Journal of Statistics & Management Systems
ISSN 0972-0510 (Print), ISSN 2169-0014 (Online)
Vol. 26 (2023), No. 2, pp. 291–308
DOI : 10.47974/JSMS-865

A new symmetric distribution on integers with applications in data modeling and zero inflated regression

Christophe Chesneau [‡]
Laboratoire de Mathématiques Nicolas Oresme
University of Caen
Caen 14032
France

Subrata Chakraborty [†]
Department of Statistics
Dibrugarh University
Dibrugarh 786004
Assam
India

Hassan S. Bakouch [§]
Department of Mathematics
Faculty of Science
Tanta 31111
Egypt

Kadir Karakaya *
Department of Statistics
Faculty of Science
Selçuk University
Konya 42130
Turkey

Abstract

In this paper, a symmetric discrete distribution on integers is proposed by using a novel method of generating new distribution on $\mathbb{Z}$. This distribution has only one parameter and possesses a very simple structure. Many important characteristics, such as, the

[‡] *E-mail:* christophe.chesneau@gmail.com
[†] *E-mail:* subrata_stats@dibru.ac.in
[§] *E-mail:* hassan.bakouch@science.tanta.edu.eg
*E-mail: kkarakaya@selcuk.edu.tr (Corresponding Author)

© Taru Publications

recurrence relation, tail length, mode, cumulative distribution function, monotonic property of log-concavity, moments and generating functions, are investigated. Maximum likelihood estimation of the model parameter is presented. An innovative zero-inflated count regression analysis is introduced based on the proposed distribution. The utility of this model is justified by considering three applications of count data modeling and regression in comparison to some established count distributions on $\mathbb{Z}$.

Subject Classification: *(2010) 60E05, 62E10, 62J02.*

Keywords: *Estimation, Symmetric distribution, Zero inflated regression.*

1. Introduction

In many real-life situations we often come across discrete data with support on $\mathbb{Z}$, which is the count variable under observation may assume integer values (both positive and negative) and also zero count. A prominent example is in modeling frequency distribution of difference in duration of increasing and decreasing exchange rate (see Inusah (2003)). Even in situations where we only observe two discrete variables taking zero and positive counts, our interest may lie in their difference, which will have support in $\mathbb{Z}$. Examples of count random variables with support in $\mathbb{Z}$ arise in many situations (see, e.g., Chakraborty (2016), Ong *et al*. (2009) and Karlis and Ntzoufras (2003)). There are not many discrete distributions having support on $\mathbb{Z}$. The earliest one that has garnered a lot of attention is the Skellam distribution (Skellam (1946)) which arises as a difference between two Poisson random variables. Other important contributions includes the discrete normal of Kemp (1997), discrete normal of Roy (2003), discrete Laplace of Inusah and Kozubowski (2006), skew Laplace distribution on integers of Kozubowski and Inusah (2006), discrete skew Laplace of Barbiero (2014), discrete alpha-skew-Laplace distribution of Harandi and Alamatsaz (2015), discrete Gumbel of Chakraborty *et al*. (2020) and discrete asymmetric Laplace of Bhati *et al*. (2020), among others. All mentioned distributions, except the discrete Laplace of Inusah and Kozubowski (2006), have two or more parameters with probability mass function involving complicated expressions. For further details on discrete distributions having support on $\mathbb{Z}$, we refer the reader to Chakraborty and Chakravarty (2016), and the references therein.

The discrete Laplace distribution with parameter $p \in (0,1)$ proposed by Inusah and Kozubowski (2006) is defined by the following probability mass function (pmf):

$$f(k;p)=\frac{1-p}{1+p}p^{|k|},\ k\in\mathbb{Z}.$$

It is derived from the general definition:

$$f(k;p)=\frac{1}{\sum_{j\in\mathbb{Z}}g(j;\sigma_p)}g(k;\sigma_p),\ k\in\mathbb{Z},$$

where $g(x;\sigma)$ denotes the pdf of the Laplace distribution, i.e., $g(x;\sigma)=(1/(2\sigma))e^{-|x|/\sigma}$, $x\in\mathbb{R}$, with $\sigma_p=-(\log p)^{-1}>0$.

As a new result, we found that the discrete Laplace distribution can also be defined as

$$f(k;p)=\frac{1}{2-h(0;p)}h(|k|;p),\ k\in\mathbb{Z},$$

where $h(k;p)$ denotes the pmf of the geometric distribution, i.e.,

$$h(k;p)=(1-p)p^{k},\ k\in\mathbb{N}=\{0,1,2,\ldots\}.$$

This simple symmetric representation can be applied to create a new distribution on $\mathbb{Z}$ based on any discrete distribution with support equal to $\mathbb{N}$.

Thus, based on this scheme, we aim to introduce a new symmetric distribution based on the Poisson-Lindley (PL) distribution with parameter $\theta>0$ defined with the following pmf:

$$f_*(k;\theta)=\frac{\theta^2(\theta+2+k)}{(\theta+1)^{k+3}},\ k\in\mathbb{N}.$$

Further details on the PL distribution can be found in Sankaran (1970).

