MODELING PROPORTION OF SUCCESS IN HIGH SCHOOL LEAVING CERTIFICATE EXAMINATION-A COMPARATIVE STUDY OF INFLATED UNIT LINDLEY AND INFLATED BETA DISTRIBUTION

SUBRATA CHAKRABORTY[1], SAHANA BHATTACHARJEE[2,*]

[1]Department of Statistics, Dibrugarh University, Assam, India

[2]Department of Statistics, Gauhati University, Assam, India

Abstract. In this article, we first introduced the inflated unit Lindley distribution considering zero or/and one inflation scenario and studied its basic distributional and structural properties. Both the proposed distributions are shown to be members of exponential family with full rank. Different parameter estimation methods are discussed along with supporting simulation studies to check their efficacy. Proportion of students passing the high school leaving certificate examination for the schools across the state of Manipur in India for the year 2020 are then modeled using the proposed distributions and compared with the inflated beta distribution to justify its benefits.

1. INTRODUCTION

In the field of applied statistics, one of the most common issues which the researchers have to deal with is data arising in terms of fractions, rates or proportions, i.e., variables which assume values in the range $(0,1)$. However, in many cases, the data arising may contain zeroes and/or

---

ones, i.e., the one observed in the intervals $[0, 1)$, $(0, 1]$ or $[0, 1]$. In such cases, continuous distributions such as Beta distribution, Kumaraswamy distribution [8] or unit-Lindley [10], all of which have support in $(0, 1)$ are not suitable for modeling the data. We need probability models which are able to capture the probability mass at 0, 1 or both. Such distributions are obtained by mixing a continuous distribution having support on $(0, 1)$ with a degenerate distribution whose probability mass is concentrated at either 0 or 1 (for data arising in $[0, 1)$ or $(0, 1]$) and with the Bernoulli distribution which assigns non-negative probability to 0 and 1 (for data arising in $[0, 1]$) [[11] and [12]].

The idea of zero inflated continuous data modelling was first reported in [1] where in the authors dealt with the case of a continuous distribution which has a non-zero probability of assuming a zero value. That is, a non-zero probability mass at zero. Such distributions are mainly termed as zero inflated version of the original distribution and are obtained as a mixture of degenerate probability mass at zero with the underlying distribution. The terminology of zero inflation is common place in the case of count distributions. Alternatively, the word zero Inflated is often replaced by zero Adjusted or zero Spiked, etc. in the literature. Examples of occurrence of zero inflated continuous variables can be seen in many areas of application (for detail see [13], [6], [9], [5] and references there in). Equally important is the situation where the observation from continuous data contains zeros and/or ones. That is, when there is situation of one-inflation or zero-one-inflation. To model such phenomenon, the probability mass at 0, at 1 or both, is included by considering mixture of continuous distribution with degenerate distribution in the case of zero or one inflation and Bernoulli in case of zero and one inflation is used. The first such was attempt seen in inflated beta distribution proposed and studied by [12].

The main objective of this article is to consider the inflated version of the recently introduced unit Lindley distribution and briefly investigate its basic distributional properties. We consider a data in (0,1) with trace of significant zero, one inflation and analyze the data with unit Lindley and its inflated variants to establish the importance of the proposed model.

High school leaving certificate (H.S.L.C.) examination is one of the most important milestones for the individual schools in particular and for the educational scenario of a state in general. The result of this examination declared in terms of the pass percentage of students is an important

indicator, reflecting the state of affair both at the micro as well as in macro level. It is not surprising, therefore, that many of the state's education department has incentive for better show in this examination while poor show often brings penalties for the schools. It's thus relevant to investigate the statistical modelling aspect of the school-wise pass proportion. We considered the data on result of H.S.L.C. examination from the state of Manipur in India for the academic session 2020 in this article. These data sets are available in the public domain ([3], [2] and [4]) and are easily accessible. In these data sets, pass percentages are given for all the schools under the Manipur secondary education board and classified with respect to three types of schools namely the, Government, Government Aided and private schools. They also provide classification of the pass results in three divisions as first division, second division and third division.

