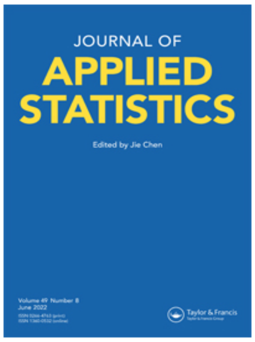

Journal of Applied Statistics

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/cjas20

An one-parameter compounding discrete distribution

Emrah Altun, Gauss M. Cordeiro & Miroslav M. Ristić

To cite this article: Emrah Altun, Gauss M. Cordeiro & Miroslav M. Ristić (2022) An one-parameter compounding discrete distribution, Journal of Applied Statistics, 49:8, 1935-1956, DOI: 10.1080/02664763.2021.1884846

To link to this article: https://doi.org/10.1080/02664763.2021.1884846

Published online: 09 Feb 2021.

JOURNAL OF APPLIED STATISTICS
2022, VOL. 49, NO. 8, 1935–1956
https://doi.org/10.1080/02664763.2021.1884846

An one-parameter compounding discrete distribution

Emrah Altun [a], Gauss M. Cordeiro [b] and Miroslav M. Ristić [c]

[a]Department of Mathematics, Bartin University, Bartin, Turkey; [b]Department of Statistics, Federal University of Pernambuco, Recife, Brazil; [c]Faculty of Sciences and Mathematics, University of Niš, Niš, Serbia

ABSTRACT

In this study, a new one-parameter discrete distribution obtained by compounding the Poisson and xgamma distributions is proposed. Some statistical properties of the new distribution are obtained including moments and probability and moment generating functions. Two methods are used for the estimation of the unknown parameter: the maximum likelihood method and the method of moments. Additionally, the count regression model and integer-valued autoregressive process of the proposed distribution are introduced. Some possible applications of the introduced models are considered and discussed.

ARTICLE HISTORY
Received 1 July 2019
Accepted 30 January 2021

KEYWORDS
Discrete distribution; compounding; Poisson distribution; xgamma distribution

MATHEMATICS SUBJECT CLASSIFICATION
62E15

1. Introduction

The Poisson distribution is the most common choice for modeling count data sets. The weakness of the Poisson distribution is that it cannot model overdispersed data sets. When the variance is higher than the mean, overdispersion occurs. To provide alternative models for overdispersed count data sets, some researchers have introduced mixed-Poisson distributions such as Shoukri *et al.* [31], Shmueli *et al.* [30], Rodríguez-Avi *et al.* [24], Mahmoudi and Zakerzadeh [21], Lord and Geedipally [19], Déniz [11], Cheng *et al.* [10], Sáez-Castillo and Conde-Sánchez [25], Zamani *et al.* [34], Gencturk and Yigiter [13], Bhati *et al.* [7], Imoto *et al.* [15], Wongrin and Bodhisuwan [33] and Altun [3,4] among others. In recent years, some researchers have shown great interest in modeling of an integer-valued autoregressive process. Time series of counts come into view in some scientific fields such as medical, sport and actuarial sciences. The monthly deaths from lung cancer, the monthly number of insurance policy and a yearly number of injured sportsman can be given as examples. McKenzie [22,23] and Al-Osh and Alzaid [1] introduced a stochastic model for the integer-valued time series data sets, known as first-order non-negative integer-valued autoregressive with Poisson innovations, shortly INARP(1). As widely documented, time series of counts display overdispersion. In this case, Poisson distribution cannot be useful any longer for INAR(1) process. Researchers have proposed different INAR(1) processes with flexible innovation distributions to overcome the overdispersion problem. Some of the important researches on overdispersed INAR(1) process can be cited as follows: INAR(1)

CONTACT Emrah Altun emrahaltun123@gmail.com Department of Mathematics, Bartin University, Bartin 74100, Turkey

process with geometric innovations by Jazi et al. [16], INAR(1) process with Poisson–Bilal innovations by Altun [5], INAR(1) process with three-parameter discrete Lindley innovations by Eliwa *et al.* [12], INAR(1) with Poisson–Lindley innovations by Lívio *et al.* [18], INAR(1) process with Katz family innovations by Kim and Lee [17], INAR(1) process with generalized Poisson and double Poisson innovations by Bourguignon *et al.* [9], INAR(1) process with geometric marginals by Borges *et al.* [8] and INAR(1) process with Skellam innovations by Andersson and Karlis [6]. Besides these important researches, Altun [4] introduced a new two-parameter discrete distribution and investigated its performance in INAR(1) process based on the binomial thinning.

We define a new discrete distribution to provide an opportunity in modeling the overdispersed count data sets. For that purpose, a new mixed-Poisson distribution, called Poisson–xgamma(PX) distribution is introduced. The important statistical properties of the PX distribution are derived. The advantages of the proposed distribution are that its probability mass and cumulative distribution functions have simple form and probability and moment generating functions have explicit expressions. The novelty of the presented study can be summarized as follows: (i) a new one-parameter discrete distribution is introduced with its statistical properties in explicit forms; (ii) the proposed distribution is applied as distribution of the innovations of the INAR(1) process to model over-dispersed time series of counts; (iii) a new regression model for discrete response variable is introduced as an alternative to the negative-binomial (NB) regression model.

It is possible to obtain several mixed-Poisson distributions by compounding the Poisson distribution with other distributions. However, it is a key point to use an appropriate distribution to introduce a flexible distribution in its shape and also a simple distribution in its statistical properties. The reason for the use of xgamma distribution as a compounding distribution of the Poisson is the simple form of the xgamma distribution which is important to derive the statistical properties of the PX distribution and estimate the unknown model parameter of the PX distribution. The new distribution has simple mathematical forms for its probability mass function (pmf) and cumulative distribution function (cdf) and can be used to model overdispersed count data sets which are widely seen in real-life data modeling. The negative-binomial distribution is the first choice by researchers to model the overdispersed count data sets because of its software support. However, when the data sets exhibit larger skewness and kurtosis than the NB distribution could model, we need more flexible count distributions. The PX distribution exhibits better modeling ability than negative binomial distribution with less parameter and complexity. Additionally, the computational codes of the developed models are available in Section 5 to reproduce the results given in this study and make the proposed models accessible and applicable by the researchers and practitioners studying in this field.

The rest of the paper is organized as follows. In Section 2, a new one-parameter discrete distribution is introduced and some of its statistical properties are investigated. In Section 3, we estimate the model parameter via maximum likelihood and method of moments. In Section 4, the finite sample performances of the maximum likelihood and method of moments estimation methods are compared for the PX distribution via a simulation study. Section 5 deals with a regression model and three real data sets are analyzed in Section 5 to prove empirically the usefulness of the proposed models against some existing models. Section 6 offers some concluding remarks.

2. The Poisson–xgamma distribution

We now introduce a new discrete distribution by compounding the Poisson and xgamma distributions. Let the random variable X follows the well-known Poisson distribution with a mean parameter $\lambda > 0$. The probability mass function (pmf) of X is

$$P(x; \lambda) = \frac{\lambda^x e^{-\lambda}}{x!}, \quad x = 0, 1, 2, \ldots$$

The mean and variance of the Poisson distribution are $E(X) = \lambda$ and $\mathrm{Var}(X) = \lambda$, respectively. So, its dispersion index is $DI(X) = \mathrm{Var}(X)/E(X) = 1$. So, the Poisson distribution does not provide any opportunity to model the overdispersion, especially observed in real-life data sets. The xgamma distribution was introduced by Sen et al. [28] following the idea of the Lindley distribution. The probability density function (pdf) of the xgamma distribution takes the form

$$f(x; \theta) = \frac{\theta^2}{1+\theta}\left(1 + \frac{\theta}{2}x^2\right) e^{-\theta x}, \quad x > 0, \tag{1}$$

where $\theta > 0$ is the shape parameter. As seen from (1), the xgamma distribution is a mixture of two distributions: the exponential distribution with the parameter θ and the gamma distribution Gamma$(3, \theta)$. Its cumulative distribution function (cdf) is

$$F(x) = 1 - \frac{1 + \theta + \theta x + \frac{\theta^2 x^2}{2}}{\theta + 1} e^{-\theta x}, \quad x \geq 0.$$

Now, we introduce the PX distribution by compounding the Poisson with xgamma distribution.

Proposition 1: *Let a random variable X (for given $\lambda > 0$) have the Poisson distribution with parameter λ. We assume that the parameter λ is a random variable having the xgamma distribution with parameter $\theta > 0$. Then, the unconditional distribution of X has the form*

$$P(X = x; \theta) = \frac{\theta^2[2(1+\theta)^2 + \theta(x+2)(x+1)]}{2(1+\theta)^{x+4}}, \quad x = 0, 1, 2, \ldots, \tag{2}$$

Proof: The random variable X has the Poisson distribution for a given fixed parameter λ. On the other hand, if the parameter λ follows the xgamma distribution with parameter $\theta > 0$, the unconditional probability of the random variable X is given by

$$\begin{aligned} P(X = x; \theta) &= \int_0^\infty P(X = x|\lambda) f(\lambda; \theta) d\lambda \\ &= \frac{\theta^2}{x!(1+\theta)} \int_0^\infty \left(\lambda^x + \frac{\theta}{2}\lambda^{x+2}\right) e^{-\lambda(1+\theta)} d\lambda \\ &= \frac{\theta^2[2(1+\theta)^2 + \theta(x+2)(x+1)]}{2(1+\theta)^{x+4}}. \end{aligned} \quad \blacksquare$$

Henceforth, we assume that the random variable X with pmf (2) has the *Poisson–xgamma* (PX for short) distribution with parameter $\theta > 0$, say $X \sim \mathrm{PX}(\theta)$. The cdf of X is

$$F(x;\theta) = P(X \le x) = 1 - \frac{x^2\theta^2 + 5x\theta^2 + 2x\theta + 2\theta^3 + 10\theta^2 + 8\theta + 2}{2(\theta+1)^{x+4}}.$$

The pmf of the PX distribution has interesting shapes which are given by the following theorem.

