Revista Colombiana de Estadística - Theoretical Statistics
January 2025, volume 48, issue 1, pp. 39–70
http://doi.org/10.15446/rce.v48n1.115319

Introducing the Discrete xLindley Distribution: A One-Parameter Model for Overdispersed Data

La distribución discreta de xLindley: un modelo de un parámetro para datos sobredispersos

João Debastiani Neto[1,a], Ricardo Puziol de Oliveira[2,b], Fernando Antonio Moala[2,c], Jorge Alberto Achcar[3,d]

[1]Department of Science, Maringá State University, Maringá, Brazil

[2]Department of Statistics, State University of São Paulo, Presidente Prudente, Brazil

[3]Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, Brazil

Abstract

In this paper, we propose a new discrete model, the discrete analog of the xLindley distribution, as an alternative for modeling overdispersed data. The model was derived using the method of infinite series, allowing us to capture complex characteristics of the data, and its properties were studied in detail. Asymptotic results are presented to validate the model parameter estimates consistency in large samples. Additionally, a Bayesian approach was considered for inference with complete and right-censored data. The performance of the Bayesian estimators was evaluated through Monte Carlo simulations, enabling a comprehensive comparison of the effectiveness and efficiency of the estimators under different scenarios. The proposed model was applied to two real datasets, demonstrating its practical utility. The practical application included the analysis of discrete events in research environments, highlighting the model's flexibility in various situations. Furthermore, a comparison with other discrete distributions was provided, showcasing the advantages of the xLindley model over existing alternatives.

***Key words*:** Count data; Discretization methods; xLindley distribution; Data dispersion; Bayesian inference; Simulation study.

[a]Ph.D. E-mail: jdneto@uem.br
[b]Ph.D. E-mail: rpuziol.oliveira@gmail.com
[c]Ph.D. E-mail: f.moala@unesp.br
[d]Ph.D. E-mail: achcar@fmrp.usp.br

Resumen

En este artículo, proponemos un nuevo modelo discreto, el análogo discreto de la distribución xLindley, como una alternativa para modelar datos sobredispersos. El modelo fue derivado utilizando el método de series infinitas, lo que nos permite capturar características complejas de los datos, y se estudiaron en detalle sus propiedades. Se presentan resultados asintóticos para validar la consistencia del modelo en grandes muestras. Además, se consideró un enfoque bayesiano para la inferencia con datos completos y censurados a la derecha. El rendimiento de los estimadores bayesianos se evaluó a través de simulaciones de Monte Carlo, lo que permitió una comparación integral de la efectividad y eficiencia de los estimadores en diferentes escenarios. El modelo propuesto se aplicó a dos conjuntos de datos reales, demostrando su utilidad práctica. La aplicación práctica incluyó el análisis de eventos discretos en entornos de investigación, destacando la flexibilidad del modelo en diversas situaciones. Además, se proporcionó una comparación con otras distribuciones discretas, mostrando las ventajas del modelo xLindley sobre alternativas existentes.

***Palabras clave*:** Datos de recuento; Métodos de discretización; Distribución xLindley; Dispersión de datos; Inferencia bayesiana; Estudio de simulación.

1. Introduction

Recently, it is observed in the literature many studies on the integration of new probabilistic models through the discretization of continuous random variables. The goal of discretization is to create probability distributions that can be applied to strictly discrete data. In survival analysis, it is common to use continuous distributions to model discrete data, for instance, the data consist of the number of cycles of a product before failure (breakage) or the number of weeks it took rats painted with a carcinogen to develop carcinoma.

The use of continuous distributions to model discrete data is a common practice in many studies, as evidenced in the work of Klein & Moeschberger (1997), where these distributions are employed in survival analyses. The choice of continuous models is due to their simplicity and broad applicability, allowing for good estimates of discrete events despite the idealizations regarding the nature of the data. This approach is emphasized by the works of Meeker & Escobar (1998) focusing on industrial reliability and Kalbfleisch & Prentice (2002) focusing on lifetime modeling in medical studies, assuming data in the presence of censoring and truncation.

Many other studies are introduced in the literature assuming continuous distributions in survival data analysis, such as the study by Lee & Wang (2003) that presents risk models applied to biomedical data with applications in chronic diseases and possible treatments and Lawless (2003) assuming exponential and Weibull distributions for failure time and reliability data in industrial applications.

Collett (2003), in turn, applies discretization to analyze survival data, especially in countable events, highlighting the use of cumulative distribution functions

to estimate probabilities associated with these events, thus providing a robust statistical framework for clinical analysis. Additionally, Hamada et al. (2008) introduce a Bayesian approach to modeling the reliability of complex systems, addressing uncertainty in parameter estimation and enabling the adaptation of continuous distributions to discrete data, thereby enhancing the understanding and application of statistical models across various fields.

An overview of discretization methods for continuous distributions and some discretized distributions is introduced by Chakraborty (2015) who explores techniques such as the infinite series method and a quantile-based approach where properties and characteristics of various discretized distributions are examined, including discrete versions of some popular probability distributions such as the normal distribution, the exponential distribution, and the lognormal distribution. This study also highlights how discretized distributions can capture details and variations in observational data not captured by continuous distributions, thus expanding the applicability of these distributions in various data analysis contexts.

The first method of discretization introduced in the literature is based on defining a probability mass function through an infinite series. This concept was first presented by Good (1953) with the discrete Good distribution to model species population frequencies.

Other authors followed this line, including Haight (1957), who developed his work focusing on queue modeling with balking, a phenomenon in which customers decide not to enter a queue if it is too long. Using the discrete Pearson III distribution, through an approach based on infinite series, Haight (1957) modeled the frequency of arrival and balking events, allowing for a more accurate analysis of the variability in waiting times and balking in service systems.

The study conducted by Siromoney (1964) introduced the Dirichlet Series distribution, emphasizing the modeling of the frequency of rainy days. The proposed approach was able to capture the variability associated with these events, providing a robust model that meets the demands of meteorological applications requiring discrete representations. This line of reasoning is supported by the work of Kemp (1997), which focused on characterizing the discrete normal distribution. The discretization of the continuous normal distribution is particularly relevant for modeling data that follows a normal distribution in discrete contexts, such as event counting.

Following this perspective, Sato et al. (1999) introduced a consistent formula for the discrete exponential distribution, applied to defect metrology in wafers, highlighting the importance of modeling frequencies of discrete events in quality control in semiconductor manufacturing processes. Complementarily, Bi et al. (2001) addressed the discrete log-normal distribution, revealing its applicability in mining massive and skewed data. The study demonstrated that this discretized version allows for a more precise analysis of phenomena following log-normal patterns in countable data sets.

Furthermore, the research of Inusah & Kozubowski (2006) on a discrete analogue of the Laplace distribution, alongside the investigation by Kozubowski & Inusah (2006), which introduced a skew discrete Laplace distribution, significantly

expanded the tools available for modeling data with asymmetry and dispersion characteristics. These developments highlight the relevance of modeling discrete distributions across various disciplines, particularly in engineering. Finally, Doray & Luong (1997) enhanced statistical inference techniques for the family of distributions proposed by Good (1953), while Kemp (2008) introduced the discrete half-normal distribution, broadening the options for modeling discrete data. Together, these investigations provide a comprehensive and dynamic view of the discretization of distributions and their practical implications in different contexts.

This discretization method by infinite series is defined as follows:

Definition 1. Let X be a continuous random variable. If X has pdf $f_X(x;\boldsymbol{\theta})$ with support on $\mathbb{R}$, then the corresponding discrete random variable Y has probability mass function (pmf) given by

$$\mathrm{P}(Y=y;\boldsymbol{\theta}) = \frac{f_X(y;\boldsymbol{\theta})}{\sum_{j=-\infty}^{\infty} f_X(j;\boldsymbol{\theta})}, \qquad y \in \mathbb{Z},$$

where $\boldsymbol{\theta}$ is the vector of parameters indexing the distribution of X.

The main goal of this paper is to derive discrete analog for the xLindley distribution, which is a one-parameter lifetime model introduced and studied by Chouia & Zeghdoudi (2021), using the infinite series method. The proposed new model can be a suitable alternative to model overdispersed count datasets. A continuous random variable X is said to have xLindley distribution if its probability density function (pdf) can be written as

$$f_X(x;\theta) = \frac{\theta^2(2+\theta+x)}{(1+\theta)^2} e^{-\theta x}, \qquad x \in \mathbb{R}_+, \tag{1}$$

where $\theta \in \mathbb{R}_+$ is the shape parameter. Chouia & Zeghdoudi (2021) shown that this model can be derived as a 2-component mixture of an Exponential distribution with mean θ^{-1} and a Gamma distribution with shape parameter 3 and scale parameter θ, with mixing proportions given by $\alpha(1+\alpha)^{-1}$ and $(1+\alpha)^{-1}$, respectively. A comprehensive discussion on the probabilistic properties of the xLindley distribution such as moments, hazard function, entropies, stochastic orderings, parameter estimation, among others is also presented on the mentioned paper. The corresponding survival function of X is given by

$$S_X(x;\theta) = \left(1 + \frac{\theta x}{(1+\theta)^2}\right) e^{-\theta x}, \qquad x \in \mathbb{R}_+, \tag{2}$$

for $\theta \in \mathbb{R}_+$. This distribution is one of several generalizations of the Lindley distribution proposed in the literature in the last years such as the power Lindley distribution introduced by Ghitany et al. (2013), the weighted Lindley distribution proposed by Ghitany et al. (2011), the quasi-Lindley distribution introduced by Shanker & Mishra (2013), the inverse Lindley distribution proposed by

Sharma et al. (2015), the transmuted Lindley distribution proposed by Merovci (2013), the inverse power Lindley proposed by Barco et al. (2017), the discrete Lindley studied by Oliveira et al. (2017), a class of bivariate Lindley distributions introduced by Oliveira et al. (2021) and many others.

Therefore, based on the discretization method presented above, we propose a new discrete distribution, named discrete xLindley distribution (DXL), with a pmf given by:

$$P(X = x; \theta) = g(\theta)\,(2 + \theta + x)\,\mathrm{e}^{-\theta\,(x+1)}, \tag{3}$$

for $\theta \in \mathbb{R}_+$ and

$$g(\theta) = \frac{\left(\mathrm{e}^{\theta} - 1\right)^2}{(2 + \mathrm{e}^{\theta})\mathrm{e}^{\theta} - (1 + \theta)}. \tag{4}$$

Note that the equation (3) is a proper pmf since,

$$\sum_{x=0}^{\infty} P(X = x; \theta) = g(\theta)\left[\sum_{x=0}^{\infty}(2 + \theta + x)e^{-\theta(x+1)}\right] = g(\theta)g^{-1}(\theta) = 1. \tag{5}$$

In Figure 1, we illustrate the behavior of pmf in (3) for different values of θ for which we can see that the proposed distribution is unimodal. Furthermore, it also satisfies the log-concave inequality $\mathrm{P}^2(X = x) \geqslant \mathrm{P}(X = x - 1)\mathrm{P}(X = x + 1)$ for $x \geqslant 1$ which implies unimodality (Keilson & Gerber, 1971).

