Applicable Analysis and Discrete Mathematics
available online at http://pefmath.etf.rs

Appl. Anal. Discrete Math. **16** (2022), 467–484.
https://doi.org/10.2298/AADM210114029B

A NON-NEGATIVE INTEGER-VALUED MODEL: ESTIMATION, COUNT REGRESSION AND PRACTICAL EXAMPLES

Hassan S. Bakouch, Kadir Karakaya[*]*, Christophe Chesneau and Yunus Akdoğan*

In this study, we propose a non-negative integer-valued model based on the sum of Poisson-Lindley and geometric distributions. We show that it corresponds to the weighted geometric distribution and also a special mixture of two negative binomial distributions with certain parameters. The main statistical properties of the new distribution are studied comprehensively, including estimation of the model parameter. A new count regression analysis is introduced by using the new distribution. Finally, we provide some applications on practical data sets.

1. INTRODUCTION

Many areas of theoretical and applied science need precise models to handle count data. In this context, there is a requirement for new discrete distributions. Due to this need, many discrete models have been introduced in the recent years. Some of these can be consulted in: [**1**], [**2**], [**3**] and [**12**].

Among the simple methods in this regard, there are discretized continuous distributions using some ideas and weighted distributions. In this paper, we introduce a discrete distribution by using a sum strategy; we consider the distribution of the sum of two random variables defined on $\mathbb{N} = \{0, 1, \ldots, \infty\}$ following the classical Poisson-Lindley (PL) distribution and geometric (G) distribution, with a special configuration on the parameters. The idea is to introduce a simple and

[*] Corresponding author. Kadir Karakaya

2020 Mathematics Subject Classification. 60E05, 62E10, 62E15.
Keywords and Phrases. Poisson-Lindley distribution, Estimation, Data analysis, Count regression.

new distribution that dominates stochastically the PL distribution, allowing the construction of alternative statistical models (fitting model, regression model ...) that outperformed models based on the classical distributions.

Let us now present the mathematical definition of the new distribution. First, let X be a random variable following the PL distribution with parameter $\theta > 0$. Then, it is defined with the following probability mass function (pmf):

$$f_X(x;\theta) = \frac{\theta^2(\theta+2+x)}{(\theta+1)^{x+3}}, \quad x \in \mathbb{N}.$$

The PL distribution is derived from the Poisson compounding scheme based on the continuous Lindley distribution by [**13**]. It was proposed by [**17**] for the modeling of count data, offering a new alternative to the Poisson (P) and negative binomial models. Mathematically, the PL distribution is proved to be unimodal, overdispersed, possesses an increasing hazard rate, and satisfies the infinite divisibility property. Also, one can prove that it is defined as a special mixture of the G and negative binomial distributions. Practically, the main difference between the PL, P and negative binomial distributions is the flexibility of skewness and kurtosis that differ over the possible numerical range of values. In particular, the skewness and kurtosis of the PL distribution can be smaller than those of the negative binomial distribution. The details on the PL distribution can be found in [**7**], [**10**] and [**15**]. Now, we consider an intermediary random variable Y following the G distribution with parameter $\theta/(1+\theta)$. That is, it is defined with the following pmf:

$$f_Y(y;\theta) = \frac{\theta}{1+\theta}\frac{1}{(1+\theta)^y}, \quad y \in \mathbb{N}.$$

We use the geometric distribution to modify the PL distribution in the following way. Assuming that X and Y are independent, we focus on the distribution of the random variable $S = X + Y$, naturally called the sum PLG (SPLG) distribution with parameter θ. By the sum structure, the SPLG distribution is decomposable and some of its mathematical properties can be directly derived from those of the PL and G distributions. Also, it is clear that $S \geq X$, implying that the SPLG distribution stochastically dominates the PL distribution as stated before. Therefore, the SPLG model constitutes a new alternative to the PL model in a stochastic sense.

The aim of this paper is to explore more deeply the new horizon of applications offered by this alternative. In the first part, we provide the closed form expressions of the corresponding pmf, cumulative distribution function (cdf), hazard rate function and other mathematical properties. Among others, we prove that the SPLG distribution can be unimodal, overdispersed and possesses an increasing hazard rate. Inference for the SPLG model is examined by four different estimation methods, and a Monte Carlo simulation study is conducted to observe the performance of estimates. As a significant contribution, a new count regression analysis is introduced based on the SPLG distribution, with three practical data. The obtained results confirm the interest of the new SPLG methodology in these settings, outperforming some standard approaches of the literature.

The paper is organized as follows: In Section 2, the presentation of the SPLG distribution is completed. The parametric estimation of the SPLG model is investigated in Section 3, with a numerical simulation study in Section 4. Two practical data analyses are presented in Section 5. In Section 6, a new count regression analysis is introduced based on the SPLG distribution, with practical data examples. Section 7 contains the paper's conclusions.

2. THE SPLG DISTRIBUTION

Hereafter, S will refer to a random variable following the SPLG distribution, that can be written with the sum structure described in the previous section. The following proposition presents the corresponding pmf of S.

Proposition 1. *The pmf of S is given by*

$$f_S(x;\theta) = \frac{\theta^3}{(\theta+1)^{x+4}}\left[\frac{x^2}{2} + \left(\theta + \frac{5}{2}\right)x + \theta + 2\right], \quad x \in \mathbb{N}.$$

Proof. The support of S is $\mathbb{N}$. Hence, for $x \in \mathbb{N}$, by the independence of X and Y, we get

$$\begin{aligned} f_S(x;\theta) = P(S = x) &= \sum_{y=0}^{x} f_X(x-y;\theta) f_Y(y;\theta) \\ &= \frac{\theta^3}{(\theta+1)^{x+4}} \sum_{y=0}^{x} (\theta + 2 + x - y) \\ &= \frac{\theta^3}{(\theta+1)^{x+4}}\left[\frac{x^2}{2} + \left(\theta + \frac{5}{2}\right)x + \theta + 2\right]. \end{aligned}$$

This completes the proof of Proposition 1. □

Some plots of the SPLG distribution for different parameter values are presented in Figure 1. From Figure 1, it is observed that the pmf is unimodal ($\theta \le 1$) and right-skewed ($\theta > 1$) when x increases.