We thus define the symmetric PL (SPL) distribution by the following pmf:

$$f(k;\theta)\ =\frac{1}{2-f_*(0;\theta)}f_*(|k|;\theta),\ k\in\mathbb{Z},$$

which on substitution of the PL pmf, reduces to

$$f(k;\theta)=\frac{\theta^2}{\theta^3+4\theta^2+6\theta+2}\frac{\theta+2+|k|}{(\theta+1)^{|k|}},\ k\in\mathbb{Z}. \tag{1.1}$$

For the purpose of this study, the distribution defined by this pmf is denoted by SPL(θ). It constitutes a new and simple discrete distribution on $\mathbb{Z}$ with exponentiated decay with respect to *k*, having some relation

with a few existing distributions (as formulated in the remark below). In this paper, we aim to study its theoretical and practical properties.

Remark 1.1 : We can write $f(k;\theta)$ as the following mixtures:

$$f(k;\theta) = a_\theta g_1(k;\theta) + b_\theta g_2(k;\theta),$$

where

$$a_\theta = \frac{\theta(\theta+2)(\theta+1)}{\theta^3+4\theta^2+6\theta+2},\ b_\theta = \frac{2(\theta+1)^2-\theta^2}{\theta^3+4\theta^2+6\theta+2},$$

$$g_1(k;\theta) = \frac{\theta}{\theta+2}\left(1-\frac{\theta}{\theta+1}\right)^{|k|},\ k\in\mathbb{Z}$$

and

$$g_2(k;\theta) = \frac{\theta^2}{2(\theta+1)^2-\theta^2}(1+|k|)\left(1-\frac{\theta}{\theta+1}\right)^{|k|},\ k\in\mathbb{Z}.$$

One can notice that $g_1(k;\theta)$ is the pmf of the discrete Laplace distribution with parameter $1/(1+\theta)$ and $g_2(k;\theta)$ is the pmf of a (non referenced) symmetric version of the negative binomial distribution with parameters 2 and $\theta/(\theta+1)$. Furthermore, as anticipated, we have $a_\theta + b_\theta = 1$. Such mixture distribution is recommended to model objects in a certain population that differ from each other. The proposed distribution takes into account the overall heterogeneity of the object to provide a more accurate data analysis.

A location version of the pmf given as Equation (1.1) can be defined by

$$f(k;\theta,\mu) = \frac{\theta^2}{\theta^3+4\theta^2+6\theta+2}\frac{\theta+2+|k-\mu|}{(\theta+1)^{|k-\mu|}},\ k\in\mathbb{Z}. \tag{1.2}$$

We denote by $\mathrm{SPL}(\theta,\mu)$ to the related distribution. One can show that the parameter μ is the mean of this distribution, which will be useful to adjust observed data with different means, among others.

2. Mathematical properties

Here, some mathematical properties of the distribution $\mathrm{SPL}(\theta)$ defined by Equation (1.1) are derived. Also, X will denote a random variable following the $\mathrm{SPL}(\theta)$ distribution. However, thanks to its flexible

properties, the SPL(θ,μ) distribution will be logically preferred for modelling aims.

2.1 *Recurrence relation*

The following recurrence relation holds. For any $k \in \mathbb{Z}$, we have

$$\frac{f(k+1;\theta)}{f(k;\theta)} = \frac{\theta+2+|k+1|}{\theta+2+|k|}(\theta+1)^{|k|-|k+1|} = \begin{cases} \left(1+\dfrac{1}{\theta+2+k}\right)\dfrac{1}{\theta+1} & \text{for } k \geq 0 \\ \left(1-\dfrac{1}{\theta+2-k}\right)(\theta+1) & \text{for } k < 0 \end{cases}.$$

From the above relation, it is clear that the distribution will have longer tails if $\theta \to 0$, and shorter if $\theta \to +\infty$.

2.2 *Mode*

The SPL(θ) distribution is symmetric about the point 0. But this mode may not always be at 0. In such situations, we can have modes of the form (-*m* and *m*) where *m* denotes a positive integer, depending on the value of the parameter θ. We can formalize this point by studying the following function: $\ell(x) = (\theta+2+x)/(\theta+1)^x$ satisfying $f(k;\theta) = [\theta^2/(\theta^3+4\theta^2+6\theta+2)]\ell(|k|)$. After analysis, the maximum on $[0,+\infty)$ with respect to *x* is given at the point

$$x_0 = \max\left(\frac{1}{\log(1+\theta)} - \theta - 2, 0\right).$$

Hence, the mode(s) of the SPL(θ) distribution is (are) given as $(-\lfloor x_0 \rfloor, \lfloor x_0 \rfloor)$ if $f(\lceil x_0 \rceil;\theta) \leq f(\lfloor x_0 \rfloor;\theta)$, or $(-\lceil x_0 \rceil, \lceil x_0 \rceil)$ otherwise, where $\lfloor x_0 \rfloor$ and $\lceil x_0 \rceil$ denote the floor and ceil functions at x_0, respectively.

