Chilean Journal of Statistics
Volume 14, Number 2 (December 2023), Pages 123–142
DOI: 10.32372/ChJS.14-02-04

Distributions Theory
Research Article

Inflated log-Lindley distribution for modeling continuous data bounded in unit interval with possible mass at boundaries

Sahana Bhattacharjee[1,*] and Subrata Chakraborty[2]

[1]Department of Statistics, Gauhati University, Guwahati, Assam, India
[2]Department of Statistics, Dibrugarh University, Dibrugarh, Assam, India
(Received: 18 August 2023 · Accepted in final form: 31 October 2023)

Abstract

The log-Lindley distribution was introduced as an alternative to the famed beta distribution to model data in the unit interval. The present article introduces inflated versions of the log-Lindley distribution. Distributional properties are investigated. Special cases of the proposed distributions are shown to be members of the exponential family. Moment and maximum likelihood estimation of the parameters are discussed with derivation of the exact expression for the information matrix. Then, the proposed distributions are compared with the inflated unit Lindley and inflated beta distributions for modeling four sets of data regarding school wise pass proportions of high school leaving examinations having zero and/or one inflation. Findings clearly suggest the superiority of the proposed distribution based on the Akaike information criteria and observed and fitted distribution function plots for all the cases considered.

Keywords: Beta distribution · Inflated distributions · Log-Lindley distribution proportion data · Numerical simulations.

Mathematics Subject Classification: Primary 60E05 · 62E15.

1. Introduction

Many researchers in the field of applied statistics often come across data on standardized rates, proportions or fractions, which assume values in the unit range (0,1). For instance, data on proportion of successful (or unsuccessful candidates in an examination), fraction of an hour required to accomplish a task by a group of workers, among others. However, it is not rare that data on proportions or fractions contain a number of zeroes and/or ones, that is, the data arising in the range [0,1), (0,1] or [0,1]. In such cases, the commonly known distributions on (0,1) such as the beta distribution (Johnson et al., 1995), Kumaraswamy distribution (Kumaraswamy, 1980), unit Lindley distribution (Mazucheli et al., 2019), and another form of the one parameter unit Lindley distribution (Mazucheli et al., 2020), fail to model the data in question as all of them allow modeling of values lying strictly between zero and one.

*Corresponding author. Email: sahana.bhattacharjee@hotmail.com

ISSN: 0718-7912 (print)/ISSN: 0718-7920 (online)

http://www.soche.cl/chjs

The above-mentioned aspects call upon the requirement to develop models which is able to capture the probability mass concentrated at the points 0, 1 or both, as the case may be. One of the most popular approaches is to mix a continuous distribution in the range (0,1) with a degenerate distribution whose probability mass is concentrated at either 0 or 1 (in case of data observed in the range [0,1) or (0,1]) and that with the Bernoulli distribution which assigns non-negative probability to both 0 and 1. These models are members of the inflated family of distributions. Zero-inflation or one-inflation occurs when in the data set, one or more observations that equal zero or one, respectively, are included. Zero inflated data occur frequently in nature and can be modeled suitably using a number of continuous distributions (Burch and Egbert, 2020; Hashimoto et al., 2019; Liu et al., 2019; Tomazella et al., 2019).The data where a high number of both zeroes and ones arise are referred to as zero-one inflated data.

The first work on development of above-mentioned models was reported in Ospina and Ferrari (2010), where the authors developed the zero-and/or inflated versions of the beta distribution and showed its utility through a real-life data set on percentage of qualified nurses in 645 Brazilian municipal districts. Cribari-Neto and Santos (2019) introduced the inflated Kumaraswamy distribution and showed its advantage over the inflated beta distribution through modeling of a data set on proportion of inhabitants in each of the 5,566 Brazilian municipalities that lived in homes with at least one bathroom and piped water in 2010. The inflated unit Lindley distribution was recently developed by Chakraborty and Bhattacharjee (2021), who studied their various statistical properties and it proved better in modeling data on proportion of successful students in high school leaving certificate examination of Manipur state of India as compared to the inflated beta distribution.

The log-Lindley distribution (Goméz-Deniz et al., 2014) was proposed as an alternative to the classical beta probability model having support in the unit interval. One of the main advantages of this distribution is that the cumulative distribution function (CDF) is obtainable in a closed form, unlike the beta distribution and also, the expression for the probability density function (PDF) and CDF does not involve any special function. It also presented as an alternative to the beta regression model which models response variable distributed in the range (0,1) so allowing more flexibility of the covariates. A reparametrized version of this model was discussed in a note by Jodrá and Jiménez-Gamero (2016). So far, no work has been attempted on developing the inflated version of the log-Lindley distribution, which may prove to be a viable alternative of the inflated beta distribution. Also, only a handful of zero-or/and-one inflated distributions are available in the literature.

Hence, this article attempts to develop the zero-or/and-one inflated log-Lindley distribution. In this article, the aim is to analyze the same data sets used in Chakraborty and Bhattacharjee (2021) by proposing a new inflated version of the log- Lindley distribution (Goméz-Deniz et al., 2014) and compare it with two models namely the inflated beta distribution and inflated unit Lindley distribution.

Section 2 introduces the zero-or-one inflated and zero-and-one inflated log-Lindley (ZOILL) distribution and explores some of the distributional properties of both the distributions. Section 3 deals with the parameter estimation of both the zero-or-one and ZOILL distribution. Findings of the simulation study to assess the performance of the maximum likelihood (ML) estimators are presented in Section 4. Data fitting applications in comparison with the inflated beta and inflated unit Lindley distribution is reported in the Section 5. Finally, concluding remarks, future scope and limitations of this study are listed in the Section 6.

2. Inflated log-Lindley distribution

In this section, the zero-or-one inflated log-Lindley distribution and the ZOILL distribution are introduced.

2.1 The zero-or-one inflated log-Lindley distribution

The log-Lindley distribution (Goméz-Deniz et al., 2014) with parameters σ and λ has a PDF stated as

$$f(y;\sigma,\lambda) = \frac{\sigma^2}{1+\lambda\sigma}\left(\lambda - \log(y)\right) y^{\sigma-1}, \quad 0 < y < 1. \tag{2.1}$$

For data arising from some real life phenomena, a number of zeroes and/or ones may be present, which is not then suitable to be modeled by any distribution having support on (0,1). Hence, it becomes necessary to incorporate a discrete component into the continuous data generating process so that the zeroes and/or ones are observed with a positive probability. Consequently, we need to look for a probability law which merges both the continuous and discrete data generating processes. One way of achieving this is by mixing two distributions: the continuous log-Lindley distribution on (0,1) and the degenerate distribution with the entire probability mass concentrated at the point c, where either $c = 0$ or $c = 1$. We then say that the data is inflated at one of the two end points of the interval (0,1). We name this mixture distribution as the inflated log-Lindley (ILL) distribution, whose CDF is given by

$$f_{\mathrm{ILL}}(y;\alpha,\sigma,\lambda) = \alpha I_{[c,1]}(y) + (1-\alpha)F(y;\sigma,\lambda). \tag{2.2}$$

where $I_A(y)$ is the indicator function which assumes the value 0 if $y \in A$ and 1 if $y \notin A$, $0 < \alpha < 1$ is the mixture parameter and F is the CDF of the log-Lindley distribution with parameters σ and λ. The random variable Y follows the log-Lindley distribution of parameters σ and λ with probability $(1-\alpha)$ and follows the degenerate distribution at c with probability α. Let Y be a random variable with CDF given by Equation (2.2). Then, the PDF of Y is given by

$$F_{\mathrm{ILL}}(y;\alpha,\theta), = \begin{cases} \alpha, & \text{if} \quad y = c; \\ (1-\alpha)f(y;\sigma,\lambda), & \text{if} \quad y \in (0,1); \end{cases} \tag{2.3}$$

where f is the log-Lindley PDF given in Equation (2.1) and $\alpha \in (0,1)$ is the probability mass at c, which shows the probability of observing 0 (when $c = 0$) or 1 (when $c = 1$). We denote this as $Y \sim \mathrm{ILL}_c(\alpha,\sigma,\lambda)$.

Let $Y \sim \mathrm{ILL}_c(\alpha,\sigma,\lambda)$.

(i) If $c = 0$, the distribution in Equation (2.3) is called the zero-inflated log-Lindley (ZILL) distribution and we specifically write $Y \sim \mathrm{ZILL}(\alpha,\sigma,\lambda)$, where $\alpha = \mathrm{P}(Y = 0)$.

(ii) If $c = 1$, the distribution in Equation (2.3) is called the one-inflated log-Lindley (OILL) distribution and we write $Y \sim \mathrm{OILL}(\alpha,\sigma,\lambda)$, where $\alpha = \mathrm{P}(Y = 1)$.

Figure 1 displays the ZILL and OILL PDFs inflated at the points $c = 0$ and $c = 1$ for different combinations of the values of σ and λ, with the value of the mixing parameter α kept fixed at 0.5. It is evident from the figure that the PDF of both the ZILL and OILL distribution behave differently for different values of σ, λ and c. σ and λ seem to control the shape of the probability curve, whose skewness decreases with an increase in the value of σ. Further, the peakedness of the curve also seem to decrease with an increase in the value of λ. In Figure 1(a)-(f), the vertical bar with a circle above represents $\alpha = 0.5$ ($\mathrm{P}(Y = 0)$ or $\mathrm{P}(Y = 1)$). The functional shape of the both the zero inflated and one inflated distribution seem to vary with the combination of σ and λ values.

