ON AN INFLATED UNIT-LINDLEY DISTRIBUTION

Sudeep R. Bapat⋆

Department of Operations Management and Quantitative Techniques
Indian Institute of Management, Indore, India
Email: sudeepb@iimidr.ac.in

Rohit Bhardwaj

Department of Decision Sciences
Indian Institute of Management, Bangalore, India
Email: rohit.bhardwaj@iimb.ac.in

SUMMARY

Modeling fractional data in various real life scenarios is a challenging task. This paper considers situations where fractional data are observed in the interval [0,1]. The unit-Lindley distribution has been discussed in the literature where its support lies between 0 and 1. In this paper, we focus on an inflated variant of the unit-Lindley distribution, where the inflation occurs at both 0 and 1. Various properties of the inflated unit-Lindley distribution are discussed and examined, including point estimation based on the maximum likelihood method and interval estimation. Finally, extensive Monte Carlo simulation and real-data analyses are carried out to compare the fit of our proposed distribution along with some of the existing ones such as the inflated beta and the inflated Kumaraswamy distributions.

Keywords and phrases: Inflated data; Inflated Beta distribution; Inflated Kumaraswamy distribution; Unit data

1   Introduction

Often we come across continuous data which lies strictly between $(0, 1)$ and practitioners need to model these using appropriate distributions. Analyzing proportions or percentages etc. fall in this category. One of the widely used distributions is the Beta distribution, which is being applied to such datasets because of its flexibility. However there is a drawback in using the Beta distribution as it is not very appropriate under some real life situations, e.g. hydrological data. Keeping this in the background, Kumaraswamy (1976) introduced another similar distribution to counter the Beta model, which is called as the Kumaraswamy distribution. An advantage here is that its distribution and quantile functions can be expressed in closed forms. On the other hand, a Lindley distribution with support on $(0, \infty)$, which was introduced by Ghitany et al. (2008) is one of the widely discussed distributions over the last decade. One may additionally refer to Nadarajah et al. (2011) or Shanker and Mishra (2013) among others, for extensions and variations of a general Lindley model. Very

---

⋆ Corresponding author
© Institute of Statistical Research and Training (ISRT), University of Dhaka, Dhaka 1000, Bangladesh.

recently, Mazucheli et al. (2019a) and Mazucheli et al. (2020a) introduced the unit-Lindley and the new-unit Lindley distributions having supports on $(0, 1)$ and $(0, 1]$ respectively, by using the transformations $Y = X/(1 + X)$ and $Y = 1/(1 + X)$ respectively, where X is a suitable Lindley distributed random variable. Some of the key advantages in using the above two models instead of a Beta model are existence of closed forms of the distribution function and the moments. Further, these two models only contain a single parameter and hence are simpler to handle. Other proposed unit-distributions include unit-Gamma (Grassia, 1977), unit-inverse Gaussian (Ghitany et al., 2019), unit-Weibull (Mazucheli et al., 2020b) etc. among many others.

On a separate note, there are abundant scenarios where such unit datasets also contain 0's, 1's or both 0's and 1's. Hence, one needs to assume values which lie between $[0, 1)$, $(0, 1]$ or $[0, 1]$, and analyze such datasets using appropriate inflated models. Such inflated models contain a mixture distribution with the continuous part being modeled through the actual distribution, whereas a degenerate distribution is fitted on 0's and 1's. In the same spirit, Ospina and Ferrari (2010) introduced the inflated beta distribution, whose density function is given by

$$f_{\text{beinf}}(y; \alpha, \gamma, \mu, \phi) = \begin{cases} \alpha(1 - \gamma) & \text{if y = 0} \\ \alpha\gamma & \text{if y = 1} \\ (1 - \alpha)f(y; \mu, \phi) & \text{if y} \in \text{(0,1),} \end{cases}$$

where α is the mixing parameter, γ is the associated Bernoulli parameter and $f(y; \mu, \phi)$ is the usual density function of a Beta$(0, 1)$ distribution. As a follow-up, Cribari and Santos (2019) introduced an inflated Kumaraswamy distribution which was seen to be better than the inflated beta model under some specific real life situations. Its density function can be written down as follows

$$f_{\text{zoik}}(y; \lambda, p, \alpha, \beta) = \begin{cases} \lambda(1 - p) & \text{if y = 0} \\ \lambda p & \text{if y = 1} \\ (1 - \lambda)g(y; \alpha, \beta) & \text{if y} \in \text{(0,1),} \end{cases}$$

where λ is the mixing parameter, p is the associated Bernoulli parameter and $g(y; \alpha, \beta)$ is the usual density function of a Kumaraswamy$(0, 1)$ distribution. In this paper, we try to extend this front by proposing a zero- and one- inflated unit-Lindley or better called as an inflated unit-Lindley (ULINF) distribution in a hope that it will be advantageous according to the earlier mentioned reasons. As for its continuous part, we will focus on the unit-Lindley distribution introduced by Mazucheli et al. (2019b) whose cumulative distribution function is given by

$$F_{\text{ulinf}}(y; \theta) = 1 - \left(1 - \frac{\theta y}{(1 + \theta)(y - 1)}\right)\exp\left(-\frac{\theta y}{1 - y}\right), \quad 0 < y < 1; \ \theta > 0.$$