Remark 2. *Thanks to Proposition 1, the pmf of S can be written as*

$$f_S(x;\theta) = cw(x;\theta) f_Y(x;\theta),$$

where $f_Y(y;\theta)$ is the pmf of the G distribution with parameter $\theta/(1+\theta)$, $w(x;\theta) = x^2/2 + (\theta + 5/2)\,x + \theta + 2$ and $c = 1/E[w(Y;\theta)] = \theta^2/(\theta+1)^3$. Thus, the pmf of S can be viewed as a particular weighted version of the pmf of Y.

Remark 3. *Also, thanks to Proposition 1, the pmf of S can be written as*

$$f_S(x;\theta) = \frac{1}{\theta+1} g(x;\theta) + \frac{\theta}{\theta+1} h(x;\theta),$$

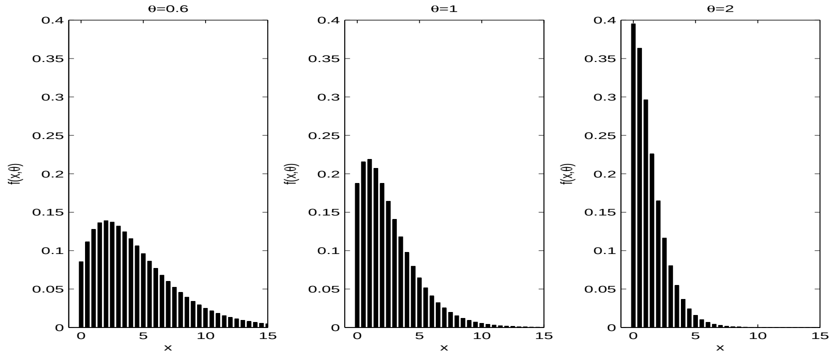

Figure 1: Plots of the pmf for selected parameter values θ

where $g(x;\theta) = \theta^3(x+2)(x+1)/[2(\theta+1)^{x+3}]$ *is the pmf of the negative binomial distribution with parameters* 3 *and* $1/(1+\theta)$ *and* $h(x;\theta) = \theta^2(x+1)/(\theta+1)^{x+2}$ *is the pmf of the negative binomial distribution with parameters* 2 *and* $1/(1+\theta)$*. Hence, the SPLG distribution is defined as the mixture of the two previous distributions, with mixing proportions* $1/(\theta+1)$ *and* $\theta/(\theta+1)$*, respectively.*

The following proportion presents the cdf of the SPLG distribution.

Proposition 4. *The cdf of* S *is given by, for any* $y \in \mathbb{N}$,

$$\begin{aligned}&F_S(y;\theta) = P(S \le y) = \\ &\frac{1}{2(\theta+1)^{y+4}}\big(2\theta^4(\theta+1)^y + 8\theta^3(\theta+1)^y - 2\theta^3 y - 4\theta^3 - \theta^2 y^2 \\ &+ 12\theta^2(\theta+1)^y - 7\theta^2 y - 12\theta^2 + 8\theta(\theta+1)^y + 2(\theta+1)^y - 2\theta y - 8\theta - 2\big).\end{aligned}$$

Proof. Thanks to Proposition 1, we can express the pmf of S as

$$\begin{aligned}&f_S(x;\theta) = \\ &\frac{\theta^3}{2(\theta+1)^6}x(x-1)\frac{1}{(\theta+1)^{x-2}} + \frac{\theta^3(\theta+3)}{(\theta+1)^5}x\frac{1}{(\theta+1)^{x-1}} + \frac{\theta^3(\theta+2)}{(\theta+1)^4}\frac{1}{(\theta+1)^x}.\end{aligned}$$

Hence, after some algebraic manipulations, we obtain

$$\begin{aligned}
F_S(y;\theta) &= \sum_{x=0}^{y} f_S(x;\theta) \\
&= \frac{\theta^3}{2(\theta+1)^6} 1_{\{y\geq 2\}} \sum_{x=2}^{y} x(x-1)\frac{1}{(\theta+1)^{x-2}} + \frac{\theta^3(\theta+3)}{(\theta+1)^5} 1_{\{y\geq 1\}} \sum_{x=1}^{y} x\frac{1}{(\theta+1)^{x-1}} \\
&+ \frac{\theta^3(\theta+2)}{(\theta+1)^4} \sum_{x=0}^{y} \frac{1}{(\theta+1)^x} \\
&= \frac{1}{2(\theta+1)^{y+4}} \big(2\theta^4(\theta+1)^y + 8\theta^3(\theta+1)^y - 2\theta^3 y - 4\theta^3 - \theta^2 y^2 \\
&+ 12\theta^2(\theta+1)^y - 7\theta^2 y - 12\theta^2 + 8\theta(\theta+1)^y + 2(\theta+1)^y - 2\theta y - 8\theta - 2\big).
\end{aligned}$$

This ends the proof of Proposition 4. □

After some operations, the hazard rate function of SPLG distribution is given by

$$h_S(x;\theta) = \frac{f_S(x;\theta)}{1 - F_S(x;\theta)} = \frac{\theta^3\,(x+1)\,(x+4+2\theta)}{2 + (2x+4)\,\theta^3 + (12 + x^2 + 7x)\,\theta^2 + (2x+8)\,\theta}.$$

The following result is about the failure rate nature of the SPLG distribution.

Theorem 5. *The hazard rate function of SPLG distribution is increasing (IFR, increasing failure rate) for all value of θ.*

Proof. Let us consider the intermediary continuous function defined by

$$u(x) = \frac{\theta^3\,(x+1)\,(x+4+2\theta)}{2 + (2x+4)\,\theta^3 + (12 + x^2 + 7x)\,\theta^2 + (2x+8)\,\theta}.$$

Then, we have

$$\begin{aligned}
w(x) = \frac{du(x)}{dx} &= \frac{2\theta^2\left(5 + 2\theta^2 x + \theta^2 x^2 + 8\theta^2 x + 8\theta x + \theta x^2\right)}{\left(2 + 2\theta^3 x + 4\theta^3 + 12\theta^2 + \theta^2 x^2 + 7\theta^2 x + 2\theta x + 8\theta\right)^2} \\
&+ \frac{2\theta^2\left(2\theta^4 + 18\theta + 2x + 11\theta^3 + 22\theta^2\right)}{\left(2 + 2\theta^3 x + 4\theta^3 + 12\theta^2 + \theta^2 x^2 + 7\theta^2 x + 2\theta x + 8\theta\right)^2}.
\end{aligned}$$

Since $\theta > 0$, we have $w(x) > 0$, implying that $u(x)$ is increasing. In particular, for any $x \in \mathbb{N}$, we have $u(x) \leq u(x+1)$, which is equivalent to $h_S(x;\theta) \leq h_S(x+1;\theta)$, implying that $h_S(x;\theta)$ is increasing with respect to x. The proof is completed. □

Some plots for different parameter values of the hazard rate function are presented in Figure 2. Also, it is concluded that Figure 2 agrees on the theory given in Theorem 5.

