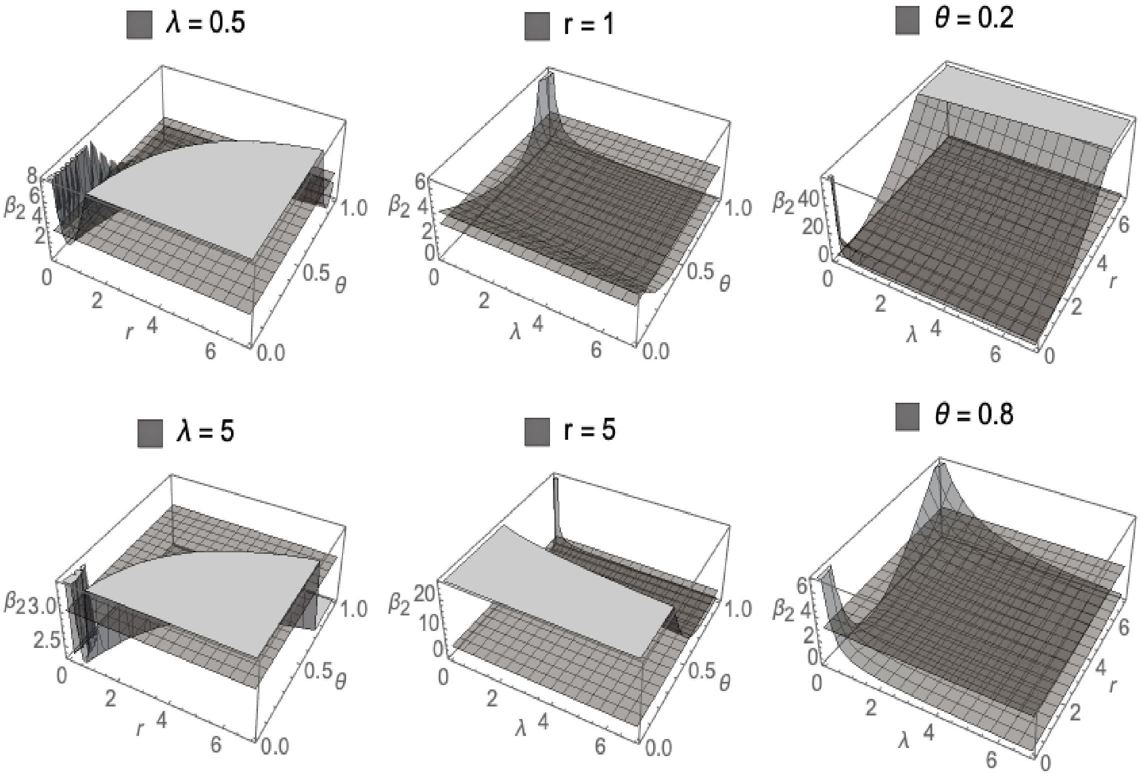

Figure 4: Kurtosis of $PoiNB(\lambda, r, \theta)$ for different choices of λ, r and θ.

indicates that the distribution can effectively model over-dispersed datasets. Similarly, for index of dispersion below 1 or equal to 1, it can effectively model under-dispersed or equi-dispersed datasets, respectively. The dispersion index of $Y \sim PoiNB(\lambda, r, \theta)$ is given by

$$DI_Y = \frac{\sigma^2}{\mu} = 1 + \frac{r(1-\theta)^2}{\theta(\lambda\theta + r(1-\theta))}. \tag{8}$$

From the expression of DI_Y above, it follows that the PoiNB distribution is equi-dispersed when $\theta = 1$ and over-dispersed for all $0 < \theta < 1$ and $r > 0$. The coefficient of variation (cv) of $PoiNB(\lambda, r, \theta)$ distribution is given by

$$cv = \frac{\sigma}{\mu} \times 100\% = \frac{\sqrt{\lambda\theta^2 + r(1-\theta)}}{\lambda\theta + r(1-\theta)} \times 100\%.$$

3.5. Reliability properties

The corresponding survival function (sf), hazard rate function (hrf), and reverse hazard rate function (rhrf) of $Y \sim PoiNB(\lambda, r, \theta)$ are

$$S_Y(y) = 1 - C\,Q(y-1), \tag{9}$$

$$h_Y(y) = \frac{C\,(1-\theta)^y\,G(y)}{1 - C\,Q(y-1)}, \tag{10}$$

$$r_Y(y) = \frac{C\,(1-\theta)^y\,G(y)}{\sum_{a=0}^{y} C\,(1-\theta)^a\,G(a)} \tag{11}$$

respectively. In the above expression, $C = e^{-\lambda}\theta^r$, $G(y) = \frac{(r+y-1)!}{(r-1)!y!}\,{}_1F_1(-y; 1-r-y; \frac{\lambda}{1-\theta})$, and $Q(y-1) = \sum_{a=0}^{y-1}(1-\theta)^a\,G(a)$.

4. Characterizations

In this section, we present our characterizations of the PoiNB distribution: (i) In terms of the conditional expectation of certain function of the random variable; (ii) Based on the reverse hazard function. The choice of the function in (i) depends on the form of the pmf. We devote a subsection to each of (i) and (ii).

4.1. Based on conditional expectation

Proposition 1. Y is a random variable with pmf given in (4) if and only if

$$E\left\{[G(y)]^{-1} \mid Y \le k\right\} = \frac{1-(1-\theta)^{k+1}}{\theta\,Q(k)} \tag{12}$$

where $G(y)$ and $Q(k)$ are given in Remark 2 and equation (6), respectively.