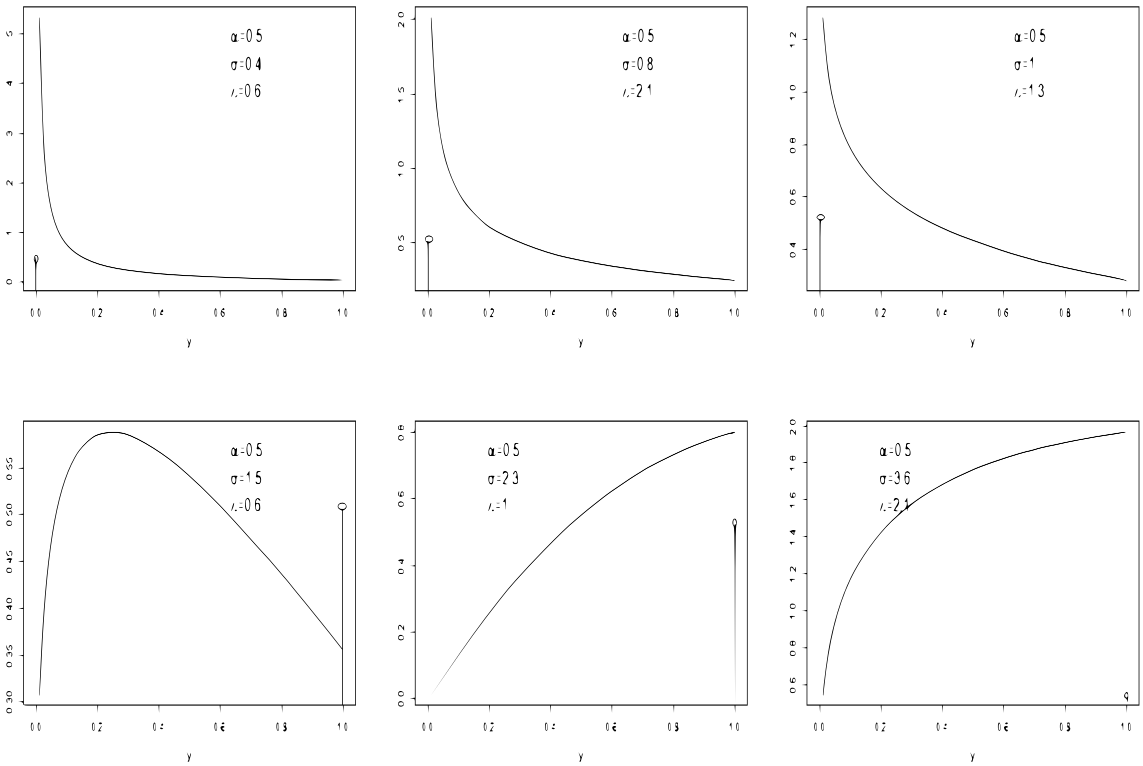


Figure 1.: PDF plots for the ZILL and OILL distributions for different values of σ and λ; $\alpha = 0.5$.

2.2 The ZOILL distribution

The zero-or-one inflated log-Lindley distribution proposed in the previous section is not suitable for modeling data that may not only contain values in the range [0,1) or (0,1], but also in [0,1]. We refer to this case as double inflation. For modeling double inflated data, a mixture of the log-Lindley distribution and Bernoulli distribution, which assigns non-zero probabilities to both the end points 0 and 1, is to be considered. The CDF of this mixture distribution, referred to as ZOILL, is given by

$$F_{\text{ZOILL}}(y; \alpha, p, \sigma, \lambda) = \alpha F_{\text{Ber}}(y; p) + (1-\alpha) F(y; \sigma, \lambda), \tag{2.4}$$

where $y \in [0,1]$, F_{Ber} denotes the Bernoulli CDF with parameter $p \in (0,1)$ and F is the CDF of the log-Lindley distribution with parameters σ and λ. Further, α is the mixing parameter which lies between 0 and 1. It follows that the PDF corresponding to the CDF in Equation (2.4) is stated as

$$f_{\text{ZOILL}}(y; \alpha, p, \sigma, \lambda) = \begin{cases} \alpha p, & \text{if} \quad y = 1; \\ \alpha(1-p), & \text{if} \quad y = 0; \\ (1-\alpha) f(y; \sigma, \lambda), & \text{if} \quad y \in (0,1). \end{cases} \tag{2.5}$$

We say that the random variable Y with PDF given in Equation (2.5) follows the ZOILL distribution and we denote it by $Y \sim \text{ZOILL}(\alpha, p, \sigma, \lambda)$. Here, $\alpha p = \text{P}(Y=1)$ and $\alpha(1-p) = \text{P}(Y=0)$. Further, $0 < \alpha p + \alpha(1-p) < 1$. Figure 2 shows the ZOILL PDF for different values of σ and λ, fixing $\alpha = 0.6$ and $p = 0.5$. It is clear that the skewness of the curve decreases with an increase in the value of both λ and σ. In Figure 2, the vertical bar with a circle above represents $\alpha p = 0.3$ and $\alpha(1-p) = 0.3$ (P($Y=0$) and P($Y=1$), respectively).

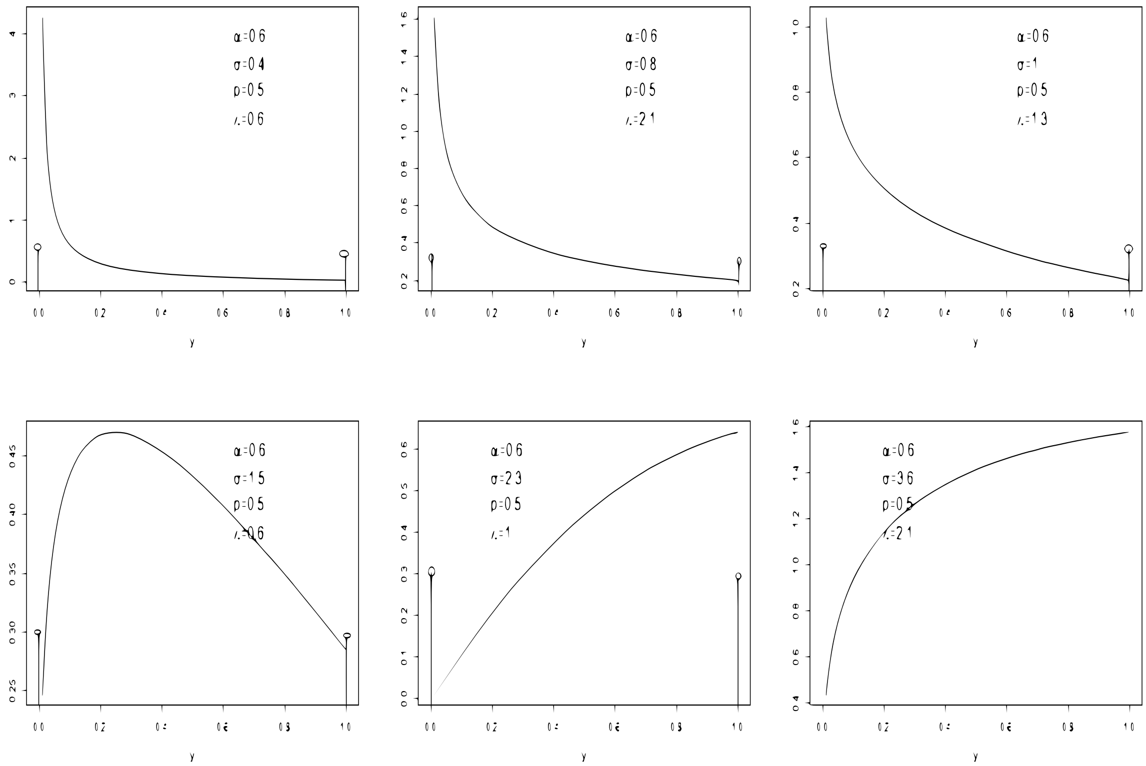


Figure 2.: Plots of ZOILL PDFs for different values of σ and λ; $\alpha = 0.6$, $p = 0.5$.

2.3 Properties

The rth raw moment of the zero-or-one inflated log-Lindley distribution defined in Equation (2.3) is given by $\mathrm{E}\,(Y^r) = \alpha c + (1-\alpha)\mu_r', \quad r = 1, 2, \ldots,$ where $\mu_r' = \sigma^2(1+\lambda(\sigma+r))/(1+\lambda\sigma)(\sigma+r)^2$ is the rth raw moment of the log-Lindley distribution. In particular, the mean and variance of Y are established as

$$\begin{aligned}
\mathrm{E}\,(Y) &= \alpha c + (1-\alpha)\frac{\sigma^2}{1+\lambda\sigma}\frac{1+\lambda(\sigma+1)}{(\sigma+1)^2};\\
\mathrm{Var}\,(Y) &= \alpha c(1-\alpha c) + \frac{\sigma^2}{1+\lambda\sigma}(1-\alpha)\bigg(\frac{1+\lambda(\sigma+2)}{(\sigma+2)^2} - \frac{1+\lambda(\sigma+1)}{(\sigma+1)^2}\\
&\quad \times\bigg(1 + (1-\alpha)\frac{\sigma^2}{1+\lambda\sigma}\frac{1+\lambda(\sigma+1)}{(\sigma+1)^2}\bigg)\bigg).
\end{aligned}$$

The rth raw moment of $Y \sim$ ZOILL is given by $\mathrm{E}\,(Y^r) = \alpha p + (1-\alpha)\mu_r'$, for $r = 1, 2, \ldots$ Then, the mean and variance of Y are formulated as

$$\begin{aligned}
\mathrm{E}\,(Y) &= \alpha p + (1-\alpha)\frac{\sigma^2}{1+\lambda\sigma}\frac{1+\lambda(\sigma+1)}{(\sigma+1)^2};\\
\mathrm{Var}\,(Y) &= \alpha p(1-\alpha p) + \frac{\sigma^2}{1+\lambda\sigma}(1-\alpha)\bigg(\frac{1+\lambda(\sigma+2)}{(\sigma+2)^2} - \frac{1+\lambda(\sigma+1)}{(\sigma+1)^2}\\
&\quad \times\bigg(1 + (1-\alpha)\frac{\sigma^2}{1+\lambda\sigma}\frac{1+\lambda(\sigma+1)}{(\sigma+1)^2}\bigg)\bigg).
\end{aligned}$$

THEOREM 2.1 A special case of ZOILL distribution defined in Equation (2.3) (specifically, for fixed value of λ) belongs to the two-parameter exponential family.