Its corresponding probability density function is given by

$$f_{\text{ulinf}}(y; \theta) = \frac{\theta^2}{1 + \theta}(1 - y)^{-3}\exp\left(-\frac{\theta y}{1 - y}\right), \quad 0 < y < 1; \ \theta > 0, \tag{1.1}$$

where $\theta > 0$, is the only unknown parameter.

Recently, Ferreira et al. (2020) has proposed a one-inflated unit-Lindley distribution as a preprint. However our current work explores more on an inflated unit-Lindley distribution (inflated at both 0 and 1). The structure of the paper is as follows: In Section 2 we introduce the inflated unit-Lindley distribution along with its properties and behavior of its density function over various parameter values. Section 3 talks about maximum likelihood estimation of the unknown parameters. We present extensive simulation studies, comparisons and analysis of practical real datasets in Sections 4 and 5 respectively. We end with brief conclusions and some limitations in Section 6.

2 The Inflated Unit-Lindley Distribution

As a build-up towards proposing an appropriate inflated model, we will present a very brief review of finite mixture models. Finite mixture models are widely used to model the data from a statistical population that consists of two or more sub-populations.

Let $\boldsymbol{y} = \{y_i, i = 1, \ldots, n\}$ be a random sample of size n from an m-component finite mixture with the following density:

$$f(y_i; \boldsymbol{\theta}) = \sum_{j=1}^{m} \alpha_j f_j(y_i; \boldsymbol{\theta_j}), \tag{2.1}$$

where $f_j(y_i; \theta)$ represent the probability densities (or masses), $\boldsymbol{\theta} = (\alpha_1, \alpha_2, \ldots, \alpha_m, \theta_1, \theta_2, \ldots, \theta_m)$ and α_j are non negative quantities such that $\alpha_1 + \alpha_2 + \cdots + \alpha_m = 1$, i.e., $0 \leq \alpha_j \leq 1$ for $j = 1, 2, \ldots, m$. For a much detailed exposition of finite mixture models, one may refer to McLachlan and Kaye (1988) or McLachlan and David (2000).

In our case, we propose a 2-component finite mixture model in which one component captures the inflated zeros and ones while another captures the values in $(0, 1)$. We use an appropriate Bernoulli distribution as the first component with PMF $f_1(.)$ and CDF $F_1(.)$ and a unit-Lindley distribution as a second component with PDF $f_2(.)$ and CDF $F_2(.)$. Hence, the probability density function (PDF) of the inflated unit-Lindley distribution (ULINF) can be written, by using (2.1), as a weighted sum of these two distributions, and its functional form is given by

$$f_{\text{ulinf}}(y; \alpha, p, \theta) = \alpha f_1(y; p) + (1 - \alpha) f_2(y; \theta), \tag{2.2}$$

while the functional form of its cumulative distribution function (CDF) is given by

$$F_{\text{ulinf}}(y; \alpha, p, \theta) = \alpha F_1(y; p) + (1 - \alpha) F_2(y; \theta),$$

where p $(0 \leq p \leq 1)$ is a Bernoulli proportion, $\theta (\theta > 0)$ is the unit-Lindley parameter and α is the mixture proportion or weight as discussed in (2.1).

Definition 2.1. Let Y be a random variable with support in the interval [0,1]. We say that Y has an inflated unit-Lindley distribution (ULINF) with parameters α, p and θ if its probability density function with respect to the measure space generated by the functional form given in equation (2.2)
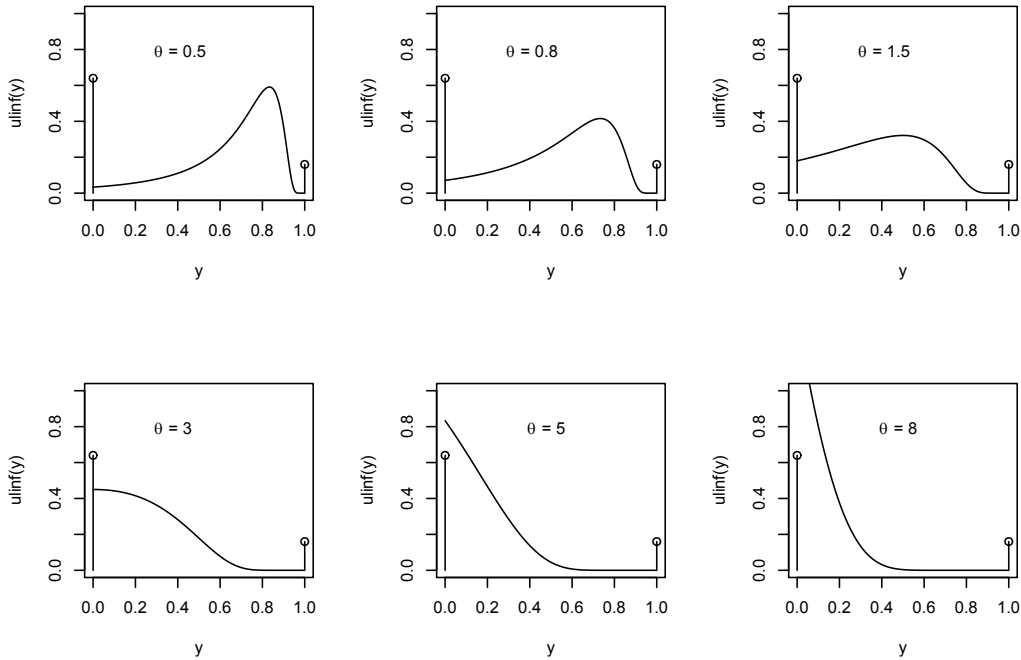
Figure 1: ULINF densities at different values of θ, with $p = 0.2$ and $\alpha = 0.8$