Proposition 2: *The pmf of the PX distribution with parameter $\theta > 0$ has the following shapes:*

(i) *If $4 - 4\theta - 15\theta^2 - 8\theta^3 < 0$, then the pmf is a decreasing function for all $x \ge 0$;*
(ii) *If $4 - 4\theta - 15\theta^2 - 8\theta^3 > 0$ and $1 - 3\theta - \theta^2 < 0$, then the pmf decreases for $x < x_1$ or $x > x_2$ and increases for $x_1 < x < x_2$. This means that a random variable X has two modes at 0 and integer number close to x_2 and the minimum at integer number close to x_1;*
(iii) *If $4 - 4\theta - 15\theta^2 - 8\theta^3 > 0$ and $1 - 3\theta - \theta^2 > 0$, then the pmf is a unimodal function with the mode at integer close to x_2. The numbers $x_1 < x_2$ are the solutions of the equation*

$$-\theta x^2 + (2 - 3\theta)x + 2(1 - 3\theta - \theta^2) = 0.$$

Proof: For deriving the shapes of the pmf, we consider

$$\frac{P(X = x+1;\theta)}{P(X = x;\theta)} - 1 = \frac{\theta[-\theta x^2 + (2 - 3\theta)x + 2(1 - 3\theta - \theta^2)]}{(1+\theta)[2(1+\theta)^2 + \theta(x+2)(x+1)]}.$$

So, the behavior depends on the sign of the function $\psi(x) = -\theta x^2 + (2 - 3\theta)x + 2(1 - 3\theta - \theta^2)$. If $4 - 4\theta - 15\theta^2 - 8\theta^3 < 0$, then equation $\psi(x) = 0$ does not have real roots, so we obtain that it is negative for all values $x \ge 0$. This implies that $P(X = x+1;\theta) < P(X = x;\theta)$, which means that the pmf is a decreasing function for all $x \ge 0$. If $4 - 4\theta - 15\theta^2 - 8\theta^3 > 0$, then equation $\psi(x) = 0$ has two roots, say $x_1 < x_2$. Now, if $1 - 3\theta - \theta^2 < 0$, then both roots x_1 and x_2 are positive. Thus we obtain that $P(X = x+1;\theta) < P(X = x;\theta)$ for $x < x_1$ or $x > x_2$, and that $P(X = x+1;\theta) > P(X = x;\theta)$ for $x_1 < x < x_2$. In the third case, we have that only one root x_2 is positive which implies that the pmf is a unimodal function. ■

Remark 1: From the previous proposition, we obtain that the pmf is a decreasing function for $\theta > 0.3733$, a unimodal function for $\theta \in (0, 0.3027)$, and a decreasing-increasing-decreasing function for $\theta \in (0.3027, 0.3733)$.

Some possible shapes of the PX distribution are displayed in Figure 1, which indicate that this distribution can be a good choice for modeling extremely right skewed and symmetric data sets.

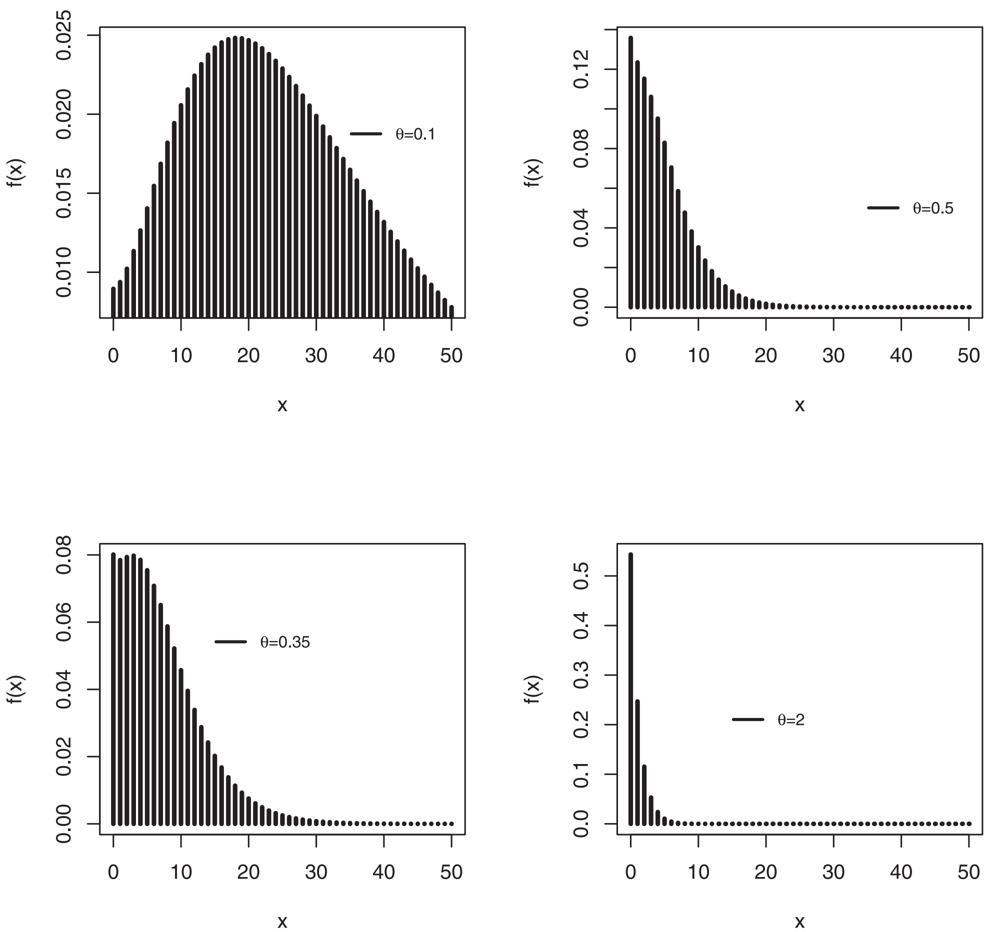

Figure 1. The pmf plots of the PX distribution.

Proposition 3: *Let a random variable X follows the PX distribution with parameter* $\theta > 0$. *Then, the probability generating function (pgf) of X is*

$$G(s;\theta) = \theta^2(1+\theta)^{-1}\left[\frac{1}{1+\theta-s} + \frac{\theta}{(1+\theta-s)^3}\right]. \tag{3}$$

Proof: First, note that the xgamma distribution has the moment generating function (mgf) given by $\Phi(t) \equiv E(e^{tX}) = \frac{\theta^2[\theta+(\theta-t)^2]}{(1+\theta)(\theta-t)^3}$. Since a random variable X for given λ has the Poisson distribution with this parameter, the probability generating function of X reads as

$$\begin{aligned} G(s;\theta) &= E[E(s^X|\lambda)] = E(e^{\lambda(s-1)}) = \Phi(s-1) \\ &= \theta^2(1+\theta)^{-1}\left[\frac{1}{1+\theta-s} + \frac{\theta}{(1+\theta-s)^3}\right]. \end{aligned}$$

■

Remark 2: Based on the previous proposition, we obtain an interesting conclusion. The pgf (3) can be rewritten as

$$G(s;\theta) = \frac{\theta}{1+\theta}\frac{\theta}{1+\theta-s} + \frac{1}{1+\theta}\left(\frac{\theta}{1+\theta-s}\right)^3.$$

This means that the PX distribution can be represented as a mixture of the geometric distribution with mean $1/\theta$ and the negative binomial distribution with mean $3/\theta$ and variance $3(1+\theta)/\theta^2$. We will use later this remark to simulate observations from the PX distribution.

The mgf of X follows from the previous proposition by setting $s = \mathrm{e}^t$

$$M(t;\theta) = \theta^2(1+\theta)^{-1}\left[\frac{1}{1+\theta-\mathrm{e}^t} + \frac{\theta}{\left(1+\theta-\mathrm{e}^t\right)^3}\right]. \tag{4}$$

Now, we consider the hazard failure rate (hrf) $r(x)$ of X defined as $r(x) = \frac{P(X=x;\theta)}{1-F(x)}$. The following proposition gives the expression of the hrf for the PX distribution.