The result Proposition 6 determines the probability generating function of the SPLG distribution.

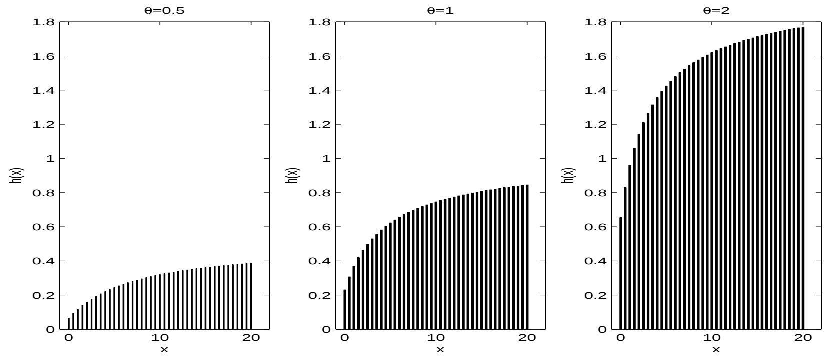

Figure 2: Plots of the hazard rate function for selected parameter values θ

Proposition 6. *The probability generating function of S is given by*

$$G_S(s;\theta) = E(s^S) = \frac{\theta^3(\theta - s + 2)}{(\theta+1)(\theta - s + 1)^3}, \quad |s| < \theta + 1.$$

Proof. Thanks to Proposition 1, we can express the pmf of S as

$$f_S(x;\theta) = \\ \frac{\theta^3}{2(\theta+1)^6} x(x-1)\frac{1}{(\theta+1)^{x-2}} + \frac{\theta^3(\theta+3)}{(\theta+1)^5} x \frac{1}{(\theta+1)^{x-1}} + \frac{\theta^3(\theta+2)}{(\theta+1)^4}\frac{1}{(\theta+1)^x}.$$

Hence,

$$\begin{aligned} G_S(s;\theta) &= \sum_{x=0}^{+\infty} s^x f_S(x;\theta) \\ &= \frac{\theta^3}{(\theta+1)^3}\left(\frac{s^2}{(\theta+1-s)^3} + \frac{(\theta+3)s}{(\theta+1-s)^2} + \frac{\theta+2}{\theta+1-s}\right) = \frac{\theta^3(\theta - s + 2)}{(\theta+1)(\theta - s + 1)^3}. \end{aligned}$$

This ends the proof of Proposition 6. □

Remark 7. *An alternative proof of Proposition 6 is as follows. We can use the definition of S. Indeed, since $S = X + Y$ with X and Y are independent, with X following the PL distribution with parameter $\theta > 0$ and Y following the G distribution with parameter $\theta/(1+\theta)$, by using [**17**], we get*

$$\begin{aligned} G_S(s;\theta) = G_X(s;\theta)G_Y(s;\theta) &= \frac{\theta^2(\theta - s + 2)}{(\theta+1)(\theta+1-s)^2} \times \frac{\theta}{\theta+1-s} \\ &= \frac{\theta^3(\theta - s + 2)}{(\theta+1)(\theta - s + 1)^3}. \end{aligned}$$

It follows from Proposition 6 that the moment generating and characteristic functions are, respectively, given by

$$M_S(t,\theta)=E(e^{tS})=\frac{\theta^3(\theta-e^t+2)}{(\theta+1)(\theta-e^t+1)^3},\quad |s|<\log(\theta+1)$$

and

$$\varphi_S(t,\theta)=E(e^{itS})=\frac{\theta^3(\theta-e^{it}+2)}{(\theta+1)(\theta-e^{it}+1)^3},\quad t\in\mathbb{R},$$

where $i=\sqrt{-1}$.

Remark 8. *The r-th derivative of $G_S(s;\theta)$ has a tractable expression; we have*

$$G_S(s;\theta)^{(r)}=\frac{(r+1)!\theta^3(2\theta+r-2s+4)}{2(\theta+1)(\theta-s+1)^{r+3}},\quad |t|<\theta+1. \tag{1}$$

In particular, one has

$$G_S(s;\theta)'=\frac{\theta^3(2\theta-2s+5)}{(\theta+1)(\theta-s+1)^4},\quad G_S(s;\theta)''=\frac{6\theta^3(\theta-s+3)}{(\theta+1)(\theta-s+1)^5}.$$

The remark above implies the following proposition.

Proposition 9. *The r-th factorial moment of S is given by*

$$\mu'_{(r)}(\theta)=E(S(S-1)\dots(S-r+1))=\frac{(r+1)!(2\theta+r+2)}{2(\theta+1)\theta^r}.$$

Proof. It follows from (1) that

$$\mu'_{(r)}(\theta)=G_S(s;\theta)^{(r)}\mid_{s=1}=\frac{(r+1)!\theta^3(2\theta+r-2s+4)}{2(\theta+1)(\theta-s+1)^{r+3}}\mid_{s=1}$$
$$=\frac{(r+1)!(2\theta+r+2)}{2(\theta+1)\theta^r}.$$

□

The next result follows from the well-known relation between the factorial and raw moments of S.

Proposition 10. *The r-th raw moment of S is given by*

$$\mu'_r(\theta)=E(S^r)=\sum_{j=0}^{r}\mathcal{S}(j,r)\frac{(j+1)!(2\theta+j+2)}{2(\theta+1)\theta^j},$$

where $\mathcal{S}(j,r)=(1/r!)\sum_{k=0}^{r-1}(-1)^k\binom{r}{k}(r-k)^j$.