Refer to the Appendix for proof of Theorem 2.1.

THEOREM 2.2 A special case of ZOILL distribution defined in Equation (2.5) (specifically, for fixed value of λ) belongs to the three-parameter exponential family.

Refer to the Appendix for proof of Theorem 2.2.

3. Estimation

In this section, the estimation of parameters through the ML method and moment estimation method of both the zero-or-one and ZOILL distribution are considered and the construction of the Fisher information matrix for ML estimators are discussed.

3.1 Context

We observe that since the parameter space of both the parameters σ and λ are unbounded, the ML estimates of these parameters do not come out to be accurate, as observed and reported in Jodrá and Jiménez-Gamero (2016). To overcome this drawback in estimation, the same authors proposed a reparametrization of the log-Lindley distribution, by introducing a parameter π in place of λ, which is bounded in the unit interval. The PDF of the log-Lindley distribution with the (σ, π) reparametrization is given by (Jodrá and Jiménez-Gamero, 2016)

$$f(y;\sigma,\pi) = \sigma\left(\pi + \sigma(\pi-1)\log(y)\right) y^{\sigma-1}; \quad 0 < y < 1, \quad \sigma > 0, \quad 0 \le \pi < 1. \tag{3.6}$$

The CDF of the reparametrized log-Lindley distribution becomes $F(y;\sigma,\pi) = \left(1 + \sigma(\pi-1)\log(y)\right) y^{\sigma}$, for $0 < y < 1$, $\sigma > 0$, and $0 \le \pi < 1$. For the rest of the work, the distribution given in Equation (3.6) is referred to as the log-Lindley distribution. It may be noted that all the results derived in the preceding sections can be worked out for this reparametrized version quite easily.

3.2 ML estimation: Reparametrized ILL distribution

The PDF of the reparametrized ILL distribution becomes

$$f_{\mathrm{ILL}}(y;\alpha,\sigma,\pi) = \begin{cases} \alpha, & \text{if} \quad y = c \\ (1-\alpha) f(y;\sigma,\pi), & \text{if} \quad y \in (0,1) \end{cases} \tag{3.7}$$

where $f(;\sigma,\pi)$ is the reparametrized log-Lindley PDF given in Equation (3.6). The ML estimates of the parameters α, σ and π of the zero-or-one inflated log-Lindley distribution are obtained by solving the ML equations corresponding to the log-likelihood function constructed from the PDF in Equation (3.7) and given by

$$l(\boldsymbol{\nu}) = \log(L(\boldsymbol{\nu}, \mathbf{y})) = l_1(\alpha; y) + l_2(\sigma, \pi; y),$$

where $\boldsymbol{\nu} = (\alpha, \sigma, \pi)$ and

$$l_1(\alpha; y) = \log(L_1(\alpha, y)) = \log(\alpha) \sum_{i=1}^{n} I_{[c]}(y_i) + \log(1-\alpha)\left(n - \sum_{i=1}^{n} I_{[c]}(y_i)\right);$$

$$\begin{aligned} l_2(\sigma,\pi;y) &= \log(L_2(\sigma,\pi,y)) \\ &= \left(n - \sum_{i=1}^{n} I_{[c]}(y_i)\right)\log(\sigma) + \sum_{\substack{i=1 \\ y_i \in (0,1)}}^{n} \log(\pi + \sigma(\pi-1)\log(y_i)) + (\sigma-1) \sum_{\substack{i=1 \\ y_i \in (0,1)}}^{n} \log(y_i). \end{aligned}$$

The score function is then obtained by differentiating the log-likelihood function and is denoted by $U(\boldsymbol{\nu}) = (U_\alpha(\alpha), U_\sigma(\sigma), U_\pi(\pi))$, where

$$\begin{aligned}
U_\alpha(\alpha) &= \frac{\partial l_1(\alpha; y)}{\partial \alpha} = \frac{1}{\alpha}\sum_{i=1}^{n} I_{[c]}(y_i) - \frac{1}{1-\alpha}\Big(n - \sum_{i=1}^{n} I_{[c]}(y_i)\Big);\\
U_\sigma(\sigma) &= \frac{\partial l_2(\sigma, \pi; y)}{\partial \sigma} = \frac{(n - \sum_{i=1}^{n} I_{[c]}(y_i))}{\sigma} + \sum_{\substack{i=1\\ y_i\in(0,1)}}^{n} \frac{(\pi-1)\log(y_i)}{\pi + \sigma(\pi-1)\log(y_i)} + \sum_{\substack{i=1\\ y_i\in(0,1)}}^{n} \log(y_i);\\
U_\pi(\pi) &= \frac{\partial l_2(\sigma, \pi; y)}{\partial \lambda} = \sum_{\substack{i=1\\ y_i\in(0,1)}}^{n} \frac{1+\sigma\log(y_i)}{\pi + \sigma(\pi-1)\log(y_i)}.
\end{aligned}$$

The ML estimators of the parameters are obtained by solving the equations stated as

$$\begin{aligned}
&U_\alpha(\alpha) = 0 \implies \widehat{\alpha} = \frac{1}{n}\sum_{i=1}^{n} I_{[c]}(y_i); \quad U_\pi(\pi) = 0 \implies \sum_{\substack{i=1\\ y_i\in(0,1)}}^{n} \frac{1+\sigma\log(y_i)}{\pi + \sigma(\pi-1)(\log y_i)} = 0;\\
&U_\sigma(\sigma) = 0 \implies \frac{(n - \sum_{i=1}^{n} I_{[c]}(y_i))}{\sigma} + \sum_{\substack{i=1\\ y_i\in(0,1)}}^{n} \frac{(\pi-1)(\log y_i)}{\pi + \sigma(\pi-1)\log(y_i)} = -\sum_{\substack{i=1\\ y_i\in(0,1)}}^{n} \log(y_i).
\end{aligned}$$

Since the second and third ML equations are non-linear in σ and π, so, numerical techniques are to be employed to get a solution to these equations. The Fisher information matrix for the zero-or-one inflated reparametrized log-Lindley distribution is

$$K(\boldsymbol{\nu}) = \begin{pmatrix} k_{\alpha\alpha} & k_{\alpha\pi} & k_{\alpha\sigma} \\ k_{\pi\alpha} & k_{\pi\pi} & k_{\pi\sigma} \\ k_{\sigma\alpha} & k_{\sigma\pi} & k_{\sigma\sigma} \end{pmatrix},$$

where $k_{\alpha\pi} = k_{\pi\alpha}$, $k_{\alpha\sigma} = k_{\sigma\alpha}$,

$$\begin{aligned}
k_{\alpha\alpha} &= -\mathrm{E}\left(\frac{\partial U_\alpha(\alpha)}{\partial \alpha}\right] = \frac{n}{\alpha(1-\alpha)}; k_{\alpha\pi} = -\mathrm{E}\left(\frac{\partial U_\pi(\pi)}{\partial \alpha}\right) = 0; k_{\alpha\sigma} = -\mathrm{E}\left(\frac{\partial U_\sigma(\alpha)}{\partial \sigma}\right) = 0;\\
k_{\pi\pi} &= -\mathrm{E}\left(\frac{\partial U_\pi(\pi)}{\partial \pi}\right) = nE_1\left(\frac{\pi}{1-\pi}\right)\left(\frac{2\pi\exp(-\frac{\pi}{\pi-1})}{(\pi-1)^3} - \frac{\exp(-\frac{\pi}{\pi-1})}{(\pi-1)} - \frac{\pi^2\exp(-\frac{\pi}{\pi-1})}{(\pi-1)^3}\right)\\
&\quad - \frac{2n\pi\exp(-\frac{2\pi}{\pi-1})}{(\pi-1)^2} + \frac{2n}{(\pi-1)} - \frac{n\exp(-\frac{\pi}{\pi-1})}{(\pi-1)}\Gamma\left(2, \frac{\pi}{1-\pi}\right);\\
k_{\pi\sigma} &= -\mathrm{E}\left(\frac{\partial U_\sigma(\sigma)}{\partial \pi}\right) = k_{\sigma\pi} = -\frac{n}{\sigma} + \frac{n\exp(-\frac{\pi}{\pi-1})}{\sigma(\pi-1)}\left(\pi E_1\left(\frac{\pi}{1-\pi}\right) - (1-\pi)\exp\left(-\frac{\pi}{\pi-1}\right)\right)\\
&\quad - \frac{n}{\sigma}\exp\left(-\frac{\pi}{\pi-1}\right)\Gamma\left(2, \frac{\pi}{1-\pi}\right) - \frac{n\pi^2\exp(-\frac{\pi}{1-\pi})}{\sigma(\pi-1)^2}E_1\left(\frac{\pi}{1-\pi}\right) + \frac{2n\pi(1-\pi)}{\sigma(\pi-1)^2};\\
k_{\sigma\sigma} &= -\mathrm{E}\left(\frac{\partial U_\sigma(\sigma)}{\partial \sigma}\right) = \frac{n(1-\alpha)}{\sigma^2} - \frac{n(\pi-1)}{\sigma^2}\exp\left(-\frac{\pi}{\pi-1}\right)\Gamma\left(2, \frac{\pi}{1-\pi}\right)\\
&\quad - \frac{n\pi^2\exp(-\frac{\pi}{1-\pi})}{\sigma^2(\pi-1)}E_1\left(\frac{\pi}{1-\pi}\right) + \frac{2n\pi}{\sigma^2}.
\end{aligned}$$