Proposition 4: *If a random variable X has the PX distribution with a parameter $\theta > 0$, then its hrf is given by*

$$r(x;\theta) = \frac{2\theta^2(1+\theta)^2 + \theta^3(x+2)(x+1)}{\theta^2x^2 + (5\theta^2+2\theta)x + 2\theta^3 + 10\theta^2 + 8\theta + 2}, \quad x = 0, 1, 2, \ldots$$

It is an increasing function for $\theta \in \left(0, \frac{1+\sqrt{5}}{2}\right)$ and a decreasing-increasing function for $\theta > \frac{1+\sqrt{5}}{2}$. For both cases, $r(x) \to \theta$ when $x \to \infty$.

Proof: The expression for the hrf $r(x)$ just follows from its definition. Let us consider now its behavior. In this sense, we derive $r(x+1) - r(x) = N(x)/D(x)$, where $N(x) = 2\theta^3(1+\theta)[\theta x^2 + (5\theta+2)x - 2\theta^2 + 2\theta + 2]$ and $D(x) = [\theta^2x^2 + (5\theta^2+2\theta)x + 2\theta^3 + 10\theta^2 + 8\theta + 2][\theta^2x^2 + (7\theta^2+2\theta)x + 2\theta^3 + 16\theta^2 + 10\theta + 2]$.

Thus the behavior depends on the nature of the function $h(x) = \theta x^2 + (5\theta+2)x - 2\theta^2 + 2\theta + 2$. If $\theta < \frac{1+\sqrt{5}}{2}$, then the function h has two negative roots, which means that it is positive for all $x \geq 0$. Thus, $r(x+1) > r(x)$ for all $x \geq 0$, so the hazard rate failure function increases for non-negative integers. If $\theta > \frac{1+\sqrt{5}}{2}$, then the function h has only one positive root $x_1 = \frac{-5\theta-2+\sqrt{8\theta^3+17\theta^2+12\theta+4}}{2\theta}$. It is negative for $0 \leq x < x_1$ and positive for $x > x_1$, which implies that the function r is decreasing for $0 \leq x < x_1$ and increasing for $x > x_1$. ■

Remark 3: The previous proposition provides the identifiability of the PX distribution with respect to its parameter θ. In fact, if we assume that $P(X = x;\theta_1) = P(X = x;\theta_2)$ for all $x \geq 0$ and $\theta_1 \neq \theta_2$, we obtain that the hrfs $r(x;\theta_1)$ and $r(x;\theta_2)$ are equal. Letting that $x \to \infty$ in both hrfs, we obtain that θ_1 and θ_2 are equal, which is not possible. Then, the PX distribution is identifiable with respect to its parameter θ.

Proposition 5: *Let a random variable X follows the PX distribution with parameter* $\theta > 0$. *Then, the factorial moments of X are*

$$\mu_{[r]} = \frac{[(r+2)(r+1)+2\theta]r!}{2\,(1+\theta)\,\theta^{r}}.$$

Proof: The proof just follows from the fact that the PX distribution is a mixture of the geometric and negative binomial distributions and their corresponding pgfs. ∎

Using the above result, the first four non-central moments of X can be expressed as

$$E\,(X) = \frac{\theta+3}{\theta\,(\theta+1)}, \quad E\left(X^2\right) = \frac{\theta^2+5\theta+12}{\theta^2\,(\theta+1)},$$
$$E\left(X^3\right) = \frac{\theta^3+9\theta^2+42\theta+60}{\theta^3\,(\theta+1)}, \quad E\left(X^4\right) = \frac{(\theta+6)\left(\theta^3+11\theta^2+54\theta+60\right)}{\theta^4\,(\theta+1.0)}.$$

Then, the variance of X is

$$\mathrm{Var}\,(X) = \frac{\theta^3+5\theta^2+11\theta+3}{\theta^2(1+\theta)^2}$$

and the dispersion index follows as

$$DI(X) = 1 + \frac{\theta^2+8\theta+3}{\theta(1+\theta)(3+\theta)}. \tag{5}$$

Equation (5) is always greater than 1 since $\theta > 0$. So, the PX distribution can be considered to model overdispersed count data sets. Using the non-central moments of the PX distribution, the skewness and kurtosis of X are obtained and given as follows:

$$S = \frac{\theta^5+8\theta^4+36\theta^3+66\theta^2+27\theta+6}{\sqrt{(\theta^3+5\theta^2+11\theta+3)^3}},$$
$$K = \frac{\theta^7+16\theta^6+124\theta^5+489\theta^4+881\theta^3+714\theta^2+306\theta+45}{(\theta^3+5\theta^2+11\theta+3)^2},$$

Figure 2 provides the plots of the mean, variance, skewness and kurtosis of the PX distribution. We note that the mean and variance decrease and the skewness and kurtosis increase when the parameter θ increases.

Further, we obtain an approximation for the density of the sample average $\overline{X} = \sum_{i=1}^{n} X_i/n$ of n independent and identically distributed (iid) random variables $X_1, \ldots, X_n$ having pmf (1) (see Tahir *et al.*, [32]). We adopt the notation $K^{(j)}(t;\theta) = \partial^j \log[M(t;\theta)/\partial t^j]$ (for $j \geq 0$) for the derivatives of the cumulant generating function (cgf) determined from (4). Clearly, $K(t;\theta) = K^{(0)}(t;\theta)$ is the cgf of X.

The density function of $\overline{X}$ can be written using the Fourier inversion integral as

$$f_{\overline{X}}(x;\theta,\tau) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \exp\Big[-\mathrm{i}\,t\,x + n\,K(\mathrm{i}\,t/n;\theta)\Big],$$

where $\mathrm{i} = \sqrt{-1}$. This equation is suitable for Daniels' saddle-point approximation. Setting $z = \mathrm{i}t/n$, the saddle-point of $K(z;\theta) - z\,x$ is $K'(\hat{z};\theta) = x$, which has an explicit solution for $\hat{z} = \hat{z}(x)$ easily found in **Mathematica** or **Maple**.

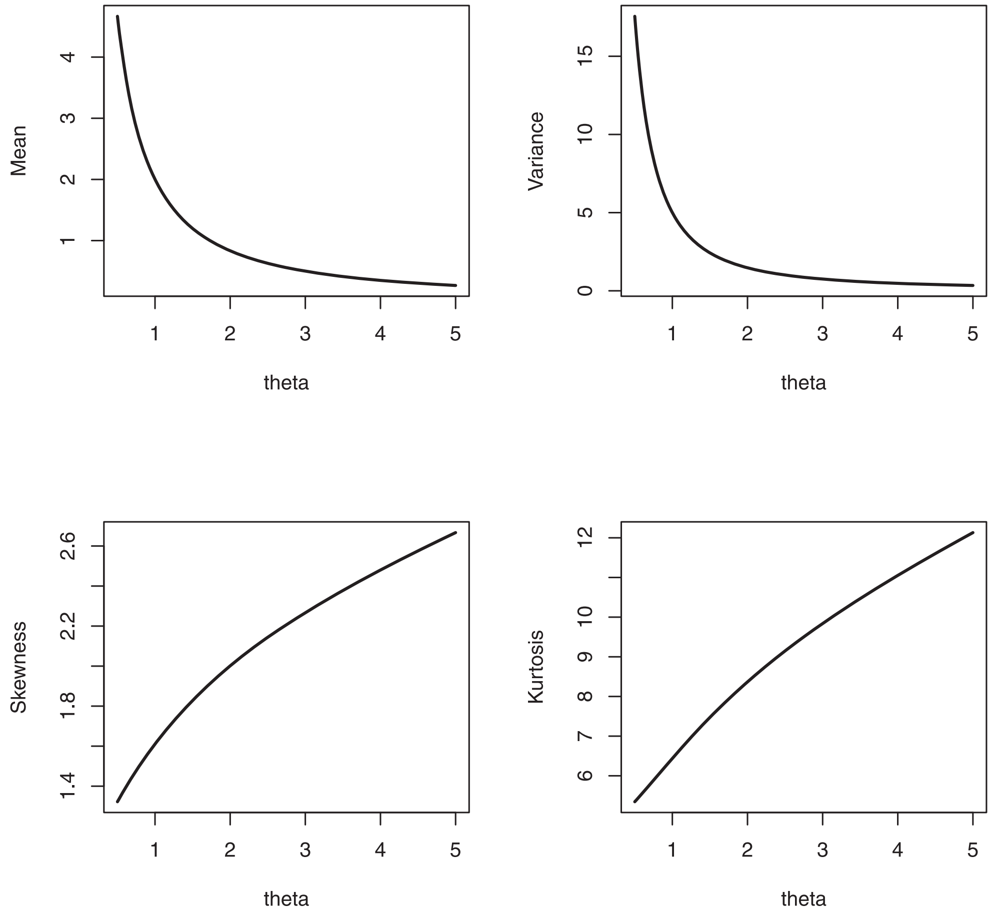

Figure 2. The plots of the statistical measures of the PX distribution.

Then, the approximate density of $\overline{X}$ has the form

$$f_{\overline{X}}(x;\theta) \simeq \left[\frac{n}{2\pi\, K^{(2)}(\hat{z};\theta)}\right]^{1/2} \exp\left\{n\left[K(\hat{z};\theta) - \hat{z}\,x\right]\right\}.$$

The approximation for $f_{\overline{X}_n}(x;\theta)$ provides a good approximation in practice.