Table 1: The first four moments of the SPLG distribution, along with the variance

θ	$E(S)$	$E\left(S^2\right)$	$E\left(S^3\right)$	$E\left(S^4\right)$	$Var(S)$
0.5	5.3333	45.3333	509.3333	7069.3333	16.8888
0.8	3.1944	17.7777	132.8819	1239.3923	7.5733
1	2.5000	11.5000	71.5000	557.5000	5.2500
1.5	1.6000	5.3333	24.1777	138.6666	2.7733
3	0.7500	1.5833	4.4722	16.1388	1.0208
5	0.4333	0.7133	1.5133	4.0893	0.5255

In particular, we can express the fourth first raw moments of S as

$$\mu_1'(\theta) = \frac{2\theta+3}{(\theta+1)\theta},$$

$$\mu_2'(\theta) = \frac{2\theta^2+9\theta+12}{\theta^2(\theta+1)},$$

$$\mu_3'(\theta) = \frac{2\theta^3+21\theta^2+60\theta+60}{\theta^3(\theta+1)}$$

and

$$\mu_4'(\theta) = \frac{2\theta^4+45\theta^3+228\theta^2+480\theta+360}{\theta^4(\theta+1)}.$$

In particular, the mean and variance of S are, respectively, given by

$$(2) \qquad \mu(\theta) = \frac{2\theta+3}{(\theta+1)\theta}, \quad \sigma(\theta)^2 = Var(S) = \frac{(2\theta+1)(\theta^2+3\theta+3)}{\theta^2(\theta+1)^2}.$$

Some numerical values of first four moments are presented in Table 1. It is concluded from Table 1 that the first four moments and variances decrease as θ increases.

Then, the index of dispersion of S is given by

$$D(\theta) = \frac{\sigma(\theta)^2}{\mu(\theta)} = \frac{(2\theta+1)(\theta^2+3\theta+3)}{\theta(\theta+1)(2\theta+3)}.$$

We have $D(\theta)-1 = (2\theta^2+6\theta+3)/[\theta(\theta+1)(2\theta+3)] > 0$, implying that $D(\theta) > 1$. Therefore, the SPLG distribution can be used to model overdispersed data sets.

Also, the skewness and kurtosis coefficients of S are given as

$$C_s(\theta) = E\left[\left(\frac{S-\mu(\theta)}{\sigma(\theta)}\right)^3\right] = \frac{\mu_3'(\theta) - 3\mu_2'(\theta)\mu(\theta) + 2\mu(\theta)^3}{\sigma(\theta)^3}$$

and

$$C_k(\theta) = E\left[\left(\frac{S-\mu(\theta)}{\sigma(\theta)}\right)^4\right] = \frac{\mu'_4(\theta) - 4\mu'_3(\theta)\mu(\theta) + 6\mu'_2(\theta)\mu(\theta)^2 - 3\mu(\theta)^4}{\sigma(\theta)^4},$$

respectively. By using the expressions of the fourth raw moments, we can express $C_s(\theta)$ and $C_k(\theta)$. For the sake of conciseness, we omit it.

3. ESTIMATION

In this section, the estimation of the model parameter is examined by some methods with checking their performance via simulation studies.

3.1 Method of maximum likelihood

Let $x_1, \ldots, x_n$ be the observations of n independent and identically random variables $X_1, \ldots, X_n$ from the SPLG distribution. Then, thanks to Proposition 1, the corresponding log-likelihood function is obtained as

$$\ell_n(\theta) = \sum_{i=1}^{n} \log[f_S(x_i;\theta)] = 3n\log(\theta) - \log(\theta+1)\sum_{i=1}^{n}(x_i+4) + \sum_{i=1}^{n}\log\left[\frac{x_i^2}{2} + \left(\theta+\frac{5}{2}\right)x_i + \theta + 2\right].$$

Then, the maximum likelihood estimator of θ, say $\widehat{\theta}$, is defined such that $\ell_n(\theta)$ is maximum with respect to the variable θ. Hence, it is the solution of the following non-linear equation: $d\ell_n(\theta)/d\theta = 0$, where

$$\frac{d\ell_n(\theta)}{d\theta} = \frac{3n}{\theta} - \frac{1}{\theta+1}\sum_{i=1}^{n}(x_i+4) + \sum_{i=1}^{n}\frac{x_i+1}{x_i^2/2 + (\theta+5/2)\,x_i + \theta + 2}.$$

There is no analytical solution for this non-linear equation. For given data, it can be solved numerically via any mathematical software. Also, under standard regularity conditions, for practical purposes, the underlying distribution of the estimator $\widehat{\theta}$ can be approximated by the normal distribution $\mathcal{N}(\theta, J_*)$, where $J_* = (-d^2\ell_n(\theta)/d\theta^2)^{-1}\mid_{\theta=\widehat{\theta}}$ and

$$\frac{d^2\ell_n(\theta)}{d\theta^2} = -\frac{3n}{\theta^2} + \frac{1}{(\theta+1)^2}\sum_{i=1}^{n}(x_i+4) - \sum_{i=1}^{n}\frac{(x_i+1)^2}{\left(x_i^2/2 + (\theta+5/2)\,x_i + \theta + 2\right)^2}.$$

Among the implications of this result, an asymptotic confidence interval for θ at the level $(1-\gamma)100\%$ with $\gamma \in (0,1)$ is given by

$$ic_\theta = \left[\theta^* - z_{\gamma/2}\sqrt{J_*},\, \theta^* + z_{\gamma/2}\sqrt{J_*}\right],$$

where $z_{\gamma/2}$ is the $(1-\gamma/2)$-quartile of the standard normal distribution $\mathcal{N}(0,1)$.

3.2 Method of moments

With the above setting, by considering the expression of $\mu(\theta)$ in Equation (2), the method of moment estimator $\widetilde{\theta}$ is given by

$$\widetilde{\theta} = \frac{2 - \overline{x} + \sqrt{\overline{x}^2 + 8\overline{x} + 4}}{2\overline{x}},$$

where $\overline{x} = (1/n)\sum_{i=1}^{n} x_i$, assuming that $\overline{x} > 0$.