Proof. If Y has pmf (4), then for $k \in \mathbb{N}^*$, the left-hand side of (12) using finite geometric sum formula, will be

$$\begin{aligned}
&E\left\{[G(y)]^{-1} \ \mid Y \le k\right\} = (F(k))^{-1}\sum_{y=0}^{k} C(1-\theta)^{y} \\
&= \frac{C\left(\frac{(1-\theta)^{k+1}-1}{(1-\theta)-1}\right)}{C\,Q(k)} = \frac{1-(1-\theta)^{k+1}}{\theta\,Q(k)}.
\end{aligned}$$

Conversely, if (12) holds, then

$$\sum_{y=0}^{k}\left\{[G(y)]^{-1} \ p(y)\right\} = F(k)\left(\frac{1-(1-\theta)^{k+1}}{\theta\,Q(k)}\right) \tag{13}$$

From (13) , we also have

$$\begin{aligned}
&\sum_{y=0}^{k-1}\left\{[G(y)]^{-1} \ p(y)\right\} = F(k-1)\left(\frac{1-(1-\theta)^{k}}{\theta\,Q(k-1)}\right) \\
&= (F(k)-p(k))\left(\frac{1-(1-\theta)^{k}}{\theta\,Q(k-1)}\right).
\end{aligned} \tag{14}$$

Now, subtracting (14) from (13) , yields

$$\begin{aligned}
&[G(k)]^{-1}\,p(k) \\
&= F(k)\left\{\left(\frac{1-(1-\theta)^{k+1}}{\theta\,Q(k)}\right) - \left(\frac{1-(1-\theta)^{k}}{\theta\,Q(k)}\right)\right\} \\
&+ p(k)\left(\frac{1-(1-\theta)^{k}}{\theta\,Q(k-1)}\right),
\end{aligned}$$

or

$$\begin{aligned}
&p(k)\left\{\frac{1}{G(k)} - \left(\frac{1-(1-\theta)^{k}}{\theta\,Q(k-1)}\right)\right\} \\
&= F(k)\left\{\left(\frac{1-(1-\theta)^{k+1}}{\theta Q(k)}\right) - \left(\frac{1-(1-\theta)^{k}}{\theta Q(k-1)}\right)\right\}.
\end{aligned}$$

From the above equality, we have

$$\begin{aligned}
\frac{p(k)}{F(k)} &= \frac{\left(\frac{1-(1-\theta)^{k+1}}{\theta\,Q(k)}\right) - \left(\frac{1-(1-\theta)^{k}}{\theta\,Q(k-1)}\right)}{\frac{1}{G(k)} - \left(\frac{1-(1-\theta)^{k}}{\theta\,Q(k-1)}\right)} \\
&= 1 - \frac{\frac{1}{G(k)} - \left(\frac{1-(1-\theta)^{k+1}}{\theta\,Q(k)}\right)}{\frac{1}{G(k)} - \left(\frac{1-(1-\theta)^{k}}{\theta\,Q(k-1)}\right)},
\end{aligned}$$

and after some simplification we arrive at

$$\frac{p(k)}{F(k)} = \frac{(1-\theta)^{k}\,G(k)}{Q(k)}$$

which is the reverse hazard function corresponding to the pmf (4), so Y has pmf (4).

4.2. Based on reverse hazard rate function

Proposition 2. Let Y be a random variable. The pmf of Y is (4) if and only if its reverse hazard rate function, r_F given in (11), satisfies the difference equation

$$\begin{aligned} & r_F(k+1) - r_F(k) \\ & = \frac{(1-\theta)^{k+1} G(k+1)}{\sum_{y=0}^{k+1} (1-\theta)^y G(y)} - \frac{(1-\theta)^k G(k)}{\sum_{y=0}^{k} (1-\theta)^y G(y)}, \quad k \in \mathbb{N}^*, \end{aligned} \tag{15}$$

with the initial condition $r_F(0) = 1$.

Proof. If Y has pmf (4) , then clearly (15) holds. Now, if (15) holds, then for every $y \in \mathbb{N}$, we have

$$\begin{aligned} & \sum_{k=0}^{y-1} \{r_F(k+1) - r_F(k)\} \\ & = \sum_{k=0}^{x-1} \left\{ \frac{\frac{1}{G(k)} - \left(\frac{1-(1-\theta)^{k+1}}{\theta\, Q(k)}\right)}{\frac{1}{G(k)} - \left(\frac{1-(1-\theta)^{k}}{\theta\, Q(k-1)}\right)} - \frac{\frac{1}{G(k+1)} - \left(\frac{1-(1-\theta)^{k+2}}{\theta\, Q(k+1)}\right)}{\frac{1}{G(k+1)} - \left(\frac{1-(1-\theta)^{k+1}}{\theta\, Q(k)}\right)} \right\}, \end{aligned}$$

or, using telescoping sum

$$r_F(y) - r_F(0) = -\frac{\frac{1}{G(y)} - \left(\frac{1-(1-\theta)^{y+1}}{\theta\, Q(y)}\right)}{\frac{1}{G(y)} - \left(\frac{1-(1-\theta)^{y}}{\theta\, Q(y-1)}\right)},$$

or in view of the initial condition $(r_F(0) = 1)$

$$r_F(y) = 1 - \frac{\frac{1}{G(y)} - \left(\frac{1-(1-\theta)^{y+1}}{\theta\, Q(y)}\right)}{\frac{1}{G(y)} - \left(\frac{1-(1-\theta)^{y}}{\theta\, Q(y-1)}\right)} = \frac{(1-\theta)^y G(y)}{Q(y)}, \quad y \in \mathbb{N}^*,$$

which is the reverse hazard function corresponding to the pmf (4).