can be written as

$$f_{\text{ulinf}}(y; \alpha, p, \theta) = \begin{cases} \alpha(1-p) & \text{if y = 0} \\ \alpha p & \text{if y = 1} \\ (1-\alpha)f(y; \theta) & \text{if y} \in \text{(0,1),} \end{cases} \tag{2.3}$$

where $f(y; \theta)$ is the unit-Lindley density as discussed in (1.1) with $\theta > 0$ and $0 \leq \alpha, p \leq 1$. We write Y $\sim$ ULINF(α, p, θ). Note that if Y $\sim$ ULINF(α, p, θ), then $P(Y = 0) = \alpha(1-p)$ and $P(Y = 1) = \alpha p$. Further, $P(Y \in (0,1)) = (1-\alpha)\int_a^b f(y|\theta)$, for $y \in (0,1)$ and $0 < a < b < 1$.

Figure 1 shows the ULINF density plots for different combinations of θ with fixed p = 0.2 and α = 0.8. Clearly, the density function takes different shapes for different values of θ and is hence significant where one observes inflated values at the end points of the unit interval.

Further, the rth moment about origin of the ULINF distribution can be written as

$$E(y^r) = \alpha p + (1-\alpha)\mu_r, \quad r = 1, 2, 3, \ldots, \tag{2.4}$$

where μ_r is the rth moment about origin of the unit-Lindley distribution. Note that E(y^r) is the weighted average of the rth moment of a Bernoulli distribution with parameter p and the corresponding r^{th} moment of the $UL(\theta)$ distribution with weights α and $(1 - \alpha)$ respectively.

Clearly, using (2.4) the mean and variance of Y become:

$$E(y) = \alpha p + (1 - \alpha)\mu_1 = \alpha p + \frac{1 - \alpha}{1 + \theta}, \text{ and}$$

$$V(y) = \alpha p + (1 - \alpha)\mu_2 - \left[\alpha p + (1 - \alpha)\mu_1\right]^2$$

$$= \alpha p + \frac{(1 - \alpha)}{1 + \theta}(\theta^2 e^\theta E_1(\theta) - \theta + 1) - \left[\alpha p + \frac{1 - \alpha}{1 + \theta}\right]^2$$

$$= \alpha p(1 - \alpha p) + \frac{(1 - \alpha)}{1 + \theta}\left(\theta^2 e^\theta E_1(\theta) - \theta + 1 - \frac{1 - \alpha}{1 + \theta} - 2\alpha p\right),$$

where $E_1(z) = \int_1^\infty z^{-a} e^{-xz} dz$ is the exponential integral function. See, Abramowitz and Stegun (1974) for more details.

Remark 1. As one may have wondered, similar models can be designed if the inflation is only one-sided i.e. only at 0 or 1. Analogous mixture density functions as seen in (2.3) can be written down to accommodate such variations. Appropriate moment derivations and maximum likelihood estimation can also be handled. The preprint by Ferreira et al. (2020) will hopefully be in this direction.

3 Estimation

In this section we present the maximum likelihood estimators of the unknown parameters for an inflated unit-Lindley distribution.