Note that $(-[x_0],[x_0])$ is reduced to 0 if $x_0 = 0$. For $\theta = 0.6,\ 0.4$ and 0.2, we have $x_0 = 0$, $x_0 = 0.7720134$, and $x_0 = 3.284815$, respectively, giving the mode(s): 0, (-1, +1), and (-3, +3), respectively.

2.3 *Cumulative distribution and hazard rate functions*

The cumulative distribution function of the SPL(θ) distribution is given by

$$F(k;\theta) = P(X \leq k) = \sum_{m=-\infty}^{k} f(m;\theta)$$
$$= \begin{cases} \dfrac{(1+\theta)^{1+k}\left(1+2\theta-k\theta+\theta^2\right)}{2+6\theta+4\theta^2+\theta^3} & \text{for } k \leq 0 \\ 1-\dfrac{(1+\theta)^{-k}\left(1+3\theta+k\theta+\theta^2\right)}{2+6\theta+4\theta^2+\theta^3} & \text{for } k > 0 \end{cases}.$$

Consequently, the hazard rate function $\lambda(k;\theta)$ can be easily obtained by using the following formula:

$$\lambda(k;\theta) = \frac{f(k;\theta)}{1-F(k-1;\theta)},\ k \in \mathbb{Z}.$$

In particular, for $k > 1$, we have

$$\lambda(k;\theta) = \frac{\theta^2(\theta+2+k)}{(1+\theta)[1+\theta(2+k+\theta)]}.$$

Thus, $\lambda(k;\theta)$ is an increasing function with respect to k and tends to $\theta/(1+\theta)$ when $k \to +\infty$ (and hence, is uniformly bounded by $\theta/(1+\theta)$).

2.4 *Log-concavity*

Log-concavity is an important characteristic of a probability distribution with many important implications. A discrete distribution with pmf $f(k;\theta)$ is said to be log-concave if it satisfies $f^2(k;\theta) \geq f(k-1;\theta)$ $f(k+1;\theta)$. Further details on log-concavity can be found in Bagnoli and Bergstrom (2005).

The SPL(θ) distribution is log-concave if $\theta \geq \sqrt{2}-1$, for all $k \in \mathbb{Z}$. Indeed, by denoting $P(k;\theta) = f(k-1;\theta)\ f(k+1;\theta)\ - f^2(k;\theta)$, if $k \leq -1$, we have

$$P(k;\theta) = -\left(\frac{\theta^2}{\theta^3+4\theta^2+6\theta+2}\right)^2 (\theta+1)^{2k} < 0.$$

On the other side, if $k \geq 1$, we have

$$P(k;\theta) = -\left(\frac{\theta^2}{\theta^3+4\theta^2+6\theta+2}\right)^2 \frac{1}{(\theta+1)^{2k}} < 0.$$

For the special case $k = 0$, if $\theta \geq \sqrt{2}-1$, we have

$$P(k;\theta)=\left(\frac{\theta^2}{\theta^3+4\theta^2+6\theta+2}\right)^2\left(\frac{(\theta+3)^2}{(\theta+1)^2}-(\theta+2)^2\right)\le 0.$$

The following properties are consequences of the log-concavity. The SPL (θ) distribution

i. has an increasing failure (hazard) rate distribution;
ii. is strongly unimodal;
iii. has moments of all orders;
iv. remains log-concave if truncated;
v. convolution with any other discrete distribution is also unimodal and log-concave.

For the concept of unimodality of discrete distribution see Keilson and Gerber (1971).

2.5 *Moments*

Now, for the sake of conciseness, we set $c_\theta=\theta^2/(\theta^3+4\theta^2+6\theta+2)$. Then, for any integer m, the $(2m+1)^{th}$ raw moment of X is given by $\mu'_{2m+1}=E(X^{2m+1})=0$. In particular, the mean of X is equal to 0. The $2m^{th}$ raw moment of X is given by

$$\begin{aligned}\mu'_{2m} &= E(X^{2m})=2\sum_{k=1}^{+\infty}k^{2m}f(k;\theta)=2c_\theta\sum_{k=1}^{+\infty}k^{2m}\frac{\theta+2+k}{(\theta+1)^k}\\ &= 2c_\theta\left[(\theta+2)Li_{-2m}\left(\frac{1}{\theta+1}\right)+Li_{-2m-1}\left(\frac{1}{\theta+1}\right)\right],\end{aligned}\tag{2.1}$$

where $Li_s(z)=\sum_{k=1}^{+\infty}z^k/k^s$, with $s\in\mathbb{R}$ and $|z|<1$, is the known Polylogarithm function which is readily available in literature and also implemented in many mathematical software's. This function is related to the zeta function, generalized Hypergeometric function, among others. Particular values of this function can be useful in deriving exact expressions. For example,