It is much more frequent in statistical applications to compute probabilities associated to $\overline{X}$. By integrating the last equation, we can write the cdf of $\overline{X}$ as

$$F_{\overline{X}}(x;\theta) \simeq \int_0^x \left[\frac{n}{2\pi\, K^{(2)}(w;\theta)}\right]^{1/2} \exp\left\{n\left[K(w;\theta) - t\,w\right]\right\} \mathrm{d}w,$$

where $t = t(w)$ is determined such that $K'(t;\theta) = w$.

By transforming variables $K^{(2)}(t;\theta)\,dt = dw$ in order to integrating with respect to the saddle-point variable t instead of w, we obtain

$$F_{\overline{X}}(x;\theta) \simeq \int_0^{t(x)} \left[\frac{nK^{(2)}(t;\theta)}{2\pi}\right]^{1/2} \exp\left\{n\left[K(t;\theta) - t\,K'(t;\theta)\right]\right\} \mathrm{d}t,$$

where $t = t(x)$ is found by solving $K'(t;\theta) = x$. This integral for $F_{\overline{X}}(x;\theta)$ is much easier to compute than the previous one since it includes explicitly the saddle-point function in the integrand. Based on this integral, Lugannani and Rice [20] derived the saddle-point approximation for the cdf of $\overline{X}$ as

$$F_{\overline{X}}(x;\theta) \simeq \Phi\left[r(x)\right] + \phi\left[r(x)\right]\left[\frac{1}{r(x)} - \frac{1}{u(x)}\right], \tag{6}$$

where $\Phi(\cdot)$ and $\phi(\cdot)$ are the cdf and pdf of the standard normal distribution,

$$r(x) = \mathrm{sgn}[t(x)]\ \{2n\ [x\,t(x) - K_X(t(x);\theta)]\}^{1/2}$$

and

$$u(x) = t(x)\ [n\,K(2)(t(x);\theta)]^{1/2}.$$

Equation (6) provides highly accurate results for the probabilities associated with the sample average of iid PX random variables.

2.1. Comparison of Poisson–xgamma and Poisson–Lindley distributions

The Poisson–Lindley (PL) distribution, proposed by Sankaran [26], is the well-known one-parameter mixed-Poisson distribution. Here, PX and PL distributions are compared with respect to the values of skewness, kurtosis and dispersion index. The required formula for the skewness, kurtosis and dispersion index values of the PL distribution can be found in Ghitany and Al-Mutairi [14].

Figure 3 displays the skewness, kurtosis and dispersion index values of the PX and PL distributions. It is clear that the both of the distributions have similar behaviors in terms of these quantities. However, the PX distribution has wider range of skewness, kurtosis and dispersion index than those of PL distribution. When the underlying data set displays high overdispersion, say $DI > 3.5$, the PX distribution could be more appropriate distribution than the PL distribution.

3. Estimation

In this section, we consider the estimation of the unknown parameter θ by two methods: the maximum likelihood (ML) method and the method of moments (MM).

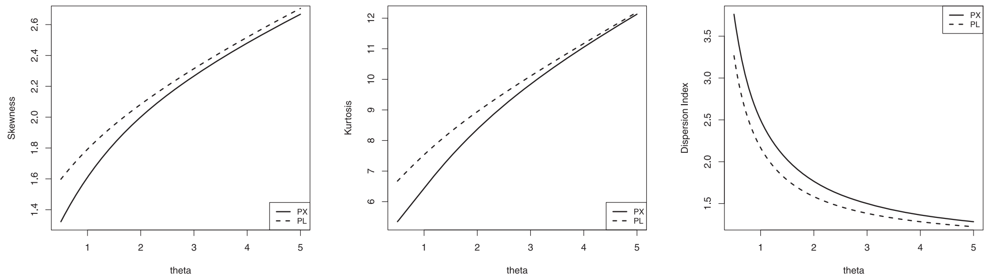

Figure 3. The comparison of the skewness, kurtosis and dispersion index of the PX and PL distributions.

3.1. Maximum likelihood estimation

Let $X_1 \ldots, X_n$ be a random sample from the PX distribution with parameter θ. The log-likelihood function for θ can be expressed as

$$l(\theta) = 2n\log\theta + \sum_{i=1}^{n} \log[2(1+\theta)^2 + \theta(X_i+2)(X_i+1)] - n\log 2 - \log(1+\theta)\left(4n + \sum_{i=1}^{n} X_i\right). \tag{7}$$

By differentiating (7) with respect to θ gives

$$\frac{\partial l(\theta)}{\partial\theta} = \frac{2n}{\theta} + \sum_{i=1}^{n} \frac{4(1+\theta) + (X_i+2)(X_i+1)}{2(1+\theta)^2 + \theta(X_i+2)(X_i+1)} - \frac{4n + \sum_{i=1}^{n} X_i}{1+\theta}. \tag{8}$$

The ML estimate of θ, say $\hat{\theta}_{ML}$, is the solution of the equation $\frac{\partial l}{\partial\theta} = 0$. This equation contains a non-linear function, and because of that the ME estimate does not have closed form. Therefore, this equation requires to be solved using numerical methods in platforms such as **R**, MATLAB or others.

A question that arises here is that ML estimate exists at all. The answer on this question is positive and it can be obtained as follows. First, note that (8) can be rewritten as

$$\frac{\partial l(\theta)}{\partial\theta} = \frac{3n}{\theta} + \frac{2(\theta^2-1)}{\theta} \sum_{i=1}^{n} \frac{1}{2(1+\theta)^2 + \theta(X_i+2)(X_i+1)} - \frac{4n + \sum_{i=1}^{n} X_i}{1+\theta}.$$

Based on the fact that $\theta^2 - 1 < (1+\theta)^2$ and $2(1+\theta)^2[2(1+\theta)^2 + \theta(X_i+2)(X_i+1)]^{-1} < 1$, we have

$$\frac{\partial l(\theta)}{\partial\theta} < \frac{4n - \theta\sum_{i=1}^{n} X_n}{\theta(1+\theta)}.$$

Thus $\frac{\partial l}{\partial\theta} < 0$ when $\theta > \frac{4}{\overline{X}_n}$ and since $\lim_{\theta\to 0} \frac{\partial l}{\partial\theta} = \infty$, it follows that there is at least one ML estimate belonging to the interval $(0, 4\overline{X}_n^{-1})$. The ML estimator of θ, $\hat{\theta}$, is consistent and asymptotically normal with $\sqrt{n}(\hat{\theta}_{MLE} - \theta) \rightarrow N(0, \mathscr{I}(\theta)^{-1})$ where $\mathscr{I}(\theta)$ is

$$\mathscr{I}(\theta) = E\left(-\frac{\partial^2}{\partial\theta^2}\ln f(x;\theta)\right).$$

So the variance estimator of $\hat{\theta}$ can be obtained as $Var(\hat{\theta}) \approx I(\hat{\theta})^{-1}$ where $I(\hat{\theta}) = -\left(\frac{\partial^2\ell}{\partial\theta^2}\right)\Big|_{\theta=\hat{\theta}}$. The second partial derivative of (7) is

$$\frac{\partial^2\ell}{\partial\theta^2} = -\frac{2n}{\theta^2} + \frac{4n + \sum_{i=1}^{n} X_i}{(\theta+1)^2} - \sum_{i=1}^{n} \frac{\left(4\theta + X_i^2 + 3X_i + 6\right)^2}{\left(2(1+\theta)^2 + \theta\,(X_i+2)\,(X_i+1)\right)^2}$$
$$+ \sum_{i=1}^{n} \frac{4}{2(1+\theta)^2 + \theta\,(X_i+2)\,(X_i+1)}.$$

The asymptotic $100(1-p)\%$ confidence intervals (CIs) for the parameter θ is

$$\widehat{\theta} \pm z_{p/2} I\left(\hat{\theta}\right)^{-1/2}$$

where $z_{p/2}$ is the upper $p/2$ quantile of the standard normal distribution.

3.2. *Method of moments*

The second method for estimating the parameter θ is the MM. In this case, the estimate of θ can be obtained from

$$\overline{X}_n = \frac{\theta+3}{\theta\,(\theta+1)}. \tag{9}$$

Solving (9) with respect to θ, we find

$$\hat{\theta}_{MM} = \frac{\sqrt{\overline{X}_n^2 + 10\overline{X}_n + 1} - \overline{X}_n + 1}{2\overline{X}_n}, \quad \overline{X}_n \neq 0.$$

The following theorem shows the behavior of $\hat{\theta}_{MM}$.

Theorem 1: *The estimator $\hat{\theta}_{MM}$ is positively biased.*

Proof: To prove this property, we will follow technique used in Ghitany and Al-Mutairi [14]. We define the function $g(t) = (\sqrt{t^2 + 10t + 1} - t + 1)/(2t)$ for $t > 0$. First, we have $\hat{\theta}_{MM} = g(\overline{X}_n)$ and $g((\theta+3)/(\theta(\theta+1))) = \theta$. Since

$$g''(t) = \frac{5t^3 + 39t^2 + 15t + \sqrt{(t^2+10t+1)^3}}{t^3\sqrt{(t^2+10t+1)^3}} > 0,$$

the function $g(t)$ is strictly convex. Using the Jensen's inequality $E(g(\overline{X}_n)) > g(E(\overline{X}_n))$ and the above results, we obtain that $E(\hat{\theta}_{MM}) > \theta$. ■

4. Simulation studies

Simulation studies are important tools to explore the differences between the several methods based on the pre-determined settings. Here, two simulation studies are presented. First one is to see the asymptotic efficiencies of the estimated methods, presented in Section 3. In the second simulation, we compare the Poisson, NB and PX distributions under different scenarios to explore the differences between these distributions.