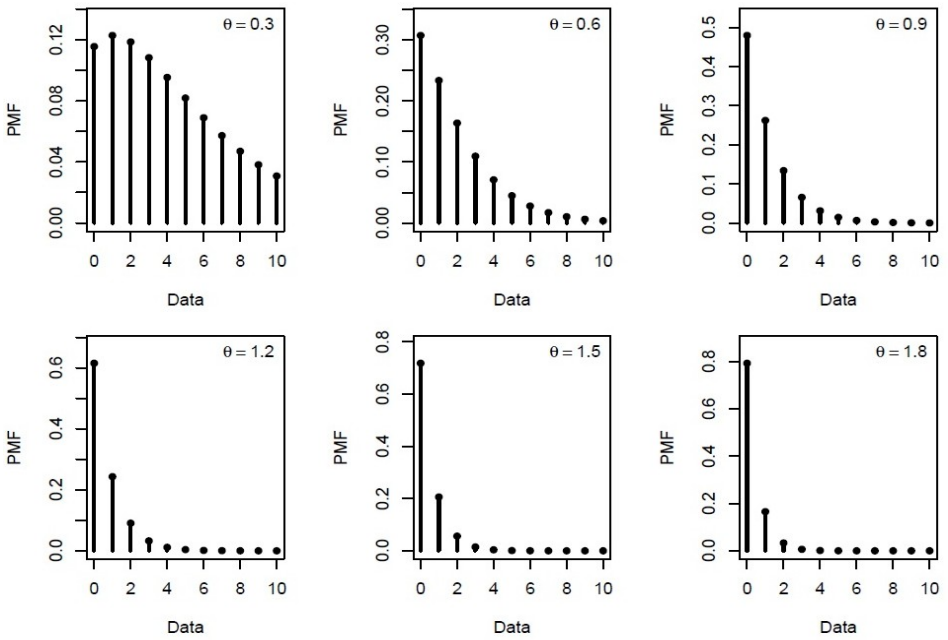

FIGURE 1: Behavior of the probability mass function for the DXL distribution assuming different values for θ ($\theta = 0.3 \to \theta = 1.8$).

It is important to point out that the discrete xLindley model proposed in this study offers a robust solution for modeling overdispersed count data, a challenge that many existing discrete distributions are not suitable for. Traditional adaptations of continuous distributions to discrete settings frequently fall short in accurately capturing overdispersion. This model bridges this gap by presenting a discrete counterpart to the xLindley distribution, specifically tailored for overdispersed count data. Its application is especially pertinent in fields such as survival analysis and reliability studies, where overdispersed discrete data are commonly encountered and demand effective modeling approaches.

The structure of this paper in as follows: in Section 2, the key probabilistic characteristics of the newly proposed DXL distribution are outlined. In Section 3, the inference methods for the model parameter are presented. In Section 5, the results of a comprehensive simulation study are shown to evaluate the properties of the newly proposed model parameters. In Section 6, two practical applications of the proposed model to real data sets highlight its usefulness. Finally, in Section 7, some concluding remarks are provided.

2. Probabilistic Properties

In this section, we present a comprehensive study of the probabilistic properties of the discrete xLindley distribution. These properties include its survival and cumulative functions (Subsection 2.1), its shape (Subsection 2.2), the hazard function (Subsection 2.3), the quantile function (Subsection 2.4), the moment properties (Subsection 2.5), the zero-modification measure (Subsection 2.6), the heavy-tail index (Subsection 2.7), stochastic orderings (Subsection 2.8), and the reliability measure (Subsection 2.9).

2.1. Survival and Cumulative Functions

Proposition 1. *The survival function of the DXL distribution is given by,*

$$P(X > x; \theta) = \left[\frac{3 + \theta + x}{(e^{\theta} - 1)^2} - \frac{(2 + \theta + x)e^{-\theta}}{(e^{\theta} - 1)^2} \right] e^{-\theta x} g(\theta), \tag{6}$$

for $\theta > 0$.

Proof. Straightforward noticing that the series $\sum_{X>x} P(X = x; \theta)$ is an convergent series and the sum converges to the expression given in Equation (6) for all $\theta > 0$. □

Remark. Since survival and cumulative functions are complementary functions, the cumulative function of the DXL distribution is given by,

$$P(X \leqslant x; \theta) = 1 - \left[\frac{3 + \theta + x}{(e^{\theta} - 1)^2} - \frac{(2 + \theta + x)e^{-\theta}}{(e^{\theta} - 1)^2} \right] e^{-\theta x} g(\theta), \tag{7}$$

for $\theta > 0$.

2.2. Mode

Proposition 2. *The mode of DXL distribution is given by,*

$$x_0 = \begin{cases} \left\lfloor \dfrac{(1+\theta)e^{\theta} - (2+\theta)}{1-e^{\theta}} \right\rfloor, & \text{if} \quad \dfrac{(1+\theta)e^{\theta} - (2+\theta)}{1-e^{\theta}} \notin \mathbb{Z}_+ \\ \dfrac{(2+\theta)e^{\theta} - (3+\theta)}{1-e^{\theta}}, & \text{if} \quad \dfrac{(2+\theta)e^{\theta} - (3+\theta)}{1-e^{\theta}} \in \mathbb{Z}_+, \end{cases} \tag{8}$$

where $\lfloor \cdot \rfloor$ *is the floor function. From Equation* (8)*, we have*

i) $P(X = x+1;\theta) < P(X = x;\theta)$ *if* $x > x_0$*;*

ii) $P(X = x+1;\theta) = P(X = x;\theta)$ *if* $x = x_0$*;*

iii) $P(X = x+1;\theta) > P(X = x;\theta)$ *if* $x < x_0$*.*

Proof. Since the DXL distribution is unimodal, its pmf satisfies the following inequalities:

$$P(X = x;\theta) \geqslant P(X = x-1;\theta), \forall x \leqslant x_0,$$

and,

$$P(X = x;\theta) \geqslant P(X = x+1;\theta), \forall x \geqslant x_0,$$

where x_0 is the mode of DXL distribution. Therefore,

$$P(X = x;\theta) \geqslant P(X = x-1;\theta) \Leftrightarrow x \leqslant \frac{(1+\theta)e^{\theta} - (2+\theta)}{1-e^{\theta}},$$

and,

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

Thus, the proof is complete since the relations are straightforward from the inequalities above. □

2.3. Hazard Function

Proposition 3. *The hazard rate of DXL distribution is an increasing function.*

Proof. The hazard rate is defined by

$$h(x;\theta) = \frac{P(X = x;\theta)}{P(X > x;\theta)} = \frac{(2+\theta+x)e^{-\theta(x+1)}(e^{\theta}-1)^2}{(3+\theta+x)e^{-\theta x} - (2+\theta+x)e^{-\theta(x+1)}},$$

for $\theta \in \mathbb{R}_+$. Taking the limit with $x \to \infty$, we have,

$$\lim_{x\to\infty} h(x;\theta) = e^{\theta} - 1,$$

that is, the hazard rate is an increasing function as θ increases which concludes the proof. □

2.4. Quantile Function

Proposition 4. *The quantile function, $Q(u)$, of the DXL distribution is given by,*

$$Q(u) = \left\lfloor \frac{2+\theta-e^{\theta}(3+\theta)}{(e^{\theta}-1)} + W_{-1}\{e^{-k(\theta)}k(\theta)(1-u)\} \right\rfloor, \tag{9}$$

where $\lfloor\cdot\rfloor$ is the floor function; $W_{-1}\{\cdot\}$ is the Lambert W function (Jodra, 2010) with negative branch and $k(\theta) = \left[\{(2+\theta)e^{\theta}-\theta-1\}\theta\right] / \left[e^{\theta}-1\right]$.

Proof. The quantile function is defined by,

$$F(Q(u)) = u \Leftrightarrow \left[\frac{(2+\theta+Q(u))e^{-\theta}}{(e^{\theta}-1)^2} - \frac{3+\theta+Q(u)}{(e^{\theta}-1)^2}\right] e^{-\theta Q(u)} g(\theta) = 1-u,$$

for $0 < u < 1$, that is,

$$\left[(2+\theta)e^{-\theta} - (3+\theta) + (e^{\theta}-1)Q(u)\right] e^{-\theta Q(u)} = \left[e^{\theta} + (1+\theta)(e^{\theta}-1)\right](1-u).$$

Setting $Z(u) = (2+\theta)e^{-\theta} - (3+\theta) + (e^{\theta}-1)Q(u)$, we have,

$$Z(u)e^{Z(u)} = \left[e^{\theta} + (1+\theta)(e^{\theta}-1)\right](1-u)e^{\left\{\frac{\theta\left[(2+\theta)e^{\theta}-(3+\theta)-1\right]}{e^{\theta}-1}\right\}}.$$

Therefore, the solution for $Z(u)$ is,

$$Z(u) = W\left\{\left[e^{\theta} + (1+\theta)(e^{\theta}-1)\right](1-u)e^{\left\{\frac{\theta\left[(2+\theta)e^{\theta}-(3+\theta)-1\right]}{e^{\theta}-1}\right\}}\right\},$$

where $W\{\cdot\}$ is the Lambet W function (Jodra, 2010; Barco et al., 2017). Now, setting $k(\theta) = \dfrac{\{(2+\theta)e^{\theta}-\theta-1\}\theta}{e^{\theta}-1}$ and inverting $Z(u)$, we have,

$$Q(u) = \left\lfloor \frac{2+\theta-e^{\theta}(3+\theta)}{(e^{\theta}-1)} + W_{-1}\{e^{-k(\theta)}k(\theta)(1-u)\} \right\rfloor, \tag{10}$$

where $\lfloor\cdot\rfloor$ is the floor function; $W_{-1}\{\cdot\}$ is the Lambert W function with negative branch. Hence, the proof. □

2.5. k^{th} Moment

Proposition 5. *The k^{th} moment of the DXL distribution is given by,*

$$E(X^k) = \frac{\left\{polylog\left(-1-k, \mathrm{e}^{-\theta}\right) + polylog\left(-k, \mathrm{e}^{-\theta}\right)(2+\theta)\right\}\left(\mathrm{e}^{\theta}+\mathrm{e}^{-\theta}-2\right)}{(2+\theta)\,\mathrm{e}^{\theta}-\theta-1}. \tag{11}$$

Proof. By definition, $\mu'_k = E(X^k)$ then

$$\mu'_k = E(X^k) = \sum_{x=0}^{\infty} x^k \cdot g(\theta) \cdot (2+\theta+x) e^{-\theta(x+1)}.$$

Assuming $g(\theta) = \dfrac{\left(\mathrm{e}^{\theta}-1\right)^2}{\mathrm{e}^{\theta}+(1+\theta)\left(\mathrm{e}^{\theta}-1\right)}$, we have

$$\mu'_k = \frac{1}{(2+\theta)e^{\theta}-\theta-1} \cdot \sum_{x=0}^{\infty} x^k \cdot (e^{\theta}-1)^2 \cdot (2+\theta+x) e^{-\theta(x+1)}.$$