$$Li_{-2}(z)=\frac{z(z+1)}{(1-z)^3},\ Li_{-3}(z)=\frac{z(z^2+4z+1)}{(1-z)^4}.$$

Also, in general, the Polylogarithm function can be expressed as a finite sum as follows

$$Li_{-n}(z) = \sum_{k=0}^{n} k! S(n+1, k+1) \left(\frac{z}{1-z} \right)^{k+1},$$

where $S(n, k)$ is the so-called Stirling's number of the second kind. Formulas for the variance and kurtosis are useful for this distribution, and are, respectively, given by

$$\mu_2' = \frac{2(1+\theta)^2(6+4\theta+\theta^2)}{\theta^2(2+6\theta+4\theta^2+\theta^3)}$$

and

$$\beta_2 = \frac{\mu_4'}{(\mu_2')^2} = \frac{(2+6\theta+4\theta^2+\theta^3)(120+168\theta+78\theta^2+16\theta^3+\theta^4)}{2(1+\theta)^2(6+4\theta+\theta^2)^2}.$$

Also, the $2m^{th}$ raw absolute moment of X is given by $\mu_{2m}^* = E(|X|^{2m}) = \mu_{2m}'$ as developed in Equation (2.1) and the $(2m+1)^{th}$ raw absolute moment of X is given by

$$\begin{aligned} \mu_{2m+1}^* &= E(|X|^{2m+1}) = 2\sum_{k=1}^{+\infty} k^{2m+1} f(k;\theta) = 2c_\theta \sum_{k=1}^{+\infty} k^{2m+1} \frac{\theta+2+k}{(\theta+1)^k} \\ &= 2c_\theta \left[(\theta+2) Li_{-2m-1}\left(\frac{1}{\theta+1} \right) + Li_{-2m-2}\left(\frac{1}{\theta+1} \right) \right]. \end{aligned}$$

As an application of the absolute moments, the Chebyshev inequality gives, for any positive integers k and m, $P(|X| \geq k) \leq k^{-2m-1} \mu_{2m+1}^*$, for instance. These kinds of moments also appear in the Lyapounov theorem.

2.6 *Probability generating function*

The probability generating function of X is obtained as

$$G(s) = E(s^X) = \sum_{k=-\infty}^{+\infty} s^k f(k;\theta) = c_\theta \sum_{k=-\infty}^{+\infty} s^k \frac{\theta+2+|k|}{(\theta+1)^{|k|}}.$$

Owing to the following geometric series expansion results:

$$\sum_{k=-\infty}^{+\infty} s^k x^{|k|} = \frac{s(1-x^2)}{(s-x)(sx-1)}, \sum_{k=-\infty}^{+\infty} |k| s^k x^{|k|} = \frac{sx(s^2x^2+x^2-4sx+s^2+1)}{(s-x)^2(sx-1)^2}$$

with $|sx| < 1$ and $|x/s| < 1$, we get

$$G(s) = \frac{s\theta^2(2-5\theta^2-4\theta^3-\theta^4-s^2(5\theta^2+4\theta^3+\theta^4-2)+s(12\theta^2+14\theta^3+6\theta^4+\theta^5-4))}{(\theta-s+1)^2(s\theta+s-1)^2(\theta^3+4\theta^2+6\theta+2)},$$

for $s \in ((\theta+1)^{-1}, \theta+1)$.

Further possible mathematical developments of the $\mathrm{SPL}(\theta,\mu)$ distribution include the discrete decomposable properties (see Artikis (2020)), and various extensions (bivariate extensions, etc). Next, we focus on the applicability of the $\mathrm{SPL}(\theta,\mu)$ distribution in several statistical scenarios of importance.

2.7 *Maximum likelihood method of parameter estimation*

In this section, we consider the maximum likelihood (ML) estimation of the parameters of the $\mathrm{SPL}(\theta,\mu)$ distribution (introducing the location parameter, to take care of the different locations of the data sets). Let $K_1, K_2, \ldots, K_n$ be n independent and identically random variables from the $\mathrm{SPL}(\theta,\mu)$ distribution, and $k_1, k_2, \ldots, k_n$ be observations of these random variables. The log-likelihood function is given by

$$\begin{aligned}\ell = \ell(\theta,\mu) = {} & 2n\log(\theta) - n\log(\theta^3+4\theta^2+6\theta+2) \\ & + \sum_{i=1}^{n}\log(\theta+2+|k_i-\mu|) - \log(\theta+1)\sum_{i=1}^{n}|k_i-\mu|.\end{aligned}$$

The ML estimates (MLEs) of the components of (θ,μ), say $(\hat{\theta},\hat{\mu})$, are obtained by

$$(\hat{\theta},\hat{\mu}) = \underset{(\theta,\mu)}{\arg\max}\left(\ell(\theta,\mu)\right).$$