In this paper, the zero-or/and -one inflated unit Lindley distribution is proposed. The paper is organized as follows - Section 2 introduces the zero-or-one inflated unit Lindley distribution and some of its distributional properties, estimation of its parameters are discussed. In Section 3, the zero-and-one inflated unit Lindley distribution is presented and some of its properties are discussed. In the next section parameter estimation for both the distributions of preceding section is presented. Section 5 evaluates the performance of the proposed estimators through extensive simulation studies. Section 6 contains empirical applications of the proposed inflated unit Lindley distribution in comparison to the inflated beta distrbution for four data sets. The paper is concluded with some final remarks presented in Section 7.

2. THE ZERO-OR-ONE INFLATED UNIT LINDLEY DISTRIBUTION

The unit Lindley distribution is a one parameter continuous distribution having support on $(0,1)$ which is obtained from the Lindley distribution through a transformation [10]. It has certain advantages over the commonly used beta distribution with two parameters defined in the range $(0,1)$ [7] such as closed form of c.d.f., quantile function and simple expressions for moments. It scores over the competing Kumaraswamy distribution with respect to the fact that there is no closed form of the moments of this distribution. This distribution also enabled the development of a new bounded regression model which is a feasible alternative to the popular Beta regression model. The unit Lindley distribution with parameter θ has the p.d.f.

(1) $$f(x|\theta) = \frac{\theta^2}{1+\theta}(1-x)^{-3}\exp\left(-\frac{\theta x}{1-x}\right), \quad \theta > 0, \ x \in (0,1)$$

The real-life data may include values such as zeroes and/or ones. In such cases, one needs to focus on incorporating a discrete component into the continuous data generating process so that the values zeroes and/or ones are observed with a positive probability. We, thus, consider a mixture of two distributions: the continuous unit Lindley distribution on $(0,1)$ and a degenerate distribution having the entire probability mass concentrated at the known point c, where $c = 0$ or $c = 1$. We refer to it as the data being inflated (having higher probability of occurrence) at one/both endpoints of the standard unit interval. The c.d.f. of the mixture distribution, known as the Inflated unit Lindley distribution is given by

(2) $$IuL_c(y;\alpha,\theta) = \alpha\mathbb{I}_{[c,1]}(y) + (1-\alpha)F(y;\theta)$$

where $\mathbb{I}_A$ is the indicator function which takes the value 0 if $y \in A$ and 1 if $y \notin A$, $0 < \alpha < 1$ is the mixture parameter and $F(;\theta)$ is the c.d.f. of the unit Lindley distribution with parameter θ. Here, the r.v. Y follows the unit Lindley distribution with parameter θ with probability $(1-\alpha)$ and the degenerate distribution at c with probability α. The p.d.f. of the Inflated unit Lindley distribution corresponding to the c.d.f in (2) is given by

(3) $$iuL_c(y;\alpha,\theta) = \begin{cases} \alpha, & \text{if } y = c \\ (1-\alpha)f(y;\theta), & \text{if } y \in (0,1) \end{cases}$$

where $f(;\theta)$ is the unit Lindley density in (1) and $\alpha \in (0,1)$ is the probability mass at c, representing the probability of observing 0 (when $c = 0$) or 1 (when $c = 1$).

Definition 2.1 Let Y be a r.v. that follows the inflated unit Lindley distribution in (3).

   (i) If $c = 0$, the distribution in (3) is called the zero-inflated unit Lindley distribution (ULZI) and we write $Y \sim ULZI(\alpha,\theta)$, where $\alpha = P(Y = 0)$.

   (ii) If $c = 1$, the distribution in (3) is called the one-inflated unit Lindley distribution (ULOI) and we write $Y \sim ULOI(\alpha,\theta)$, where $\alpha = P(Y = 1)$.

The rth raw moment of Y is

$$E(Y^r) = \alpha c + (1-\alpha)\mu_r'; \quad r = 1,2,...$$

where $\mu'_r = \dfrac{r}{1+\theta} \displaystyle\int_0^1 \left[\dfrac{x^{r-1}(1-\theta+x)}{(1-x)} \exp\left(-\dfrac{\theta x}{1-x}\right) \right] dx$ is the rth raw moment of the unit Lindley distribution. In particular, the mean and variance of Y are

$$(4) \qquad E(Y) = \alpha c + (1-\alpha)\mu'_1 = \alpha c + \frac{(1-\alpha)}{(1+\theta)}$$

$$(5) \qquad Var(Y) = \alpha c (1-\alpha c) + \frac{1-\alpha}{(1+\theta)} \left[\left\{ \theta^2 e^\theta Ei(1,\theta) - \theta + 1 \right\} - 2\alpha c - \frac{1-\alpha}{(1+\theta)} \right]$$

where $Ei(1,\theta)$ represents the exponential integral function [10].