4.1. Comparison of MLE and MM methods

This section deals with the finite sample performances of ML and MM estimators of the parameter θ of PX distribution. First, we describe how to generate the random variables from PX(θ). Since the xgamma distribution is the special mixture of exponential and gamma distributions, the below algorithm is used to generate random variables from $PX(\theta)$.

Algorithm 1

(1) Set the parameter θ
(2) Generate u_i from $U\,(0,1)$ distribution
(3) If $U_i \leq \frac{\theta}{\theta+1}$ generate $\lambda_i \sim \text{Exp}\,(\theta)$, otherwise, generate $\lambda_i \sim \text{Gamma}\,(3,\theta)$
(4) Generate $Y_i \sim Poisson\,(\lambda_i)$
(5) Repeat steps 2, 3 and 4 n times

The simulation results are obtained by R software. The simulation study is carried out with $N = 10,000$ replications for $\theta = (0.35, 0.50, 2)$ and $n = (20, 50, 100, 200, 500)$. The following measures are calculated to assess the simulation results:

$$Bias = \sum_{j=1}^{N} \frac{\hat{\theta}_j - \theta}{N}, \quad MRE = \sum_{j=1}^{N} \frac{\hat{\theta}_j \big/ \theta}{N} \quad \text{and} \quad MSE = \sum_{j=1}^{N} \frac{\left(\hat{\theta}_j - \theta\right)^2}{N}.$$

The simulation results are reported in Table 1. We expect to see that estimated biases and MSEs should be near the zero for sufficiently high sample sizes. Also, the estimated MREs should be near the one. When we analyze the results in Table 1, it is seen that the estimated biases and MSEs approach the zero when the sample size increases for the MLE method. The estimated MREs are also near the intended value, one. These results confirm the consistency property of the MLE method. The similar results are also obtained for the MM method. Therefore, we could conclude that MM and ML estimators are equally efficient. The MM and MLE methods work well for estimating the parameter θ.

4.2. Comparison of Poisson, NB and PX models

In this section, we compare the Poisson, NB and PX models via simulation study. We generate the random variables from the contaminated process, given as $\alpha\text{Poisson}(\theta) +$

Table 1. Estimated biases, MSEs and MREs of θ based on the MLE and MM estimation methods.

		Bias		MSE		MRE	
Parameters	Sample size	MLE	MM	MLE	MM	MLE	MM
$\theta = 0.35$	20	1.0119	0.9642	0.3927	0.3909	1.0289	1.0276
	50	0.4261	0.4144	0.1429	0.1438	1.0122	1.0118
	100	0.2010	0.1912	0.0656	0.0657	1.0057	1.0055
	200	0.1257	0.1208	0.0344	0.0345	1.0036	1.0035
	500	0.1063	0.1049	0.0126	0.0128	1.0030	1.0030
$\theta = 0.50$	20	1.4769	1.4365	0.8816	0.8824	1.0295	1.0287
	50	0.4986	0.4781	0.3077	0.3065	1.0100	1.0096
	100	0.1668	0.1599	0.1580	0.1580	1.0033	1.0032
	200	0.1779	0.1765	0.0696	0.0697	1.0036	1.0035
	500	0.0773	0.0761	0.0291	0.0291	1.0015	1.0015
$\theta = 2$	20	12.2859	12.5570	41.3636	41.5663	1.0614	1.0628
	50	5.6770	5.8009	12.9938	13.0658	1.0284	1.0290
	100	4.3709	4.3996	6.3636	6.3806	1.0219	1.0220
	200	1.7794	1.8451	2.7720	2.7842	1.0089	1.0092
	500	0.0938	0.1110	1.0109	1.0107	1.0005	1.0006

* The results of the biases and MSEs are multiplied by 100.

$(1-\alpha)\mathrm{NB}(r,p)$ where $\alpha \in (0,1)$. Two scenarios are studied. These are given below.

$$\text{Scenario I} \rightarrow 0.90\ \text{Poisson}\ (0.5) + (1-0.90)\ \text{NB}\ (2, 0.5)$$

$$\text{Scenario II} \rightarrow 0.80\ \text{Poisson}\ (2) + (1-0.80)\ \text{NB}\ (0.5, 0.5)$$

The contaminated process is used to generate overdispersion in the generated data. For reach generated data, we calculate the root mean square errors (RMSEs) and mean absolute errors (MAEs). The formulate for the RMSEs and MAE are given, respectively, by

$$RMSE = \sqrt{\sum_{i=1}^{n} \frac{(\hat{y}_i - y_i)^2}{n}}$$

$$MAE = \frac{\sum_{i=1}^{n} |\hat{y}_i - y_i|}{n}$$

where $\hat{y}_i$ and y_i are the fitted and observed frequencies, respectively. The simulation replication is determined as $N = 10,000$. The used sample sizes are $n = 30$, 50, 100 and 500. The means of the RMSEs and MAEs for each fitted distributions are reported in Table 2. As expected, when the sample size increases, the estimated RMSEs and MAEs decrease for all distributions. However, the PX distribution has the lowest values of the RMSEs and MAEs for both scenarios and all sample sizes. Therefore, we conclude that the PX distribution provides better results than the Poisson and NB distributions under the used simulation scenario.

5. PX regression model

It is well known fact that the Poisson and negative binomial regression models are the common choices for modeling the discrete dependent variable with covariates. Here, an alternative model to these count regression models is introduced based on the PX distribution. Considering the re-parametrization $\theta = (2\mu)^{-1}(\sqrt{\mu^2 + 10\mu + 1} - \mu + 1)$, the PX

Table 2. Simulation results of the Poisson, NB and PX models.

		Scenario I				Scenario II			
Metrics	Models	$n=30$	$n=50$	$n=100$	$n=500$	$n=30$	$n=50$	$n=100$	$n=500$
RMSE	Poisson	4.9784	4.0784	3.3573	2.5686	5.3242	4.5289	3.8425	3.0295
MAE		4.3115	3.3607	2.6413	1.7857	4.4925	3.6610	2.9881	2.0987
RMSE	NB	2.8591	2.0254	1.5437	0.8975	2.9244	2.2828	1.7378	0.9854
MAE		1.8481	1.4421	1.1301	0.5704	2.1146	1.7110	1.2802	0.6342
RMSE	PX	2.6828	2.0028	1.4849	0.7866	2.8558	2.2642	1.6732	0.8576
MAE		1.8159	1.4193	1.0855	0.4827	2.0474	1.6635	1.2348	0.5409

The results are multiplied by 100.

density can be expressed in terms of the mean $E(Y)=\mu>0$ as

$$P\left(Y=y;\mu\right)=\left[\begin{array}{c}2\left\{(2\mu)^{-1}\left(\sqrt{\mu^2+10\mu+1}-\mu+1\right)\right\}^2\\ \left(1+(2\mu)^{-1}\left(\sqrt{\mu^2+10\mu+1}-\mu+1\right)\right)^2\\ +(2\mu)^{-1}\left(\sqrt{\mu^2+10\mu+1}-\mu+1\right)\left(y+2\right)\left(y+1\right)\end{array}\right]$$
$$\times\left[2\left(1+(2\mu)^{-1}\left(\sqrt{\mu^2+10\mu+1}-\mu+1\right)\right)^{y+4}\right]^{-1},\quad y=0,1,\ldots. \tag{10}$$

The explanatory variables are related to the ith mean by the log-link function, namely

$$\mu_i=E(Y_i)=\exp\left(\boldsymbol{x}_i^T\boldsymbol{\beta}\right),\quad i=1,\ldots,n, \tag{11}$$

where $\boldsymbol{x}_i^T=(x_{i1},\ldots,x_{ik})$ is the vector of explanatory variables and $\boldsymbol{\beta}=(\beta_0,\beta_1,\ldots,\beta_k)^T$ is the unknown vector of regression coefficients. The log-likelihood function expressed in terms of the means of the observations takes the form

$$\ell\left(\boldsymbol{\beta}\right)=\sum_{i=1}^{n}\log\left[\begin{array}{c}2\left\{(2\mu_i)^{-1}\left(\sqrt{\mu_i^2+10\mu_i+1}-\mu_i+1\right)\right\}^2\\ \times\left(1+\left\{(2\mu_i)^{-1}\left(\sqrt{\mu_i^2+10\mu_i+1}-\mu_i+1\right)\right\}\right)^2\\ +\left\{(2\mu)^{-1}\left(\sqrt{\mu_i^2+10\mu_i+1}-\mu_i+1\right)\right\}^3\left(y_i+2\right)\left(y_i+1\right)\end{array}\right]$$
$$\sum_{i=1}^{n}\left(y_i+4\right)\log\left[2\left(1+\left\{(2\mu_i)^{-1}\left(\sqrt{\mu_i^2+10\mu_i+1}-\mu_i+1\right)\right\}\right)\right], \tag{12}$$

where μ_i is a function of $\boldsymbol{\beta}$ thorough (11). The unknown parameter vector, $\boldsymbol{\beta}$, can be determined by maximizing (12) with the **nlm** function of the **R** software. The asymptotic distribution of $(\widehat{\boldsymbol{\beta}}-\boldsymbol{\beta})$ is multivariate normal with zero mean and variance–covariance matrix, $K(\boldsymbol{\beta})^{-1}$. Here, $K(\boldsymbol{\beta})$ is Fisher information matrix. It is possible to replace the Fisher information with observed information matrix whose elements can be numerically calculated by using the **hessian** function of the **R** software. The inverse of the observed information matrix is used to obtain the asymptotic standard errors of $\hat{\boldsymbol{\beta}}$.