They can be determined by investigating the components of the score function given as

$$\frac{\partial\ell}{\partial\theta} = \frac{2n}{\theta} - \frac{n(3\theta^2+8\theta+6)}{\theta^3+4\theta^2+6\theta+2} + \sum_{i=1}^{n}\frac{1}{\theta+2+|k_i-\mu|} - \frac{1}{\theta+1}\sum_{i=1}^{n}|k_i-\mu|$$

and, assuming that $k_i \neq \mu$,

$$\frac{\partial\ell}{\partial\mu} = \sum_{k_i>\mu}\frac{1}{\theta+2+k_i-\mu} - \sum_{k_i<\mu}^{n}\frac{1}{\theta+2-k_i+\mu} + \log(\theta+1)\left(\sum_{k_i>\mu}1 - \sum_{k_i<\mu}1\right).$$

By solving the equations corresponding to the above component equals 0, simultaneously with respect to θ and μ, we get $\hat{\theta}$ and $\hat{\mu}$.

Since no analytical formula is possible, numerical maximization of $\ell(\theta, \mu)$ with respect to θ and μ is carried out using Mathematica **NMaximize** function.

Under standard regularity conditions, the properties of the MLEs can be used for the construction of statistical procedures. In particular, the asymptotic distribution of a MLE is normal with well-identified parameters. We refer the reader to Casella and Berger (1990) for more details on these properties, and also to Eidous and Abu-Hawwas (2021) for recent developments on the practical approximations of the normal distribution.

3. Data modeling

For the purpose of data modeling, we consider the following three well-known distributions defined over $\mathbb{Z}$ for comparison with the proposed SPL (θ, μ) distribution.

1. Discrete Logistic (DLOG) distribution defined by Chakraborty and Chakravarty (2016) and having the following pmf:

$$f(k;\theta,\mu) = \frac{(1-\theta)\theta^{k-\mu}}{(1+\theta^{k-\mu})(1+\theta^{k-\mu-1})},\ k \in \mathbb{Z},$$

with $\theta > 0$ and $\mu \in \mathbb{R}$.

2. Discrete Laplace (DL) distribution is defined by the following pmf:

$$f(k;\theta) = \frac{1-\theta}{1+\theta}\theta^{|k|},\ k \in \mathbb{Z},$$

with $\theta \in (0,1)$ and $\mu \in \mathbb{R}$.

3. Skellam (SK) distribution is defined by the following pmf:

$$f(k;\theta,\mu) = \exp[-(\theta+\mu)](\theta/\mu)^{k/2} I_{[k]}(2\sqrt{\theta\mu}),\ k \in \mathbb{Z},$$

with $\theta > 0$ and $\mu \in \mathbb{R}$, where $I_{[n]}(z)$ is the modified Bessel function of the first kind.

For the data fitting and modeling, we investigate two data sets and their description is as follows. The first data set is taken from Chakraborty and Chakravarty (2016), and it is about the difference between two physicians in counting the number of palpable lymph nodes in 32

randomly selected participants from a prospective study of male sexual contacts of men with AIDS or an AIDS related condition (ARC), assessed for presence or absence of generalized lymphadenopathy. The data set was also used by Rosner (2006).

The second data set is taken from Harandi and Alamatsaz (2015), and it is about the difference between the desired and the existing state values in a study to investigate the extent of success of Iranian universities in transferring technology to industry and their effective factors.

For the two data sets, the observed mean was 2 and 3, respectively. Thus, the SPL (θ, μ) distribution is fitted for the two data sets.

Performances of the distributions are compared with the model selection criteria named as AIC (Akaike information criterion). The AIC is defined as AIC $= -2\hat{\ell} + 2k$, where k is the number of parameter(s) and $\hat{\ell}$ is the maximum of the log-likelihood function. Also, Chi-Square and Kolmogorov-Smirnov (KS) tests of goodness of fit are used with their p-values, for a given data set. In computing the test statistic, the total expected frequency is first adjusted to ensure equality with total observed frequency.

The results corresponding to the first data set are collected in Table 1.

Table 1 shows that the SPL distribution has the smallest AIC and largest p-values. Hence, for the first data set, the SPL model is the best choice with respect to all criteria.

Now, the results corresponding to the second data set are presented in Table 2.

Table 1

MLEs, Chi-Square and KS tests related to the considered models for the first data set.

		DLOG (θ, μ)	DL (θ)	SK (θ, μ)	SPL (θ, μ)
MLE	θ	0.4982	0.7299	4.7570	0.9859
	μ	3.2578	---	2.1320	2.9999
$\hat{\ell}$		-75.4575	-90.9096	-75.4322	**-74.9372**
AIC		154.9150	183.8192	154.8644	**153.8744**
Chi-Square		9.6505	48.7968	10.7551	**8.7639**
(p-value)		0.2093	1.8×10^{-7}	0.1497	**0.1873**
KS		0.1079	0.4420	0.1563	**0.0897**
(p-value)		0.5116	2.2×10^{-16}	0.1964	**0.6620**

Table 2

MLEs, Chi-Square and KS tests related to the considered models for the second data set.