Figure (1) shows the ULZI and ULOI densities inflated at $c=0$ and $c=1$ for different values of θ and with the value of the mixing parameter α fixed at 0.5. It can be observed that the p.d.f. of both the ULZI and ULZO distribution assume different shapes for different values of θ and c. θ is seen to control the shape of the probability curve, whose skewness increases with an increase in the value of θ. In all the sub plots in figure (1), the vertical bar with a circle above represents $\alpha = 0.5$ ($P(Y=0)$ or $P(Y=1)$). Further, both the ULZI and ULOI distributions have the same functional shape in $(0,1)$, the only difference being in the mass point, which is 0 for the ULZI and 1 for the ULOI distribution.
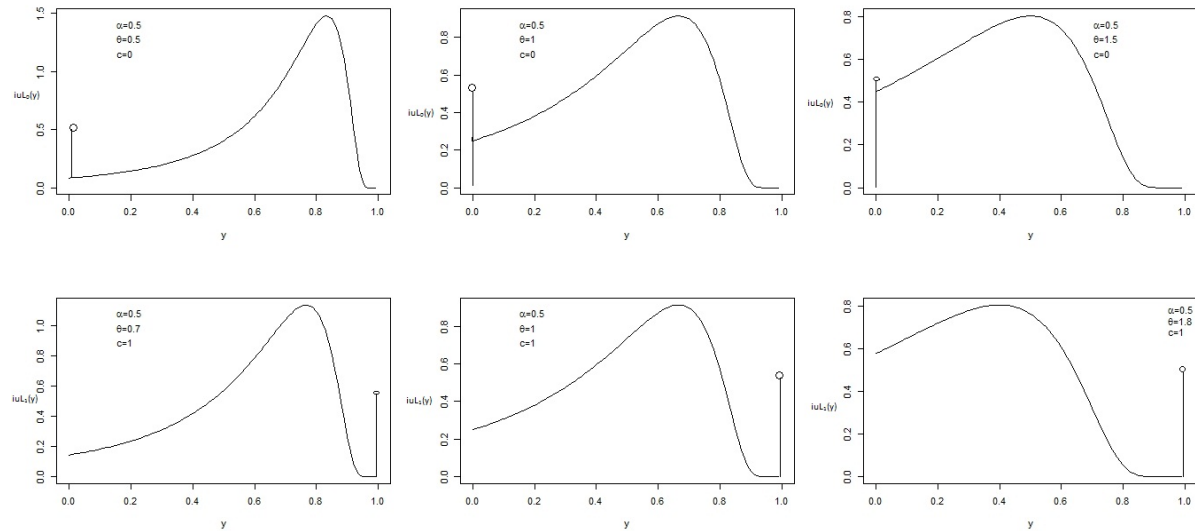

FIGURE 1. ULZI and ULOI densities for different values of θ and c; $\alpha = 0.5$

Result1 The zero- and one-inflated unit Lindley distribution in (3) are member of two-parameter exponential family distributions of full rank.

Proof: Denoting

$$T_1(y) \;=\; I_c(y) = \begin{cases} 1, & \text{if } y = c \\ 0, \text{if } y \in (0,1) \end{cases}$$

$$T_2(y) \;=\; \begin{cases} 1, & \text{if } y = c \\ 0, \text{if } y \in (0,1) \end{cases}$$

$$\eta_1 \;=\; \log \frac{\alpha}{1-\alpha} + \log \frac{\theta^2}{1+\theta}$$

$$\eta_2 \;=\; \theta$$

We can rewrite the p.d.f. of the zero- and one-inflated unit Lindley distribution given in the equation (3) as

$$iuL_c(y; \eta_1 \eta_2) \;=\; (1 - y\mathbb{I}_{[c,1]}(y))^{-3} \exp\left[\{\eta_1 T_1(y) + \eta_2 T_2(y)\} \right.$$
$$\left. - \log\left\{1 + \exp\left(\eta_1 - \log\left(\frac{1+\eta_2}{\eta_2^2}\right)\right)\right\}\right]$$