Using some properties of the summations, in addition to the application of distributive and potentiating properties in the summation, we obtain

$$\begin{aligned}
\mu'_k &= \frac{1}{(2+\theta)\,\mathrm{e}^{\theta}-\theta-1}\Bigg[\theta\,\mathrm{e}^{\theta}\,polylog\left(-k,\mathrm{e}^{-\theta}\right)+\mathrm{e}^{\theta}\,polylog\left(-1-k,\mathrm{e}^{-\theta}\right)\\
&+ 2\,\mathrm{e}^{\theta}\,polylog\left(-k,\mathrm{e}^{-\theta}\right)-2\,\theta\,polylog\left(-k,\mathrm{e}^{-\theta}\right)-2\,polylog\left(-1-k,\mathrm{e}^{-\theta}\right)\\
&- 4\,polylog\left(-k,\mathrm{e}^{-\theta}\right)+\mathrm{e}^{-\theta}\theta\,polylog\left(-k,\mathrm{e}^{-\theta}\right)+\mathrm{e}^{-\theta}\,polylog\left(-1-k,\mathrm{e}^{-\theta}\right)\\
&+ 2\,\mathrm{e}^{-\theta}\,polylog\left(-k,\mathrm{e}^{-\theta}\right)\Bigg],
\end{aligned}$$

where *polylog* is the general polylogarithm function (Xu et al., 2016). Thus,

$$E(X^k) = \frac{\left\{polylog\left(-1-k,\mathrm{e}^{-\theta}\right)+polylog\left(-k,\mathrm{e}^{-\theta}\right)(2+\theta)\right\}\left(\mathrm{e}^{\theta}+\mathrm{e}^{-\theta}-2\right)}{(2+\theta)\,\mathrm{e}^{\theta}-\theta-1},$$

and the proof is complete. □

Proposition 6. *Let X be a discrete random variable according to a DXL distribution with parameter $\theta \in \mathbb{R}_+$ and $h_1(\theta) = (2+\theta)\,e^{\theta}-\theta-1$ and $h_2(\theta) = \left[e^{\theta}(1+\theta)-\theta\right](1+\theta)\,e^{\theta}$. The equations of the mean (μ), variance (σ^2), coefficient of variation (γ), skewness ($\sqrt{\beta_1}$) and kurtosis (β_2) are given, respectively, by*

$$\begin{aligned}
\mu &= \frac{(3+\theta)\,\mathrm{e}^{\theta}-\theta-1}{\left(\mathrm{e}^{\theta}-1\right)h_1(\theta)};\\
\sigma^2 &= \frac{(3+\theta)\,h_1(\theta)\mathrm{e}^{2\,\theta}-h_2(\theta)}{\left(\mathrm{e}^{\theta}-1\right)^2\left[h_1(\theta)\right]^2};\\
\gamma &= \frac{\left(\mathrm{e}^{\theta}-1\right)h(\theta)}{(3+\theta)\,\mathrm{e}^{\theta}-\theta-1}\sqrt{\frac{(3+\theta)\,h(\theta)\mathrm{e}^{2\,\theta}-h_2(\theta)}{\left(\mathrm{e}^{\theta}-1\right)^2\left[h(\theta)\right]^2}};
\end{aligned}$$

$$\begin{aligned}\sqrt{\beta_1} &= \left\{\frac{(3\,\theta+17)\,\mathrm{e}^{2\,\theta}+(3+\theta)\,\mathrm{e}^{3\,\theta}+(-3\,\theta+5)\,\mathrm{e}^{\theta}-1-\theta}{\left(\mathrm{e}^{\theta}-1\right)^{3}h_1(\theta)}\right\}\\ &\times \left\{\frac{(3+\theta)\,h_1(\theta)\mathrm{e}^{2\theta}-h_2(\theta)}{\left(\mathrm{e}^{\theta}-1\right)^{2}\left[h_1(\theta)\right]^{2}}\right\}^{-\frac{3}{2}};\\ \beta_2 &= -\frac{\left[(-\theta-3)\mathrm{e}^{4\,\theta}-(10\theta+46)\,\mathrm{e}^{3\,\theta}-66\,\mathrm{e}^{2\,\theta}+(10\,\theta\,-6)\,\mathrm{e}^{\theta}+\theta+1\right]}{\left\{-(3+\theta)\,h_1(\theta)\mathrm{e}^{2\,\theta}+\left[h_2(\theta)\right]\right\}^{2}\left[h_1(\theta)\right]^{-3}}.\end{aligned}$$

Proof. By the simple definition of mean (μ), variance (σ^2), coefficient of variation (γ), skewness ($\sqrt{\beta_1}$) and kurtosis (β_2). □

A normalized measure of dispersion can be obtained by using the variance-to-mean relationship. This measure is the well-known index of dispersion (ID) which, in this case, is given by

$$ID = \frac{\sigma^2}{\mu} = \frac{(3+\theta)\,h_1(\theta)\mathrm{e}^{2\theta}-h_2(\theta)}{\left(\mathrm{e}^{\theta}-1\right)h_1(\theta)\left[h_1(\theta)+e^{\theta}\right]}. \tag{12}$$

where $h_j(\theta), j = 1, 2$ are defined in Proposition 6.

2.6. Zero-Modification Measure

Another useful measure is the zero-modification (ZM) index which is defined based on the Poisson distribution. This index is interpreted as:

- $ZM > 0$ indicates zero-inflation.
- $ZM < 0$ indicates zero-deflation.
- $ZM = 0$ indicates no zero-modification.

For the xLindley distribution with parameter $\theta \in \mathbb{R}_+$, the ZM index is given by the expression:

$$ZM = 1 + \left(e^{\theta}-1\right)^{3}\left\{g(\theta)\right\}^{-1}\left\{\frac{\ln\left[(2+\theta)e^{-\theta}g(\theta)\right]}{k(\theta)}\right\} \tag{13}$$

where $k(\theta) = (3+\theta)\,e^{\theta}-\theta-1$. From (13), we observe that ZM $\to 0$ as $\theta \to \infty$ and ZM $\to 1$ as $\theta \to 0$. This implies, besides the usual case (ZM $= 0$), that the DXL distribution is suitable to deal with zero-inflation but is not indicated to model zero-deflated datasets. Further, it is clear that the coefficient of skewness and the coefficient of kurtosis are decreasing as θ decreases. The asymmetry degree and the flatness of a distribution can be measured by its coefficients of skewness and kurtosis, respectively. These coefficients are essential to characterize the shape of any distribution. Simultaneous large values for the mean, variance and ID (index of dispersion) are obtained when θ is small. Table 1 summarizes, for selected values of θ, the nature and the behavior of these coefficients along with the measures previously presented.

TABLE 1: Theoretical descriptive statistics under DXL distribution.

	Measures						
θ	Mean	Variance	ID	CV	ZM	Skewness	Kurtosis
0.30	4.9962	20.8170	4.1665	0.9132	0.5681	6.1262	21.7198
0.60	1.9228	4.6218	2.4037	1.1181	0.3865	5.1048	18.6762
0.90	1.0071	1.8141	1.8012	1.3373	0.2713	4.5517	17.2863
1.20	0.6009	0.9042	1.5048	1.5825	0.1933	4.2363	16.7269
1.50	0.3848	0.5139	1.3353	1.8627	0.1390	4.0811	16.7868
1.80	0.2574	0.3166	1.2300	2.1861	0.1006	4.0545	17.4372

2.7. Heavy-Tail Index

Proposition 7. *The DXL distribution has heavy tails when θ tends to zero.*

Proof. The heavy-tail (HT) index is defined by

$$\mathrm{HT} = \lim_{x\to\infty} \frac{\mathrm{P}\left(X = x+1;\theta\right)}{\mathrm{P}\left(X = x;\theta\right)},$$

for $\theta \in \mathbb{R}_+^2$. For the DXL distribution, one can easily obtain $\mathrm{HT} = e^{\theta}$. A discrete distribution is said to have heavy tails if $\mathrm{HT} \to 1$ when $x \to \infty$. Hence,

$$\lim_{\theta\to 0} \mathrm{HT} = \lim_{\theta\to 0} e^{\theta} = 1,$$

which concludes the proof. □

2.8. Stochastic Orderings

Proposition 8. *Let X be a random variable with probability mass function given by in Equation (3), and let Y be a geometric random variable with the probability mass function $P_Y(X = x;\theta) = e^{-\theta x}(1 - e^{-\theta})$ then the likelihood ratio order is $Y \leqslant_{lr} X$ and $L(x;\theta) = P_X(X = x;\theta)/P_Y(X = x;\theta)$ is an increasing function in x. Also, the stochastic order is $Y \leqslant_{st} X$, the hazard rate order is $Y \leqslant_{hr} X$ and the expectation order is $Y \leqslant_{E} X$.*

Proof. Since,

$$L(x;\theta) = \frac{(2+\theta+x)(e^{\theta}-1)}{(2+\theta)e^{\theta} - (\theta+1)},$$

we have $L(x) \leqslant L(x+1), \forall\theta > 0$. Therefore, the proof is complete. □

2.9. Reliability Measure

Proposition 9. *Suppose X and Y are independent DXL random variables with parameters θ_1 and θ_2, respectively. The stress-strength parameter, $R = P(X < Y)$, is given by,*

$$\begin{aligned} R &= \Big\{A(\theta_1,\theta_2+1)e^{3\theta_2+2\theta_1} - A(\theta_1,\theta_2)e^{2\theta_2+2\theta_1} + B(\theta_1,\theta_2)e^{\theta_1+\theta_2} \\ &- B(\theta_1,\theta_2+1)e^{\theta_1+2\theta_2} + (1+\theta_1)C(\theta_2)\Big\}\frac{(e^\theta-1)^2}{(e^{\theta_1+\theta_2}-1)^3\,C(\theta_1)C(\theta_2)} \end{aligned}$$

where $A(\theta_1,\theta_2) = (2+\theta_1)(2+\theta_2)$, $B(\theta_1,\theta_2) = (2\theta_2+3)\theta_1 + 3\theta_2 + 3$ *and* $C(\theta_i) = (2+\theta_i)e^{\theta_i} - \theta_i - 1$.

Proof. The stress-strength parameter is defined by

$$\begin{aligned} R &= \sum_{x=0}^{\infty} P(X=x;\theta_1)P(X>x;\theta_2) \\ &= \sum_{x=0}^{\infty} \frac{\left\{(3+\theta_2+x)e^{-\theta_2 x} - (2+\theta_2+x)e^{-\theta_2(x+1)}\right\}\{2+x+\theta_1\}e^{\theta_1(x+1)}}{\{(2+\theta_2)e^{\theta_2}-\theta_2-1\}\{g(\theta_1)\}^{-1}}, \end{aligned}$$

for $\theta_1, \theta_2 \in \mathbb{R}_+^2$. Setting

- $A(\theta_1,\theta_2) = (2+\theta_1)(2+\theta_2)$.
- $B(\theta_1,\theta_2) = (2\theta_2+3)\theta_1 + 3\theta_2 + 3$
- $C(\theta_i) = (2+\theta_i)e^{\theta_i} - \theta_i - 1$

hence the proof. □

3. Inference Methods

3.1. Complete Data

Let $\mathbf{X} = (X_1,\ldots,X_n)$ be a random sample of size n from the DXL distribution and $\boldsymbol{x} = (x_1,\ldots,x_n)$ its observed values. The log-likelihood function of θ can be expressed as

$$\ell_n(\theta;\boldsymbol{x}) = n\ln[g(\theta)] - n\theta(\overline{x}+1) + \sum_{i=1}^{n}\ln(2+\theta+x_i), \tag{14}$$

where $\overline{x}$ is the sample mean and $g(\theta)$ is defined in (3). The MLE $\widehat{\theta}$ of θ can be obtained by direct maximization of the log-likelihood function (14). Hence, the component of the score vector, U_θ, is given by

$$U_\theta = \frac{\partial \ell_n(\theta;\boldsymbol{x})}{\partial\theta} = \frac{ng'(\theta)}{g(\theta)} - n(\overline{x}+1) + \sum_{i=1}^{n}\frac{1}{2+\theta+x_i}.$$