Table 3. MLEs, observed and fitted values and χ^2 values.

The number of chromatid aberrations	Observed frequencies	Expected frequencies				
		Poisson	PL	NB	GPL	PX
0	268	231.36	257.02	270.34	269.24	259.42
1	87	126.67	93.39	78.53	78.70	90.36
2	26	34.67	32.76	29.79	30.86	32.49
3	9	6.33	11.21	12.18	12.55	11.61
4	4	0.87	3.77	5.16	5.13	4.06
5	2	0.09	1.25	2.23	2.09	1.38
6	1	0.01	0.41	0.98	0.85	0.46
7	3	0.00	0.13	0.43	0.35	0.15
Total	400	400	400	400	400	400
		Estimates				
	λ	0.547 (0.109)	–	–	1.576 (0.259)	–
	θ	–	2.379 (0.169)	–	0.473 (0.159)	2.803 (0.188)
	r	–	–	0.619 (0.126)	–	–
	p	–	–	0.531 (0.056)	–	–
χ^2		39.146	6.283	2.410	2.940	4.862
df		2	3	2	2	3
p-value		< 0.001	0.098	0.299	0.229	0.182
$\hat{\ell}$		−439.514	−399.857	−403.455	−400.553	−398.041
AIC		881.027	801.714	810.910	805.106	798.081
BIC		881.629	802.316	812.114	813.089	798.683

6. Some illustrative examples

Three real data sets are analyzed to prove empirically the usefulness of the proposed models defined under the PX distribution. The developed computational codes are accessible in https://github.com/emrahaltun/PX-paper-computational-codes.

6.1. Chromatid aberrations

We consider the data set from Shanker and Fesshaye [29] related to the number of chromatid aberrations (0.2 g chinon 1, 24 hours). Shanker and Fesshaye [29] used the Poisson and PL distributions to model the data set. Since the data set displays overdispersion, we believe that the PX distribution could be more appropriate choice than the PL distribution. We compare the performance of the PX distribution on this data set with Poisson, PL, generalized Poisson–Lindley (GPL) and NB distributions. The MLEs and their standard errors (SEs), maximized $\hat{\ell}$, χ^2 test and corresponding p-values, Akaike Information Criteria (AIC) and Bayesian Information Criteria (BIC) are reported in Table 3 for the fitted distributions. The computational results are obtained using the **R** software. The lower values of these criteria indicate the better fitted model to the data.

Table 3 lists the estimated parameters of the fitted distributions and model selection criteria such as AIC, BIC, estimated χ^2 value and its p-value. The standard errors (SEs) of the estimated parameters are given in parentheses. To calculate the χ^2 value, the expected frequencies less than 5 are merged for both observed and expected frequencies. The χ^2 test statistic and corresponding p-value indicate that the PX, PL, GPL and NB distributions provide adequate fit to the current data, except the Poisson distribution. However,

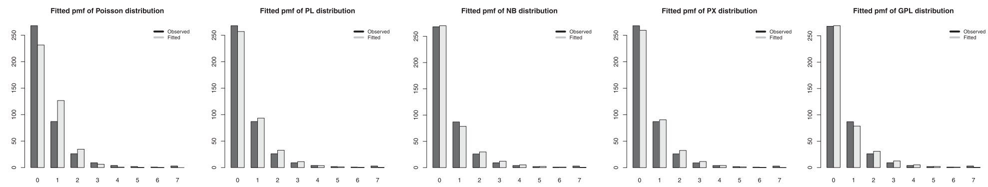

Figure 4. The estimated pmfs of the fitted distributions.

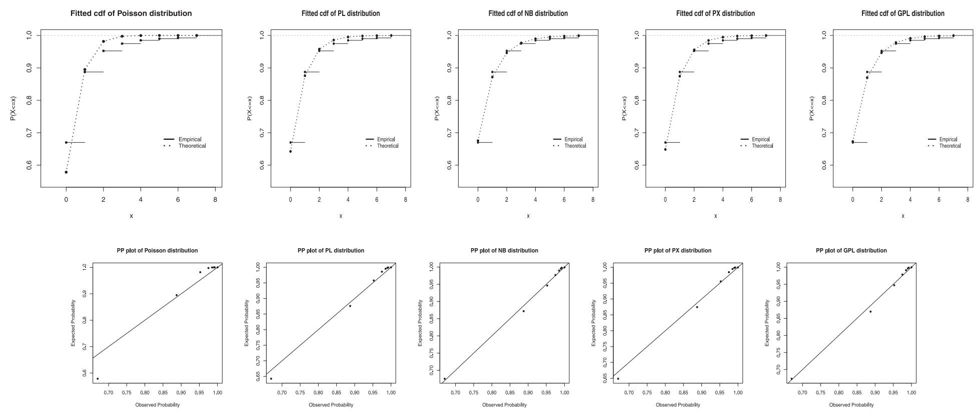

Figure 5. The estimated cdfs and PP plots of the fitted distributions.

the PX distribution has the lowest values of the AIC and BIC statistics. Therefore, the PX distribution provides better fits than Poisson, PL, GPL and NB distributions to these data.

Figures 4 and 5 display the estimated pmfs, cdfs and probability–probability (PP) plots of the fitted distributions. As seen from these figures, the Poisson distribution has under-estimation problem for the zeros in the current data. The representation of the probability of zeros is inadequate in the Poisson distribution. However, the probability of zeros is sufficiently represented in the NB, PL, GPL and PX distributions.

6.2. Length of hospital stay

In this section, we compare the PX regression model with the Poisson, PL, NB and Poisson-transmuted exponential (PTE), introduced by Bhati et al. [7], regression models by means of the AZPRO data. The detail information on the PL regression can be found in Altun [2]. The data set is given in the **COUNT** package of the **R** software. The data come from the 1991 Arizona cardiovascular patient files. Besides, Altun [5] used the same data to compare the different count data models. The length of the hospital stay of patients y_i is modeled by the following covariates: cardiovascular procedure $(x_{1i})(1 = CABG, 0 = PTCA)$, sex $(x_{2i})(1 = male, 0 =$ female), type of admission $(x_{3i})(1 = urgent, 0 =$ elective) and age $x_{4i}(1 = \text{age} > 75, 0 = \text{age} \leq 75)$. The systematic components of the regression models is defined by

$$\mu_i = \exp\left(\beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i}\right).$$

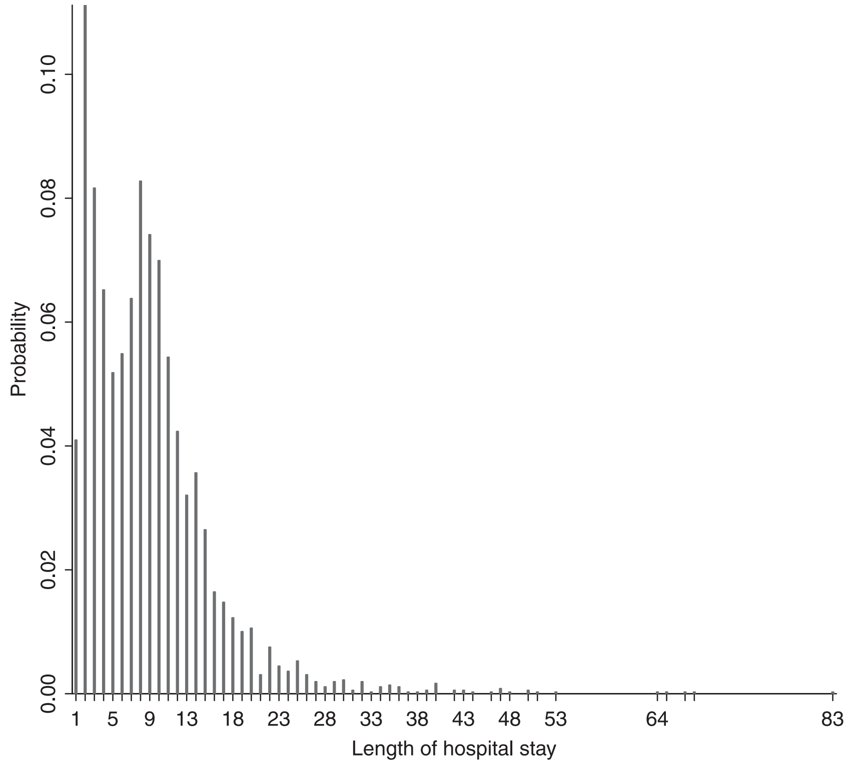

Figure 6. The distribution of length of stay of patients.