5. Parameter Estimation

Let $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)$ denote a random sample of size n drawn from the $PoiNB(\lambda, r, \theta)$ distribution, and $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ denote a realization on $\mathbf{Y}$. The objective of this section is to estimate the parameters λ, r and α based on the available data $\mathbf{y}$. We present the maximum likelihood methods of estimation. Using the pmf of $Y \sim PoiNB(\lambda, r, \theta)$ in (2), the log-likelihood function of the parameters λ, r and θ can easily be found as

$$l(\lambda, q, \alpha; \boldsymbol{y}) = \sum_{i=1}^{n} \log\left(e^{-\lambda}\theta^r(1-\theta)^{y_i}\frac{(r+y_i-1)!}{(r-1)!y_i!}\ {}_1F_1(-y_i; 1-r-y_i; \frac{\lambda}{1-\theta})\right). \tag{16}$$

Here are some functions that will be used in the score functions :

$$
\begin{aligned}
&\psi_n(z) = \frac{\partial^{n+1}}{\partial z^{n+1}} ln(\Gamma z) = \frac{\partial^n}{\partial z^n}\psi_0(z),\\
&{}_1F_1^{(0,n,0)}(a;b;z) = \sum_{k=0}^{\infty} \frac{(a)_k\, z^k}{k!}\frac{\partial^n}{\partial b^n}\frac{1}{(b)_k},\\
&{}_1F_1^{(0,0,n)}(a;b;z) = \frac{(a)_n}{(b)_n}\,{}_1F_1(a+n;b+n;z),\\
&{}_1F_1^{(0,1,0)}(a;b;z) = \psi(b)\,{}_1F_1(a;b;z) - \sum_{k=0}^{\infty}\frac{(a)_k\, z^k\, \psi(b+k)}{k!\,(b)_k},\\
&{}_1F_1^{(0,1,1)}(a;b;z) = \frac{a}{b^2}{}_1F_1(a+1;b+1;z) - \frac{a}{b}\sum_{k=0}^{\infty}\frac{(a+1)_k\, z^k}{k!}\frac{\partial}{\partial b}\frac{1}{(b+1)_k}.
\end{aligned}
$$

Let $A(y_i) = {}_1F_1(-y_i;1-r-y_i;\dfrac{\lambda}{1-\theta})$, $B_j(y_i) = {}_1F_1^{(0,\,j,\,0)}(-y_i;1-r-y_i;\dfrac{\lambda}{1-\theta})$, and $C_{jk}(y_i) = {}_1F_1^{(0,\,i,\,j)}(-y_i;1-r-y_i;\dfrac{\lambda}{1-\theta})$ for $j,\, k\, =\, 1,2\ldots n$. Differentiating (16), with respect to parameters the parameters λ, r, and θ, we get the score functions as

$$
\frac{\partial}{\partial\lambda} l(\lambda,r,\theta;\boldsymbol{y}) = -n - \sum_{i=1}^{n}\frac{y_i\, A(y_i-1)}{(1-r-y_i)(1-\theta)A(y_i)}, \tag{17}
$$

$$
\frac{\partial}{\partial r} l(\lambda,r,\theta;\boldsymbol{y}) = n(\log\theta - \psi_0(r)) + \sum_{i=1}^{n}\psi_0(y_i+r) - \sum_{i=1}^{n}\frac{B_1(y_i)}{A_1(y_i)}, \tag{18}
$$

$$
\frac{\partial}{\partial\theta} l(\lambda,r,\theta;\boldsymbol{y}) = \frac{nr}{\theta} - \sum_{i=1}^{n}\frac{y_i}{1-\theta} - \sum_{i=1}^{n}\frac{\lambda\, y_i\, A(y_i-1)}{(1-\theta)^2(1-r-y_i)A(y_i)}. \tag{19}
$$

The maximum likelihood estimators are ideally derived by simultaneously solving the equations resulting from setting the right-hand sides of (17), (18), and (19) to zero. However, the structural complexity of these equations makes it challenging to obtain explicit expressions for the maximum likelihood estimators. Therefore, we choose to employ numerical optimization techniques to directly maximize the log-likelihood function with respect to the parameters. Let $\hat{\lambda}_{ML}$, $\hat{r}_{ML}$, and $\hat{\theta}_{ML}$ represent the maximum likelihood estimates (MLE) for λ, r, and θ" respectively.

In order to derive the information matrix, it is necessary to calculate the second-order partial derivatives of the log-likelihood function concerning the parameters λ, r, and θ. However, obtaining precise expressions for all these second-order partial derivatives can be cumbersome and challenging. The second-order partial derivatives of the log-likelihood function for $PoiNB(\lambda, r, \theta)$ are given as

$$\frac{\partial^2 l(\lambda,r,\theta;\boldsymbol{y})}{\partial\lambda^2} = \sum_{i=1}^{n}\left(\frac{y_i(y_i-1)A(y_i-2)}{(1-\theta)^2(1-r-y_i)(2-r-y_i)A(y_i)} - \frac{y_i^2\,A(y_i-1)^2}{(1-\theta)^2(1-r-y_i)^2\,A(y_i)^2}\right) \tag{20}$$

$$\frac{\partial^2 l(\lambda,r,\theta;\boldsymbol{y})}{\partial r^2} = -n\,\psi_1(r) + \sum_{i=1}^{n}\left[\psi_1(y_i+r) - \frac{A(y_i)^2}{A(y_i)^2} + \frac{B_2(y_i)}{A(y_i)}\right] \tag{21}$$

$$\frac{\partial^2 l(\lambda,r,\theta;\boldsymbol{y})}{\partial\lambda\partial r} = \sum_{i=1}^{n}\frac{y_i\left(-A(y_i-1)/(1-r-y_i)+B_1(y_i-1)-(A(y_i-1)A(y_i))/A(y_i)\right)}{(1-\theta)(1-r-y_i)A(y_i)} \tag{22}$$

$$\begin{aligned}\frac{\partial^2 l(\lambda,r,\theta;\boldsymbol{y})}{\partial\theta^2} &= -\frac{nr}{\theta^2} + \sum_{i=1}^{n}\left(-\frac{y_i}{(1-\theta)^2} - \frac{2\lambda\,y_iA(y_i-1)}{(1-\theta)^3(1-r-y_i)A(y_i)}\right)\\ &\quad + \sum_{i=1}^{n}\left(\frac{\lambda^2\,y_i(y_i-1)\,A(y_i-2)}{(1-\theta)^4(1-r-y_i)(2-r-y_i)\,A(y_i)} - \frac{\lambda^2\,y_i^2\,A(y_i-1)^2}{(1-\theta)^4(1-r-y_i)^2\,A(y_i)^2}\right)\end{aligned} \tag{23}$$