		DLOG (θ, μ)	DL (θ)	SK (θ, μ)	SPL (θ, μ)
MLE	θ	0.7634	0.8595	26.1131	0.3590
	μ	4.4796	--	22.4014	3.7457
$\hat{\ell}$		-368.8065	-389.0533	-373.1016	**-368.6788**
AIC		741.6130	780.1066	750.2032	**741.3756**
Chi-Square (p-value)		32.8654 0.0834	39.8666 0.0523	**28.5691** **0.3310**	30.4645 0.1698
KS (p-value)		**0.0420** **0.9026**	0.2835 2.2×10^{-16}	0.0925 0.1845	0.0631 0.6225

Hence, from Table 2, for the second data set, in terms of AIC value, the SPL model is the best, followed by the DLOG model. In terms of p-value of Chi-square, the SK model is the best, followed by the SPL model, and, in terms of p-value of the KS test, the DLOG model is the best, followed by the SPL model. So, in aggregate, we conclude that the SPL model is the best.

4. Zero inflated SPL regression model

4.1 *Definition*

Regression analysis plays a vital role in the data modeling in presence of covariates. In the case where the response variable is of a count nature, the ordinary regression can not be applied. The most widely used Poisson and negative binomial count regression models do not fit the situation when one encounters zero inflated count responses. In this section, a zero-inflated SPL count regression model is provided using the SPL distribution with the pmf given in Equation (1.2). By mixing a degenerate probability mass with a SPL distribution, the zero-inflated SPL distribution, is defined by the following pmf:

$$f_o(k;\theta,\mu,\pi)=\begin{cases}\pi+(1-\pi)\dfrac{\theta^2}{\theta^3+4\theta^2+6\theta+2}\dfrac{\theta+2+|\mu|}{(\theta^2+1)^{|\mu|}}, k=0\\ \\ (1-\pi)\dfrac{\theta^2}{\theta^3+4\theta^2+6\theta+2}\dfrac{\theta+2+|k-\mu|}{(\theta^2+1)^{|k-\mu|}}, k\in\mathbb{Z}-\{0\}\end{cases}, \tag{4.1}$$

where $\pi\in(0,1)$ and it is expressed by the zero-inflated SPL model and denoted by ZISPL(θ,μ,π).

Let $K_i,(i=1,2,\ldots,n)$ be a ZISPL(θ,μ_i,π_i) random variable with pmf given in Equation (4.1) and $K_1,K_2,\ldots,K_n$ be n independent random variables, and $k_1,k_2,\ldots,k_n$ be observations of these random variables. Then, the ZISPL count regression model is given by the following quantities:

$$\mu_i=x_i^{\mathrm{T}}\beta, \tag{4.2}$$

and

$$\log\left(\frac{\pi_i}{1-\pi_i}\right)=z_i^{\mathrm{T}}\gamma, \tag{4.3}$$

where $\beta=(\beta_1,\beta_2,\ldots,\beta_p)^{\mathrm{T}}\in\mathbb{R}^p$ and $\gamma=(\gamma_1,\gamma_2,\ldots,\gamma_q)^{\mathrm{T}}\in\mathbb{R}^q$, $p+q+1<n$, $x_i^{\mathrm{T}}=(x_{i1},x_{i2},\ldots,x_{ip})$ and $z_i^{\mathrm{T}}=(z_{i1},z_{i2},\ldots,z_{iq})$ are known values of covariates. Also, in this study, it is assumed that the matrices $\boldsymbol{X}=(x_1,x_2,\ldots,x_p)^{\mathrm{T}}$ and $\boldsymbol{Z}=(z_1,z_2,\ldots,z_q)^{\mathrm{T}}$ have ranks p and q, respectively.

4.2 *Maximum likelihood estimation on the ZISPL model parameters*

Let $K_i,(i=1,2,\ldots,n)$ be independent ZISPL(θ,μ_i,π_i) random variable with pmf given in Equation (4.1). Let $k_1,k_2,\ldots,k_n$ be observations of these random variables. Using Equations (4.2) and (4.3) and the pmf given by Equation (4.1), the log-likelihood function can be obtained as

$$\begin{aligned}\ell(\Xi)\propto &\sum_{i\in\{j:k_j=0,\ j=1,2,\ldots,n\}}\log\left(\pi_i+(1-\pi_i)\frac{\theta^2}{\theta^3+4\theta^2+6\theta+2}\frac{\theta+2+|\mu_i|}{(\theta^2+1)^{|\mu_i|}}\right)+\\ &\sum_{i\in\{j:k_j\in\mathbb{Z}^+-\{0\},\ j=1,2,\ldots,n\}}\log(1-\pi_i)+\log\left(\frac{\theta^2}{\theta^3+4\theta^2+6\theta+2}\frac{\theta+2+|k_i-\mu_i|}{(\theta^2+1)^{|k_i-\mu_i|}}\right),\end{aligned} \tag{4.4}$$