3.3 Least and weighted least squares methods

Let $x_{(1)} < x_{(2)} < \cdots < x_{(n)}$ denote the ordered observations from the SPLG distribution. The least squares error estimator of θ can be obtained by minimizing the following function with respect to θ:

$$L(\theta) = \sum_{i=1}^{n} \left(F_S\left(x_{(i)}; \theta\right) - \frac{i}{n+1} \right)^2 .$$

Also, the weighted least squares estimator of θ is achieved by minimizing the following function with respect to θ:

$$WL(\theta) = \sum_{i=1}^{n} \frac{(n+2)(n+1)^2}{i(n-i+1)} \left(F_S\left(x_{(i)}; \theta\right) - \frac{i}{n+1} \right)^2 .$$

All minimization problem can be done via some numerical methods such as Nelder-Mead or BFGS. The BFGS is an iterative method used to solve unconstrained nonlinear optimization problems and was proposed by [**6**]. Some studies using this method can be found in [**18**] and [**19**].

3.4 Method of proportions

The method of proportions is proposed by [**11**] to estimate the parameters of discrete Weibull distribution. Here, we use the same method for estimating the parameter of the SPLG distribution. First, let us define the indicator function

$$\upsilon(x_i) = \begin{cases} 1, & x_i = 0 \\ 0, & x_i > 0 \end{cases} .$$

Then $\Upsilon = (1/n)\sum\limits_{i=1}^{n} \upsilon(x_i)$ denotes the proportion of 0's in the sample. It is clear that the random version of the empirical proportion Υ is a consistent and unbiased estimator of the probability $f_S(0;\theta) = \theta^3(2+\theta)/(\theta+1)^4$. Therefore, the proportion estimator of the parameter θ is obtained from the solution of the equation given by

$$\frac{\theta^3(2+\theta)}{(\theta+1)^4} = \Upsilon. \tag{3}$$

Numerically, Equation (3) can be solved numerically using Newton-Raphson method.

4. SIMULATION STUDY

To get information about the performance of the previous estimators, we conduct an appropriate simulation study. Hence, 5000 trials are used to estimate the bias and mean squares errors (MSEs) of the maximum likelihood estimates (MLEs), least square estimates (LSEs), weighted least square estimates (WLSEs), proportion estimates (PEs), and moment estimates (MEs). Different sample sizes and four parameter settings are considered. The results are given in Table 2. From this table, it is concluded that bias and MSEs of all estimates decrease when n increases as expected, noting that the bias converges to zero for its negative and positive values. Moreover, the MLEs and PEs are the best estimates in terms of bias and MSE.

5. MODELING PRACTICAL DATA WITH ANALYSIS

In this section, two practical data examples are carried out to show the applicability of the SPLG model compared to other well-established models. Especially, the Poisson-Lindley (PL) (see [**17**]), discrete-Weibull (DW) (see [**16**]), Discrete additive Perks-Weibull (DAPW) (see [**20**]) and G models are used to fit two real-life data sets. In order to specify the best model, we calculate the log-likelihood values (ℓ), Akaike information criterion (AIC), Bayesian information criterion (BIC), Kolmogorov-Smirnov (KS) goodness-of-fit statistic and the related p-value for all models. Method of ML is used in practical data applications and ML estimates are shown as $\widehat{p}_i$, $i = 1, 2, 3, 4$. Computations of the estimates are obtained by the **optim** routine, and all goodness-of-fit statistics are calculated by the **goftest** routine in the R introduced by [**5**].

Data set 1: The first data set consists of the 2003 final examination marks of 48 slow space students in mathematics in the Indian Institute of Technology at Kanpur. The data set is taken from [**9**] and is given in Table 3.

Data set 2: The second data set is taken from [**21**] and represents the number of menstrual cycles to pregnancy. The data were obtained retrospectively, starting from pregnancy in each case. Reference [**21**] analyzed fecundability data for a total of 586 women, contributing to a total of 1844 cycles. For this data set, the data have been combined for 12 or more cycles. The data are given in Table 5.

According to the results in Tables 4 and 6, the SPLG distribution is more flexible in fitting the considered data than PL, DW, DAPW and G distributions. In both tables, we note that the G distribution is not fitted to the corresponding data set (based on the p-value) but we use it for comparison purposes and demonstrating the quality of the SPLG distribution as a modification of G. Furthermore, the modelling ability of the DW distribution for the first data set and the DAPW distribution for the second data set was examined, and the corresponding p-values in

Table 2: Bias and MSE of the estimates for some sample sizes and parameter values