$$\begin{aligned}\frac{\partial^2 l(\lambda,r,\theta;\boldsymbol{y})}{\partial\lambda\partial\theta} &= \sum_{i=1}^{n}\left(-\frac{y_iA(y_i-1)}{(1-\theta)^2(1-r-y_i)\,A(y_i)} + \frac{\lambda\,y_i(y_i-1)\,A(y_i-2)}{(1-\theta)^3(1-r-y_i)(2-r-y_i)\,A(y_i)}\right)\\ &\quad - \sum_{i=1}^{n}\frac{\lambda\,y_i^2\,A(y_i-1)}{(1-\theta)^3(1-r-y_i)^2\,A(y_i)}\end{aligned} \tag{24}$$

$$\frac{\partial^2 l(\lambda,r,\theta;\boldsymbol{y})}{\partial r\partial\theta} = \frac{n}{\theta} - \sum_{i=1}^{n}\frac{\lambda}{(1-\theta)^2}\left(\frac{y_i\,A(y_i-1)B_1(y_i)}{(1-r-y_i)A(y_i)^2} + \frac{C_{11}(y_i)}{A(y_i)}\right). \tag{25}$$

The Fisher's information matrix for (λ, q, α) is

$$I_Y(\lambda,r,\theta) = \begin{pmatrix} -E\left(\frac{\partial^2 l(\lambda,\,r,\,\theta;\boldsymbol{y})}{\partial\lambda^2}\right) & -E\left(\frac{\partial^2 l(\lambda,r,\theta;\boldsymbol{y})}{\partial\lambda\partial r}\right) & -E\left(\frac{\partial^2 l(\lambda,\,r,\,\theta;\boldsymbol{y})}{\partial\lambda\partial\theta}\right) \\ -E\left(\frac{\partial^2 l(\lambda,\,r,\,\theta;\boldsymbol{y})}{\partial\lambda\partial r}\right) & -E\left(\frac{\partial^2 l(\lambda,\,r,\,\theta;\boldsymbol{y})}{\partial r^2}\right) & -E\left(\frac{\partial^2 l(\lambda,\,r,\,\theta;\boldsymbol{y})}{\partial r\partial\theta}\right) \\ -E\left(\frac{\partial^2 l(\lambda,\,r,\,\theta;\boldsymbol{y})}{\partial\lambda\partial\theta}\right) & -E\left(\frac{\partial^2 l(\lambda,\,r,\,\theta;\boldsymbol{y})}{\partial r\partial\theta}\right) & -E\left(\frac{\partial^2 l(\lambda,\,r,\,\theta;\boldsymbol{y})}{\partial\theta^2}\right)\end{pmatrix}.$$

This can be approximated by

$$\widehat{I}_Y(\lambda,\,r,\,\theta) \approx \begin{pmatrix} -\frac{\partial^2 l(\lambda,\,r,\,\theta;\boldsymbol{y})}{\partial\lambda^2} & -\frac{\partial^2 l(\lambda,r,\theta;\boldsymbol{y})}{\partial\lambda\partial r} & -\frac{\partial^2 l(\lambda,\,r,\,\theta;\boldsymbol{y})}{\partial\lambda\partial\theta} \\ -\frac{\partial^2 l(\lambda,\,r,\,\theta;\boldsymbol{y})}{\partial\lambda\partial r} & -\frac{\partial^2 l(\lambda,\,r,\,\theta;\boldsymbol{y})}{\partial r^2} & -\frac{\partial^2 l(\lambda,\,r,\,\theta;\boldsymbol{y})}{\partial r\partial\theta} \\ -\frac{\partial^2 l(\lambda,\,r,\,\theta;\boldsymbol{y})}{\partial\lambda\partial\theta} & -\frac{\partial^2 l(\lambda,\,r,\,\theta;\boldsymbol{y})}{\partial r\partial\theta} & -\frac{\partial^2 l(\lambda,\,r,\,\theta;\boldsymbol{y})}{\partial\theta^2}\end{pmatrix}_{(\lambda,\,r,\,\theta)=(\hat{\lambda}_{ML},\hat{r}_{ML},\hat{\theta}_{ML})}.$$

For a large value of n, the maximum likelihood estimators $\hat{\lambda}_{ML}$, $\hat{r}_{ML}$, and $\hat{\theta}_{ML}$ exhibit consistency and asymptotic normality. The distribution of $\sqrt{n}(\hat{\lambda}_{ML}-\lambda,\,\hat{r}_{ML}-r,\hat{\theta}_{ML}-\theta)$ is trivariate normal with zero means and the dispersion matrix $\hat{I}^{-1}=[d_{ij}]_{3\times3}$, where $d_{ij}=d(\hat{\lambda}_{ML},\hat{r}_{ML},\hat{\theta}_{ML};y)$. The dispersion matrix $\hat{I}^{-1}$ includes the variances of $\hat{\lambda}_{ML}$, $\hat{r}_{ML}$, and $\hat{\theta}_{ML}$, denoted by d_{11}, d_{22}, and d_{33}, respectively. Let z_α represent the $(1-\alpha)$-th quantile of the standard normal distribution. The asymptotic $(1-\alpha)\times 100\%$ confidence interval for the parameters λ, r and θ are given by

$$\left(\hat{\lambda}_{ML} - z_{\alpha/2}\sqrt{d_{11}}\ ,\ \hat{\lambda}_{ML} + z_{\alpha/2}\sqrt{d_{11}}\right), \left(\hat{r}_{ML} - z_{\alpha/2}\sqrt{d_{22}}\ ,\ \hat{r}_{ML} + z_{\alpha/2}\sqrt{d_{22}}\right)$$

and

$$\left(\hat{\theta}_{ML} - z_{\alpha/2}\sqrt{d_{33}}\ ,\ \hat{\theta}_{ML} + z_{\alpha/2}\sqrt{d_{33}}\right).$$

6. PoiNB regression model

In this section, we introduce a reparametrized version of the $PoiNB(\lambda, r, \theta)$ model and subsequently employ the generalized linear model (GLM) approach to establish a novel count regression model rooted in the proposed distribution, referred to as $PoiNB_{GLM}$. Let μ and ϕ denote the mean and index of dispersion of $Y \sim PoiNB(\lambda, r, \theta)$, given in (3) and (8), respectively.