Now taking

$$T(y) \;=\; (T_1(y), T_2(y)),$$

$$\eta \;=\; (\eta_1, \eta_2),$$

$$B^*(\eta) \;=\; \log\{1 + \exp(\eta_1 - B(\eta_2))\},$$

$$B(\eta_2) \;=\; \log\frac{1+\eta_2}{\eta_2^2} \text{ and}$$

$$h(y) \;=\; (1 - y\mathbb{I}_{[c,1]}(y))^{-3}$$

the p.d.f. can finally be expressed as

$$\exp\left[\eta' T(y) - B^*(\eta)\right] h(y).$$

Note that the functions $B^*(\eta)$ is a real valued function of $\eta_1, \eta_2, h(y)$ is a positive real valued function. The transformation from (α, θ) to (η_1, η_2) is obviously one-one from $(0,1) \times R^+$ to $R \times R^+$. Hence the p.d.f. in (3) belongs to a two parameter exponential family distribution of full rank.

3. THE ZERO-AND-ONE-INFLATED UNIT LINDLEY DISTRIBUTION

The zero-or-one inflated unit Lindley distribution introduced in the previous section is suitable for modeling data which has data inflation on either of the two end points of the standard unit interval $(0,1)$, but not on both the end points. To model data arising in the interval $[0,1]$, we need a mixture of the unit Lindley distribution and the Bernoulli distribution, which assigns non-negative probability to the points 0 and 1. The c.d.f. of the mixture distribution, known as the zero-and-one inflated unit Lindley distribution (ULZOI) is given by

$$(6) \qquad ULZOI(y; \alpha, p, \theta) = \alpha Ber(y; p) + (1 - \alpha) F(y; \theta)$$

where $y \in [0,1]$, $Ber(; p)$ represents the c.d.f. of a Bernoulli r.v. and $F(; \theta)$ is the c.d.f. of the unit Lindley distribution with parameter θ. Further, α is the mixing parameter which lies between 0 and 1. We say that a r.v. Y assuming values in $[0,1]$ has a ULZOI distribution with parameters α, p and θ if its density function with respect to the c.d.f. in (6) is given by

$$(7) \qquad ulzoi(y; \alpha, p, \theta) = \begin{cases} \alpha p, & \text{if } y = 1 \\ \alpha(1-p), & \text{if } y = 0 \\ (1-\alpha) f(y; \theta), & \text{if } y \in (0,1) \end{cases}$$

where $0 < \alpha, p < 1,$ $, \theta > 0$ and $f(y; \theta)$ is the p.d.f. of the unit Lindley distribution in (1). We write $Y \sim ULZOI(\alpha, p, \theta)$ and in that case, $\alpha p = P(Y = 1)$ and $\alpha(1 - p) = P(Y = 0)$. Also, the rth raw moment of Y is

$$E(Y^r) = \alpha p + (1 - \alpha) \mu_r'; r = 1, 2, \ldots$$

Consequently, the mean and variance of Y are

$$(8) \qquad E(Y) = \alpha p + (1 - \alpha) \mu_1' = \alpha p + \frac{(1 - \alpha)}{(1 + \theta)}$$

$$(9) \quad Var(Y) = \alpha p (1 - \alpha p) + \frac{1-\alpha}{(1+\theta)} \left[\left\{ \theta^2 e^\theta Ei(1,\theta) - \theta + 1 \right\} - 2\alpha p - \frac{1-\alpha}{(1+\theta)} \right]$$