Suppose $\widehat{\boldsymbol{\nu}} = (\widehat{\alpha}, \widehat{\sigma}, \widehat{\pi})$ denote the ML estimators of $\boldsymbol{\nu}$. In large samples, $\sqrt{n}\,(\widehat{\boldsymbol{\nu}} - \boldsymbol{\nu})$ is asymptotically normally distributed, that is, $\sqrt{n}(\widehat{\boldsymbol{\nu}} - \boldsymbol{\nu}) \xrightarrow{D} N_3(\mathbf{0}, k(\boldsymbol{\nu})^{-1})$, where $k(\boldsymbol{\nu})$ is the Fisher information matrix. This result can be used to approximate the confidence intervals for the parameters α, σ and π. Let $\delta \in (0, 0.5)$. Then $(1-\delta) \times 100\%$ asymptotic confidence intervals for α, σ and π are given, respectively, by $\widehat{\alpha} \pm Z_{1-\delta/2}\mathrm{SE}(\widehat{\alpha})$, $\widehat{\sigma} \pm Z_{1-\delta/2}\mathrm{SE}(\widehat{\sigma})$, and $\widehat{\pi} \pm Z_{1-\delta/2}\mathrm{SE}(\widehat{\pi})$, where SE denotes the corresponding standard error and $Z_{1-\delta/2}$ is the quantile of the standard normal distribution at δ level of significance.

3.3 ML estimation: Reparametrized ZOILL distribution

The PDF of the zero-and-one inflated reparametrized log-Lindley distribution becomes

$$f_{\mathrm{ZOILL}}(y; \alpha, p, \sigma, \pi) = \begin{cases} \alpha p, & \text{if} \quad y = 1; \\ \alpha(1-p), & \text{if} \quad y = 0; \\ (1-\alpha) f(y; \sigma, \pi), & \text{if} \quad y \in (0,1); \end{cases} \tag{3.8}$$

where $f(;\sigma,\pi)$ is the reparametrized log-Lindley PDF given in Equation (3.6). The ML estimates of the parameters α, σ and π of the ZOILL distribution are obtained by solving the ML equations corresponding to the log-likelihood function constructed from the PDF in Equation (3.8) stated as $l(\boldsymbol{\nu}) = \log(L(\boldsymbol{\nu}, \mathbf{y})) = l_1(\alpha; y) + l_2(p; y) + l_3(\sigma, \pi; y)$, where

$$l_1(\alpha; y) = \log(\alpha) \sum_{i=1}^{n} I_{\{0,1\}}(y_i) + \log(1-\alpha)\Big(n - \sum_{i=1}^{n} I_{\{0,1\}}(y_i)\Big);$$

$$l_2(p; y) = \log(p) \sum_{i=1}^{n} y_i I_{\{0,1\}}(y_i) + \log(1-p)\Big(\sum_{i=1}^{n} I_{\{0,1\}}(y_i) - \sum_{i=1}^{n} y_i I_{\{0,1\}}(y_i)\Big);$$

$$l_3(\sigma, \pi; y) = \Big(n - \sum_{i=1}^{n} I_{\{0,1\}}(y_i)\Big) \log(\sigma) + \sum_{\substack{i=1 \\ y_i \in (0,1)}}^{n} \log(\pi + \sigma(\pi - 1)\log(y_i))$$
$$+(\sigma - 1) \sum_{\substack{i=1 \\ y_i \in (0,1)}}^{n} \log(y_i).$$

Then, the corresponding score function is given by $U(\boldsymbol{\nu}) = (U_\alpha(\alpha), U_p(p), U_\sigma(\sigma), U_\pi(\pi))$, where

$$U_\alpha(\alpha) = \frac{1}{\alpha} \sum_{i=1}^{n} I_{\{0,1\}}(y_i) - \frac{1}{1-\alpha}\Big(n - \sum_{i=1}^{n} I_{\{0,1\}}(y_i)\Big);$$

$$U_p(p) = \frac{1}{p} \sum_{i=1}^{n} y_i I_{\{0,1\}}(y_i) - \frac{1}{1-p}\Big(\sum_{i=1}^{n} I_{\{0,1\}}(y_i) - \sum_{i=1}^{n} I_{[1]}(y_i)\Big);$$

$$U_\sigma(\sigma) = \frac{\Big(n - \sum_{i=1}^{n} I_{\{0,1\}}(y_i)\Big)}{\sigma} + \sum_{\substack{i=1 \\ y_i \in (0,1)}}^{n} \frac{(\pi - 1)\log(y_i)}{\pi + \sigma(\pi - 1)\log(y_i)} + \sum_{\substack{i=1 \\ y_i \in (0,1)}}^{n} \log(y_i);$$

$$U_\pi(\pi) = \sum_{\substack{i=1 \\ y_i \in (0,1)}}^{n} \frac{1 + \sigma \log(y_i)}{\pi + \sigma(\pi - 1)\log(y_i)}.$$

The ML estimates of the parameters are obtained by solving the equations established as

$$
\begin{aligned}
&U_\alpha(\alpha)=0 \implies \widehat{\alpha}=\frac{1}{n}\sum_{i=1}^{n} I_{\{0,1\}}(y_i);\, U_p(p)=0 \implies \widehat{p}=\frac{\sum_{i=1}^{n} I_{[1]}(y_i)}{\sum_{i=1}^{n} I_{\{0,1\}}(y_i)}\\
&U_\sigma(\sigma)=0 \implies \frac{n-\sum_{i=1}^{n} I_{\{0,1\}}(y_i)}{\sigma}+\sum_{\substack{i=1\\ y_i\in(0,1)}}^{n}\frac{(\pi-1)\log(y_i)}{\pi+\sigma(\pi-1)\log(y_i)}=-\sum_{\substack{i=1\\ y_i\in(0,1)}}^{n}\log(y_i);\\
&U_\pi(\pi)=0 \implies \sum_{\substack{i=1\\ y_i\in(0,1)}}^{n}\frac{1+\sigma\log(y_i)}{\pi+\sigma(\pi-1)\log(y_i)}=0.
\end{aligned}
$$

Since the last two ML equations are non-linear in σ and π, they are solved using numerical computational techniques. The corresponding Fisher information matrix for the ZOILL distribution is given by

$$
K(\boldsymbol{\nu})=\begin{pmatrix} k_{\alpha\alpha} & k_{\alpha p} & k_{\alpha\pi} & \kappa_{\alpha\sigma}\\ k_{p\alpha} & k_{pp} & k_{p\pi} & k_{p\sigma}\\ k_{\pi\alpha} & k_{\pi p} & k_{\pi\pi} & k_{\pi\sigma}\\ k_{\sigma\alpha} & k_{\sigma p} & k_{\sigma\pi} & k_{\sigma\sigma}\end{pmatrix},
$$

where

$$
\begin{aligned}
k_{\alpha\alpha}&=\frac{n}{\alpha(1-\alpha)};\\
k_{\alpha p}&=k_{\alpha\pi}=k_{\alpha\sigma}=k_{p\alpha}=k_{p\pi}=k_{p\sigma}=k_{\pi\alpha}=k_{\pi p}=k_{\sigma\alpha}=k_{\sigma p}=0;\\
k_{pp}&=\frac{n\alpha}{p(1-p)};\\
k_{\pi\pi}&=nE_1\left(\frac{\pi}{1-\pi}\right)\left(\frac{2\pi\exp(-\frac{\pi}{\pi-1})}{(\pi-1)^3}-\frac{\exp(-\frac{\pi}{\pi-1})}{(\pi-1)}-\frac{\pi^2\exp(-\frac{\pi}{\pi-1})}{(\pi-1)^3}\right)\\
&\quad-\frac{2n\pi\exp(-\frac{2\pi}{\pi-1})}{(\pi-1)^2}+\frac{2n}{(\pi-1)}-\frac{n\exp(-\frac{\pi}{\pi-1})}{(\pi-1)}\Gamma\left(2,\frac{\pi}{1-\pi}\right);\\
k_{\pi\sigma}&=k_{\sigma\pi}=-\frac{n}{\sigma}+\frac{n\exp(-\frac{\pi}{\pi-1})}{\sigma(\pi-1)}\left(\pi E_1\left(\frac{\pi}{1-\pi}\right)-(1-\pi)\exp\left(-\frac{\pi}{\pi-1}\right)\right)\\
&\quad-\frac{n}{\sigma}\exp\left(-\frac{-\pi}{\pi-1}\right)\Gamma\left(2,\frac{\pi}{1-\pi}\right)-\frac{n\pi^2\exp\left(-\frac{\pi}{1-\pi}\right)}{\sigma\,(\pi-1)^2}E_1\left(\frac{\pi}{1-\pi}\right)+\frac{2n\pi(1-\pi)}{\sigma(\pi-1)^2};\\
k_{\sigma\sigma}&=\frac{n(1-\alpha)}{\sigma^2}-\frac{n(\pi-1)}{\sigma^2}\exp\left(-\frac{\pi}{\pi-1}\right)\Gamma\left(2,\frac{\pi}{1-\pi}\right)\\
&\quad-\frac{n\pi^2\exp\left(-\frac{\pi}{1-\pi}\right)}{\sigma^2(\pi-1)}E_1\left(\frac{\pi}{1-\pi}\right)+\frac{2n\pi}{\sigma^2}.
\end{aligned}
$$