where $\Xi=(\beta,\gamma,\mu)$,

$$\pi_i = \frac{\exp(z_i^{\mathrm{T}}\gamma)}{1+\exp(z_i^{\mathrm{T}}\gamma)}$$

and

$$\mu_i = x_i^{\mathrm{T}}\beta,$$

for $i = 1,\ldots,n$. The MLE of $\boldsymbol{\Xi}$, denoted by $\widehat{\Xi}$, can be defined by the following argmax expression:

$$\widehat{\Xi} = \arg\max_{\Xi}\Big(\ell(\Xi)\Big).$$

It can be obtained by **optim** function in R with different algorithms, such as Nelder-Mead, BFGS, CG or SANN.

4.3 *Practical example with the ZISPL model*

In this subsection, the data taken from Crawley (2012), correspond to the count of infected blood cells (per mm2) on microscope slides arranged from $n = 511$ randomly selected persons are examined. The proportion of zeros (about 61.5%) leading to the use of the proposed ZISPL model. The blood cells data were also analyzed via other zero inflated models, such as zero-inflated Poisson, zero-inflated negative binomial, zero-inflated Bell (introduced by Lemonte *et al*. (2019)) and the zero-inflated CosPois (introduced by Bakouch *et al*. (2021)) models. Here, we compare the mentioned models with the ZISPL model considering the blood cells data.

It is noticed that the R functions **zeroinfl** can be used to perform these count regression analyses. The response variable represents the count of infected blood cells, the first explanatory variable is the smoking status of the subject (0: yes; 1: no), and the second explanatory variable is gender (0: female; 1:male). More information about the blood cells data can be found in Lemonte *et al*. (2019). Tables 3 and 4 indicate the maximum likelihood estimates of all models, including an intercept. The $\hat{\ell} = \ell(\widehat{\Xi})$ and AIC are shown in these tables to compare the models. Tables 3 and 4 show that the ZISPL regression model outperforms the other well-known count regression models with respect to all the criteria. According to the ZISPL model results, the gender explanatory variable has no statistical effect on the infected blood cell response variable (p-value $>$ 0.01). However, the explanatory variable smoking has statistical effects on the infected blood cell response variable (p-value $<$ 0.01). From Table 4, it can also be noted that the infected blood cells in smokers is 1.9999, since it is -0.9996 units

Table 3

The MLEs with standard errors (SEs) of zero-inflated SPL, Poisson and negative binomial model parameters based on the blood cell data

	ZISPL model				Zero-inflated Poisson model				Zero-inflated negative binomial model		
	Estimate (SE)	z value	p value		Estimate (SE)	z value	p value		Estimate (SE)	z value	p value
(Intercept)	1.9999(0.0116)	172.405	0.0000		0.9540(0.1179)	8.092	0.0000		0.8740(0.1511)	5.784	0.0000
Smoking	-0.9996(0.0114)	-87.684	0.0000		-0.4982(0.1242)	-4.011	0.0001		-0.5251(0.1476)	-3.558	0.0004
Gender	0.0001(0.0110)	0.009	0.9927		-0.0173(0.1239)	-0.140	0.8890		0.0003(0.1480)	0.002	0.9984
Other model parameter (θ)	1.5809(0.1213)	13.033	0.0000						1.6903(0.6142)	2.752	0.0059
		Zero-inflation model coefficients (binomial with logit link):									
(Intercept)	-0.8099(0.2822)	-2.870	0.0041		-0.7738(0.292)	-2.650	0.0080		-1.053(0.416)	-2.531	0.0114
Smoking	1.6391(0.2756)	5.947	0.0000		1.5446(0.288)	5.363	0.0000		1.672(0.375)	4.459	0.0000
Gender	-0.4053(0.2326)	-1.742	0.0814		-0.4251(0.252)	-1.687	0.0916		-0.447(0.290)	-1.541	0.1232
$\hat{\ell}$	-513.50				-612.62				-610.07		
AIC	1041.00				1237.24				1234.14		

Table 4
The MLEs with SEs of zero-inflated CosPois and Bell model parameters based on the blood cell data

	Zero-inflated CosPois model			Zero inflated Bell model		
	Estimate (SE)	z value	p value	Estimate (SE)	z value	p value
(Intercept)	0.9540 (0.1179)	8.092	0.0000	0.7166 (0.2068)	3.465	0.0005
Smoking	-0.4982 (0.1242)	-4.011	0.0001	-0.6117 (0.2185)	2.800	0.0051
Gender	-0.0173 (0.1239)	-0.140	0.8890	0.0363 (0.2055)	0.177	0.8598
Other model parameter (θ)	3.1415 (0.0799)	39.318	0.0000			
(Intercept)	-0.7738 (0.2921)	-2.649	0.0081	-1.9522 (0.8838)	-2.209	0.0272
Smoking	1.5445 (0.2882)	5.359	0.0000	2.1763 (0.8603)	2.530	0.0114
Gender	-0.4251 (0.2518)	-1.688	0.0914	-0.4958 (0.4444)	-1.116	0.2646
$\hat{\ell}$	-612.60			-610.32		
AIC	1239.00			1232.65		

less in smokers, then the infected blood cells in non-smoker is 1.0003. Therefore, we can conclude that there is a decrease in the expected value of infected blood cells by a unit of (approximately) 1 in non-smokers. Also, non-smoker patients have 5.1505 [exp(1.6391)] times greater odds of not having infected blood cells than smoker patients.