There is no closed-form solution for the MLE of θ, and therefore, standard optimization algorithms such as Newton-Raphson, BFGS or Nelder-Mead based

methods may be used to obtain numerical estimates. On other hand, under suitable regularity conditions (Lehmann & Casella, 1998), the asymptotic distribution of the MLE $\widehat{\theta}$ is a univariate Normal distribution with mean θ and variance $\widehat{\sigma}_\theta$, which can be consistently estimated by the inverse of the observed Fisher's information given by,

$$\mathcal{I}_0(\theta) = \left[\; U_{\theta\theta} \; \right],$$

where

$$U_{\theta\theta} = n\left\{\frac{g''(\theta)}{g(\theta)} - \left(\frac{g'(\theta)}{g(\theta)}\right)^2\right\} - \sum_{i=1}^{n} \frac{1}{(2+\theta+x_i)^2}.$$

However, under the maximum likelihood theory, a consistent estimator for the variance of $\widehat{\theta}$ is obtained by the inverse of the Fisher information of θ, evaluated at $\theta = \widehat{\theta}$, i.e., $\widehat{\sigma}_{\boldsymbol{\theta}} = \mathcal{I}_E^{-1}(\theta)$. Thus, assuming complete data, the expected Fisher's information for the proposed DXL model is given by,

$$\mathcal{I}_E(\theta) = \left[\; n\left\{\left(\frac{g'(\theta)}{g(\theta)}\right)^2 - \frac{g''(\theta)}{g(\theta)}\right\} + g(\theta)\left\{\Phi\left(\frac{1}{e^\theta}, 1, \theta\right) e^\theta - \frac{e^\theta}{\theta} - \frac{1}{1+\theta}\right\} \; \right],$$

where $\Phi(\cdot)$ is the Lerch transcendent function (Hassani, 2007; Ferreira et al., 2017). Finally, in order to obtain interval estimates for the parameter θ, one can use large sample approximations to get the $100 \times (1-\alpha)\,\%$ two-sided CIs as $\widehat{\theta} \pm z_{1-\alpha/2}\sqrt{\mathcal{I}_E^{-1}(\theta)}$ where $z_{1-\alpha/2}$ is the upper $(\alpha/2)^{th}$ percentile of the standard Normal distribution.

3.2. Right-Censored Data

Let us consider the situation when the lifetime, X_i, is not completely observed and may be subject to right censoring and let C_i be the censoring time for the ith individual. From a sample of size n, it is observed $X_i = \min\{X_i, C_i\}$ and an indicator variable $\delta_i = I(X_i < C_i)$, where

- $\delta_i = 1$ if X_i is a complete observed lifetime.
- $\delta_i = 0$ if it is a right censored lifetime.

In this case, the log-likelihood function assuming the DXL distribution is given by,

$$\begin{aligned} \ell_n(\theta; \boldsymbol{x}) &= \{nr + n(1-r)\}\ln[g(\theta)] - \{nr + n\overline{x}\}\theta + 2n(1-r)\ln\left(e^\theta - 1\right) \\ &+ \sum_{i=1}^{n}(1-\delta_i)\ln\left(3+\theta+x_i - (2+\theta+x_i)e^{-\theta}\right) \\ &+ \sum_{i=1}^{n}\delta_i \ln(2+\theta+x_i), \end{aligned} \tag{15}$$

where $\overline{x}$ is the sample mean and $r = \sum_{i=1}^{n}\delta_i$ is the number of uncensored observations.

In the same way as in complete data, the MLE $\widehat{\theta}$ for the unknown parameter θ is obtained by maximizing the log-likelihood function defined in (15) from which we can see that there is no closed-form for the MLE. In addition, under suitable regularity conditions, the observed Fisher's information is given by,

$$\mathcal{I}_0(\theta) = \left[\; U_{\theta\theta} \; \right],$$

where

$$\begin{aligned} U_{\theta\theta} &= \{nr + n(1-r)\} \left\{ \frac{g''(\theta)}{g(\theta)} - \left(\frac{g'(\theta)}{g(\theta)} \right)^2 \right\} - \frac{2n(1-r)e^{\theta}}{\left(e^{\theta}-1\right)^2} \\ &+ \sum_{i=1}^{n} \frac{(1-\delta_i)\left\{1 + e^{-2\theta} + \left[\theta^2 + (2x_i+5)\theta + x_i^2 + 5x_i + 2\right] e^{-\theta}\right\}}{\{(2+\theta+x_i)e^{-\theta} - (x+\theta+3)\}^2} \\ &- \sum_{i=1}^{n} \frac{\delta_i}{(2+\theta+x_i)^2}. \end{aligned}$$

In this case, a consistent estimator for the variance of $\widehat{\theta}$ is also obtained by the inverse of the Fisher information of θ, evaluated at $\theta = \widehat{\theta}$, i.e., $\widehat{\sigma}_{\boldsymbol{\theta}} = \mathcal{I}_E^{-1}(\theta)$. Thus, assuming the right-censored data, the expected Fisher's information for the proposed DXL model is given by,

$$\begin{aligned} \mathcal{I}_E(\theta) &= \{nr + n(1-r)\} \left\{ \left(\frac{g'(\theta)}{g(\theta)} \right)^2 - \frac{g''(\theta)}{g(\theta)} \right\} + \frac{2n(1-r)e^{\theta}}{\left(e^{\theta}-1\right)^2} \\ &+ nrg(\theta) \left\{ \Phi\left(\frac{1}{e^{\theta}}, 1, \theta \right) e^{\theta} - \frac{e^{\theta}}{\theta} - \frac{1}{1+\theta} \right\} \\ &+ (n-r)g(\theta) \left\{ \frac{(2+\theta)\left\{e^{2\theta} + 1 + \left(\theta^2 + 5\theta + 2\right) e^{\theta}\right\}}{\left[(3+\theta)\, e^{\theta} - \theta - 2\right]^2 e^{\theta}} \right\} \\ &\times {}_6\mathrm{F}_5\left\{ \left[1, 3+\theta, A(\theta), A(\theta), -B(\theta)+\theta+\frac{7}{2}, B(\theta)+\theta+\frac{7}{2}\right], \right. \\ &\qquad \left. \left[\theta+2, C(\theta), C(\theta), B(\theta)+\theta+\frac{5}{2}, -B(\theta)+\theta+\frac{5}{2}\right], \frac{1}{e^{\theta}} \right\}, \end{aligned}$$

where $\Phi(\cdot)$ is the Lerch transcendent function and ${}_pF_q$ is the generalized hypergeometric function (Virchenko et al., 2001) where p is number of operands of the first argument, q is number of operands of the second argument, and,

$$\begin{aligned} A(\theta) &= \frac{(3+\theta)\,\mathrm{e}^{\theta} - \theta - 2}{\mathrm{e}^{\theta} - 1} \\ B(\theta) &= \frac{\sqrt{\left[17e^{\theta} - 4(e^{2\theta}+1)\right]\mathrm{e}^{-\theta}}}{2} \\ C(\theta) &= \frac{(4+\theta)\,\mathrm{e}^{\theta} - \theta - 3}{\mathrm{e}^{\theta} - 1}. \end{aligned}$$

Similarly to the uncensored case, one can use large sample approximations to get the $100 \times (1-\alpha)\,\%$ two-sided CIs as $\widehat{\theta} \pm z_{1-\alpha/2}\sqrt{\mathcal{I}_E^{-1}(\theta)}$ where $z_{1-\alpha/2}$ is the upper $(\alpha/2)^{th}$ percentile of the standard Normal distribution.

3.3. Bayesian Inference

The Byesian paradigm is based on specifying a probability model for the observed data D, given a vector of unknown parameters η (assuming η is a random variable) and provides a rational method for updating the new information using the Bayes' rule and prior distributions for the uncertainty about η. Thus, in this work, we have adopted a squared loss function, $L(\eta, a) = (\eta - a)^2$, to determine the Bayesian estimators for complete and right-censored data. In addition, in presence of covariates, we have adopted the following regression structure for θ:

$$\theta_i = \exp\left(\beta_0 + \sum_{j=1}^{n} \beta_j x_{ji}\right), \tag{16}$$

where $\boldsymbol{\beta}$ is the vector of regression parameters. Since in some cases, there is no expert's information available to justify the choice of informative priors for the model parameters, we have to specify prior distributions that represent weak information, such as proper distributions with large variance. For this work, the weak prior distributions adopted are given by,

$$\begin{aligned}
\theta &\sim \text{Gamma}(0.001, 0.001) \text{ (Complete and Right-censored Data - Scenario 1).}\\
\theta &\sim \text{Uniform}(0, 10) \text{ (Complete and Right-censored Data - Scenario 2).}\\
\theta &\sim \text{Jeffreys's prior (Complete and Right-censored Data - Scenario 3).}\\
\boldsymbol{\beta} &\sim \text{Normal}_{q+1}\left(\mathbf{0}, 10^2\boldsymbol{\mathcal{I}}_q\right) \text{ (Regression Data).}
\end{aligned}$$

where $\boldsymbol{\beta}$ is the vector of regression parameters and $\boldsymbol{\mathcal{I}}_{q+1}$ is an identity matrix of order $q+1$. The Bayesian approach for the estimation of parameters from DXL distribution can be considered by writing the unnormalized joint posterior distribution of the vector $\boldsymbol{\zeta}$ of all parameters as,

$$\begin{aligned}
\pi(\boldsymbol{\zeta}; \boldsymbol{t}) &\propto \exp\left\{\ell(\boldsymbol{\zeta}; \boldsymbol{t})\right\}\pi(\theta) \quad \text{(Complete Data).}\\
\pi(\boldsymbol{\zeta}; \boldsymbol{t}, \boldsymbol{\delta}) &\propto \exp\left\{\ell(\boldsymbol{\zeta}; \boldsymbol{t}, \boldsymbol{\delta})\right\}\pi(\theta) \quad \text{(Right-censored Data).}\\
\pi(\boldsymbol{\zeta}; \boldsymbol{t}, \boldsymbol{\delta}) &\propto \exp\left\{\ell(\boldsymbol{\zeta}; \boldsymbol{t}, \boldsymbol{\delta})\right\}\pi(\boldsymbol{\beta}) \quad \text{(Regression Data).}
\end{aligned} \tag{17}$$

Inferences for the components of the vector $\boldsymbol{\zeta}$ are entirely based on the marginal posterior densities, which can be obtained by integrating the expressions in Equation (17), apart from the normalizing constant. However, deriving analytical expressions for these densities is infeasible, mainly due to the complexity of the associated log-likelihood function. In this case, we may resort to usual Markov Chain Monte Carlo (MCMC) methods (Gelfand & Smith, 1990; Chib & Greenberg, 1995) to drawn pseudo-random samples for the marginal *posterior* densities. In this work, we have adopted the Metropolis-within-Gibbs (MwG) sampler, whose steps are described in the following.

- Step 1: Choose an initial value $\boldsymbol{\zeta}^{(0)}$ for $\boldsymbol{\zeta}$. Denote $\boldsymbol{\zeta}$ at the kth step as $\boldsymbol{\zeta}^{(k)}$;
- Step 2: Generate $\theta^{(k)}$ from $\pi(\theta; \boldsymbol{t})$ for complete data model;
- Step 3: Generate $\theta^{(k)}$ from $\pi(\theta; \boldsymbol{t}, \boldsymbol{\delta})$ for right-censored data model;
- Step 4: Generate $\boldsymbol{\beta}^{(k)}$ from $\pi(\boldsymbol{\beta}; \boldsymbol{t}, \boldsymbol{\delta})$ for regression model;
- Step 5: Repeat Steps 2-4 N times;
- Step 6: Compute the component-wise Monte Carlo *posterior* estimate of $\boldsymbol{\zeta}$ as

$$\widehat{\boldsymbol{\zeta}} = \frac{1}{N-b} \sum_{k=b}^{N-1} \boldsymbol{\zeta}^{(k+1)},$$

where b is the burn-in period.