3.1 The likelihood function

Consider a random sample $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$ from the density function given in (2.3) with parameter space Λ, where Λ is a 3 length vector given by $(\alpha, p, \theta)'$. The corresponding likelihood function is thus given by

$$L(\Lambda|\boldsymbol{y}) = \prod_{i=1}^n f_{\text{ulinf}}(y_i|\alpha, p, \theta) \tag{3.1}$$
$$= L_1(\alpha|\boldsymbol{y}) \times L_2(p|\boldsymbol{y}) \times L_3(\theta|\boldsymbol{y}),$$

where

$$L_1(\alpha|\boldsymbol{y}) = \prod_{i=1}^n \alpha^{\mathbb{1}_{\{0,1\}}(y_i)}(1 - \alpha)^{1 - \mathbb{1}_{\{0,1\}}(y_i)} \tag{3.2}$$
$$= \alpha^{\sum_{i=1}^n \mathbb{1}_{\{0,1\}}(y_i)}(1 - \alpha)^{n - \sum_{i=1}^n \mathbb{1}_{\{0,1\}}(y_i)},$$

$$L_2(p|\boldsymbol{y}) = \prod_{i=1}^n \left[p^{y_i}(1 - p)^{1 - y_i}\right]^{\mathbb{1}_{\{0,1\}}(y_i)} \tag{3.3}$$
$$= p^{\sum_{i=1}^n y_i \mathbb{1}_{\{0,1\}}(y_i)}(1 - p)^{\sum_{i=1}^n (1 - y_i)\mathbb{1}_{\{0,1\}}y_i}$$
$$= p^{\sum_{i=1}^n \mathbb{1}_{\{1\}}(y_i)}(1 - p)^{\left[\sum_{i=1}^n \mathbb{1}_{\{0,1\}}(y_i) - \sum_{i=1}^n \mathbb{1}_{\{1\}}(y_i)\right]},$$

$$L_3(\theta|\boldsymbol{y}) = \prod_{i=1}^{n} f(y_i|\theta) \quad for \ \ y_i \in (0,1)$$

$$= \prod_{i=1}^{n} \frac{\theta^2}{1+\theta}(1-y_i)^{-3} exp\left(-\frac{\theta y_i}{1-y_i}\right). \tag{3.4}$$

From equation (3.1), (3.2), (3.3),and (3.4) we can write the log-likelihood function for $\Lambda = (\alpha, p, \theta)'$ as

$$\ell(\Lambda|\boldsymbol{y}) = \ell_1(\alpha|\boldsymbol{y}) + \ell_2(p|\boldsymbol{y}) + \ell_3(\theta|\boldsymbol{y}),$$

where

$$\ell_1(\alpha|\boldsymbol{y}) = \ln(\alpha) \sum_{i=1}^{n} \mathbb{1}_{\{0,1\}}(y_i) + \ln(1-\alpha)\left[n - \sum_{i=1}^{n} \mathbb{1}_{\{0,1\}}(y_i)\right], \tag{3.5}$$

$$\ell_2(p|\boldsymbol{y}) = \ln(p) \sum_{i=1}^{n} \mathbb{1}_{\{1\}}(y_i) + \ln(1-p)\left[\sum_{i=1}^{n} \mathbb{1}_{\{0,1\}}(y_i) - \sum_{i=1}^{n} \mathbb{1}_{\{1\}}(y_i)\right], \tag{3.6}$$

$$\ell_3(\theta|\boldsymbol{y}) = 2n \ln \theta - n \ln(1+\theta) - \theta t(\boldsymbol{y}) - 3 \sum_{i=1}^{n} \ln(1-y_i), \tag{3.7}$$

where $t(\boldsymbol{y}) = \sum_{i=1}^{n} \frac{y_i}{1-y_i}$ and $\mathbb{1}_S(\boldsymbol{y})$ is an indicator variable that takes a value 1 when y $\in S$ and 0 when y $\notin S$.

3.2   Maximum likelihood estimators

In order to find out the maximum likelihood estimators, one has to differentiate (3.5), (3.6), (3.7) with respect to the unknown parameters, and equate them to zero. Following are the first derivatives:

$$\frac{\partial \ell_1(\alpha|y)}{\partial \alpha} = \frac{1}{\alpha} \sum_{i=1}^{n} \mathbb{1}_{\{0,1\}}(y_i) - \frac{1}{1-\alpha}\left[n - \sum_{i=1}^{n} \mathbb{1}_{\{0,1\}}(y_i)\right],$$

$$\frac{\partial \ell_2(p|y)}{\partial p} = \frac{1}{p} \sum_{i=1}^{n} \mathbb{1}_{\{1\}}(y_i) - \frac{1}{1-p}\left[\sum_{i=1}^{n} \mathbb{1}_{\{0,1\}}(y_i) - \sum_{i=1}^{n} \mathbb{1}_{\{1\}}(y_i)\right],$$

$$\frac{\partial \ell_3(\theta|y)}{\partial \theta} = \frac{2n}{\theta} - \frac{n}{1+\theta} - t(\boldsymbol{y}). \tag{3.8}$$