5. Conclusion

A new symmetric discrete distribution on $\mathbb{Z}$ with a very simple structure is proposed by using an innovative scheme. Some of its important structural properties are investigated. The model fitting as well as the inflated regression analysis showed that the proposed model has provided adequate evidence in its favor as a promising new symmetric count data model. The new method of generating discrete symmetric models in $\mathbb{Z}$ proposed here has the potential to be used to derive other models by considering different baseline distributions. These will be taken up in future follow up studies. An innovative zero inflated regression model

is also presented. The applicability of the new zero inflated regression analysis is demonstrated based on practical data.

References

[1] Artikis, C.T. (2020). Characterizing discrete self-decomposable distributions by incorporating an integral equation for probability generating functions, *Journal of Interdisciplinary Mathematics* 23(3), 701-706.

[2] Bagnoli, M. and Bergstrom, T. (2005). Log-concave probability and its applications, Econom. Theory 26, 445-469.

[3] Bakouch, H.S., Chesneau, C., Karakaya, K. and Kus, C. (2021). The Cos-Poisson model with a novel count regression analysis, *Hacettepe Journal of Mathematics and Statistics* 50 (2), 559-578.

[4] Barbiero, A. (2014). An alternative discrete skew Laplace distribution. Statistical Methodology 16, 47-67.

[5] Bhati, D., Chakraborty, S. and Lateef, S.G. (2020). A discrete probability model suitable for both symmetric and asymmetric count data. FILOMAT 34(8), 2559-2572.

[6] Casella, G. and Berger, R. L. (1990). Statistical Inference, Brooks/Cole Publishing Company: Bel Air, CA, USA.

[7] Chakraborty, S. and Chakravarty, D. (2016). A new discrete probability distribution with integer support on $(-\infty, \infty)$. Communications in Statistics: Theory and Methods 45(2), 492-505.

[8] Chakraborty, S., Chakravarty, S., Mazucheli, J. and Bertoli, W. (2020). A discrete analog of Gumbel distribution: properties, parameter estimation and applications. *Journal of Applied Statistics* 48(4), 712-737.

[9] Crawley, M.J. (2012). The R Book, 2nd ed., JohnWiley and Sons, New York.

[10] Eidous, E. and Abu-Hawwas, J. (2021). An accurate approximation for the standard normal distribution function. *Journal of Information and Optimization Sciences* 42(1), 17-27.

[11] Harandi, S.S. and Alamatsaz, M.H. (2015). Discrete alpha-skew-Laplace distribution. SORT 39(1), 71-84

[12] Inusah, S. (2003). Discrete Laplace distributions, Thesis, Department of Mathematics and Statistics, University of Nevada at Reno.

[13] Inusah, S. and Kozubowski, T.J. (2006). A discrete analogue of the Laplace distribution. *Journal of Statistical Planning and Inference* 136, 1090-1102.

[14] Karlis, D. and Ntzoufras, I. (2003). Analysis of sports data using bivariate Poisson models. Statistician 52, 381-393.

[15] Keilson, J. and Gerber, H. (1971) Some results for discrete unimodality. J. Am. Stat. Assoc. 66, 386-389.

[16] Kozubowski, T.J. and Inusah, S. (2006). A skew Laplace distribution on integers. Ann. Inst. Stat. Math. 58, 555-571.

[17] Lemonte, A.J., Moreno-Arenas, G. and Castellares, F. (2019). Zero-inflated Bell regression models for count data. *Journal of Applied Statistics* 47(2), 265-286.

[18] Ong, S.H., Shimizu, K. and Choung, M.N. (2008). A class of distribution arising from difference of two random variables. Comput. Stat. Data Anal. 52, 1490-1499.

[19] Roy, D. (2003). The discrete normal distribution. Commun. Stat. Theory Methods 32(10), 1871-1883.

[20] Rosner, B. (2000). Fundamentals of Biostatistics (5th ed.). Belmont, CA, Duxbury Press.

[21] Sankaran, M. (1970). The discrete Poisson-Lindley distribution. Biometrics 26, 145-149.

[22] Skellam, J.G. (1946). The frequency distribution of the difference between two Poisson variates belonging to different populations. J. R. Stat. Soc. Ser. A 109, 296.

Received May, 2021

Revised September, 2021