4. Model Diagnostics

4.1. Gelman-Rubin Diagnostic

The Gelman-Rubin diagnostic, also known as the $\widehat{R}$ statistic, is a commonly used method for assessing the convergence of Markov Chain Monte Carlo (MCMC) simulations. It evaluates convergence by comparing the variance within multiple chains to the variance between chains. When multiple MCMC chains are initiated from different starting values, this diagnostic examines whether the chains are converging to the same stationary distribution. In this way, for each parameter, two types of variances are calculated:

- **Within-chain variance** (W): This measures the variability within each individual chain.
- **Between-chain variance** (B): This measures the variability between the means of the different chains.

The Gelman-Rubin statistic $\widehat{R}$ is the square root of the ratio of the estimated variance of the parameter (a weighted average of within- and between-chain variances) to the within-chain variance:

$$\widehat{R} = \sqrt{\frac{\widehat{V}}{W}},$$

where $\widehat{V}$ is an estimate of the total variance, combining both between-chain and within-chain components. For this measure, if $\widehat{R} \approx 1$, it indicates that the between-chain and within-chain variances are approximately equal, suggesting that the chains have likely converged. Otherwise, if $\widehat{R} > 1.1$, it implies that the chains may not have fully converged and may still be exploring different regions of the parameter space.

4.2. Geweke Diagnostic

The Geweke diagnostic is another convergence assessment tool for MCMC simulations, focusing on the behavior of a single chain. It is based on comparing the means of different sections of the chain to detect whether the chain has reached equilibrium. In this case, the chain is divided into two segments:

- **Early segment**: Typically the first 10% of the chain.
- **Late segment**: Typically the last 50% of the chain.

The Geweke diagnostic computes a z-score-like statistic by comparing the mean of the parameter in the early and late segments:

$$Z = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\text{Var}(\overline{X}_1) + \text{Var}(\overline{X}_2)}},$$

where $\overline{X}_1$ and $\overline{X}_2$ represent the means of the early and late portions of the chain, and $\text{Var}(\overline{X}_1)$ and $\text{Var}(\overline{X}_2)$ are their variances. For this measure, if the chain has converged, the Geweke statistic Z should be approximately normally distributed with mean 0 and variance 1, as the early and late means should not differ significantly.

4.3. Heidelberger-Welch Diagnostic

The Heidelberger-Welch diagnostic is a widely-used tool for assessing the convergence of Markov Chain Monte Carlo (MCMC) simulations. It evaluates whether a chain has reached stationarity and tests if the samples can be treated as coming from a stationary distribution. This diagnostic operates in two phases: a stationarity test and a half-width test. That is,

- **Stationarity Test:** The stationarity test checks if the MCMC chain has converged to its stationary distribution. This is done by applying a Cramér-von Mises test to the null hypothesis that the first portion of the chain is drawn from the same distribution as the later portion. The test iteratively discards early samples (burn-in period) until the remaining sequence passes the test, indicating convergence.
- **Half-Width Test:** After passing the stationarity test, the half-width test assesses the accuracy of the estimates based on the remaining samples. It calculates the mean and confidence interval (CI) for each parameter and verifies whether the half-width of the confidence interval is within a predetermined tolerance (relative to the mean). This ensures that the sample size is sufficient for reliable estimation.

4.4. Deviance Information Criteria (DIC)

There are many methods for Bayesian model selection that are useful for comparing competing models. The most popular method is the Deviance Information Criterion (DIC), which works simultaneously to measure the model's fit and complexity. The DIC criterion is defined as

$$\text{DIC} = \mathbb{E}_{\boldsymbol{\theta}}\left[\text{D}\left(\boldsymbol{\theta}\right)\right] + p_{\text{D}} = \underline{\text{D}}\left(\boldsymbol{\theta}\right) + p_{\text{D}},$$

where $\text{D}(\boldsymbol{\theta}) = -2\ell(\boldsymbol{\theta}; \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z})$ is the deviance function and $p_{\text{D}} = \underline{\text{D}}(\boldsymbol{\theta}) - \text{D}(\widehat{\boldsymbol{\theta}})$ is the effective number of model parameters, where $\widehat{\boldsymbol{\theta}}$ is the posterior expected value. Noticeably, we are not able to compute the expectation of $\text{D}(\boldsymbol{\theta})$ over $\boldsymbol{\theta}$ analytically. Therefore, an approximate Monte Carlo estimator for such a measure is

$$\widehat{\underline{\text{D}}}\left(\boldsymbol{\theta}\right) = -\frac{2}{B}\sum_{i=1}^{B}\ell\left(\boldsymbol{\theta}_i; \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{z}\right),$$

and so the DIC can be estimated by

$$\widehat{\text{DIC}} = 2\widehat{\underline{\text{D}}}\left(\boldsymbol{\theta}\right) - \text{D}(\widehat{\boldsymbol{\theta}}).$$

4.5. Cox-Snell Residuals

Model validation procedures play an essential role when evaluating the suitability of any fitted model. In general, residual metrics are widely used in such a context, being those measures responsible for indicating departures from the underlying model assumptions by quantifying the data variability that the fitted model did not accommodate. In this way, we will consider here a popular residual metric proposed by Cox & Snell (1968), which can be straightforwardly used in our context to assess the appropriateness of the proposed model when used in the analysis of real datasets. The Cox-Snell residuals are defined by

$$r_i = -\log\left[S(t_i)\right].$$

If the obtained model fit is adequate, then the Cox-Snell residuals should follow an Exponential distribution with mean equals 1 (if T has survival distribution $S(t)$, then $-\log[S(T)] \sim \exp(1)$. However, if a survival time is right-censored, then the corresponding Cox-Snell residual, say r_i^+, is lower than r_i, which was obtained from an uncensored observation with the same lifetime. These modified residuals were derived by assuming that the difference between the cumulative hazard functions, $H(t_i)$ and $H(t_i^+)$, also follow exp(1) distributions. Thus, the modified Cox-Snell residuals for censored observations are defined by

$$r_i^+ = 1 - \log\left[S(t_i)\right] \quad \text{or} \quad r_i^+ = \log(2) - \log\left[S(t_i)\right].$$

5. Monte Carlo Simulation Study

In this section, by using $B = 10,000$ Monte Carlo simulation, we evaluated the bias (B) and the mean squared error (MSE) of the Bayesian estimators of $\widehat{\theta}$ of the DXL distribution under complete data. To run the simulation, we have considered $\theta = 0.3, \ldots (0.3) \ldots, 1.8$ for sample sizes ranging from 10 to 100 by 10 and the prior distributions specified in Section 3.3 for θ. The inverse-transform method for discrete distributions (Devroye, 2006) was implemented to generate the pseudo-random samples according to the steps:

- Step 1: Generate $U \sim Uniform(0, 1)$.
- Step 2: Define X by $F(X-1) = \sum_{i<X} p_i < U \leq \sum_{i \leq X} p_i = F(X)$ where $P(X = i) = p_i$. Set $X = 0$ and $S = p_0$.
- Step 3: While $U > S$, do $X = X + 1$ and $S = S + p_X$.
- Step 4: Return X.

The simulation process was performed `R` software (R Development Core Team, 2017) coupled with `JAGS` software via using `R2jags` package (Su & Yajima, 2012). The quantities of interest were estimated by,

- $\text{BIAS}(\widehat{\theta}) = B^{-1} \sum_{i=1}^{B} (\widehat{\theta}_i - \theta)$;
- $\text{MSE}(\widehat{\theta}) = B^{1/2} \sum_{i=1}^{B} (\widehat{\theta}_i - \theta)^2$;

The boxplot of the biases and mean squared error for the Bayesian estimators are presented in Figures 2, 4, 6 for biases, and Figures 3, 5, 7 for mean squared error. From the obtained results, in each scenario, we observe that the bias of $\widehat{\boldsymbol{\theta}}$ converges to zero when the sample size increases, as well as, the mean squared error of $\widehat{\boldsymbol{\theta}}$, independent of prior choice. These result indicates a good performance for the Bayesian estimators for parameter θ.

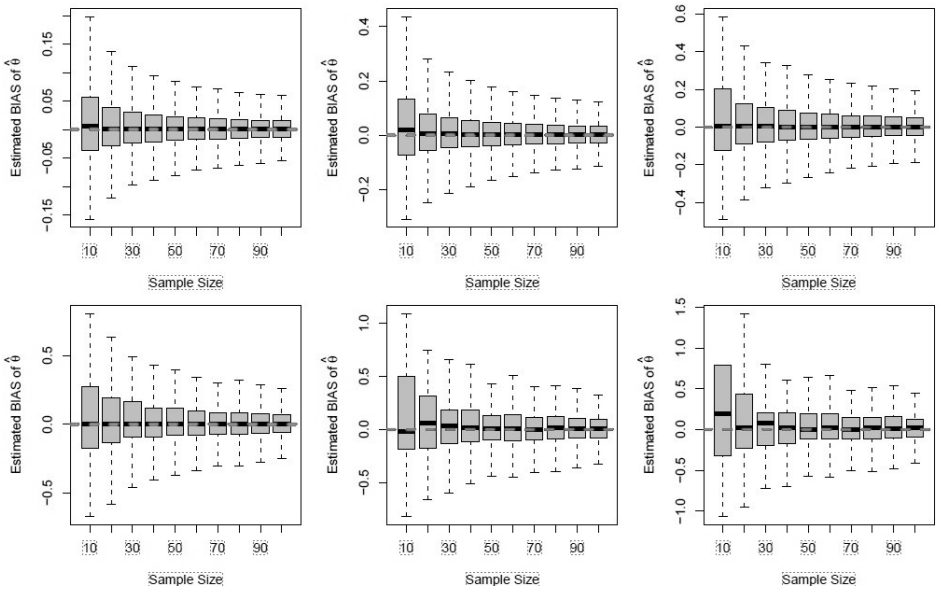

FIGURE 2: Estimated bias for θ ($\theta = 0.3 \rightarrow \theta = 1.8$) for discrete DXL distribution under the assumption of $Gamma(0.001, 0.001)$ prior distribution.

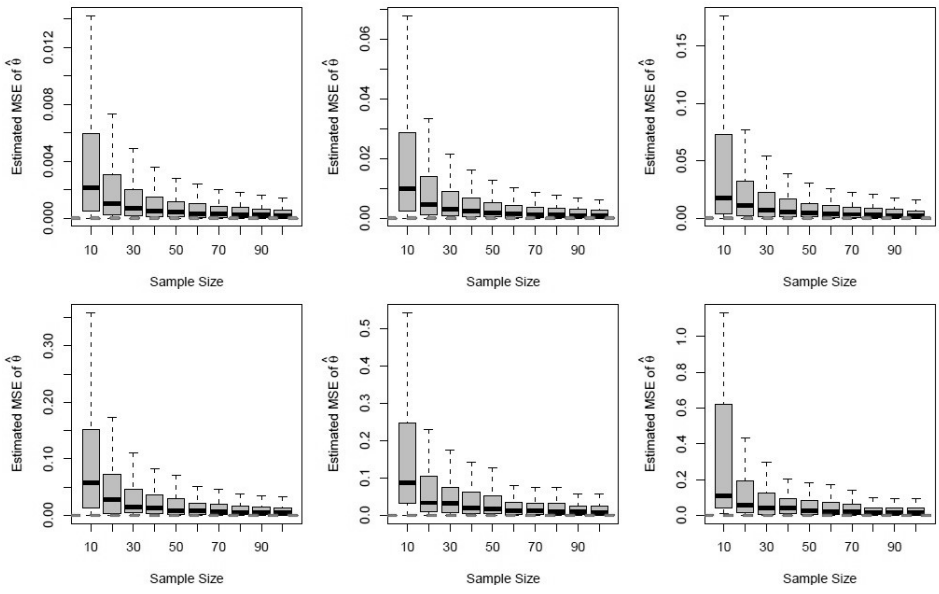

FIGURE 3: Estimated mse for θ ($\theta = 0.3 \rightarrow \theta = 1.8$) for DXL distribution under the assumption of $Gamma(0.001, 0.001)$ prior distribution.