One can easily see that the maximum likelihood estimators of α and p are hence, $\hat{\alpha} = T_1/n$ and $\hat{p} = T_2/T_1$ (0/0 being regarded as 0) where $T_1 = \sum_{i=1}^{n} \mathbb{1}_{\{0,1\}}(y_i)$ and $T_2 = \sum_{i=1}^{n} \mathbb{1}_{\{1\}}(y_i)$. Here, $\hat{\alpha}$ is the probability of zeros and ones in the observed sample and $\hat{p}$ is the probability of zeros among the observations that equal zero or one. Since $\hat{\alpha}$ is an unbiased estimator of α and is a function of the complete sufficient statistic, it is also the UMVUE of α, (Corollary 1.6.16 and Theorem 1.6.22 Lehmann and Casella, 1998) and its variance can be written as $\text{Var}(\hat{\alpha}) = \alpha(1-\alpha)/n$. On the other hand, the maximum likelihood estimator for θ can be obtained by equating (3.8) to zero and is given by

$$\hat{\theta} = \frac{1}{2t(\boldsymbol{y})}\left[n - t(\boldsymbol{y}) + \sqrt{t(\boldsymbol{y})^2 + 6nt(\boldsymbol{y}) + n^2}\right].$$

The Fisher information matrix for the parameters of the inflated unit-Lindley distribution can be written as

$$K(\Lambda) = \begin{pmatrix} k_{\alpha\alpha} & 0 & 0 \\ 0 & k_{pp} & 0 \\ 0 & 0 & k_{\theta\theta} \end{pmatrix}, \tag{3.9}$$

where

$$k_{\alpha\alpha} = \frac{n}{\alpha(1-\alpha)}, \qquad k_{pp} = \frac{n\alpha}{p(1-p)}, \qquad k_{\theta\theta} = \frac{2n}{\theta^2} - \frac{n}{(1+\theta)^2}.$$

Here, α, p, and θ are orthogonal parameters and hence the respective components of the score vector are uncorrelated.

Proposition 3.1. *The inflated unit-Lindley distribution belongs to the exponential family of distributions having full rank. Hence, for large samples, it follows that*

$$\sqrt{n}(\hat{\Lambda} - \Lambda) \xrightarrow{d} N_3(0, K(\Lambda)^{-1}),$$

where $\Lambda = (\alpha, p, \theta)'$ and $K(\Lambda)$ is a 3 × 3 fisher information matrix which is as given in (3.9).

Now, let m(Λ) be a continuous and differentiable function of Λ. Then, the asymptotic distribution of m($\hat{\Lambda}$), which is the ML estimator of m(Λ), can be obtained using the delta method (Sec. 1.9 Lehmann and Casella, 1998). We hence get,

$$\sqrt{n}(m(\hat{\Lambda}) - m(\Lambda)) \xrightarrow{d} N_3(0, b(\Lambda)),$$

where $b(\Lambda) = \dot{m}(\Lambda)^T K(\Lambda)^{-1} \dot{m}(\Lambda)$ with $\dot{m}(\Lambda) = \partial m(\Lambda)/\partial\Lambda$.

Following this, one can construct approximate confidence intervals based on the asymptotic normality $\hat{\Lambda}$. Using the limiting distribution, we can construct the confidence intervals for α, p, and θ. Hence, (1-α^*) × 100% asymptotic confidence intervals for α, p, and θ are given by $\hat{\alpha} \pm z_{(1-\alpha^*/2)}$ se($\hat{\alpha}$), $\hat{p} \pm z_{(1-\alpha^*/2)}$ se($\hat{p}$), and $\hat{\theta} \pm z_{(1-\alpha^*/2)}$ se($\hat{\theta}$) respectively, where α^* is the required significance level.

4 Analysis from Simulations

In this section, we carry out extensive simulations to gauge the behaviour of the parameters underlying our proposed inflated unit-Lindley distribution. We focus on point estimation, and Tables 1 and 2 contain the relative biases and mean squared errors of the maximum likelihood estimators of θ and p, and also of the mean ($\hat{\mu}_y$) and variances ($\hat{\sigma}_y^2$) of the observed sample data. We fix the values of $\alpha = 0.25, \theta = 1.5$ and $p = 0.4$ in Table 1, whereas $\alpha = 0.5, \theta = 1.5$ and $p = 0.4$ in Table 2. We report results for varying sample sizes from small (50) to large (1000) and run 10,000 simulations. All the analysis was carried out using R (R Core Team, 2020) and used the LindleyR package developed by Mazucheli et al. (2019a). The R Code for the simulation study can be found in Web-Appendix A.