$$\mu = \lambda + \frac{r(1-\theta)}{\theta},$$
$$\phi = 1 + \frac{r(1-\theta)^2}{\theta(r + \lambda\theta - r\theta)}.$$

To achieve the reparametrization of the $PoiNB(\lambda, r, \theta)$ distribution, one may use the inverse transformations $\lambda = \mu - \sqrt{r\mu(\phi-1)}$ and $\theta = \left(1 + \sqrt{\mu(\phi-1)/r}\right)^{-1}$ in the pmf of PoiNB distribution given in (2)

$$p_Y(y) = \frac{\left(\sqrt{\frac{\mu(\phi-1)}{r}}\right)^y}{\left(1+\sqrt{\frac{\mu(\phi-1)}{r}}\right)^{r+y}} \exp\left(-\mu + \sqrt{r\mu(\phi-1)}\right) \frac{(r+y-1)!}{(r-1)!y!}$$
$$ {}_1F_1\left(-y; 1-r-y; r\left(1+\sqrt{\frac{\mu(\phi-1)}{r}}\right)\left(\sqrt{\frac{\mu}{r(\phi-1)}}-1\right)\right). \tag{26}$$

This PoiNB model has one location parameter μ, one dispersion parameter ϕ and the integer-valued index r can be considered as a scaling parameter.

Consider an observed sample of size n, denoted as y_1, y_2, ..., y_n, drawn from the $PoiNB(\lambda, r$
$, \theta)$ model as defined in (26). In this context, y_i represents the response variable associated with a specific set of covariates $\boldsymbol{x_i'}$ for each $i = 1, 2, ..., n$. Moreover, we assume that the mean of the response variable y_i is linked to the covariates with a log link function given by

$$\mu_i = e^{\boldsymbol{x'_i\beta}}, \quad i = 1, 2, ...n. \tag{27}$$

Here, $\boldsymbol{x_i'} = (1, x_{i1}, x_{i2}, ..., x_{ip})$ represents the covariate vector corresponding to the i-th observation, where $i = 1, 2, ..., n$, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, ..., \beta_p)'$ is the vector of unknown regression coefficients. By substituting the expression for μ_i from (27) into (26), we express the probability mass function (pmf) of $y_i|\boldsymbol{x_i'} \sim$

$PoiNB(\mu_i, \phi, r)$ as:

$$P_Y(y_i|\boldsymbol{x_i'}) = \frac{\left(\sqrt{\dfrac{e^{\boldsymbol{x_i'\beta}}(\phi-1)}{r}}\right)^{y_i}}{\left(1+\sqrt{\dfrac{e^{\boldsymbol{x_i'\beta}}(\phi-1)}{r}}\right)^{r+y_i}} \exp\left(-e^{\boldsymbol{x_i'\beta}} + \sqrt{re^{\boldsymbol{x_i'\beta}}(\phi-1)}\right) \frac{(r+y_i-1)!}{(r-1)!y_i!}$$
$$ {}_1F_1\left(-y_i; 1-r-y_i; r\left(1+\sqrt{\frac{e^{\boldsymbol{x_i'\beta}}(\phi-1)}{r}}\right)\left(\sqrt{\frac{e^{\boldsymbol{x_i'\beta}}}{r(\phi-1)}}-1\right)\right). \qquad (28)$$

Using (28), we obtain the log-likelihood function of $\boldsymbol{\delta} = (\boldsymbol{\beta},\ \phi,\ r)'$ for given $y_1, y_2, ..., y_n$ and fixed $\boldsymbol{x_1'}, \boldsymbol{x_2'}, ..., \boldsymbol{x_n'}$ as

$$l(\boldsymbol{\delta}) = \sum_{i=1}^{n}\left(-e^{\boldsymbol{x_i'\beta}} + \sqrt{r\beta(x_i,\phi)}\right) + \sum_{i=1}^{n} y_i \log\left(\sqrt{\frac{\beta(x_i,\phi)}{r}}\right) - \sum_{i=1}^{n}(y_i+r)\log\left(1+\sqrt{\frac{\beta(x_i,\phi)}{r}}\right) +$$
$$\sum_{i=1}^{n}\log\left(\frac{(r+y_i-1)!}{(r-1)!y_i!}{}_1F_1\left(-y_i; 1-r-y_i; r\left(1+\sqrt{\frac{\beta(x_i,\phi)}{r}}\right)\left(\sqrt{\frac{\beta(x_i,\phi)}{r(\phi-1)^2}}-1\right)\right)\right), \qquad (29)$$

where $\beta(x_i,\phi) = e^{\boldsymbol{x_i'\beta}}(\phi-1)$. We may use numerical methods directly to maximize $l(\boldsymbol{\delta})$ and obtain the maximum likelihood estimates of $\boldsymbol{\beta}, \phi$ and r.

7. Data analysis

In this section, we demonstrate the suitability of the PoiNB distribution by applying it to two real-life datasets. The first data set (Dataset I) relates to the count of journeys undertaken by Dutch families with at least one car during a specific survey week in 1989. The initial analysis of this dataset was conducted by Ophem (2000). The second dataset (Dataset II) pertains to the number of ticks observed on sheep, as examined by Fisher (1941). Table 1 presents important summary measures of the response variables. The dispersion indices (DI) for both datasets exceed unity, indicating the presence of over-dispersion. We assess the fitted values of the PoiNB distribution in comparison to several established discrete distributions such as the Poisson (P) distribution, geometric (G) distribution, negative binomial (NB) distribution, COM-Poisson (CMP) distribution, and $BerG$ distribution. The comparison of the fitted models is based on conventional model selection criteria such as the negative log-likelihood (-LL), the Akaike information criterion (AIC), the Bayesian information criterion (BIC), the Chi-square goodness of fit test (χ^2) and the resulting p-value.

Table 1: Descriptive statistics for the data sets.