Suppose $\widehat{\boldsymbol{\nu}}=(\widehat{\alpha},\widehat{p},\widehat{\sigma},\widehat{\pi})$. In large samples, $\sqrt{n}(\widehat{\boldsymbol{\nu}}-\boldsymbol{\nu})$ is asymptotically normally distributed, that is, $\sqrt{n}(\widehat{\boldsymbol{\nu}}-\boldsymbol{\nu}) \xrightarrow{D} N_4(0,k(\boldsymbol{\nu})^{-1})$ where $k(\boldsymbol{\nu})$ is the Fisher information matrix. Let $\delta\in(0,0.5)$. Then $(1-\delta)\times 100\%$ asymptotic confidence intervals for α, p, σ and π are given, respectively, by $\widehat{\alpha}\pm Z_{1-\delta/2}\mathrm{SE}(\widehat{\alpha})$, $\widehat{p}\pm Z_{1-\delta/2}\mathrm{SE}(\widehat{p})$,$\widehat{\sigma}\pm Z_{1-\delta/2}\mathrm{SE}(\widehat{\theta})$; and $\widehat{\pi}\pm Z_{1-\delta/2}\mathrm{SE}(\widehat{\pi})$.

3.4 Method of moments

The moment method consists in estimating the parameters by equating the expressions for $\mathrm{E}(Y^r)$ to the corresponding sample moments and then proceeding to solve the resulting equations for those parameters. Let us denote the first two sample moments by m_1' and m_2', respectively. After reparametrizing the log-Lindley PDF and then the ZOILL distribution, the rth moment of the ILL distribution becomes, for $c = 0$ or $c = 1$, the expression formulated as

$$\mathrm{E}(Y^r) = \alpha c + (1-\alpha)\mu_r'^{*}, \quad r = 1, 2, \dots,$$

where $\mu_r'^{*} = \sigma(\sigma + r\pi)/(\sigma + r)^2$ is the rth raw moment of the reparametrized log-Lindley distribution and the rth moment of the ZOILL distribution is given by

$$\mathrm{E}(Y^r) = \alpha p + (1-\alpha)\frac{\sigma(\sigma + r\pi)}{(\sigma + r)^2}, \quad r = 1, 2, \dots$$

To obtain the moment estimates of the parameters of the reparametrized ILL distribution, we must solve the equations presented as

$$\alpha c + (1-\alpha)\frac{\sigma(\sigma+\pi)}{(\sigma+1)^2} = m_1', \quad \alpha c + (1-\alpha)\frac{\sigma(\sigma+2\pi)}{(\sigma+2)^2} = m_2'.$$

Substituting α by $1/n\sum_{i=1}^n I_{[c]}(y_i)$, we get the moment estimates of π and σ as

$$\widehat{\pi} = \frac{\sigma^2(\bar{y}_1 - 1) + 2\sigma\bar{y}_1 + \bar{y}_1}{\sigma}, \quad \widehat{\sigma} = \frac{2\bar{y}_1 - \bar{y}_2 \pm \sqrt{3\bar{y}_2(2\bar{y}_1 - \bar{y}_2) + 2(\bar{y}_1 - 2\bar{y}_2)}}{\bar{y}_2 - 2\bar{y}_1 + 1},$$

where

$$\bar{y}_1 = \frac{m_1' - \alpha c}{1-\alpha}, \quad \bar{y}_2 = \frac{m_2' - \alpha c}{1-\alpha}.$$

To estimate the parameters of the reparametrized ZOILL distribution using the moment method, we must solve the equations stated as

$$\alpha p + (1-\alpha)\frac{\sigma(\sigma+\pi)}{(\sigma+1)^2} = m_1', \quad \alpha p + (1-\alpha)\frac{\sigma(\sigma+2\pi)}{(\sigma+2)^2} = m_2'.$$

Substituting α by $1/n\sum_{i=1}^n I_{[c]}(y_i)$ and p by $\sum_{i=1}^n I_{[1]}(y_i)/\sum_{i=1}^n I_{\{0,1\}}(y_i)$ and after some simplifications, the moment estimates of σ and λ can be obtained similarly as in the previous section arriving at

$$\widehat{\pi} = \frac{\sigma^2(\bar{y}_1 - 1) + 2\sigma\bar{y}_1 + \bar{y}_1}{\sigma}, \quad \widehat{\sigma} = \frac{2\bar{y}_1 - \bar{y}_2 \pm \sqrt{3\bar{y}_2(2\bar{y}_1 - \bar{y}_2) + 2(\bar{y}_1 - 2\bar{y}_2)}}{\bar{y}_2 - 2\bar{y}_1 + 1}$$

$$\bar{y}_1 = \frac{m_1' - \alpha p}{1-\alpha}, \quad \bar{y}_2 = \frac{m_2' - \alpha p}{1-\alpha}$$

As observed and reported in Jodrá and Jiménez-Gamero (2016), the moment method for the log-Lindley distribution has serious limitations and is not advisable. As expected, we have observed the same for the proposed inflated version. Hence, we have not considered the moment method in our simulation study in the next section.

4. Simulation Study

In this section, two separate Monte-Carlo simulation studies, one for the ZILL distribution and the other one for the ZOILL are conducted to generate random variables from the respective distributions. Thereafter, ML estimates based on the generated samples are calculated and their performance is assessed with the help of average bias (AB) and root mean square error (RMSE). To simulate n observations from zero-inflated log-Lindley distribution, we use Algorithm 1.

Algorithm *1* Algorithm to generate random numbers from ZILL(α, σ, π) distribution.

1: n random numbers were generated from the uniform distribution in $(0, 1)$, say U_i, for $i = 1, \ldots, n$.
2: If $U_i < \alpha$, then $y_i = 0$.
3: If $U_i \geq \alpha$, then
$$y_i = \exp\left(\frac{1}{\sigma(1-\pi)}\right)\exp\left(\frac{1}{\sigma}\right)W_{-1}\left(\frac{\exp\left(\frac{1}{\pi-1}\right)}{\pi-1}\right)U_i.$$
4: Repeat steps 1 to 3 until the required data are generated.

Observations are simulated from the ZOILL(α, p, σ, π) distribution using Algorithm 2.

Algorithm *2* Algorithm to generate random numbers from ZILL(α, σ, π) distribution.

1: n random numbers from the uniform distribution in $(0, 1)$ were generated first, say U_i, for $i = 1, \ldots n$.
2: If $U_i \leq \alpha p$, then $y_i = 0$.
3: If $U_i \leq \alpha$, then we assign $y_i = 1$; otherwise we assign
$$y_i = \exp\left(\frac{1}{\sigma(1-\pi)}\right)\exp\left(\frac{1}{\sigma}\right)W_{-1}\left(\frac{\exp\left(\frac{1}{\pi-1}\right)}{\pi-1}\right)U_i,$$
where W_{-1} denotes the negative branch of the Lambert W function.
4: Repeat steps 1 to 3 until the required data are generated.

The calculation of the ML estimate have been done using the `R` software with its package `maxLik`. To choose the initial values for implementation of the `maxBFGS` command under the `maxLik` package, the same procedure as mentioned in Jodrá and Jiménez-Gamero (2016) has been adopted. Accordingly, the initial guess value of σ is taken as $\sigma = (\widehat{\sigma}_0 + \widehat{\sigma}_1)/2$ and that of π is taken as 0.5. Here, $\widehat{\sigma}_0 = m_1'/(1 - m_1')$ and $\widehat{\sigma}_1 = (m_1' + \sqrt{m_1'})/(1 - m_1')$, m_1' being the first order sample moment.

From Table 1 and Table 2, it is seen that for both ZILL and ZOILL distributions, the inflation parameters are consistently estimated with rapid decrease in the RMSE with increase in the sample size, where as the other parameters, the decrease in RMSE is slower. For higher inflation, it is observed that the other parameters are estimated with lesser precision.

The `R` programming language and the packages `GoFKernel`, `maxLik` and `lamW` have been used for simulating from the proposed distribution and computing the AB and RMSE of the ML estimates of the parameters. For finding the ML estimates of the parameters from the zero inflated log-Lindley distribution, the starting values of σ and π are taken to be 0.8 and 0.5, respectively. For simulating from the ZOILL distribution, the starting value of both p and π has been taken to be 0.5. As for the starting value of σ, the formula for calculation of the same is presented in Section 3.3.

Table 1.: Simulation results for ZILL distribution for different set of values of α, σ and π.