Table 1: Simulation results for ULINF(α, p, θ) distribution with $\alpha = 0.25$, $\theta = 1.5$, p = 0.4, E(Y) = $\mu_Y = 0.4$ and V(Y) = $\sigma_Y = 0.0925$

Estimator	n	Bias	MSE	Estimator	n	Bias	MSE
$\hat{p}$	50	-0.0018	0.0202	$\hat{\mu}_Y$	50	-0.0010	0.0202
	100.	0.0023	0.0095		100	0.0000	0.0095
	200	-0.0006	0.0047		200	-0.0006	0.0047
	500	-0.0001	0.0019		500	0.0000	0.0019
	1000	0.0004	0.0010		1000	0.0001	0.0010
$\hat{\theta}$	50	0.0205	0.0413	$\hat{\sigma}_Y$	50	-0.0033	0.1136
	100	0.0120	0.0195		100	-0.0005	0.1054
	200	0.0076	0.0095		200	-0.0004	0.0989
	500	0.0008	0.0036		500	-0.0001	0.0963
	1000	0.0009	0.0018		1000	0.0000	0.0958

In both the tables, the mean squared error is computed as the average squared difference between the estimated value and the true value of the parameter and one can clearly see that the MSE decreases as the sample size increases. We also observed that the estimates are converging at a slow rate as we increase the sample size from 50 to 1000. The convergence for the estimate of p approaches the true value from below whereas for θ, it approaches to the true value from above. For example, for $n = 50$, the estimates of p and θ are 0.3968 and 1.5292, respectively whereas for $n = 1000$, the estimates are 0.3996 and 1.5010 respectively.

5   Analysis from Real Data

In this section, we present real life applications of our proposed distribution to illustrate its potentiality under different types of data sets. We compare our proposed model with the inflated beta and the inflated Kumaraswamy distributions and use AIC and BIC as model selection criteria. Computations for fitting the inflated beta and the inflated Kumaraswamy models were done by using the `Rfast` and `Rfast2` packages, proposed by Papadakis et al. (2019) and Papadakis et al. (2020) respectively, which can be found in the CRAN packages repository.

5.1   Data set I

The first data set contains the percent of newborn elephants who had their heads up at halfway during pregnancy, among different herd sizes. The data is taken from `https://stackoverflow.com/questions/55984490/arcsine-transformation-of-percentage-data`.

Table 2: Simulation results for ULINF(α, p, θ) distribution with $\alpha = 0.5$, $\theta = 1.5$, p = 0.4, E(Y) = $\mu_Y = 0.4$ and V(Y) = $\sigma_Y = 0.1417$

Estimator	n	Bias	MSE	Estimator	n	Bias	MSE
$\hat{p}$	50	-0.0027	0.0098	$\hat{\mu}_Y$	50	-0.0029	0.0098
	100	-0.0003	0.0048		100	-0.0005	0.0048
	200	0.0011	0.0025		200	0.0004	0.0025
	500	0.0002	0.0009		500	-0.0000	0.0010
	1000	0.0000	0.0004		1000	0.0000	0.0005
$\hat{\theta}$	50	0.0423	0.0636	$\hat{\sigma}_Y$	50	-0.0027	0.0751
	100	0.0155	0.0288		100	-0.0012	0.0714
	200	0.0072	0.0139		200	-0.0005	0.0698
	500	0.0041	0.0057		500	-0.0002	0.0678
	1000	0.0009	0.0028		1000	-0.0001	0.0672

The data set contains 27 observations with 2 zeros and 6 ones, and is as follows:

0.0000, 1.0000, 0.8000, 0.2500, 0.5714, 1.0000, 0.0000, 0.2500, 0.5000, 1.0000, 1.0000, 0.7000, 1.0000, 0.1429, 0.2667, 1.0000, 0.5000, 0.4000, 0.6765, 0.4359, 0.0541, 0.4490, 0.4150, 0.6923, 0.1429, 0.0707, 0.0605.

Table 3: Likelihood-based Statistics

Model	Estimate	LL	AIC	BIC
Inflated unit-Lindley (ULINF)	$\hat{\theta} = 1.4446$	-17.8201	37.6402	38.9361
Inflated beta (BEINF)	$\hat{\mu} = 1.4065$	-17.8949	39.7897	42.3814
	$\hat{\phi} = 2.2685$			
Inflated Kumaraswamy (ZOIK)	$\hat{\alpha} = 1.3514$	-17.8760	39.7519	42.3436
	$\hat{\beta} = 2.3707$			