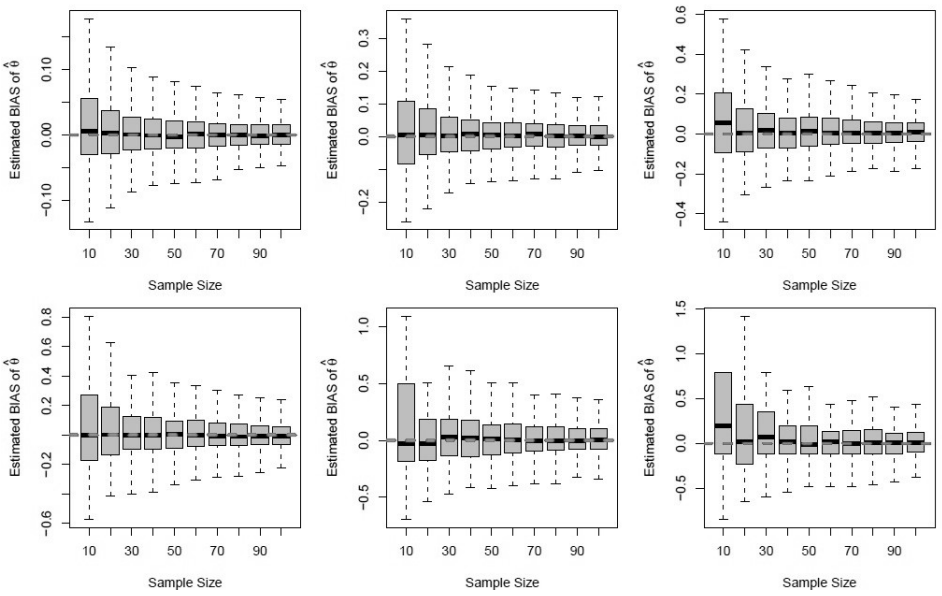

FIGURE 4: Estimated bias for θ ($\theta = 0.3 \rightarrow \theta = 1.8$) for discrete DXL distribution under the assumption of $Uniform(0, 10)$ prior distribution.

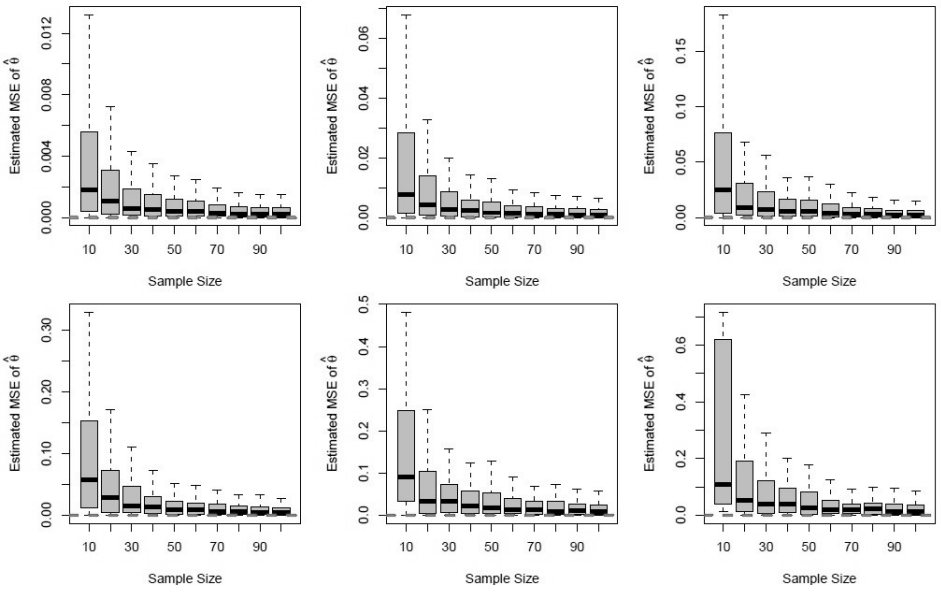

FIGURE 5: Estimated mse for θ ($\theta = 0.3 \rightarrow \theta = 1.8$) for DXL distribution under the assumption of $Uniform(0, 10)$ prior distribution.

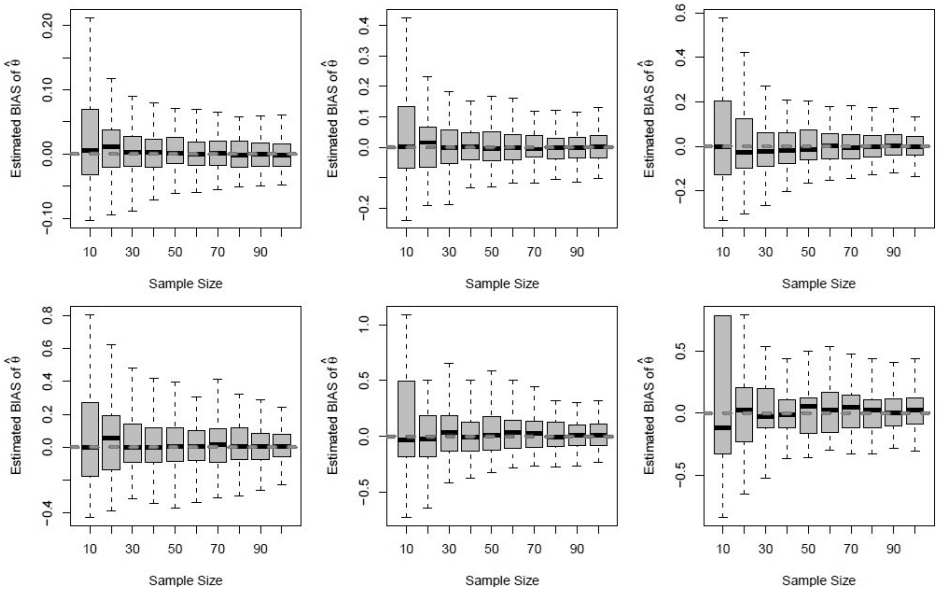

FIGURE 6: Estimated bias for θ ($\theta = 0.3 \rightarrow \theta = 1.8$) for discrete DXL distribution under the assumption of Jeffrey's prior distribution.

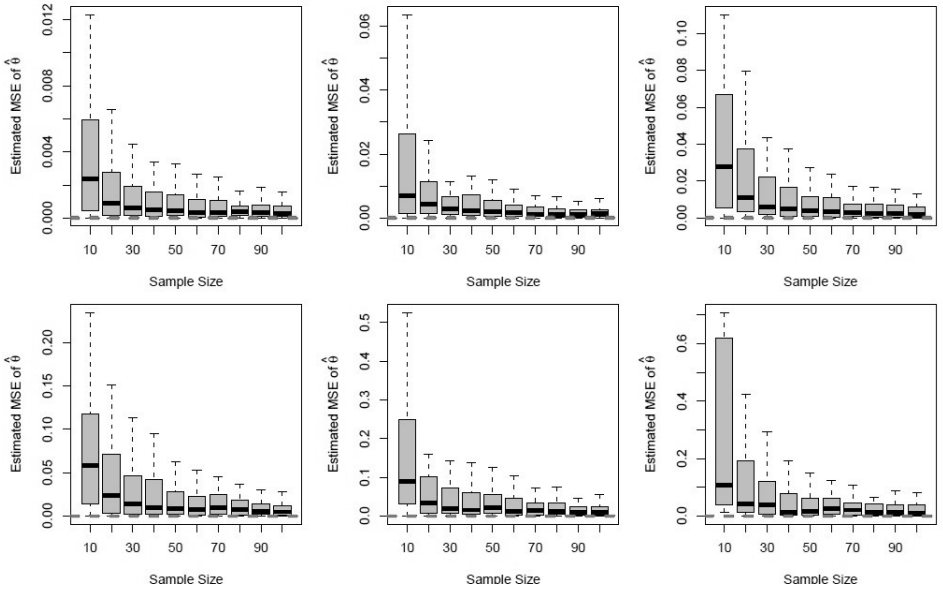

FIGURE 7: Estimated mse for θ ($\theta = 0.3 \rightarrow \theta = 1.8$) for DXL distribution under the assumption of Jeffrey's prior distribution.

6. Real Data Applications

In this section, we present some applications using real dataset as a way to show that the proposed model may be attractive alternatives to some standard existing discrete distributions. We consider here just the Bayesian approach to estimate the model parameters. All computations were performed using `R2jags` (Su & Yajima, 2012) package from the `R` (R Development Core Team, 2017) software. To obtain the posterior marginal distributions and corresponding summaries of interest we adopted the MwG algorithm for MCMC sampling. For each generated sample, three chains with $N = 10,000$ values was generated for each component of $\boldsymbol{\zeta}$ considering a burn-in period of 5% of the chain's size. To obtain pseudo-independent samples from the *posterior* distribution (17), one value out of every 10 generated values was kept in the chain, resulting in chains of length 1000 for each parameter.

6.1. Alberta Fires

The first application studied is based on data about the monthly number of fires that occurred in a 67,000 km^2 region of boreal forest in northeastern Alberta, Canada over a seven year period from 1996 to 2002. The data was collected from the Alberta government's Historical Wildfire Database. Information tracked for each fire includes: cause, size, location (latitude and longitude, legal land description, and forest area), time and duration, weather conditions, staffing and physical resources used to suppress the fire, and area burned.

For the statistical analysis, the number of fires was used as the response variable and the parameters of the DXL distribution were estimated using a Bayesian approach assuming three different scenarios for prior distributions assumptions for the parameter θ: an approximately non-informative gamma prior – $Gamma(0.001, 0.001)$, an approximately non-informative uniform prior – $U(0, 10)$, and the Jeffrey's prior distribution. The `MwG` algorithm was run and tests for convergence were performed, revealing stationary (DXL model passed on Heidelberger-Welch diagnostic) of the generated chains after burn-in. The obtained results are presented in Table 2, with the fit of the DXL distribution compared with the fits of the Geometric (with PMF given by $P(X = x) = (1 - \theta)^x\theta$ and $\theta \sim U(0, 1)$) and Poisson (with PMF given by $P(X = x) = (e^{-\theta}\theta^x)/x!$ and $\theta \sim U(0, 100)$) distributions. As a discrimination criteria, we have adopted DIC criterion which is also presented in Table 2 where we can observe the DXL model is better fitted by the data (smaller value).

The estimated mean (μ) and 95% credible intervals for each model were also shown in the Table 2, indicating that the estimated mean for the DXL distribution was close to the empirical mean, suggesting a good fit for the data. The best fitted model was concluded to be the DXL distribution, using the DIC criteria.