	$\alpha = 0.1, \sigma = 0.5, \pi = 0.2$								
	α			σ			π		
n	ML estimate	AB	RMSE	ML estimate	AB	RMSE	ML estimate	AB	RMSE
50	0.1006	0.0006	0.0047	0.6550	0.1550	0.0219	0.1123	-0.0876	0.0179
100	0.1016	0.0016	0.0023	0.6526	0.1526	0.0152	0.1062	-0.0937	0.0109
250	0.0995	-0.0009	0.0009	0.6508	0.1508	0.0095	0.1066	-0.0933	0.0061
500	0.1006	0.0006	0.0004	0.6500	0.1500	0.0067	0.1052	-0.0947	0.0042
	$\alpha = 0.2, \sigma = 1.5, \pi = 0.7$								
	α			σ			π		
n	ML estimate	AB	RMSE	ML estimate	AB	RMSE	ML estimate	AB	RMSE
50	0.2012	0.0012	0.0063	3.1915	1.6915	0.2392	0.3076	-0.3423	0.0561
100	0.2002	0.0002	0.0031	3.1870	1.6870	0.1687	0.3000	-0.3999	0.0399
250	0.1992	-0.0007	0.0012	3.1646	1.6646	0.1052	0.3021	-0.3978	0.0251
500	0.2009	0.0009	0.0006	3.1687	1.6687	0.0746	0.3046	-0.3953	0.0176

Table 2.: Simulation results for ZOILL distribution for different set of values of α, σ, p and π.

	$\alpha = 0.1, \sigma = 0.5, p = 0.3, \pi = 0.2$					
	α			σ		
n	ML estimate	AB	RMSE	ML estimate	AB	RMSE
50	0.2017	0.0017	0.0063	0.7759	0.2759	0.0390
100	0.2003	0.0003	0.0032	0.7740	0.2740	0.0274
250	0.2003	0.0003	0.0012	0.7730	0.2730	0.0172
500	0.1997	-0.0002	0.0006	0.7743	0.2743	0.0122
	π			p		
n	ML estimate	AB	RMSE	ML estimate	AB	RMSE
50	0.0838	-0.1161	0.0191	0.2896	-0.0103	0.0169
100	0.0798	-0.1201	0.0127	0.2974	-0.0025	0.0082
250	0.0786	-0.1213	0.0077	0.2973	-0.0026	0.0034
500	0.0770	-0.1229	0.0055	0.2999	-1.9175	0.0016
	$\alpha = 0.6, \sigma = 1.5, p = 0.8, \pi = 0.7$					
	α			σ		
n	ML estimate	AB	RMSE	ML estimate	AB	RMSE
50	0.2998	-0.0001	0.0071	4.4580	2.9580	0.4183
100	0.3017	0.0017	0.0035	4.4005	2.9005	0.2905
250	0.3003	0.0003	0.0014	4.3881	2.8881	0.1826
500	0.2998	-0.0001	0.0007	4.3626	2.8626	0.1280
	π			p		
n	ML estimate	AB	RMSE	ML estimate	AB	RMSE
50	0.2645	-0.5354	0.0758	0.6985	-0.0014	0.0144
100	0.2678	-0.5321	0.0532	0.7027	0.0027	0.0068
250	0.2712	-0.5287	0.0334	0.7000	0.0003	0.0026
500	0.2726	-0.5273	0.0235	0.6991	-0.0008	0.0013

5. Data analysis: Applications

In this section, we illustrate our proposed model through some real-life data sets with four real-life data sets on High School Leaving Examination results of the State of Manipur, India, for the year 2020 presented in BSA (2020a), BSA (2020b), and BSA (2020c).

5.1 Exploratory data analysis

The variable is the proportion of who have passed the examination in 2020. Data set-I is from in 67 aided schools, data set-II is for 305 government schools, data set-III is from 523 private schools and the data set-IV is the universe of all schools data. The proposed model is compared with inflated beta (Ospina and Ferrari, 2010) and inflated unit Lindley distribution (Chakraborty and Bhattacharjee, 2021).

Table 3 shows the descriptive statistics of the data sets on proportion of students studying in government schools, private schools, aided schools and that of the combined data set of all schools of Manipur, India who have passed the High School Leaving Certificate (H.S.L.C.) Examination, 2020:

Table 3.: Descriptive statistics.

Data set	Minimum	Maximum	Mean	1st quartile	Median	3rd quartile
I (aided schools)	0	1	0.4219	0.0238	0.3250	0.7456
II (government schools)	0	1	0.3632	0.0476	0.3143	0.6154
III (private schools)	0	1	0.6705	0.4706	0.7520	0.9189
IV (all schools combined)	0	1	0.5471	0.2500	0.6000	0.8621

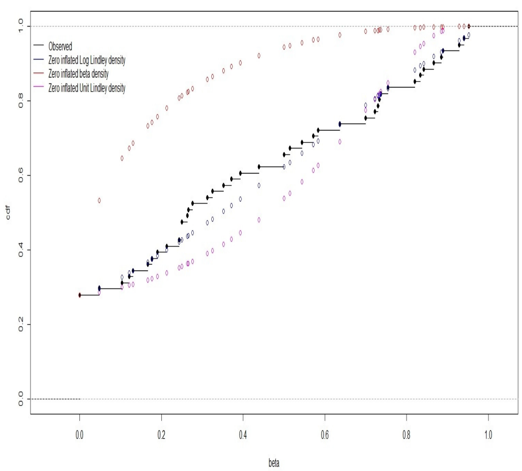
5.2 Parameter estimation

The ZILL, zero inflated unit Lindley (ZIUL) and zero inflated beta (ZIB) distributions have been fitted to these above mentioned data sets as well as the combined data set of all schools together, where the parameters have been estimated using the ML method. For the purpose of optimization of the log-likelihood function to obtain the ML estimates of the parameters, the moment estimates of the parameters can be considered as the initial values. For example, for the 'Aided schools' data set, the moment estimates of σ and π have come out to be 0.4622 and 0.0147, respectively. These values of the parameters have been used as initial guess values for the optimization routine. It can be similarly done for all the remaining data sets. The log-likelihood and Akaike information criterion (AIC) measures have been employed as the goodness of fit criteria to compare the fit of each of the distribution to the data sets. Table 4 displays the ML estimates and standard errors of the parameters, log-likelihood and AIC measures of the zero inflated log-Lindley, ZIUL and zero inflated beta which have been fitted to the four data sets.

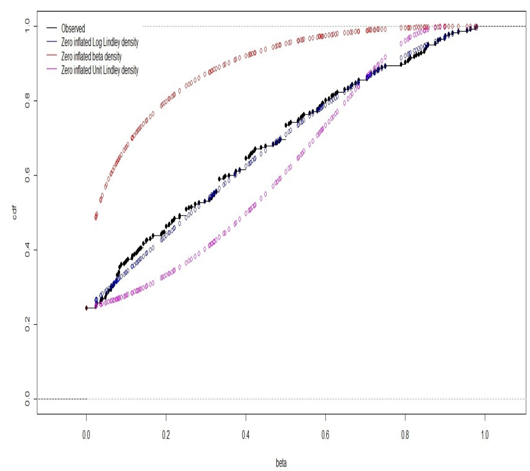
Table 4.: ML estimates, standard errors, log-likelihood and AIC values for fitting zero inflated models.

Data set	Distribution	ML estimate (SE) of the parameters	LL	AIC
I	ZILL	$\alpha = 0.2785(0.0574), \pi = 0.0338(0.1230), \sigma = 3.3403(0.3224)$	-6.1218	18.2436
	ZIB	$\alpha = 0.2785(0.0574), \mu = 0.5142(0.0389), \phi = 2.5074(0.4595)$	-35.3275	76.6551
	ZIUL	$\alpha = 0.2785(0.0574), \theta = 0.5273(0.0459)$	-15.1632	34.3264
II	ZILL	$\alpha = 0.2438(0.0255), \pi = 0.00002(0.0347), \sigma = 2.5048(0.1029)$	-25.3057	56.6114
	ZIB	$\alpha = 0.2438(0.0255), \mu = 0.5142(0.0389), \phi = 2.5034(0.2096)$	-146.3012	298.6023
	ZIUL	$\alpha = 0.2438(0.0255), \theta = 0.7617(0.0328)$	-96.2335	195.467
III	ZILL	$\alpha = 0.0440(0.0093), \pi = 0.2082(0.0437), \sigma = 5.3119(0.1658)$	-1.2036	8.4072
	ZIB	$\alpha = 0.0440(0.0093), \mu = 0.6599(0.0112), \phi = 2.6921(0.1591)$	-7.9643	21.9287
	ZIUL	$\alpha = 0.0440(0.0094), \theta = 0.2288(0.0065)$	-4.7269	13.4538
IV	ZILL	$\alpha = 0.1303(0.0117), \pi = 0.2057(0.0333), \sigma = 3.7290(0.0921)$	-20.2751	46.5502
	ZIB	$\alpha = 0.1303(0.0117), \mu = 0.5771(0.0099), \phi = 2.1898(0.0988)$	-289.3462	584.6923
	ZIUL	$\alpha = 0.1303(0.0117), \theta = 0.3011(0.0067)$	-187.3517	378.7034

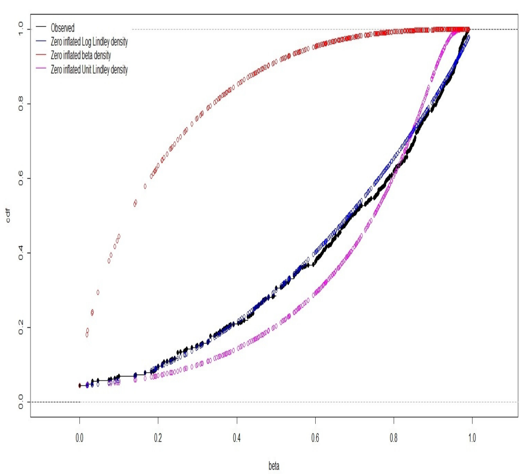
Figure 3 (a)-(d) exhibit the plot of the empirical CDF of the four datasets and the distribution functions of the fitted ZILL, ZIUL and ZIB distributions. It is clear from visual scrutiny of the plots in Figure 3 (a)-(d) that the proposed zero inflated log-Lindley distribution is a quite better fit to all the four datasets in comparison to the well established zero inflated beta and ZIUL distribution. It is clear from Table 4 that the AIC value for all the four data sets is the least corresponding to the zero inflated log-Lindley distribution. Thus, it can be concluded that the zero inflated log-Lindley distribution is able to model each of the data set better than the other zero inflated distribution considered.