Dataset	Variable	n	Mean	Median	Var	std.dev	DI(I_Y)	CV(%)
I	Number of trips	1839	3.04	3	3.41	1.84	1.12	60.80
II	Number of ticks	82	6.56	5	34.77	5.89	5.29	89.87

In Tables 2 and 3, the expected frequencies are derived from estimated probabilities, which, in turn, are computed using maximum likelihood estimates of the relevant parameters. The goodness of fit is assessed using the chi-square (χ^2) statistic, along with associated p-values and additional model selection criteria such as the negative log-likelihood (-LL), the Akaike information criterion (AIC), and the Bayesian information criterion (BIC). From the analysis, it can be observed that the PoiNB distribution performs better compared to the other distributions as it has the least values for -LL, AIC, BIC and χ^2. Note that the p-values corresponding to the χ^2 goodness of fit tests for Dataset II is well above the nominal level in case of PoiNB

Table 2: Goodness of fit results for Dataset I.

		Expected Frequency					
y	freq	Poi	Geo	NB	CMP	$BerG$	PoiNB
0	75	88.19	455.50	102.32	99.99	75.05	95.60
1	312	267.88	342.68	281.25	275.00	556.96	280.30
2	384	406.83	257.80	399.41	398.89	381.10	411.89
3	421	411.91	193.95	391.89	397.94	260.77	404.97
4	307	312.80	145.91	297.35	304.40	178.43	300.36
5	183	190.03	109.77	186.17	189.50	122.09	179.99
6	77	96.20	82.58	100.10	99.70	83.54	91.50
7	47	41.74	62.12	47.50	45.50	57.16	41.23
8	15	15.85	46.74	20.29	18.35	39.11	17.33
$\geq$9	18	07.57	141.95	12.17	09.72	84.76	15.83
	1839	1839	1839	1839	1839	1839	1839
Estimated		$\hat{\lambda}$=3.04	$\hat{\theta}$=0.25	$\hat{r}$=28.87	$\hat{\lambda}$=2.75	$\hat{\pi}$=0.87	$\hat{\lambda}$=2.91
				$\hat{\theta}$=0.90	$\hat{r}$=0.92	$\hat{\theta}$=2.18	$\hat{r}$=0.03
							$\hat{\theta}$=0.22
(χ^2,df)		(29.99,8)	(1008,8)	(23.33,7)	(26.24,7)	(399,7)	**(14.46,6)**
$p-$value		0.0002	<0.0001	0.0014	<0.0001	<0.0001	0.0248
−LL		3615.52	4156.28	3610.09	3613.52	3797.14	**3596.37**
AIC		7233.04	8314.5	7224.18	7231.04	7598.28	**7198.74**
BIC		7238.55	8320.01	7235.21	7242.09	7609.31	**7215.29**

Table 3: Goodness of fit results for Dataset II.

		Expected Frequency					
y	freq	Poi	Geo	NB	CMP	$BerG$	PoiNB
0	04	00.11	10.84	05.25	07.08	04.00	02.83
1	05	00.76	09.40	07.34	07.54	11.31	07.28
2	11	02.49	08.16	08.03	07.54	09.67	10.09
3	10	05.45	07.08	07.96	07.26	08.27	10.31
4	09	08.95	06.14	07.48	06.82	07.07	08.99
5	11	11.75	05.33	06.80	06.28	06.04	07.32
6	03	12.84	04.62	06.04	05.68	05.16	05.86
7	05	12.04	04.01	05.28	05.07	04.41	04.72
8−10	07	21.80	09.13	11.76	11.76	09.76	09.67
11−14	09	05.50	07.46	08.68	09.17	07.58	06.98
$\geq$15	08	00.26	09.77	07.36	07.76	08.69	07.89
	82	82	82	82	82	82	82
Estimated		$\hat{\lambda}$=6.56	$\hat{\theta}$=0.13	$\hat{r}$=1.77	$\hat{\lambda}$=1.06	$\hat{\pi}$=0.66	$\hat{\lambda}$=1.99
				$\hat{\theta}$=0.21	$\hat{r}$=0.09	$\hat{\theta}$=5.89	$\hat{r}$=0.67
							$\hat{\theta}$=0.13
(χ^2,df)		(94.39, 5)	(17.09, 8)	(9.12, 8)	(12.25, 8)	(9.10, 6)	**(4.01, 5)**
$p-$value		<0.0001	0.0292	0.3317	0.1402	0.1677	**0.5478**
−LL		325.11	242.20	237.96	239.45	238.66	**236.18**
AIC		652.22	486.40	479.92	482.90	481.32	**478.36**
BIC		654.62	488.80	484.73	487.71	486.13	**485.58**

model. Moreover, as depicted in Figure 5 and Figure 6, it is clear that the expected frequencies generated by the proposed distribution offer the most accurate approximation to the observed frequencies for both datasets.