Figure 6 displays the distribution of the length of stay data. The mean and variance of the response variable are 8.831 and 47.973, respectively, which provides a clear evidence for overdispersion.

The models are compared based on the minimized negative log-likelihoods, AIC and BIC values. The obtained results such as parameter estimates and information criteria are given in Table 4. These results show that the PX regression model is the best choice for the current data since its AIC and BIC are lower than other models. So, we conclude that PX model provides reasonable results in case of overdispersion.

Using the parameter estimates of the PX model, we obtain the following results. The lengths of stay increase when the individuals receive CABG procedure, have urgency admission, and are older than 75. The average length of stay in hospital is longer for women than for man patients.

6.3. *Weekly number of syphilis cases*

The proposed distribution can be very helpful for modeling the integer-valued time series data. We consider the PX distribution for the distribution of the innovations of a classical integer-valued autoregressive model of the first order (INAR(1)) given by

$$X_t = \alpha \circ X_{t-1} + \varepsilon_t, \quad t \in \mathbb{Z},$$

Table 4. The results of fitted count regression models.

	Poisson		NB		PL		PTE		PX	
Covariates	Estimate	*p*-value	Estimate	*p*-value	Estimate	*p*-value	Estimate	*p*-value	Estimate	*p*-value
β_0	1.4558	< 0.001	1.0780	< 0.001	1.4122	< 0.001	1.4778	< 0.001	1.3996	< 0.001
	(0.0158)		(0.0298)		(0.0372)		(0.0441)		(0.0349)	
β_1	-0.9606	< 0.001	1.0866	< 0.001	0.9844	< 0.001	0.9363	< 0.001	0.9721	< 0.001
	(0.0122)		(0.0243)		(0.0291)		(0.0356)		(0.0270)	
β_2	-0.1240	< 0.001	0.0724	0.003	-0.1265	< 0.001	-0.2150	< 0.001	-0.1269	< 0.001
	(0.0118)		(0.0249)		(0.0304)		(0.0368)		(0.0280)	
β_3	0.3266	< 0.001	0.5319	< 0.001	0.1193	< 0.001	0.0568	0.156	0.1201	< 0.001
	(0.0121)		(0.0249)		(0.0323)		(0.0400)		(0.0298)	
β_4	0.1224	< 0.001	0.3161	< 0.001	0.3837	< 0.001	0.4360	< 0.001	0.3732	< 0.001
	(0.0124)		(0.0273)		(0.0302)		(0.0357)		(0.0280)	
ℓ	-11189.9		-10578.9		-10625.6		-11176.4		-10569.8	
AIC	22389.8		21169.8		21261.18		22364.8		21149.6	
BIC	22420.7		21206.9		21292.11		22401.91		21180.6	

where $0 \leq \alpha < 1$ and $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ represent the innovations which is a sequence of iid integer-valued random variables having the PX distribution with parameter $\theta > 0$. The innovations ε_t are independent of X_{t-k} for all $k \geq 1$ and all counting series incorporated in the binomial thinning $\alpha \circ X_t$. The binomial thinning operator is defined as

$$\alpha \circ X_{t-1} = \sum_{j=1}^{X_{t-1}} W_j,$$

where $\{W_j\}_{j\geq 1}$ is a sequence of iid Bernoulli random variables with probability of success α. For $\alpha \in [0,1)$, the INAR(1) process is stationary (see Al-Osh and Alzaid [1] for more details).

As mentioned before, we consider the PX distribution as the distribution of the innovations $\{\varepsilon_t\}$. Thus let $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ be a sequence of iid random variables having the PX distribution given by (1). The process with these innovations is called as the INARPX(1) process. Some of its properties are obtained as follows. The one-step transition probability of the INARPX(1) model is given by

$$\Pr\left(X_t = k \middle| X_{t-1} = l\right) = \sum_{i=0}^{\min(k,l)} \binom{l}{i} \alpha^i (1-\alpha)^{l-i} \times \frac{\theta^2[2(1+\theta)^2 + \theta\,(k-i+2)\,(k-i+1)]}{2(1+\theta)^{k-i+4}}. \tag{13}$$

The mean and variance of the INARPX(1) process are, respectively, given by

$$\mu_X = \frac{\theta+3}{\theta\,(\theta+1)\,(1-\alpha)},$$
$$\sigma_X^2 = \frac{\alpha\theta\,(\theta+1)\,(\theta+3) + \theta^3 + 5\theta^2 + 11\theta + 3}{\theta^2(1+\theta)^2\left(1-\alpha^2\right)}.$$

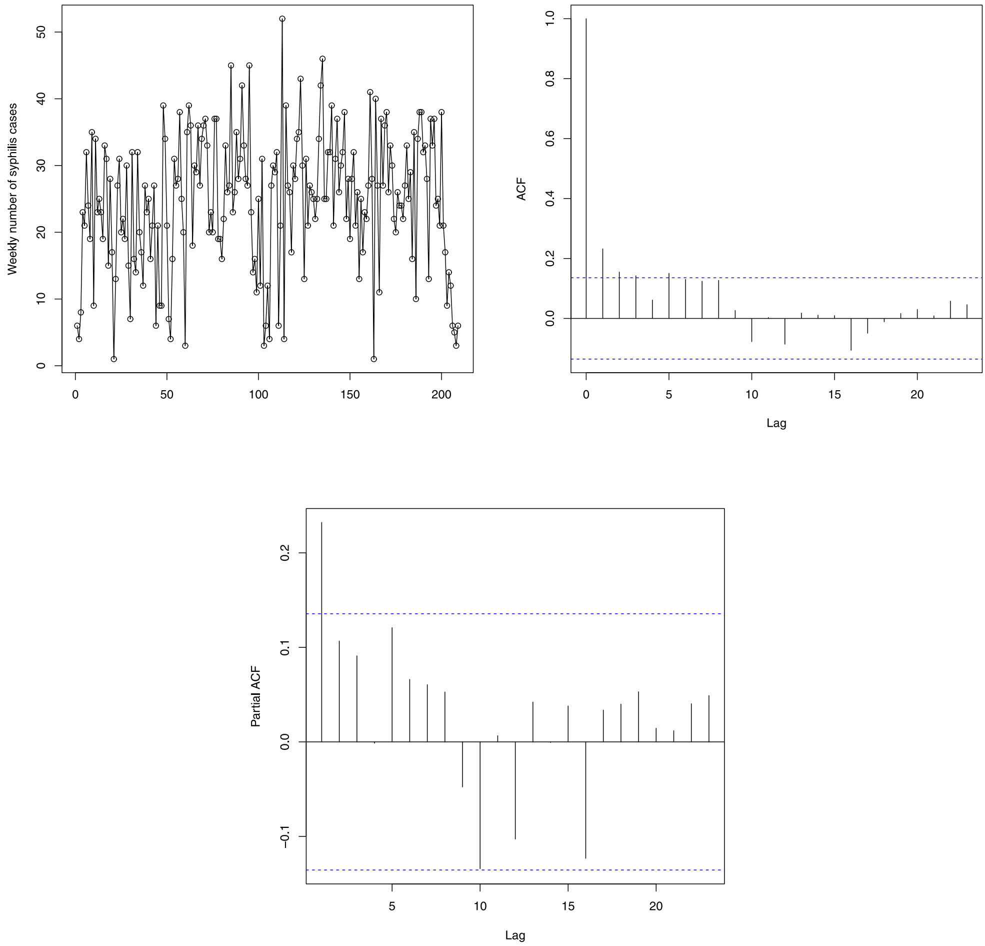

Figure 7. The ACF and PACF plots of the weekly number of syphilis cases.

The dispersion index of the INARPX(1) process is

$$DI_X = 1 + \frac{3 + \theta(8 + \theta)}{\theta(1 + \theta)(3 + \theta)(1 + \alpha)},$$

which indicates overdispersion.

The model parameters, α and θ, have to be estimated from the observations of the process, $X_1, \ldots, X_T$. Many simulations have been performed by some authors which support the use of the conditional maximum likelihood (CML) method to estimate the parameters of the INAR(1) process (see Bourguignon *et al.*, [9] and Lívio *et al.* [18]). Therefore, we consider the CML method to obtain the parameters of the INARPX(1) model. The conditional log-likelihood function for this model takes the form

$$\ell\left(p, \theta\right) = \sum_{t=2}^{T} \log\left[\Pr\left(X_t = k \vert\, X_{t-1} = l\right)\right], \tag{14}$$

Table 5. The CML estimators of the fitted INAR(1) processes with model selection criteria for the weekly number of syphilis cases.