TABLE 2: Posterior summaries for the parameters θ and the mean (μ) of DXL, Geometric (G) and Poisson (P) distributions for the number of fires in northeastern of Alberta, Canada.

Model	Param.	Post. Mean	Std. Dev.	95% Cred. Int.	Rhat	DIC
DXL (Gamma Prior)	θ	0.1401	0.0032	(0.1340, 0.1467)	1.001	6309.4
	μ	12.4460	0.3247	(11.7989, 13.0815)	1.002	
DXL (Uniform Prior)	θ	0.1402	0.0033	(0.1338, 0.1467)	1.009	6309.1
	μ	12.4351	0.3337	(11.8009, 13.1058)	1.009	
DXL (Jeffrey's Prior)	θ	0.1403	0.0033	(0.1341, 0.1468)	1.005	6309.9
	μ	12.4273	0.3278	(11.7880, 13.0674)	1.005	
G	θ	0.9255	0.0024	(0.9211, 0.9303)	1.007	6332.3
	μ	13.4457	0.4198	(12.6292, 14.2928)	1.007	
P	θ	12.4411	0.1213	(12.2142, 12.6881)	1.010	11222.7
	μ	12.4411	0.1213	(12.2142, 12.6881)	1.010	

Given that the choice of prior is interchangeable (see Table 2), Figures 8 and 9 present the diagnostics plots for the Gelman-Rubin and Geweke diagnostics applied to the DXL model, assuming only the uniform prior (selected by the DIC criteria). The Gelman-Rubin diagnostic plots show the $\widehat{R}$ values for each parameter, where values close to 1 indicate convergence. From this, we can conclude that the parameters of the DXL model exhibit convergence across chains if $\widehat{R} \leq 1.1$. Meanwhile, the Geweke diagnostic plots compare early and late segments of each chain, showing Z-scores for each parameter. Values within an acceptable range confirm stationarity within the chain. Together, these diagnostics suggest that the DXL model has reached a stable posterior distribution, assuming both tests fall within acceptable convergence thresholds.

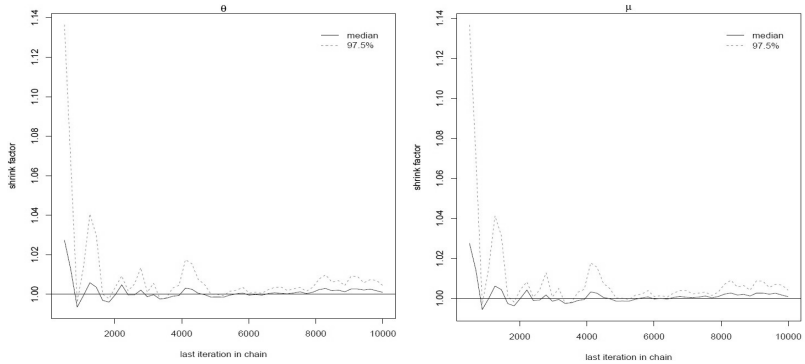

FIGURE 8: Gelman-Rubin diagnostic plot for the DXL model parameters (left-panel is for θ and right-panel is for μ), displaying the $\widehat{R}$ values for each parameter, with values approaching 1 indicating convergence across multiple MCMC chains.

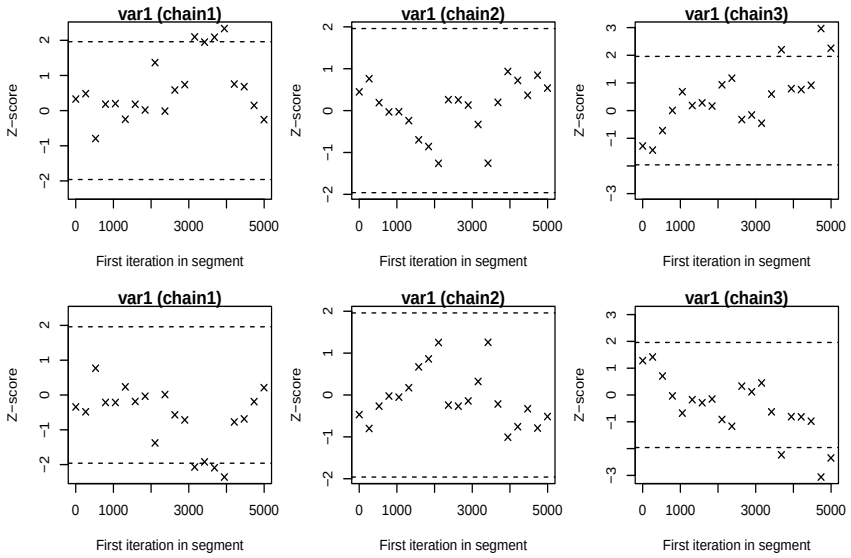

FIGURE 9: Geweke diagnostic plot for the DXL model parameters (upper-panel is for θ and lower-panel is for μ), illustrating the Z-scores comparing the means of the early and late segments of each MCMC chain.

Figure 10 presents the traceplots for the DXL model parameters. Each traceplot displays the MCMC samples over iterations for a given parameter, providing a visual assessment of convergence. Since the traceplots for the proposed exhibit stable patterns with no visible trends or drifts, indicating that the MCMC chains for the DXL model have converged successfully. This suggests that the chains are well-mixed, providing reliable estimates from the posterior distribution.

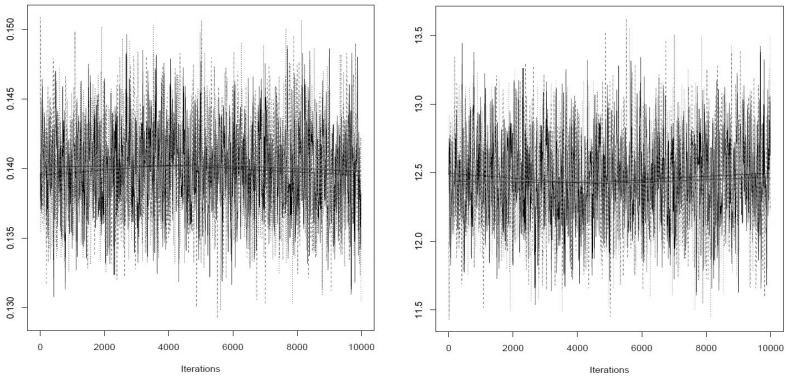

FIGURE 10: Traceplots for the DXL model parameters (left-panel is for θ and right-panel is for μ) showing MCMC samples over iterations (black is for chain 1, red is for chain 2, and blue is for chain 3).

6.2. Breast Cancer

Breast cancer is the second most common cancer diagnosed in women globally, particularly in the US, as pointed by Mayo Clinic. In a second study, a dataset of 81 breast cancer patients from the Department of Breast Surgery, Cancer Institute Hospital of Japanese Foundation for Cancer Research was used for the analysis. The survival times of the patients, in complete months, were assumed to follow the DXL distribution under a Bayesian approach. In addition, the following regression model was assumed for the i-th patient:

$$\theta_i = \exp\left(\beta_0 + \sum_{j=1}^{4} \beta_j x_{ji}\right), \tag{18}$$

where

- x_{1i}: Presenting Symptom (incidental imaging finding, inflammation, lump, or nipple inversion/blood discharge PDO);
- x_{2i}: Cancer Grade (Grade 1, Grade 2, or Grade 3);
- x_{3i}: Vascular Invasion (Present, or Absent);
- x_{4i}: Surgical Procedure (Breast conserving surgery, or Mastectomy);

Approximately non-informative normal prior distributions were used for the regression parameters. The results of the MwG algorithm are shown in Table 3. The generated chains were found to be stationary after the burn-in period. The analysis of the results showed that none of the four covariates had a significant impact on the patients' survival times, as evidenced by the fact that zero is included in the 95% credible intervals for each covariate's regression parameters. The Half-Normal probability plot in Figure 11 also suggests the good fit of the DXL model for the data, as the estimated Cox & Snell (1968) residuals are within the simulated envelope and there are no severe violations of the model assumptions.

TABLE 3: Posterior summaries for the parameters of DXL regression model.

Model	Param.	Post. Mean	Std. Dev.	95% Cred. Int.	Rhat	DIC
DXL Regression Model	β_0	-3.7009	0.5874	(-4.8517, -2.5618)	1.009	796.7
	β_1	0.0104	0.1626	(-0.3168, 0.3230)	1.012	
	β_2	-0.0902	0.1095	(-0.3023, 0.1253)	1.001	
	β_3	-0.1321	0.1805	(-0.4960, 0.2202)	1.004	
	β_4	-0.0691	0.1664	(-0.3937, 0.2609)	1.003	

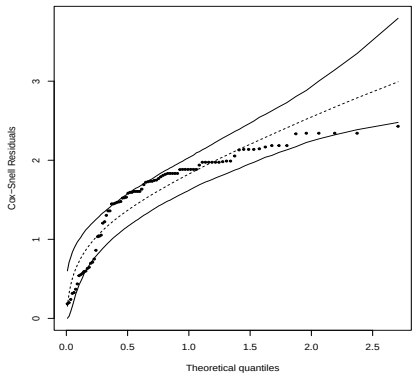

FIGURE 11: Half-Normal plot with simulated envelope for the Cox-Snell residuals.

Figure 12 presents the traceplots for the DXL regression model parameters. Each traceplot displays the MCMC samples over iterations for a given parameter, providing a visual assessment of convergence. Since the traceplots for the proposed exhibit stable patterns with no visible trends or drifts, indicating that the MCMC chains for the DXL model have converged successfully. This suggests that the chains are well-mixed, providing reliable estimates from the posterior distribution.

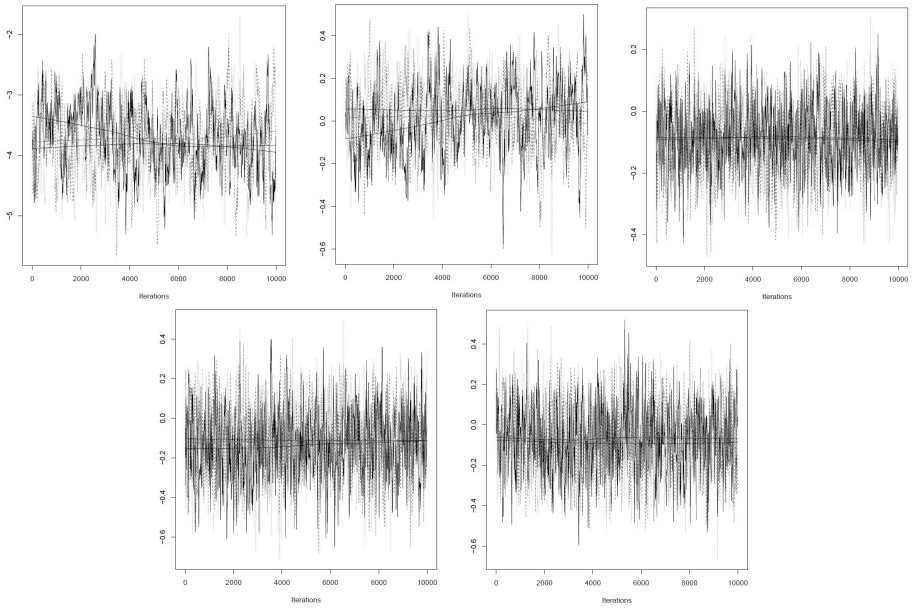

FIGURE 12: Traceplots for the DXL regression model parameters (left-panel is for θ and right-panel is for μ) showing MCMC samples over iterations (black is for chain 1, red is for chain 2, and blue is for chain 3).