where $Ei(1,\theta)$ represents the exponential integral function. Figure (2) presents the ULZOI density for different values of θ for $\alpha = 0.7$ and $p = 0.5$. It is evident from this plot that as the value of θ increases, the skewness also increases and also, for higher values of θ, the probability curve of ULZOI distribution approaches the reverse sigmoid curve.
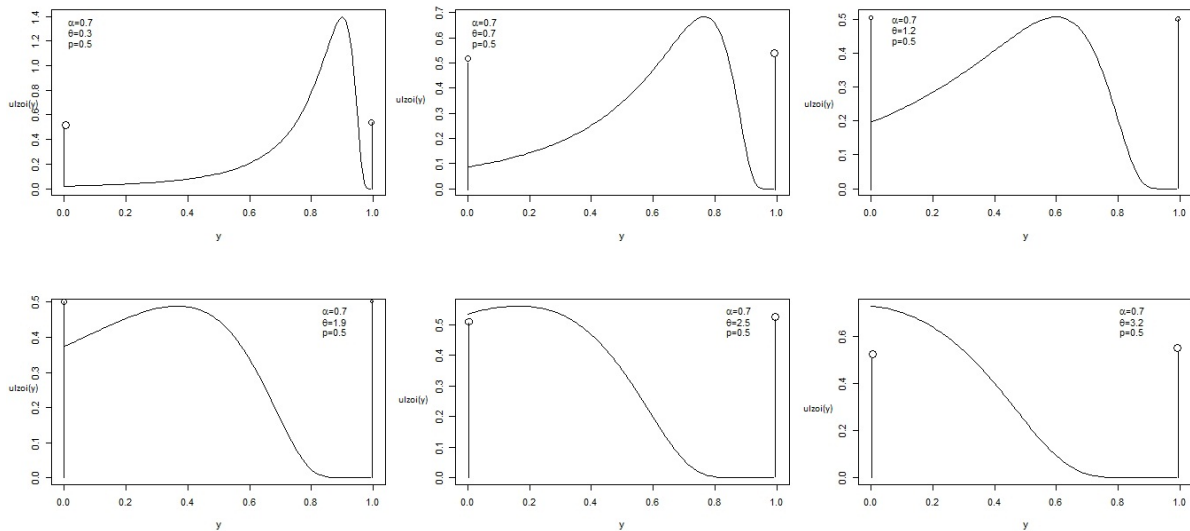
FIGURE 2. ULZOI density for different values of θ; $\alpha = 0.7$ and $p = 0.5$

Result2 The zero-and-one-inflated unit Lindley distribution in (3) is a three parameter exponential family distribution of full rank.

Proof: Denoting

$$T_1(y) = I_{\{0,1\}}(y) = \begin{cases} 1, & \text{if } y = 0 \text{ or } 1 \\ 0, & \text{if } y \in (0,1) \end{cases}$$

$$T_2(y) = y I_{\{0,1\}}(y) = \begin{cases} 1, & \text{if } y = 1 \\ 0, & \text{if } y = 0 \end{cases}$$

$$T_3(y) = \frac{y}{1-y}$$

$$\eta_1 \;=\; \log\frac{\alpha}{1-\alpha}+\log\frac{1+\theta}{\theta^2}-\log\frac{1}{1-p}$$

$$\eta_2 \;=\; \log\frac{1}{1-p}$$

$$\eta_3 \;=\; \theta$$

We can rewrite the p.d.f. of the zero- and one-inflated unit Lindley distribution given in the equation (7) as

$$iuL_c\,(y;\eta_1\eta_2,\eta_3) \;=\; (1-y\mathbb{I}_{[c,1]}\,(y))^{-3}\exp\left[\{\eta_1 T_1(y)+\eta_2 T_2(y)+\eta_3 T_3(y)\}\right.$$
$$\left.-\log\left\{1+\exp\left(\eta_1+\log\left(1+e_2^\eta\right)-\log\left(\frac{1+\eta_3}{\eta_3^2}\right)\right)\right\}\right]$$

$$T(y) \;=\; (T_1(y),T_2(y),T_3(y)),$$

$$\eta \;=\; (\eta_1,\eta_2,\eta_3),$$

$$B^*(\eta) \;=\; \log\{1+\exp(\eta_1+M(\eta_2)-N(\eta_3))\},$$

$$M(\eta_2) \;=\; \log(1+e^{\eta_2})$$

$$N(\eta_3) \;=\; \log\frac{1+\eta_3}{\eta_3^2}\ \text{and}$$

$$h(y) \;=\; (1-y\mathbb{I}_{[0,1]}\,(y))^{-3}$$

the p.d.f. can finally be expressed as

$$\exp\left[\eta'T(y)-B^*(\eta)\right]h(y).$$

Note that the functions $B^*(\eta)$ is a real valued function of (η_1,η_2,η_3),$h(y)$ is a positive real valued function. The transformation from (α,p,θ) to (η_1,η_2,η_3) is obviously one-one from $(0,1)\times(0,1)\times R^+$ to $R\times R^+\times R^+$. Also neither the Ts nor the ηs are linearly related. Hence the p.d.f. in (7) belongs to a three parameter exponential family distribution of full rank.