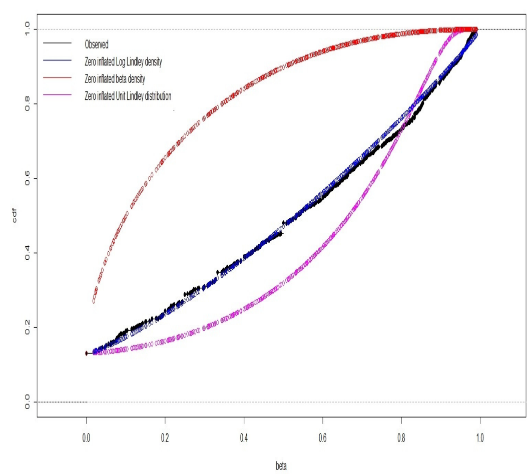
(a) data set I



(b) data set II



(c) data set III



(d) data set IV

Figure 3.: Observed PDF and CDF plot of the fitted ZILL, ZIUL and ZIB distributions for proportion of students passing the H.S.L.C. exam 2020, Manipur for indicated data set.

In a similar manner, the fitting of ZOILL distribution is compared with the zero-and-one inflated versions of the beta (ZOIB) and unit Lindley (ZOIUL) distributions, where the parameter estimation is carried out using the ML method and the AIC value is applied to compare the goodness-of-fit of the distributions. Table 5 exhibits the ML estimates and standard errors of the parameters, log-likelihood and AIC measures of the ZOILL, ZOIB and ZOIUL models in the four data sets.

Table 5 also indicates that the ZOILL distribution is a better fit to each of the four data sets as compared to the zero-and-one inflated version of the beta and unit Lindley distribution as the AIC value corresponding to the zero-and-one log-Lindley distributions is the least. Hence, it can be concluded that the zero-and-one log-Lindley distribution has a better modeling potential as compared to the other competing distributions.

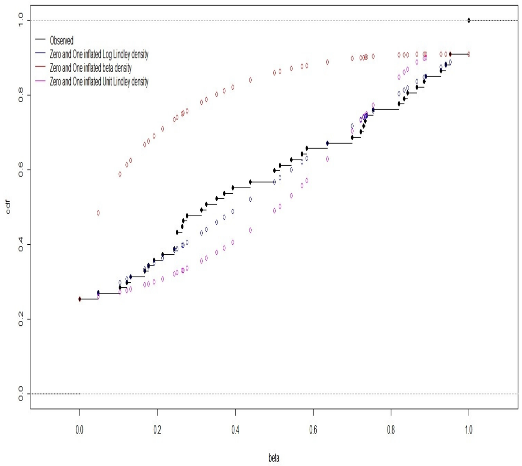
Table 5.: ML estimates, standard errors, log-likelihood and AIC values for fitting zero and one inflated models.

Data set	Distribution	ML estimate (SE) of the parameters	LL	AIC
I	ZOILL	$\alpha = 0.3432(0.0580), \pi = 0.0338(0.1230), \sigma = 3.3403(0.3224), p = 0.2608(0.0915)$	-14.0787	36.1574
	ZOIB	$\alpha = 0.3432(0.0580), \mu = 0.5142(0.0389), \phi = 2.5074(0.4595), p = 0.2608(0.0915)$	-55.5281	119.0562
	ZOIUL	$\alpha = 0.3432(0.0580), \theta = 0.9615(0.0943), p = 0.2608(0.0915)$	-30.2491	66.4982
II	ZOILL	$\alpha = 0.2983(0.0262), \pi = 0.00002(0.0347), \sigma = 2.5048(0.1029), p = 0.2417(0.0448)$	-76.7908	169.5816
	ZOIB	$\alpha = 0.2983(0.0262), \mu = 0.5142(0.0389), \phi = 2.5034(0.2096), p = 0.2417(0.0448)$	-225.3318	458.6637
	ZOIUL	$\alpha = 0.2983(0.0262), \theta = 0.7617(0.0387), p = 0.2416(0.0490)$	-100.6649	207.3298
III	ZOILL	$\alpha = 0.1281(0.0146), \pi = 0.2082(0.0437), \sigma = 5.3119(0.1658), p = 0.6855(0.0566)$	-44.8977	97.7954
	ZOIB	$\alpha = 0.1281(0.0146), \mu = 0.6599(0.0112), \phi = 2.6921(0.1591), p = 0.6855(0.0566)$	-163.7026	335.4052
	ZOIUL	$\alpha = 0.1281(0.0146), \theta = 0.2288(0.0076), p = 0.6855(0.0566)$	-75.5236	157.0472
IV	ZOILL	$\alpha = 0.2022(0.0134), \pi = 0.2057(0.0333), \sigma = 3.7290(0.0921), p = 0.4088(0.0365)$	-115.5306	239.0612
	ZOIB	$\alpha = 0.2022(0.0134), \mu = 0.5771(0.0099), \phi = 2.1898(0.0988), p = 0.4088(0.0365)$	-544.6631	1097.326
	ZOIUL	$\alpha = 0.2022(0.0134), \theta = 0.3011(0.0080), p = 0.4088(0.0365)$	-316.8217	639.6434

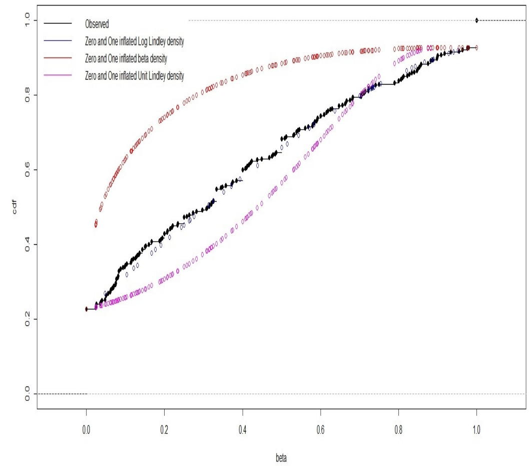
Figure 4 (a)-(d) exhibit the plot of the empirical CDF of the four datasets and the distribution functions of the fitted ZOILL, ZOIUL and ZOIB distributions. The plots in Figure 4 (a)-(d) clearly show that the proposed ZOILL distribution is a much better fit to all the four datasets as compared to the more popular ZOIB distribution and the recently developed ZOIUL distribution.

6. Conclusions, limitations and future research

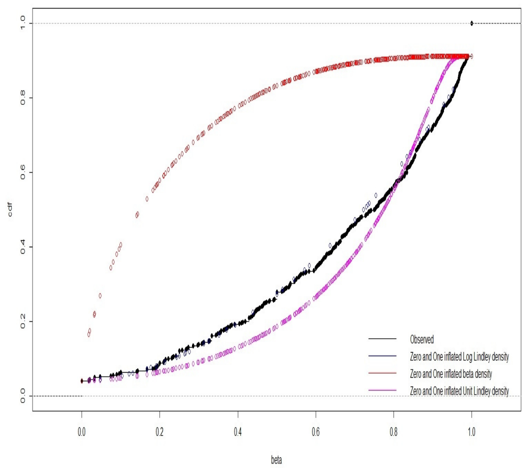
The idea behind this article emerged from a work on inflated version of the beta distribution, which is able to model data having observations equal to 0 or 1, apart from those lying in the interval $(0, 1)$. A very few alternatives to this distribution are available in the literature. This motivated us to pursue this work, whose chief contributions are (i) introducing a inflated version of the log-Lindley distribution to deal with bounded data in unit interval with possible mass at 0 and/or 1, (ii) assessing the performance of the different parameter estimation procedures through simulation studies and (iii) analysis of real life pass proportion data sets using programming code in the `R` software. The proposed distribution (both the zero-or-one and zero-and-one versions) are found to be members of the exponential family. This article further has established the proposed model as a viable alternative to well known inflated beta distribution and the recently proposed inflated unit Lindley distribution. One of the limitations of the present study is that simulation has been performed for only two sets of combination of the parameters for both the zero inflated and zero-and-one inflated case. Future works will include among other the study of regression modeling by considering available covariates and related inferences.



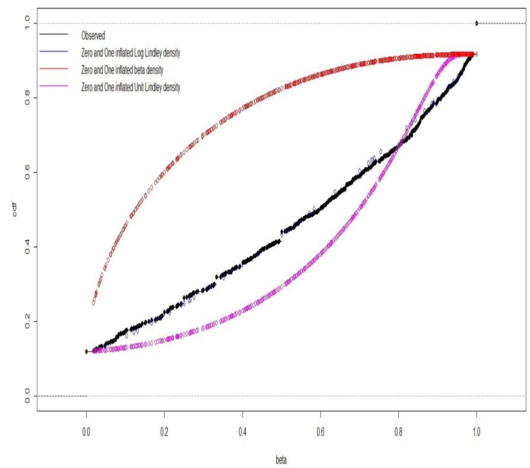
(a) data set I



(b) data set II



(c) data set III



(d) data set IV

Figure 4.: Observed PDF and CDF plot of the fitted ZOILL, ZOIUL and ZOIB distributions for proportion of students passing the H.S.L.C. exam 2020, Manipur for indicated data set.

Appendix

This section contains the proofs of the Theorems 2.1 and 2.2 presented in Section 2.3 of the article.