For this data, we fit the inflated unit-Lindley distribution, ULINF(α, p, θ) and compare it with the inflated beta, BEINF($\alpha, \gamma, \mu, \phi$) and the inflated Kumaraswamy distribution, ZOIK($\lambda, p, \alpha, \beta$). The maximum likelihood estimates of the parameters of the ULINF distribution are $\hat{\alpha} = 0.2963$, $\hat{p} = 0.75$ and $\hat{\theta} = 1.4446$, for the BEINF distribution are $\hat{\alpha} = 0.2963$, $\hat{\gamma} = 0.75$, $\hat{\mu} = 1.4065$, $\hat{\phi} = 2.2685$
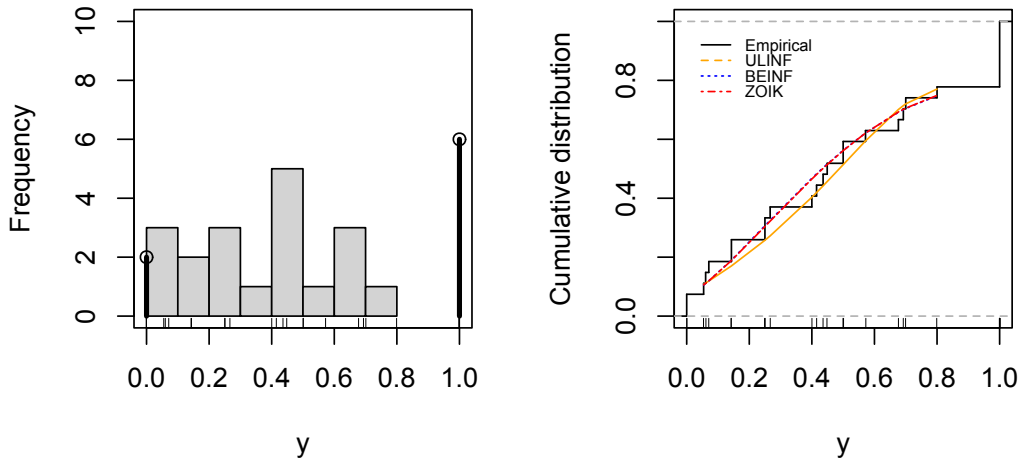
Figure 2: Frequency histogram and estimated cumulative distribution functions

and for the ZOIK distribution are $\hat{\lambda} = 0.2963$, $\hat{p} = 0.75$, $\hat{\alpha} = 1.3514$, $\hat{\beta} = 2.3707$. Table 3 contains these values along with the log-likelihood (LL), AIC and BIC values for the three distributions.

Figure 2 shows the histogram and the empirical distribution of the data along with the estimated cumulative distribution functions for the inflated unit-Lindley, inflated beta and inflated Kumaraswamy. The plot clearly shows that the inflated unit-Lindley is a better fit for the data as compared to the other two distributions.

5.2   Data set II

In spirit of providing a valid illustration having a higher sample size, we generate a pseudo-real data set containing 300 observations with 30 zeros, 50 ones and the remaining 220 observations lying strictly between 0 and 1. Table 4 shows a few summary statistics of the dataset, just for comparison. A suitable R code for generating the hypothetical data can be found in Web-Appendix B.

Table 4: Summary Statistics

Minimum	1st Quantile	Median	Mean	3rd Quantile	Maximum
0.0000	0.1700	0.4708	0.4664	0.6777	1.0000

The maximum likelihood estimates of the parameters of ULINF, BEINF and ZOIK with their corresponding log-likelihood, AIC, and BIC value for the hypothetical data is shown in Table 5. The ML estimates of the Bernoulli parameter p and the mixing parameters α, λ are equal for all the distributions used in the analysis and are $\hat{p} = 0.625$ and $\hat{\alpha} = \hat{\lambda} = 0.2667$ respectively. We observe

that the AIC and BIC values for the inflated unit-Lindley distribuiton are less than the corresponding values for the inflated beta and the inflated Kumaraswamy distributions. This signifies that our proposed model is better than the other two.

Table 5: Likelihood-based Statistics

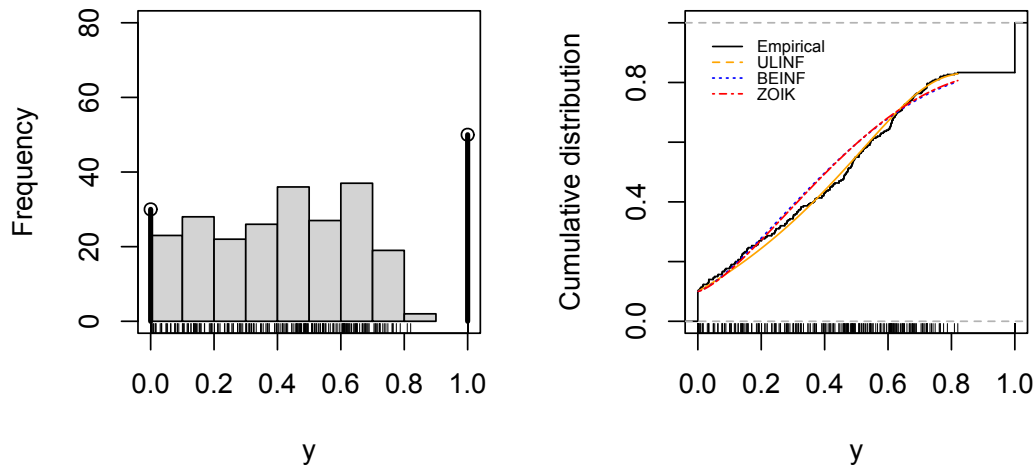| Model	Estimate	LL	AIC	BIC
Inflated unit-Lindley (ULINF)	$\hat{\theta} = 1.4228$	-131.0674	264.1348	265.4306
Inflated beta (BEINF)	$\hat{\mu} = 1.3793$	-148.6740	301.3479	303.9396
	$\hat{\phi} = 2.1292$			
Inflated Kumaraswamy (ZOIK)	$\hat{\alpha} = 1.3751$	-147.4528	298.9057	301.4974
	$\hat{\beta} = 2.3080$			

Figure 3: Frequency histogram and estimated cumulative distribution functions

Figure 3 depicts a histogram of the dataset and the empirical distribution function along with the fitted distributions namely, ULINF, BEINF and ZOIK for the hypothetical data. Visually, it is evident that among others, the inflated unit-Lindley distribution fits the dataset better.