4. ESTIMATION AND RELATED

In this section, the maximum likelihood estimation (MLE) of the parameters and construction of the Fisher Information Matrix is considered.

4.1. MLE: Zero or One inflated ULD. The likelihood function for $v = (\alpha, \theta)$ based on the random sample $y = (y_1, y_2, \ldots, y_n)'$ from iuL_c is given by

$$
\begin{aligned}
L(v, y) &= \prod_{i=1}^{n} iuL_c (y_i; \alpha, \theta) \\
&= \prod_{i=1}^{n} \left[\left\{ \alpha^{\mathbb{I}_{[c]}(y_i)} (1-\alpha)^{1-\mathbb{I}_{[c]}(y_i)} \right\} \left\{ f(y, \theta)^{1-\mathbb{I}_{[c]}(y_i)} \right\} \right] \\
&= L_1(\alpha, y) \times L_2(\theta, y)
\end{aligned}
$$

It can be seen that the likelihood function is factorized into two terms L_1 and L_2 where L_1 depends on α and L_2 only on θ. Now,

$$
\begin{aligned}
L_1(\alpha, y) &= \prod_{i=1}^{n} \alpha^{\mathbb{I}_{[c]}(y_i)} (1-\alpha)^{1-\mathbb{I}_{[c]}(y_i)} \\
&= \alpha^{\sum_{i=1}^{n} \mathbb{I}_{[c]}(y_i)} (1-\alpha)^{n-\sum_{i=1}^{n} \mathbb{I}_{[c]}(y_i)}
\end{aligned}
$$

and

$$
L_2(\theta, y) = \prod_{\substack{i=1 \\ y_i \in (0,1)}}^{n} f(y_i; \theta)^{1-\mathbb{I}_{[c]}(y_i)}
$$

The log-likelihood function of the Zero-or-One Inflated Unit Lindley distribution is given by $l(v) = \log L(v, y) = l_1(\alpha; y) + l_2(\theta; y)$ where

$$
\begin{aligned}
l_1(\alpha; y) &= \log L_1(\alpha, y) \\
&= \log \alpha \sum_{i=1}^{n} \mathbb{I}_{[c]}(y_i) + \log(1-\alpha) \left\{ n - \sum_{i=1}^{n} \mathbb{I}_{[c]}(y_i) \right\}
\end{aligned}
$$

and

$$
\begin{aligned}
l_2(\theta; y) &= \log L_2(\theta, y) \\
&= \left\{ n - \sum_{i=1}^{n} \mathbb{I}_{[c]}(y_i) \right\} \log \left(\frac{\theta^2}{1+\theta} \right) - 3 \sum_{\substack{i=1 \\ y_i \in (0,1)}}^{n} \log(1-y_i) - \theta \sum_{\substack{i=1 \\ y_i \in (0,1)}}^{n} \frac{y_i}{1-y_i}
\end{aligned}
$$

The score function is then obtained by differentiating the log-likelihood function and is denoted by $U(v) = [U_\alpha(\alpha), U_\theta(\theta)]$, where

$$
\begin{aligned}
U_\alpha(\alpha) &= \frac{\partial l_1(\alpha; y)}{\partial \alpha} \\
&= \frac{1}{\alpha} \sum_{i=1}^{n} \mathbb{I}_{[c]}(y_i) - \frac{1}{1-\alpha} \left\{ n - \sum_{i=1}^{n} \mathbb{I}_{[c]}(y_i) \right\} \\
U_\theta(\theta) &= \frac{\partial l_2(\theta; y)}{\partial \theta} \\
&= \frac{2+\theta}{\theta(1+\theta)} \left\{ n - \sum_{i=1}^{n} \mathbb{I}_{[c]}(y_i) \right\} - \sum_{\substack{i=1 \\ y_i \in (0,1)}}^{n} \frac{y_i}{1-y_i}
\end{aligned}
$$