		Bias					MSEs				
θ	n	MLEs	LSEs	WLSEs	PEs	MEs	MLEs	LSEs	WLSEs	PEs	MEs
0.5	25	0.0108	0.0301	0.0301	-0.0511	0.0215	0.0059	0.0084	0.0084	0.0770	0.8249
	50	0.0051	0.0151	0.0151	-0.0154	0.0043	0.0028	0.0037	0.0037	0.0295	0.0044
	75	0.0051	0.0111	0.0111	-0.0091	0.0051	0.0018	0.0023	0.0023	0.0170	0.0018
	100	0.0024	0.0070	0.0070	-0.0043	0.0025	0.0013	0.0016	0.0016	0.0119	0.0013
	150	0.0017	0.0045	0.0045	-0.0046	0.0017	0.0008	0.0010	0.0010	0.0074	0.0008
	200	0.0013	0.0035	0.0035	-0.0024	0.0013	0.0006	0.0008	0.0008	0.0056	0.0006
	300	0.0010	0.0024	0.0024	0.0001	-0.0028	0.0004	0.0005	0.0005	0.0037	0.0067
	400	0.0007	0.0018	0.0018	-0.0016	0.0007	0.0003	0.0004	0.0004	0.0027	0.0003
	500	0.0006	0.0015	0.0015	-0.0013	-0.0007	0.0003	0.0003	0.0003	0.0022	0.0023
0.8	25	0.0253	0.0578	0.0578	0.0006	0.0376	0.0194	0.0277	0.0277	0.0899	0.1607
	50	0.0124	0.0283	0.0283	0.0012	0.0108	0.0086	0.0113	0.0113	0.0384	0.0123
	75	0.0064	0.0163	0.0163	-0.0011	0.0051	0.0055	0.0068	0.0068	0.0243	0.0085
	100	0.0063	0.0131	0.0131	0.0001	0.0064	0.0042	0.0052	0.0052	0.0188	0.0042
	150	0.0028	0.0077	0.0077	-0.0008	0.0028	0.0027	0.0034	0.0034	0.0121	0.0027
	200	0.0022	0.0059	0.0059	-0.0001	0.0022	0.0020	0.0024	0.0024	0.0088	0.0020
	300	0.0023	0.0049	0.0049	0.0019	0.0023	0.0014	0.0016	0.0016	0.0057	0.0014
	400	0.0012	0.0032	0.0032	0.0018	0.0012	0.0010	0.0012	0.0012	0.0044	0.0010
	500	0.0010	0.0025	0.0025	-0.0001	0.0010	0.0008	0.0009	0.0009	0.0035	0.0008
1	25	0.0297	0.0690	0.0690	0.0086	0.0301	0.0327	0.0472	0.0472	0.1098	0.0328
	50	0.0167	0.0361	0.0361	0.0091	0.0169	0.0157	0.0204	0.0204	0.0523	0.0158
	75	0.0100	0.0219	0.0219	0.0076	0.0101	0.0098	0.0120	0.0120	0.0331	0.0098
	100	0.0069	0.0160	0.0160	0.0012	0.0070	0.0075	0.0092	0.0092	0.0246	0.0075
	150	0.0050	0.0117	0.0117	0.0037	0.0051	0.0048	0.0057	0.0057	0.0166	0.0048
	200	0.0041	0.0088	0.0088	0.0020	0.0025	0.0036	0.0044	0.0044	0.0128	0.0074
	300	0.0037	0.0066	0.0066	0.0017	0.0033	0.0023	0.0028	0.0028	0.0081	0.0033
	400	0.0026	0.0047	0.0047	0.0008	0.0026	0.0017	0.0020	0.0020	0.0059	0.0017
	500	0.0016	0.0035	0.0035	0.0017	0.0016	0.0014	0.0016	0.0016	0.0050	0.0014
1.5	25	0.0584	0.1164	0.1164	0.0506	0.0586	0.1112	0.1543	0.1543	0.2417	0.1127
	50	0.0308	0.0575	0.0575	0.0192	0.0312	0.0463	0.0595	0.0595	0.1081	0.0464
	75	0.0167	0.0358	0.0358	0.0092	0.0169	0.0286	0.0366	0.0366	0.0699	0.0286
	100	0.0145	0.0290	0.0290	0.0104	0.0146	0.0218	0.0275	0.0275	0.0519	0.0219
	150	0.0094	0.0188	0.0188	0.0083	0.0095	0.0141	0.0166	0.0166	0.0327	0.0142
	200	0.0067	0.0129	0.0129	0.0069	0.0068	0.0102	0.0124	0.0124	0.0244	0.0102
	300	0.0044	0.0085	0.0085	0.0029	0.0045	0.0069	0.0083	0.0083	0.0168	0.0069
	400	0.0026	0.0058	0.0058	0.0002	0.0027	0.0053	0.0063	0.0063	0.0127	0.0053
	500	0.0018	0.0038	0.0038	-0.0004	0.0018	0.0040	0.0046	0.0046	0.0094	0.0040

Table 3: The first data set

29 25 50 15 13 27 15 18 7 7 8 19 12 18 5 21 15 86 21 15 14 39 15 14
70 44 6 23 58 19 50 23 11 6 34 18 28 34 12 37 4 60 20 23 40 65 19 31

Table 4: Some results for the first data set

	SPLG	G	PL	DAPW
ℓ	-197.8903	-206.8961	-198.5976	-198.8223
-2ℓ	395.7806	413.7922	397.1951	397.6446
AIC	397.7806	415.7922	399.1951	405.6446
BIC	399.6518	417.6634	401.0663	413.1294
KS	0.1068	0.2223	0.1108	0.1077
p-value	0.6435	0.0145	0.5966	0.6333
$\widehat{p}_1$	0.1119	0.0372	0.0745	13.9925
$\widehat{p}_2$				0.0009
$\widehat{p}_3$				0.0029
$\widehat{p}_4$				1.7010

both cases are significantly less than 0.05, so we do not include the DW distribution in Table 4 and the DAPW distribution in Table 6. Empirical and SPLG cdfs are provided in Figures 3 and 4; and it is observed that there is a good fit between the cdfs, that is, the new model fits the two practical data sets.

6. COUNT REGRESSION ANALYSIS

Let X be the response variable and y be associated $p \times 1$ vector of covariates. We consider that the response variable X follows the SPLG distribution with mean $\mu(y)$. Furthermore, the mean of the response variable is supposed to be linked with the explanatory variables by log-linear form, i.e., $\mu_i = \exp\left(\beta \mathbf{y}_i^T\right)$ where $\beta = (\beta_1, \beta_2, \ldots, \beta_p)$ and $\mathbf{y}_i = (1, y_{1i}, y_{2i}, \ldots, y_{pi})$. By replacing θ with $\left(2 - \mu_i + \sqrt{4 + 8\mu_i + \mu_i^2}\right)/(2\mu_i)$, we obtain the re-parameterized pmf:

$$f_*(x;\beta) = \frac{\left(2 - \mu_i + \sqrt{4 + 8\mu_i + \mu_i^2}\right)^3 / (2\mu_i)^3}{\left[(2 - \mu_i + \sqrt{4 + 8\mu_i + \mu_i^2})/(2\mu_i) + 1\right]^{x+4}} \times$$

$$\left[\frac{x^2}{2} + \left(\frac{2 - \mu_i + \sqrt{4 + 8\mu_i + \mu_i^2}}{2\mu_i} + \frac{5}{2}\right)x + \frac{2 - \mu_i + \sqrt{4 + 8\mu_i + \mu_i^2}}{2\mu_i} + 2\right].$$

The corresponding log-likelihood equation is given as

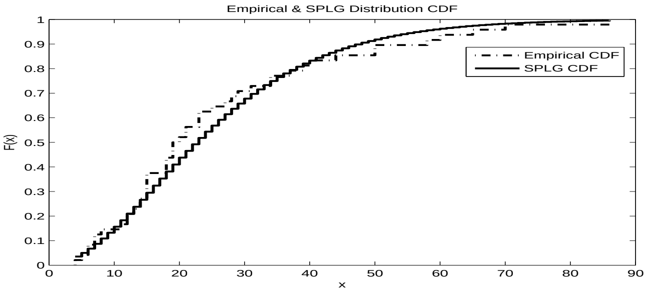

Figure 3: Empirical and estimated cdfs of the SPLG distribution based on the first data set.