7. Concluding Remarks

This study presents a discrete adaptation of the xLindley distribution, named the Discrete xLindley (DXL) distribution, to address the modeling of count datasets exhibiting overdispersion. The DXL model was formulated using the method of infinite series, and its fundamental probabilistic properties—such as the mean, variance, moment-generating function, and coefficients of variation, skewness, and kurtosis—were rigorously derived. This distribution demonstrates suitability for zero-inflated datasets, evaluated through the Zero-Modified (ZM) measure. Additionally, the log-likelihood function, score function, and asymptotic interval estimators for the parameters were established.

A Monte Carlo simulation study validated the applicability of the DXL distribution, and empirical testing was conducted using two real datasets, with parameters estimated through a Bayesian approach employing the Metropolis-within-Gibbs (MwG) algorithm. Model selection was guided by the Deviance Information Criterion (DIC), with results indicating the DXL model's competitive performance compared to standard discrete models, such as the Poisson and Geometric distributions. An `R` package is currently under development to provide comprehensive tools for fitting the DXL model, and the scripts used in model fitting are available upon request from the authors.

[Received: May 2024 — Accepted: November 2024]

References

Barco, K. V. P., Mazucheli, J. & Janeiro, V. (2017), 'The inverse power Lindley distribution', *Communications in Statistics-Simulation and Computation* **46**(8), 6308–6323.

Bi, Z., Faloutsos, C. & Korn, F. (2001), The dgx distribution for mining massive, skewed data, *in* 'Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining', ACM, San Francisco, CA, USA, pp. 17–26.

Chakraborty, S. (2015), 'Generating discrete analogues of continuous probability distributions - A survey of methods and constructions', *Journal of Statistical Distributions and Applications* **2**(1), 1–30.

Chib, S. & Greenberg, E. (1995), 'Understanding the Metropolis-Hastings algorithm', *The American Statistician* **49**(4), 327–335.

Chouia, S. & Zeghdoudi, H. (2021), 'The xlindley distribution: Properties and application', *Journal of Statistical Theory and Applications* **20**(2), 318–327.

Collett, D. (2003), *Modelling survival data in medical research*, 2 edn, Chapaman and Hall, New York.

Cox, D. R. & Snell, E. J. (1968), 'A general definition of residuals', *Journal of the Royal Statistical Society: Series B (Methodological)* **30**(2), 248–265.

Devroye, L. (2006), Nonuniform random variate generation, *in* S. G. Henderson & B. L. Nelson, eds, 'Handbooks in Operations Research and Management Science', Vol. 13, Elsevier, Amsterdam, chapter 3, pp. 83–121.

Doray, L. G. & Luong, A. (1997), 'Efficient estimators for the Good family', *Communications in Statistics - Simulation and Computation* **26**(3), 1075–1088.

Ferreira, E., Kohara, A. & Sesma, J. (2017), 'New properties of the Lerch's transcendent', *Journal of Number Theory* **172**, 21–31.

Gelfand, A. E. & Smith, A. F. (1990), 'Sampling-based approaches to calculating marginal densities', *Journal of the American Statistical Association* **85**(410), 398–409.

Ghitany, M., Alqallaf, F., Al-Mutairi, D. & Husain, H. (2011), 'A two-parameter weighted Lindley distribution and its applications to survival data', *Mathematics and Computers in Simulation* **81**(6), 1190–1201.

Ghitany, M. E., Al-Mutairi, D. K., Balakrishnan, N. & Al-Enezi, L. (2013), 'Power Lindley distribution and associated inference', *Computational Statistics & Data Analysis* **64**, 20–33.

Good, I. J. (1953), 'The population frequencies of species and the estimation of population parameters', *Biometrika* **40**(3-4), 237–264.

Haight, F. A. (1957), 'Queueing with balking', *Biometrika* **44**(3/4), 360–369.

Hamada, M. S., Wilson, A. G., Reese, C. S. & Martz, H. F. (2008), *Bayesian reliability*, Springer Series in Statistics, Springer, New York.

Hassani, M. (2007), 'Approximation of the dilogarithm function', *Journal of Inequalities in Pure and Applied Mathematics* **8**(3), 1–7.
`https://emis.de/journals/JIPAM/article901.html`

Inusah, S. & Kozubowski, T. J. (2006), 'A discrete analogue of the Laplace distribution', *Journal of Statistical Planning and Inference* **136**(3), 1090–1102.

Jodra, P. (2010), 'Computer generation of random variables with Lindley or Poisson–Lindley distribution via the Lambert–W function', *Mathematics and Computers in Simulation* **81**(4), 851–859.

Kalbfleisch, J. D. & Prentice, R. L. (2002), *The statistical analysis of failure time data*, 2nd edn, Wiley, New York.

Keilson, J. & Gerber, H. (1971), 'Some results for discrete unimodality', *Journal of the American Statistical Association* **66**(334), 386–389.

Kemp, A. W. (1997), 'Characterizations of a discrete Normal distribution', *Journal of Statistical Planning and Inference* **63**(2), 223–229.

Kemp, A. W. (2008), *The discrete Half–Normal distribution*, Birkhäuser Boston, Boston, pp. 353–360. In Advances in Mathematical and Statistical Modeling.

Klein, J. P. & Moeschberger, M. L. (1997), *Survival analysis: Techniques for censored and truncated data*, Springer-Verlag, New York.

Kozubowski, T. J. & Inusah, S. (2006), 'A skew Laplace distribution on integers', *Annals of the Institute of Statistical Mathematics* **58**(3), 555–571.

Lawless, J. F. (2003), *Statistical models and methods for lifetime data*, 2 edn, John Wiley & Sons, Hoboken, N.J.

Lee, E. T. & Wang, J. W. (2003), *Statistical methods for survival data analysis*, 3 edn, John Wiley & Sons, Hoboken, NJ.

Lehmann, E. J. & Casella, G. (1998), *Theory of Point Estimation*, Springer Verlag.

Meeker, W. Q. & Escobar, L. A. (1998), *Statistical methods for reliability data*, John Wiley & Sons, New York.

Merovci, F. (2013), 'Transmuted Lindley distribution', *International Journal of Open Problems in Computer Science & Mathematics* **6**.

Oliveira, R. P., Achcar, J. A., Mazucheli, J. & Bertoli, W. (2021), 'A new class of bivariate lindley distributions based on stress and shock models and some of their reliability properties', *Reliability Engineering & System Safety* **211**, 107528.

Oliveira, R. P. d., Mazucheli, J. & Achcar, J. A. (2017), 'A comparative study between two discrete lindley distributions', *Ciência e Natura* **39**(3), 539–552.

R Development Core Team (2017), *R: A language and environment for statistical computing*, R Foundation for Statistical Computing, Vienna, Austria.

Sato, H., Ikota, M., Sugimoto, A. & Masuda, H. (1999), 'A new defect distribution metrology with a consistent discrete exponential formula and its applications', *IEEE Transactions on Semiconductor Manufacturing* **12**(4), 409–418.

Shanker, R. & Mishra, A. (2013), 'A quasi Lindley distribution', *African Journal of Mathematics and Computer Science Research* **6**(4), pp–64.

Sharma, V. K., Singh, S. K., Singh, U. & Agiwal, V. (2015), 'The inverse Lindley distribution: A stress-strength reliability model', *Journal of Industrial and Production Engineering* **32**(3), 162–173.

Siromoney, G. (1964), 'The general Dirichlet's Series distribution', *Journal of the Indian Statistical Association* **2-3**(2), 1–7.

Su, Y.-S. & Yajima, M. (2012), *R2jags: A Package for Running JAGS from R*. R package version 0.03-08. `https://cran.r-project.org/package=R2jags`

Virchenko, N., Kalla, S. & Al-Zamel, A. (2001), 'Some results on a generalized hypergeometric function', *Integral Transforms and Special Functions* **12**(1), 89–100.

Xu, C., Yan, Y. & Shi, Z. (2016), 'Euler sums and integrals of polylogarithm functions', *Journal of Number Theory* **165**, 84–108.

Appendix A. Jeffrey's Prior Distribution

The Jeffrey's prior distribution is characterized by its non-informative nature, specifically designed to offer a Bayesian framework that accommodates uncertainty in the parameters. This prior is typically expressed in terms of the likelihood function of the model, ensuring that it remains invariant under reparameterization. For the proposed DXL model, assuming complete data, the Jeffrey's prior $\pi(\theta)$ for the parameter θ can be formulated as:

$$\pi(\theta) = \left\{ n \left\{ \left(\frac{g'(\theta)}{g(\theta)} \right)^2 - \frac{g''(\theta)}{g(\theta)} \right\} + g(\theta) \left\{ \Phi\left(\frac{1}{e^\theta}, 1, \theta \right) e^\theta - \frac{e^\theta}{\theta} - \frac{1}{1+\theta} \right\} \right\}^{1/2}$$

where $\Phi(\cdot)$ is the Lerch transcendent function (Hassani, 2007; Ferreira et al., 2017).

Appendix B. Posterior Distribution

In this section, we will derive the posterior distribution for the proposed DXL model only under complete data assumption, since for right-censored and regression follows the same procedure. The derivation of the posterior distribution in a Bayesian framework involves applying Bayes' theorem, which combines the prior distribution with the likelihood of the observed data. Given a parameter θ, the posterior distribution $\pi(\theta \mid x)$, where x represents the observed data, is expressed as follows:

$$\pi(\theta \mid x) \propto L(\theta) \cdot \pi(\theta)$$

where $L(\theta)$ is the likelihood function, and $\pi(\theta)$ is the prior distribution, which captures our beliefs about θ before observing the data.

In this paper, we considered three classes of prior distributions: gamma, uniform and Jeffrey's. Here, we will assume only the Jeffrey's prior for the calculation, since for the other classes are straightforward. Then, substituting the Jeffrey's prior into the equation above, we have:

$$\pi(\theta \mid x) \propto L(\theta) \cdot \sqrt{I(\theta)}$$

Then, given the likelihood of the proposed DXL for complete data, the unnormalized posterior distribution is:

$$\begin{aligned} \pi(\theta \mid x) &= \exp\left\{ n \ln\left[g(\theta)\right] - n\theta(\overline{x}+1) + \sum_{i=1}^{n} \ln(2+\theta+x_i) \right\} \\ &\times \left\{ n \left\{ \left(\frac{g'(\theta)}{g(\theta)} \right)^2 - \frac{g''(\theta)}{g(\theta)} \right\} + g(\theta) \left\{ \Phi\left(\frac{1}{e^\theta}, 1, \theta \right) e^\theta - \frac{e^\theta}{\theta} - \frac{1}{1+\theta} \right\} \right\}^{1/2} \end{aligned}$$

where $\theta \in \mathbb{R}_+$, $\Phi(\cdot)$ is the Lerch transcendent function (Hassani, 2007), and

$$g(\theta) = \frac{\left(\mathrm{e}^{\theta} - 1\right)^2}{(2 + \mathrm{e}^{\theta})\mathrm{e}^{\theta} - (1 + \theta)}. \tag{A1}$$

To obtain the posterior distribution in a usable form, it is necessary to normalize the posterior by integrating over all possible values of θ, that is, it is necessary to divide the unnormalized posterior above for $\int L(\theta') \cdot \sqrt{I(\theta')}\, d\theta'$. In this case, this integral acts as the normalization constant, ensuring that the posterior distribution is correctly scaled to integrate to one. For the proposed DXL model, however, to solve this integral numerical methods are required, since it cannot be solved by analytical methods.