6 Limitations and Conclusions

In this paper, we introduced an inflated version of the unit-Lindley distribution to model datasets which are inflated at both 0 and 1. Appropriate density and distribution functions were derived along with expressions for moments and maximum likelihood estimates for the parameters. Extensive

simulation and real data analyses were carried out to compare our proposed model with some of the existing ones such as the inflated beta and the inflated Kumaraswamy distributions, where it was observed that the inflated unit-Lindley performed better than the other two.

We would like to point out a few limitations in terms of data acquisition for our proposed model. The model is suitable for situations where the parameter θ of the unit-Lindley distribution takes values more than 1.3 approximately, i.e., there are very few data points between the third and fourth quantiles namely, 0.75 and 1. Such right skewed unit-Lindley works well considering any values of parameters α and p. Observing such kind of data in real life may be a challenging task, but in such case inflated unit-Lindley is the appropriate model to fit.

References

Abramowitz, M. and Stegun, I. (1974), *Handbook of Mathematical Functions with Formulas, graphs and Mathematical Tables*, Dover Publications, Incorporated, New York.

Cribari, N. and Santos, J. (2019), "Inflated Kumaraswamy distributions," *Annals of the Brazilian Academy of Sciences*, 91.

Ferreira, A., Mazucheli, J., and Puziol de Oliveira, R. (2020), "The one-inflated unit-Lindley distribution," DOI: 10.13140/RG.2.2.29162.34243.

Ghitany, M., Atieh, B., and Nadarajah, S. (2008), "Lindley distribution and its application," *Mathematics and Computers in Simulation*, 78, 493–506.

Ghitany, M., Mazucheli, J., Menezes, A. F. B., and Alqallaf, F. (2019), "The unit-inverse Gaussian distribution: A new alternative to two-parameter distributions on the unit interval," *Communications in Statistics - Theory and Methods*, 48, 3423–3438.

Grassia, A. (1977), "On a family of distributions with argument between 0 and 1 obtained by transformation of the Gamma distribution and derived compound distributions," *Australian Journal of Statistics*, 19, 108–114.

Kumaraswamy, P. (1976), "Sine power probability density function," *Journal of Hydrology*, 31, 181–184.

Lehmann, E. and Casella, G. (1998), *Theory of point estimation*, Springer, 2nd ed.

Mazucheli, J., Bapat, S., and Menezes, A. (2020a), "A new one-parameter unit-Lindley distribution," *Chilean Journal of Statistics*, 11, 53–67.

Mazucheli, J., Fernandes, L., and Oliveira, R. (2019a), "LindleyR: The Lindley Distribution and its Modifications," *R package version 1.1.0*.

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

Mazucheli, J., Menezes, A. F. B., Fernandes, L. B., de Oliveira, R. P., and Ghitany, M. E. (2020b), "The unit-Weibull distribution as an alternative to the Kumaraswamy distribution for the modeling of quantiles conditional on covariates," *Journal of Applied Statistics*, 47, 954–974.

McLachlan, G. J. and David, P. (2000), *Finite mixture models*, John Wiley and Sons, New York.

McLachlan, G. J. and Kaye, E. B. (1988), *Mixture models. Inference and applications to clustering*, vol. 1, Dekker, New York.

Nadarajah, S., Hassan, B., and Rasool, T. (2011), "A generalized Lindley distribution," *Sankhya B*, 73.

Ospina, R. and Ferrari, S. (2010), "Inflated beta distributions," *Stat Papers*, 51, 111–126.

Papadakis, M., Tsagris, M., Dimitriadis, M., Fafalios, S., Tsamardinos, I., Fasiolo, M., Borboudakis, G., Burkardt, J., Zou, C., Lakiotaki, K., and Chatzipantsiou, C. (2020), *Rfast: A Collection of Efficient and Extremely Fast R Functions*, R package version 2.0.0.

Papadakis, M., Tsagris, M., Fafalios, S., and Dimitriadis, M. (2019), *Rfast2: A Collection of Efficient and Extremely Fast R Functions II*, r package version 0.0.5.

R Core Team (2020), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.

Shanker, R. and Mishra, A. (2013), "A quasi Lindley distribution," *African Journal of Mathematics and Computer Science Research*, 6, 64–71.