Model	Parameters	Estimate	SE	AIC	BIC	μ_X	σ_X^2	DI_X
INARPX(1)	α	0.214	0.037	1605.189	1611.874	24.571	189.348	7.706
	θ	0.142	0.009					
INARPL(1)	α	0.249	0.037	1630.809	1637.493	24.736	226.287	9.148
	θ	0.103	0.007					
INARP(1)	α	0.148	0.026	2016.540	2023.220	24.720	24.720	1.000
	λ	21.063	0.709					
Empirical						24.630	105.680	4.290

where $\Pr(X_t = k|\, X_{t-1} = l)$ is given by (13). Since it is not possible to obtain explicit forms of the CML estimators of the parameters of the INARPX(1) process, direct maximization of (14) by using a statistical software, such as **R**, S-Plus, Matlab, is needed to obtain the CML estimators of (α, θ). The standard errors of the estimated parameters are obtained by means of observed information matrix evaluated at $(\widehat{\alpha}, \widehat{\theta})$.

We provide an application to real data to prove empirically the usefulness of the INARPX(1) (INAR PX) model in case of overdispersion as compared to the INARP(1) (INAR Poisson) and INARPL(1) (INAR Poisson–Lindley) models. The one-step translation probabilities for these models can be found in Altun [4]. The data used refers to the weekly number of syphilis cases in the United States from 2007 to 2010 in New York. The data set can be found in **ZIM** package of the **R** software.

First, we investigate the possible overdispersion in the data used by means of a hypothesis test introduced by Schweer and Weiß [27]. We calculate the mean, variance and dispersion index. These are 24.631, 105.676 and 4.290, respectively. Then, we perform a hypothesis test of Schweer and Weiß. The obtained p-value is $0 < 0.001$ which indicates that the data display overdispersion. In this case, it is clear that the more flexible distribution than the Poisson is needed to model the overdispersion in the data.

Figure 7 displays the times series plot, sample autocorrelation function (ACF) and partial ACF (PACF) of the weekly number of syphilis cases. There is a clear cut-off after the first lag at the ACF plot which indicates that AR(1) model can be suitable for the data.

The information on the fitted INAR(1) processes such as estimated parameters, AIC and BIC values are given in Table 5. It is obvious that the INARPX(1) is more appropriate model than other competitive model for the data since its AIC and BIC values are the lowest.

7. Conclusion

We introduce a new one-parameter discrete distribution called the PX distribution. Some statistical properties of the new distribution are studied comprehensively. A new regression model for non-negative discrete response variable is defined and applied to a real data set. Additionally, INAR(1) process with PX innovations are introduced. Empirical findings show that the PX distribution provides acceptable results for the over-dispersed data sets.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Emrah Altun http://orcid.org/0000-0001-5065-2523
Gauss M. Cordeiro http://orcid.org/0000-0002-3052-6551

References

[1] M.A. Al-Osh and A.A. Alzaid, *First-order integer-valued autoregressive (INAR (1)) process*, J. Time Ser. Anal. 8 (1987), pp. 261–275.
[2] E. Altun, *A new zero-inflated regression model with application*, J. Stat. Stat. Actuarial Sci. 11 (2018), pp. 73–80.
[3] E. Altun, *A new model for over-dispersed count data: Poisson quasi-Lindley regression model*, Math. Sci. 13 (2019a), pp. 241–247.
[4] E. Altun, *A new generalization of geometric distribution with properties and applications*, Commun. Stat. Simul. Comput. 49 (2019b), pp. 1–15.
[5] E. Altun, *A new one-parameter discrete distribution with associated regression and integer-valued autoregressive models*, Math. Slovaca 70 (2020), pp. 979–994.
[6] J. Andersson and D. Karlis, *A parametric time series model with covariates for integers in $\mathbb{Z}$*, Stat. Model. 14 (2014), pp. 135–156.
[7] D. Bhati, P. Kumawat, and E. Gómez-Déniz, *A new count model generated from mixed Poisson transmuted exponential family with an application to health care data*, Commun. Stat. Theor. Meth.46 (2017), pp. 11060–11076.
[8] P. Borges, M. Bourguignon, and F.F. Molinares, *A generalised NGINAR (1) process with inflated-parameter geometric counting series*, Aust. N. Z. J. Stat. 59 (2017), pp. 137–150.
[9] M. Bourguignon, J. Rodrigues, and M. Santos-Neto, *Extended Poisson INAR (1) processes with equidispersion, underdispersion and overdispersion*, J. Appl. Stat. 46 (2019), pp. 101–118.
[10] L. Cheng, S.R. Geedipally, and D. Lord, *The Poisson–Weibull generalized linear model for analyzing motor vehicle crash data*, Saf. Sci. 54 (2013), pp. 38–42.
[11] E.G. Déniz, *A new discrete distribution: properties and applications in medical care*, J. Appl. Stat. 40 (2013), pp. 2760–2770.
[12] M.S. Eliwa, E. Altun, M. El-Dawoody, and M. El-Morshedy, *A new three-parameter discrete distribution with associated INAR (1) process and applications*, IEEE Access. 8 (2020), pp. 91150–91162.
[13] Y. Gencturk and A. Yigiter, *Modelling claim number using a new mixture model: negative binomial gamma distribution*, J. Stat. Comput. Simul. 86 (2016), pp. 1829–1839.
[14] M.E. Ghitany and D.K. Al-Mutairi, *Estimation methods for the discrete Poisson–Lindley distribution*, J. Stat. Comput. Simul. 79 (2009), pp. 1–9.
[15] T. Imoto, C.M. Ng, S.H. Ong, and S. Chakraborty, *A modified Conway–Maxwell–Poisson type binomial distribution and its applications*, Commun. Stat.-Theor. Meth. 46 (2017), pp. 12210–12225.
[16] M.A. Jazi, G. Jones, and C.D. Lai, *Integer valued AR (1) with geometric innovations*, J. Iranian Stat. Soc. 11 (2012), pp. 173–190.
[17] H. Kim and S. Lee, *On first-order integer-valued autoregressive process with Katz family innovations*, J. Stat. Comput. Simul. 87 (2017), pp. 546–562.
[18] T. Lívio, N. Mamode Khan, M. Bourguignon, and H. Bakouch, *An INAR (1) model with Poisson–Lindley innovations*, Econom. Bulletin 38 (2018), pp. 1505–1513.
[19] D. Lord and S.R. Geedipally, *The negative binomial-Lindley distribution as a tool for analyzing crash data characterized by a large amount of zeros*, Accident Anal. Prevent. 43 (2011), pp. 1738–1742.
[20] R. Lugannani and S. Rice, *Saddle point approximation for the distribution of the sum of independent random variables*, Adv. Appl. Probab. 12 (1980), pp. 475–490.
[21] E. Mahmoudi and H. Zakerzadeh, *Generalized Poisson–Lindley distribution*, Commun. Stat. Theor. Meth. 39 (2010), pp. 1785–1798.

[22] E. McKenzie, *Some simple models for discrete variate time series 1*, JAWRA J. Amer Water Res Assoc. 21 (1985), pp. 645–650.
[23] E. McKenzie, *Autoregressive moving-average processes with negative binomial and geometric marginal distributions*, Adv. Appl. Probab. 18 (1986), pp. 679–705.
[24] J. Rodríguez-Avi, A. Conde-Síánchez, A.J. Sáez-Castillo, M.J. Olmo-Jiménez, and A.M. Martínez-Rodríguez, *A generalized waring regression model for count data*, Comput. Stat. Data. Anal. 53 (2009), pp. 3717–3725.
[25] A.J. Sáez-Castillo and A. Conde-Sánchez, *A hyper-Poisson regression model for overdispersed and underdispersed count data*, Comput. Stat. Data. Anal. 61 (2013), pp. 148–157.
[26] M. Sankaran, *The discrete Poisson–Lindley distribution*, Biometrics 26 (1970), pp. 145–149.
[27] S. Schweer and C.H. Weiß, *Compound Poisson INAR (1) processes: stochastic properties and testing for overdispersion*, Comput. Stat. Data. Anal. 77 (2014), pp. 267–284.
[28] S. Sen, S.S. Maiti, and N. Chandra, *The xgamma distribution: statistical properties and application*, J. Mod. Appl. Stat. Meth. 15 (2016), pp. 38–0.
[29] R. Shanker and H. Fesshaye, *On Poisson–Lindley distribution and its applications to biological sciences*, Biom. Biostat. Int. J. 2 (2015), pp. 103–107.
[30] G. Shmueli, T.P. Minka, J.B. Kadane, S. Borle, and P. Boatwright, *A useful distribution for fitting discrete data: revival of the Conway–Maxwell–Poisson distribution*, J. R. Stat. Soc.: Ser. C (Appl. Stat.) 54 (2005), pp. 127–142.
[31] M.M. Shoukri, M.H. Asyali, R. VanDorp, and D. Kelton, *The Poisson inverse Gaussian regression model in the analysis of clustered counts data*, J. Data. Sci. 2 (2004), pp. 17–32.
[32] M.H. Tahir, M.A. Hussain, G.M. Cordeiro, M. El-Morshedy, and M.S. Eliwa, *A new Kumaraswamy generalized family of distributions with properties, applications, and bivariate extension*, Math. 8 (2020), pp. 1989.
[33] W. Wongrin and W. Bodhisuwan, *Generalized Poisson–Lindley linear model for count data*, J. Appl. Stat. 44 (2017), pp. 2659–2671.
[34] H. Zamani, N. Ismail, and P. Faroughi, *Poisson-weighted exponential univariate version and regression model with applications*, J. Math. Stat. 10 (2014), pp. 148–154.