The maximum likelihood estimator of α is $\hat{\alpha} = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}_{[c]}(y_i)$, i.e., the proportion of n values that are equal to c and the maximum likelihood estimator of θ is

$$
\hat{\theta} = \frac{1}{2 \sum_{\substack{i=1 \\ y_i \in (0,1)}}^{n} \frac{y_i}{1-y_i}} \left(n - \sum_{\substack{i=1 \\ y_i \in (0,1)}}^{n} \mathbb{I}_{[c]}(y_i) - \sum_{\substack{i=1 \\ y_i \in (0,1)}}^{n} \frac{y_i}{1-y_i} + \right.
$$

$$
\left. \left[\left(\sum_{\substack{i=1 \\ y_i \in (0,1)}}^{n} \frac{y_i}{1-y_i} \right)^2 + \left\{ n - \sum_{i=1}^{n} \mathbb{I}_{[c]}(y_i) \right\}^2 - 6 \sum_{\substack{i=1 \\ y_i \in (0,1)}}^{n} \frac{y_i}{1-y_i} \left(n - \sum_{i=1}^{n} \mathbb{I}_{[c]}(y_i) \right) \right]^{\frac{1}{2}} \right)
$$

The Fisher Information matrix for the Zero or One inflated Unit Lindley law is

$$
K(v) = \begin{bmatrix} k_{\alpha\alpha} & k_{\alpha\theta} \\ k_{\theta\alpha} & k_{\theta\theta} \end{bmatrix}
$$

where

$$
\begin{aligned}
k_{\alpha\alpha} &= -E\left[\frac{\partial U_\alpha(\alpha)}{\partial \alpha}\right] \\
&= \frac{n}{\alpha(1-\alpha)}
\end{aligned}
$$

$$k_{\alpha\theta} = -E\left[\frac{\partial U_\theta(\theta)}{\partial\alpha}\right]$$

$$= 0$$

$$k_{\theta\alpha} = -E\left[\frac{\partial U_\alpha(\alpha)}{\partial\theta}\right]$$

$$= 0$$

$$k_{\theta\theta} = -E\left[\frac{\partial U_\theta(\theta)}{\partial\theta}\right]$$

$$= \frac{n(1-\alpha)\left(\theta^2+4\theta+2\right)}{\theta^2(1+\theta)^2}$$

Suppose $\hat{v} = (\hat{\alpha},\hat{\theta})$ denote the m.l.e. of $v = (\alpha,\theta)$. In large samples, $\hat{v}$ is asymptotically normally distributed, i.e., $\hat{v} \xrightarrow{D} N_2\left(v,k(v)^{-1}\right)$ where $k(v)$ is the Fisher Information Matrix. Using this result, approximate confidence intervals for the parameters α and θ can be constructed. Let $\delta \in (0,0.5)$. Then $(1-\delta) \times 100\%$ asymptotic confidence intervals for α and θ are given respectively by

$$\hat{\alpha} \pm Z_{1-\frac{\alpha}{2}} SE(\hat{\alpha}) \text{ and } \hat{\theta} \pm Z_{1-\frac{\alpha}{2}} SE(\hat{\theta})$$

where $SE(.)$ denotes standard error and $Z_{1-\frac{\alpha}{2}}$ is the standard normal quantile.

4.2. MLE: Zero and One inflated ULD.
The p.d.f. of the Zero and One inflated unit Lindley distribution can be re-written as

$$ulzoi(y;\alpha,p,\theta) = \left[\alpha p^y(1-p)^{1-y}\right]^{\mathbb{I}_{\{0,1\}}(y)}\left[(1-\alpha)f(y;\theta)\right]^{1-\mathbb{I}_{\{0,1\}}(y)}$$

$$= \left[\alpha^{\mathbb{I}_{\{0,1\}}(y)}(1-\alpha)^{1-\mathbb{I}_{\{0,1\}}(y)}\right]\left[p^y(1-p)^{1-y}\right]^{\mathbb{I}_{\{0,1\}}(y)}$$

$$\left[f(y;\theta)\right]^{1-\mathbb{I}_{\{0,1\}}(y)}$$

where $\mathbb{I}_{\{0,1\}}(y)$ is an indicator function such that

$$\mathbb{I}_{\{0,1\}}(y) = \begin{cases} 1, \text{if } y \in \{0,1\} \\ 0, \text{if } y \notin \{0,1\} \end{cases}$$