Proof of Theorem 2.1 The PDF in Equation (2.3) can be re-written as

$$f\left(y;\alpha,\sigma,\lambda\right)=\alpha^{I_c(y)}\left(1-\alpha\right)^{1-I_c(y)}\left(\frac{\sigma^2}{1+\lambda\sigma}\left(\lambda-\log(y)\right)y^{\sigma-1}\right)^{1-I_c(y)}, \tag{1}$$

where

$$I_c(y)=\begin{cases}1, & \text{if} \quad y=c, \quad \text{with} \quad c=0 \quad \text{or} \quad c=1;\\ 0, & \text{if} \quad y\in(0,1)\,.\end{cases}$$

By denoting $T_1(y) = I_c(y)$, $T_2(y) = \log(y)(1 - I_c(y))$, $\eta_1 = \log(\alpha(1+\lambda\sigma))/(\sigma^2(1-\alpha))$ and $\eta_2 = \sigma$, the PDF in Equation (1) can be arranged as

$$f\left(y;\eta_1,\eta_2\right) = \exp\Bigg(I_c(y)\bigg(\log\left(\frac{\alpha}{1-\alpha}\right) - \log\left(\frac{\sigma^2}{1+\lambda\sigma}\right)\bigg) + \left(1 - I_c(y)\right)\sigma\log(y) \\ + \bigg(\log\left(1-\alpha\right) + \log\left(\frac{\sigma^2}{1+\lambda\sigma}\right)\bigg)\Bigg)\left(\frac{\lambda - \log(y)}{y}\right)^{1-I_c(y)}.$$

Now, by defining $T(\mathbf{y}) = (T_1(y), T_2(y))$, $\boldsymbol{\eta} = (\eta_1, \eta_2)$, $B(\boldsymbol{\eta}) = \log(\sigma^2(1-\alpha)/(1+\lambda\sigma))$, and $h(y) = ((\lambda - \log(y))/y)^{1-I_c(y)}$, the PDF can be expressed as

$$f\left(y;\eta_1,\eta_2\right) = \exp\left(\boldsymbol{\eta}^\top T\left(\mathbf{y}\right) + B\left(\boldsymbol{\eta}\right)\right) h\left(y\right),$$

where the function $B(\boldsymbol{\eta})$ is a real valued function of η_1, η_2, and $h(y)$ is a positive real valued function. The transformation from (α, σ) to (η_1, η_2) is obviously one-one from $(0,1)\times R^+$ to $R\times R^+$. Hence, the PDF presented in Equation (2.3) belongs to a two parameter exponential family distribution when λ is fixed.

Proof of Theorem 2.2 The PDF stated in Equation (2.5) can be re-written as

$$f\left(y;p,\alpha,\sigma,\lambda\right) = \alpha^{I_{\{0,1\}}(y)}\left(1-\alpha\right)^{1-I_{\{0,1\}}(y)} p^{yI_{\{0,1\}}(y)}\left(1-p\right)^{I_{\{0,1\}}(y) - yI_{\{0,1\}}(y)} \\ \times\left(\left(1-\alpha\right)\frac{\sigma^2}{1+\lambda\sigma}\left(\lambda - \log(y)\right)y^{\sigma-1}\right)^{1-I_{\{0,1\}}(y)}.$$

where

$$I_{\{0,1\}}(y) = \begin{cases} 1, & \text{if} \quad y = 0 \text{ or } 1; \\ 0, & \text{if} \quad y \in (0,1)\,. \end{cases}$$

By denoting $T_1(y) = I_{\{0,1\}}(y)$, $T_2(y) = 1 - I_{\{0,1\}}(y)$, $T_3(y) = yI_{\{0,1\}}(y)$, $\eta_1 = \log\left((\alpha\left(1-p\right)\left(1+\sigma\lambda\right))/(\sigma^2\left(1-\alpha\right)^2)\right)$, $\eta_2 = \sigma$, and $\eta_3 = \log\left(p/(1-p)\right)$, the PDF in Equation (2) can be written as

$$f\left(y;\eta_1,\eta_2,\eta_3\right) = \exp\left(I_{\{0,1\}}(y)\left(\log\left(\tfrac{\alpha(1-p)(1+\lambda\sigma)}{(1-\alpha)^2\sigma^2}\right)\right) + \left(1 - I_{\{0,1\}}(y)\right)\sigma\log(y)\right. \\ \left. + yI_{\{0,1\}}(y)\log\left(\tfrac{p}{1-p}\right) + \log\left(\tfrac{(1-\alpha)^2\sigma^2}{(1+\lambda\sigma)}\right)\right)\left(\tfrac{\lambda-\log(y)}{y}\right)^{1-I_{\{0,1\}}(y)}.$$

Now, by defining $B^*\left(\boldsymbol{\eta}\right) = \log\left(\sigma^2\left(1-\alpha\right)^2/(1+\lambda\sigma)\right)$, $T(\mathbf{y}) = (T_1(y), T_2(y), T_3(y))$, $\boldsymbol{\eta} = (\eta_1, \eta_2, \eta_3)$, and $h(y) = ((\lambda - \log(y))/y)^{1-I_{\{0,1\}}(y)}$, the PDF can be expressed as

$$f\left(y;\eta_1,\eta_2,\eta_3\right) = \exp\left(\boldsymbol{\eta}^\top T\left(\mathbf{y}\right) + B^*\left(\boldsymbol{\eta}\right)\right) h\left(y\right),$$

where the functions $B^*(\boldsymbol{\eta})$ is a real valued function of (η_1, η_2, η_3), $h(y)$ is a positive real valued function. Further, the transformation from (α, p, σ) to (η_1, η_2, η_3) is one-one from $(0,1)\times(0,1)\times\mathrm{R}^+$ to $\mathrm{R}\times\mathrm{R}^+\times\mathrm{R}$. Hence, the PDF given in Equation (2.5) belongs to a three parameter exponential family distribution when λ is fixed.

Author Contributions

Conceptualization and methodology: S.C.; data preparation: S.B.; formal analysis: S.B. and S.C.; software and validation: S.B. and S.C.; supervision: S.C.; writing-original draft preparation: S.B.; writing-review and editing: S.C. and S.B. All authors have read and agreed to the published version of the article.

Funding

No funding was availed to carry out this research work.

Conflicts of interest

The authors declare no conflict of interest.

Acknowledgements

We would like to express our heartfelt gratitude to the anonymous referees, the Associate Editor and the Editors-in-chief for their constructive comments and valuable suggestions which significantly improved the earlier version of the article.

References

Board of Secondary Education, Manipur., 2020a. School-wise pass percentage for H.S.L.C. Examination, 2020: Aided. http://e-paolive.net/download/2020/HSLC/7_{PC}_(AIDED).pdf.

Board of Secondary Education, Manipur., 2020b. School-wise pass percentage for H.S.L.C. Examination, 2020: Government. http://e-paolive.net/download/2020/HSLC/6_{PC}_(GOVT).pdf.

Board of Secondary Education, Manipur., 2020c. School-wise pass percentage for H.S.L.C. Examination, 2020: Private. http://e-paolive.net/download/2020/HSLC/8_{PC}_(PVT).pdf.

Burch, B. and Egbert, J., 2019. Zero-inflated beta distribution applied to word frequency and lexical dispersion in corpus linguistics. Journal of Applied Statistics, 47(2), 337–353.

Chakraborty, S. and Bhattacharjee, S., 2021. Modeling proportion of success in high school leaving certificate examination-a comparative study of inflated unit Lindley and inflated beta distribution. Journal of Mathematical and Computational Science, 11, 7170–7198.

Cribari-Neto, F. and Santos, J., 2019. Inflated Kumaraswamy distributions. Annals of the Brazilian Academy of Sciences, 91, e20180955.

Goméz-Déniz, E., Sordo, M.A., and Calderín-Ojeda, E., 2014. The log-Lindley distribution as an alternative to the beta regression model with applications in insurance. Insurance: Mathematics and Economics, 54, 49–57.

Hashimoto, E.M., Ortega, E.M, Cordeiro, G.M., Cancho, V.G., and Klauberg, C., 2019. Zero-spiked regression models generated by gamma random variables with application in the resin oil production. Journal of Statistical Computation and Simulation, 89, 52–70.

Jodrá, P. and Jiménez-Gamero, M.D., 2016. A note on the log-Lindley distribution. Insurance: Mathematics and Economics, 71, 189–194.

Johnson, N.L., Kotz, S., and Balakrishnan, N., 1995. Continuous univariate distributions. Wiley, New York, USA.

Kumaraswamy, P., 1980. A generalized probability density function for double-bounded random processes. Journal of Hydrology, 46, 79–88.

Liu, L., Shih, Y.C.T., Strawderman, R.L., Zhang, D., Johnson, B.A., and Chai, H., 2019. Statistical analysis of zero-inflated nonnegative continuous data: A review. Statistical Science, 34, 253–279.

Mazucheli, J., Menezes, A.F.B., and Chakraborty, S., 2019. On the one parameter unit-Lindley distribution and its associated regression model for proportion data. Journal of Applied Statistics, 46, 700–714.

Mazucheli, J., Bapat, S.R., and Menezes, A.F.B., 2020. A new one-parameter unit-Lindley distribution. Chilean Journal of Statistics, 11(1), 53–67.

Ospina, R. and Ferrari, S.L.P., 2010. Inflated beta distributions. Statistical Papers, 51, 111–126.

Tomazella, V., Pereira, G.H., Nobre, J.S., and Santos-Neto, M., 2019. Zero-adjusted reparameterized Birnbaum-Saunders regression model. Statistics and Probability Letters, 149, 142–145.