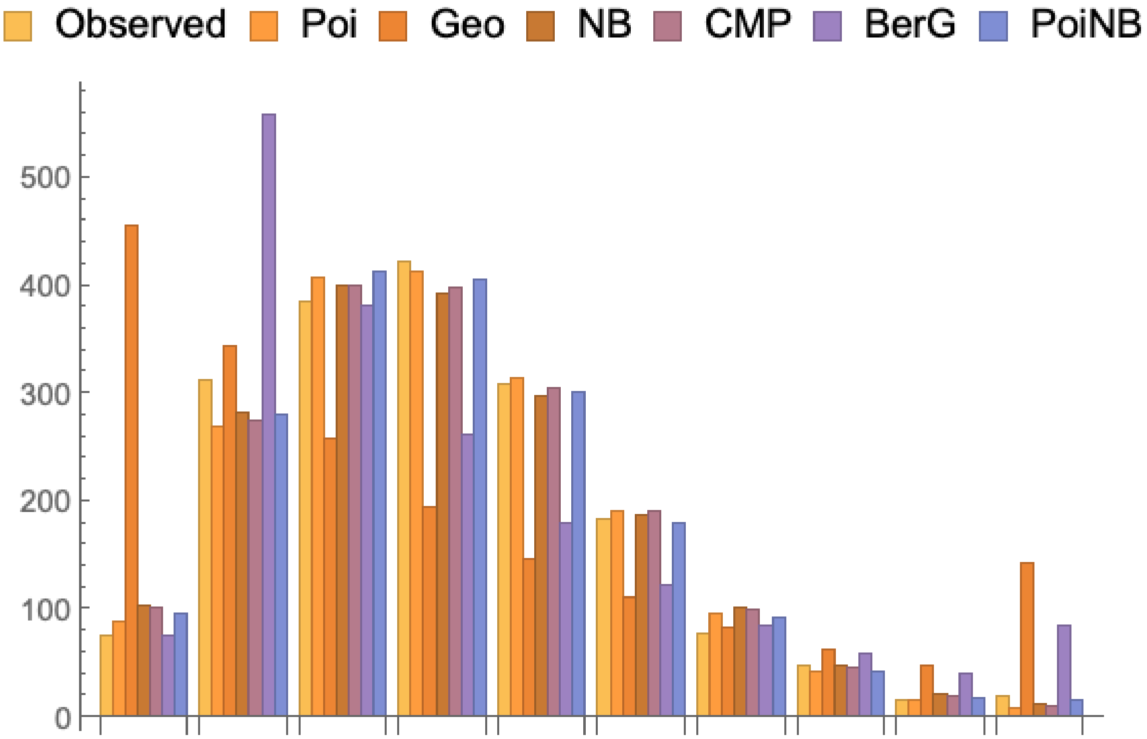

Figure 5: Observed and fitted distributions for Dataset I

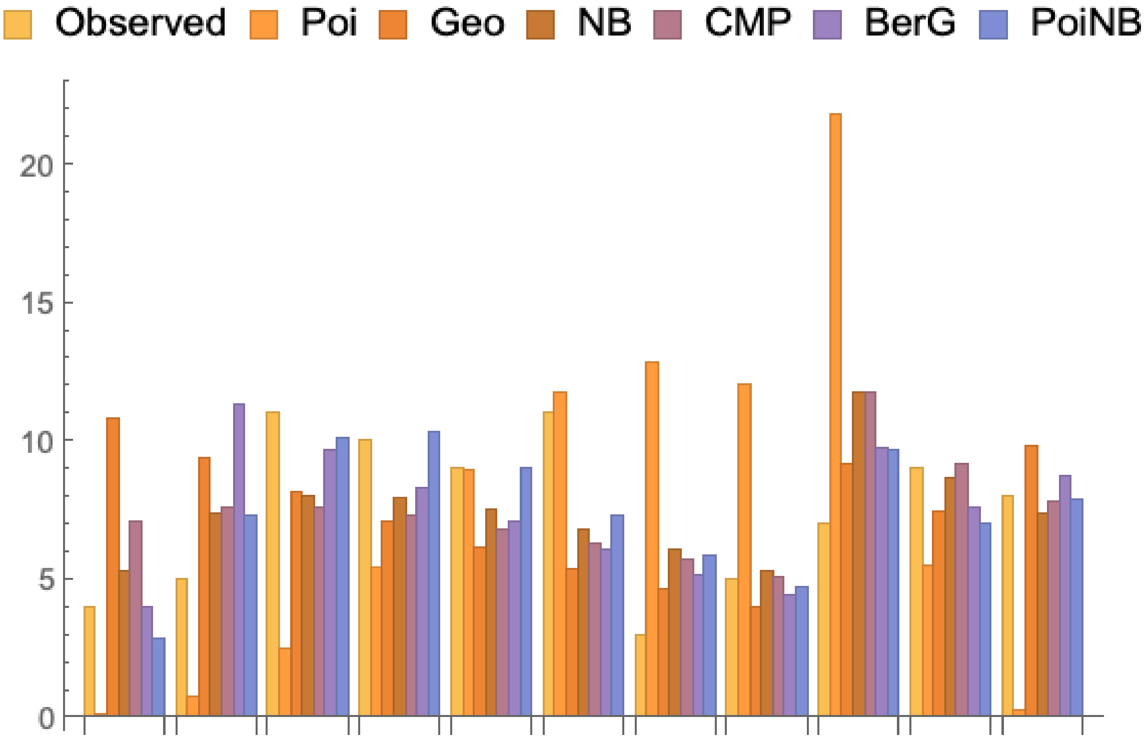

Figure 6: Observed and fitted distributions for Dataset II

8. Discussion

In the current work, we have introduced the PoiNB distribution, thoroughly studied and fitted it to real life datasets. The simple structural properties of the proposed distribution should prove to be useful for the practitioners. From the application point of view, the proposed model is easy to use for modelling over-dispersed data. Despite the availability of several other over-dispersed count models, the proposed model may find wide applications due to the interpretability of its parameters. The maximum likelihood estimation method is utilized to estimate the unknown parameters. Results of the two real- life datasets show that the PoiNB distribution exhibits a better fit compared to the popular count models such as the Poisson, the geometric, the negative binomial, the COM-Poisson and the BerG model. Moreover, we have introduced the PoiNB regression model through the generalized linear model approach. Further investigation into the application of this regression model is necessary. Additionally, attention should be directed towards exploring the usage and suitability of this versatile new distribution for count time series analysis.

References

E. Altun. A new generalization of geometric distribution with properties and applications. *Communications in Statistics-Simulation and Computation*, 49(3):793–807, 2020.

S. K. Bar-Lev and Ad Ridder. Exponential dispersion models for overdispersed zero-inflated count data. *Communications in Statistics-Simulation and Computation*, pages 1–19, 2021.

G. E. Bardwell and E. L. Crow. A two parameter family of hyper-Poisson distributions. *Journal of the American Statistical Association*, 59:133–141, 1964.

M. Bourguignon and C. H. Weiß. An INAR (1) process for modeling count time series with equidispersion, underdispersion and overdispersion. *Test*, 26(4):847–868, 2017.

M. Bourguignon, D. I. Gallardo, and R. M. R. Medeiros. A simple and useful regression model for underdispersed count data based on Bernoulli–Poisson convolution. *Statistical Papers*, 63(3):821–848, 2022.