Table 5: The second data set

Cycles	1	2	3	4	5	6	7	8	9	10	11	12
Number of women	227	123	72	42	21	31	11	14	6	4	7	28

$$\begin{aligned}\ell_n(\beta) &= \sum_{i=1}^{n} f_*(x_i;\beta)\\ &= 3\sum_{i=1}^{n}\log\left(\frac{2-\mu_i+\sqrt{4+8\mu_i+\mu_i^2}}{2\mu_i}\right)-\sum_{i=1}^{n}(x_i+4)\log\left(\frac{2+\mu_i+\sqrt{4+8\mu_i+\mu_i^2}}{2\mu_i}\right)\\ &+\sum_{i=1}^{n}\log\left[\frac{x_i^2}{2}+\left(\frac{2-\mu_i+\sqrt{4+8\mu_i+\mu_i^2}}{2\mu_i}+\frac{5}{2}\right)x_i+\frac{2+3\mu_i+\sqrt{4+8\mu_i+\mu_i^2}}{2\mu_i}\right].\end{aligned}$$

Then, the MLE of β, say $\widehat{\beta}$, is defined such that $\ell_n(\beta)$ is maximum with respect to the vector β. Hence, it is the solution of the following non-linear equations: $d\ell_n(\beta)/d\beta_j = 0$ with $j = 1, \ldots, p$. The vectorial solution is not in closed form and cannot be solved explicitly. Some numerical methods can be used to achieve solutions (see [**14**]). Therefore, one parameter regression models are chosen in regression applications.

We examine an application of the proposed method to analyze the number of stays after hospital admission in the USA among the elderly population, aged 65 years or more, taking into account the data obtained from [**4**]. This data application is reviewed by [**8**] for uniform Poisson (UP) and P regression models. The number of stays after hospital admission (HOSP) are taken as the response variable, the description of the explanatory variables are presented in Table 7. The regression model is fitted by SPLG, UP, and P regression models. Table 8 presents the MLE of the SPLG, UP, and P models.

Based on Table 8, we can strongly conclude that the proposed SPLG regression model is preferred over the P and UP models based on the log-likelihood ($\ell_{\max}$)

Table 6: Some results for the second data set

	SPLG	G	PL	DW
ℓ	-1291.1853	-1342.3429	-1304.3567	-1291.7149
-2ℓ	2582.3706	2684.6858	2608.7134	2583.4298
AIC	2584.3706	3686.6858	2610.7134	2587.4298
BIC	2588.7439	2691.0591	2615.0867	2596.1764
KS	0.1447	0.2225	0.1483	0.1483
p-value	0.2587	0.0172	0.0654	0.2415
$\widehat{p}_1$	0.8109	0.7588	0.5227	0.1332
$\widehat{p}_2$				1.3951

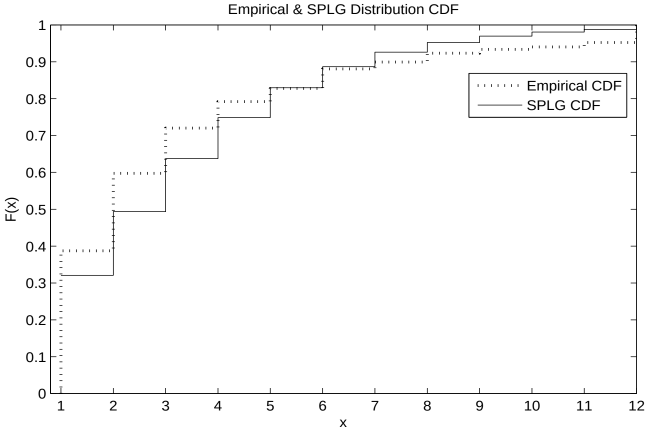

Figure 4: Empirical and estimated cdfs of the SPL distribution based on the second data set.

and AIC. According to the SPLG regression results, MALE, MARRIED, FAMINC, EMPLOYED, PRIVINS and MEDICAID explanatory variables have no statistical effect on the HOSP response variable ($p\ values > 0.01$).

However, other explanatory variables EXCLHLTH, POORHLTH, NUMCHRON and AGE have statistical effects on the HOSP response variable ($p\ values < 0.01$). Also, the number of stays after hospital admission for the self-perception of individuals with poor health is 0.6120 times higher than the self-perception of individuals with excellent health. It is seen that individuals with chronic diseases and conditions have 0.2716 times higher number of stays after hospital admission than individuals without chronic diseases and conditions. Finally, with the advancing age (aged 65 years or more), individuals stay longer after admission to the hospital. That is, age is directly proportional to the number of hospitalizations.

Table 7: Explanatory variables description

EXCLHLTH	A dummy variable which takes the value 1 if self-perceived health is excellent (x_{i1})
POORHLTH	A dummy variable which takes the value 1 if self-perceived health is poor (x_{i2})
NUMCHRON	A count variable giving the number of chronic diseases and condition (x_{i3})
AGE	Age divided by 10 (x_{i4})
MALE	A dummy variable which takes the value 1 if the patient is male (x_{i5})
MARRIED	A dummy variable for marital status (x_{i6})
FAMINC	Family income in \$10,000 (x_{i7})
EMPLOYED	A dummy variable which takes the value 1 if the patient is employed (x_{i8})
PRIVINS	A dummy variable which takes the value 1 if the patient is covered by private (x_{i9}) health insurance (x_{i10})
MEDICAID	A dummy variable which takes the value 1 if the patient is covered by Medicaid (x_{i11})

Table 8: The MLEs with standard errors (SEs) of SPLG, P and UP model parameters based on [**4**] data