N. L. Campbell, L. J. Young, and G. A. Capuano. Analyzing over-dispersed count data in two-way cross-classification problems using generalized linear models. *Journal of Statistical Computation and Simulation*, 63(3):263–281, 1999.

S. Chakraborty. On some distributional properties of the family of weighted generalized poisson distribution. *Communications in Statistics—Theory and Methods*, 39(15):2767–2788, 2010.

S. Chakraborty and D. Bhati. Transmuted geometric distribution with applications in modeling and regression analysis of count data. *SORT-Statistics and Operations Research Transactions*, pages 153–176, 2016.

S. Chakraborty and R. D. Gupta. Exponentiated geometric distribution: another generalization of geometric distribution. *Communications in Statistics-Theory and Methods*, 44(6):1143–1157, 2015.

S. Chakraborty and S. H. Ong. Mittag-leffler function distribution-a new generalization of hyper-Poisson distribution. *Journal of Statistical distributions and applications*, 4(1):1–17, 2017.

J. Del Castillo and M. Pérez-Casany. Weighted poisson distributions for overdispersion and underdispersion situations. *Annals of the Institute of Statistical Mathematics*, 50(3):567–585, 1998.

B. Efron. Double exponential-families and their use in generalized linear-regression. *Journal of the American Statistical Association*, 81:709–721, 1986.

P. Fisher. Negative binomial distribution. *Annals of Eugenics*, 11:182–787, 1941.

R. A. Fisher, A. S. Corbet, and C. B. Williams. The relation between the number of species and the number of individuals in a random sample of an animal population. *The Journal of Animal Ecology*, pages 42–58, 1943.

E. Gómez-Déniz. Another generalization of the geometric distribution. *Test*, 19(2):399–415, 2010. doi: 10.1007/s11749-009-0169-3.

F. Hassanzadeh and I. Kazemi. Analysis of over-dispersed count data with extra zeros using the Poisson log-skew-normal distribution. *Journal of Statistical Computation and Simulation*, 86(13):2644–2662, 2016.

P. G. Hoel. On indices of dispersion. *The Annals of Mathematical Statistics*, 14(2):155–162, 1943.

G. C. Jain and P. C. Consul. A generalized negative binomial distribution. *SIAM Journal on Applied Mathematics*, 21(4):501–513, 1971. doi: 10.1137/0121056.

T. Kim, D. S. Kim, J. Kwon, H. Lee, and S. H. Park. Some properties of degenerate complete and partial bell polynomials. *Advances in Difference Equations*, 2021(1):1–12, 2021.

J. Makcutek. A generalization of the geometric distribution and its application in quantitative linguistics. *Romanian Reports in Physics*, 60(3):501–509, 2008.

A. Moghimbeigi, M. R. Eshraghian, K. Mohammad, and B. Mcardle. Multilevel zero-inflated negative binomial regression modeling for over-dispersed count data with extra zeros. *Journal of Applied Statistics*, 35(10):1193–1202, 2008.

M. Moqaddasi Amiri, L. Tapak, and J. Faradmal. A mixed-effects least square support vector regression model for three-level count data. *Journal of Statistical Computation and Simulation*, 89(15):2801–2812, 2019.

A Nandi, S Chakraborty, and A Biswas. A new over-dispersed count model based on poisson-geometric convolution. *Communications in Statistics - Simulation and Computation, doi - 10.1080/03610918.2024.2329997*, 2024.

V. Nekoukhou, M. H. Alamatsaz, and H. Bidram. A discrete analogue of the generalized exponential distribution. *Communications in Statistics - Theory and Methods*, 41(11):2000–2013, 2012. doi: 10.1080/03610926.2011.555044.

H. v. Ophem. Modeling selectivity in count-data models. *Journal of Business & Economic Statistics*, 18(4): 503–511, 2000.

A. N. Philippou, C. Georghiou, and G. N. Philippou. A generalized geometric distribution and some of its properties. *Statistics and Probability Letters*, 1(4):171–175, 1983. doi: 10.1016/0167-7152(83)90025-1.

M. Rodrigues-Motta, H. P. Pinheiro, E. G. Martins, M. S. Araújo, and S. F. dos Reis. Multivariate models for correlated count data. *Journal of Applied Statistics*, 40(7):1586–1596, 2013.

F. Sarvi, A. Moghimbeigi, and H. Mahjub. GEE-based zero-inflated generalized Poisson model for clustered over or under-dispersed count data. *Journal of Statistical Computation and Simulation*, 89(14):2711–2732, 2019.

K. F. Sellers and G. Shmueli. A flexible regression model for count data. *The Annals of Applied Statistics*, pages 943–961, 2010.

L. Tapak, O. Hamidi, P. Amini, and G. Verbeke. Random effect exponentiated-exponential geometric model for clustered/longitudinal zero-inflated count data. *Journal of Applied Statistics*, 47(12):2272–2288, 2020.

R. C. Tripathi, R. C. Gupta, and T. J. White. Some generalizations of the geometric distribution. *Sankhya Ser. B*, 49(3):218–223, 1987.

F. Tüzen, S. Erbaş, and H. Olmuş. A simulation study for count data models under varying degrees of outliers and zeros. *Communications in Statistics-Simulation and Computation*, 49(4):1078–1088, 2020.

S. Wang, N.G. Cadigan, and H.P. Benoît. Inference about regression parameters using highly stratified survey count data with over-dispersion and repeated measurements. *Journal of Applied Statistics*, 44(6): 1013–1030, 2017.

Y. Wang, L. J. Young, and D. E. Johnson. A UMPU test for comparing means of two negative binomial distributions. *Communications in Statistics-Simulation and Computation*, 30(4):1053–1075, 2001.

W. Wongrin and W. Bodhisuwan. Generalized Poisson–Lindley linear model for count data. *Journal of Applied Statistics*, 44(15):2659–2671, 2017.

A. Zarzo, J. S. Dehesa, and R. J. Yañez. Distribution of zeros of gauss and kummer hypergeometric functions: A semiclassical approach. *Annals Numer. Math*, 2:457–472, 1995.