	SPLG model			P model			UP model		
	Estimate (SE)	z value	p value	Estimate (SE)	z value	p value	Estimate (SE)	z value	p value
(Intercept)	-3.5499(0.3814)	-9.308	0.0000	-3.3762(0.34)	-9.930	0.0000	-3.530(0.37)	-9.541	0.0000
EXCLHLTH	-0.7172(0.1812)	-0.958	0.0001	-0.7267(0.17)	–4.275	0.0000	-0.725(0.18)	-4.028	0.0001
POORHLTH	0.6120(0.0764)	8.010	0.0000	0.6187(0.06)	10.312	0.0000	0.627(0.07)	8.957	0.0000
NUMCHRON	0.2716(0.0210)	12.933	0.0000	0.2636(0.02)	13.180	0.0000	0.274(0.02)	13.700	0.0000
AGE	0.2009(0.0481)	4.177	0.0000	0.1787(0.04)	4.468	0.0000	0.197(0.04)	4.925	0.0000
MALE	0.1548(0.0687)	2.253	0.0242	0.1317(0.06)	2.195	0.0282	0.154(0.06)	2.567	0.0103
MARRIED	-0.0445(0.0718)	-0.620	0.5354	-0.0391(0.06)	-0.652	0.5146	-0.043(0.07)	-0.614	0.5390
FAMINC	0.0051(0.0107)	0.477	0.6336	0.0072(0.01)	0.720	0.4715	0.005(0.01)	0.500	0.6171
EMPLOYED	0.0251(0.1140)	0.220	0.8257	0.0220(0.10)	0.220	0.8259	0.023(0.11)	0.209	0.8344
PRIVINS	0.1870(0.0860)	2.174	0.0297	0.1970(0.07)	2.814	0.0049	0.200(0.08)	2.500	0.0124
MEDICAID	0.2262(0.1120)	2.020	0.0434	0.2359(0.09)	2.621	0.0088	0.227(0.11)	2.064	0.0391
ℓ	-2915.422			-3042.828				-2951.330	
AIC	5852.845			6107.657				5924.660	

7. CONCLUSIONS

This study proposes a new discrete distribution with support on $\mathbb{N}$. Some distributional properties are studied, such as moments, index of dispersion, coefficients of skewness and kurtosis. Four estimators are examined to estimate the one model parameter. Extensive simulation studies for four parameter settings are conducted. The flexibility of the new model is demonstrated using two practical datasets. A new count regression model is introduced based on the new distribution. Then, it is applied to a medical data set and it is observed that the new model is a competitive to the current models. In a future study, the proposed distribution will be used to construct integer-valued time series models.

REFERENCES

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

2. D. Bhati, H. S. Bakouch: *A new infinitely divisible discrete distribution with applications to count data modeling.* Communications in Statistics - Theory and Methods, **48** (6) (2019), 1401–1416.

3. C. Chesneau, H. S. Bakouch, Y. Akdogan, K. Karakaya: *The Binomial-Discrete Poisson-Lindley Model: Modeling and Applications to Count Regression.* Commun. Math. Res., **38** (1) (2022), 28–51.

4. P. Deb, P.K. Trivedi: *Demand for medical care by the elderly: A finite mixture approach.* J. Appl. Econ., **12** (3) (1997),313–336.

5. J. Faraway, G. Marsaglia, J. Marsaglia, A. Baddeley: *goftest: Classical goodness-of-fit tests for univariate distributions.* R package version 1-0, (2017).

6. R. Fletcher: *Practical methods of optimization.* John Wiley and Sons, 2013.

7. M.E. Ghitany, D. K. Al-Mutairi: *Estimation methods for the discrete Poisson-Lindley distribution.* Journal of Statistical Computation and Simulation, **79** (2009),1–9.

8. E. Gomez-Deniz: *Another generalization of the geometric distribution.* Test, **19** (2010), 399–415.

9. R.D. Gupta, D. Kundu: *A new class of weighted exponential distributions.* Statistics, **43** (6) (2009), 621–630.

10. D. Karlis, E. Xekalaki: *Mixed Poisson distributions.* International Statistical Review, **73** (2005), 35–58.

11. M. S. A. Khan, A. Khalique, A. M. Abouammoh: *On estimating parameters in a discrete Weibull distribution.* IEEE Trans Reliab., **38** (3) (1989), 348–350.

12. C. Kus, Y. Akdogan, A. Asgharzadeh, I. Kinaci, K. Karakaya: *Binomial-discrete Lindley distribution.* Communications Faculty of Sciences University of Ankara Series A1 Mathematics and Statistics, **68** (1) (2018), 401–411.

13. D. V. Lindley: *Fiducial distributions and Bayes' theorem.* Journal of the Royal Statistical Society. Series B (Methodological), (1958), 102–107.

14. Y. Ma, W. Gui: *Pivotal inference for the inverse Rayleigh distribution based on general progressively Type-II censored samples.* Journal of Applied Statistic, **46** (2019), 71–797.

15. M. Mohammadpour, H. S. Bakouch, M. Shirozhan: *Poisson-Lindley INAR(1) model with applications.* Brazilian Journal of Probability and Statistics, **32** (2018), 262–280.

16. T. Nakagawa, S.Osaki: *Discrete Weibull distribution.* IEEE Trans Reliab., **24** (1975), 300–301.

17. M. Sankaran: *The discrete Poisson-Lindley distribution.* Biometrics, **26** (1970), 145–149.

18. C. Tanis: *On Transmuted power function distribution: characterization, risk measures, and estimation.* Journal of New Theory, **34** (2021), 72–81.

19. C. Tanis, B. Saracoglu, C. Kus, A. Pekgor: *Transmuted complementary exponential power distribution: properties and applications.* Cumhuriyet Science Journal, **41** (2) (2020), 419–432.

20. A. Tyagi, N. Choudhary, B. Singh: *Discrete additive Perks–Weibull distribution: properties and applications.* Life Cycle Reliability and Safety Engineering **8** (2019), 183–199.

21. C.R. Weinberg, B.C. Gladen *The beta-geometric distribution applied to comparative fecundability studies.* Biometrics **42** (1986), 547–560.

(Received 14. 01. 2021.)
(Revised 26. 09. 2022.)

Hassan S. Bakouch
Department of Mathematics,
Faculty of Science, Tanta University,
Tanta, Egypt,
Department of Mathematics,
College of Science, Qassim University,
Buraydah, Saudi Arabia
E-mail: *hassan.bakouch@science.tanta.edu.eg*
hnbakouch@gmail.com

Kadir Karakaya
Department of Statistic,
Faculty of Science, Selçuk University,
Konya, Turkey
E-mail: *kkarakaya@selcuk.edu.tr*

Christophe Chesneau
LMNO,
Université de Caen,
Campus II, Science 3, Caen, France
E-mail: *christophe.chesneau@unicaen.fr*

Yunus Akdoğan
Department of Statistic,
Faculty of Science, Selçuk University,
Konya, Turkey
E-mail: *yakdogan@selcuk.edu.tr*