Here, it can observed that the first term depends only on α, the second term depends only on p and the third term depends only on θ. The likelihood function for $v = (\alpha,p,\theta)$ on the random

sample $(y_1, y_2, \ldots, y_n)$ from the $ulzoi\,(y; \alpha, p, \theta)$ distribution is given by

$$
\begin{aligned}
L(v, y) &= \prod_{i=1}^{n} ulzoi\,(y_i; \alpha, p, \theta) \\
&= L_1(\alpha; y) \times L_2(p; y) \times L_3(\theta; y)
\end{aligned}
$$

where

$$
\begin{aligned}
L_1(\alpha; y) &= \prod_{i=1}^{n} \left[\alpha^{\mathbb{I}_{\{0,1\}}(y_i)} (1-\alpha)^{1-\mathbb{I}_{\{0,1\}}(y_i)} \right] \\
&= \alpha^{\sum_{i=1}^{n} \mathbb{I}_{\{0,1\}}(y_i)} (1-\alpha)^{n-\mathbb{I}_{\{0,1\}}(y_i)} \\
L_2(p; y) &= \prod_{i=1}^{n} \left[p^{y_i} (1-p)^{1-y_i} \right]^{\mathbb{I}_{\{0,1\}}(y_i)} \\
&= p^{\sum_{i=1}^{n} y_i \mathbb{I}_{\{0,1\}}(y_i)} (1-p)^{\sum_{i=1}^{n} (1-y_i)\mathbb{I}_{\{0,1\}}(y_i)} \\
L_3(\theta; y) &= \prod_{\substack{i=1 \\ y_i \in (0,1)}}^{n} f(y_i; \theta) \\
&= \exp\left(-\theta \sum_{i=1}^{n} \frac{y_i}{1-y_i} \right) \prod_{i=1}^{n} \left(\frac{\theta^2}{1+\theta} \right) (1-y_i)^{-3}
\end{aligned}
$$

The corresponding log-likelihood function is $l(v) = \log L(v, y) = l_1(\alpha; y) + l_2(p; y) + l_3(\theta; y)$

where

$$
\begin{aligned}
l_1(\alpha; y) &= \log \alpha \sum_{i=1}^{n} \mathbb{I}_{\{0,1\}}(y_i) + \log(1-\alpha) \left\{ n - \sum_{i=1}^{n} \mathbb{I}_{\{0,1\}}(y_i) \right\} \\
l_2(p; y) &= \log p \sum_{i=1}^{n} y_i \mathbb{I}_{\{0,1\}}(y_i) + \log(1-p) \left[\sum_{i=1}^{n} \mathbb{I}_{\{0,1\}}(y_i) - \sum_{i=1}^{n} y_i \mathbb{I}_{\{0,1\}}(y_i) \right] \\
l_3(\theta; y) &= \sum_{\substack{i=1 \\ y_i \in (0,1)}}^{n} \log\left(\frac{\theta^2}{1+\theta} \right) - 3 \sum_{\substack{i=1 \\ y_i \in (0,1)}}^{n} \log(1-y_i) - \theta \sum_{\substack{i=1 \\ y_i \in (0,1)}}^{n} \frac{y_i}{1-y_i}
\end{aligned}
$$

The score function is denoted by

$$
U(v) = [U_\alpha(\alpha), U_p(p), U_\theta(\theta)]
$$

where

$$U_\alpha(\alpha) = \frac{\partial l_1(\alpha; y)}{\partial \alpha}$$

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

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

$$U_\theta(\theta) = \frac{2+\theta}{\theta(1+\theta)}\left[n - \sum_{i=1}^{n} \mathbb{I}_{\{0,1\}}(y_i)\right] - \sum_{\substack{i=1 \\ y_i \in (0,1)}}^{n} \frac{y_i}{1-y_i}$$

The m.l.e of α is $\hat{\alpha} = \frac{1}{n}\sum_{\substack{i=1 \\ y_i \in (0,1)}}^{n} \mathbb{I}_{\{0,1\}}(y_i)$, which is the proportion of discrete values in the

sample. The m.l.e of p is $\hat{p} = \dfrac{\sum_{i=1}^{n} \mathbb{I}_{\{1\}}(y_i)}{\sum_{i=1}^{n}\mathbb{I}_{\{0,1\}}(y_i)}$. The m.l.e of θ is

$$\hat{\theta} = \frac{1}{2\sum_{\substack{i=1 \\ y_i \in (0,1)}}^{n} \frac{y_i}{1-y_i}}\left(n - \sum_{\substack{i=1 \\ y_i \in (0,1)}}^{n} \mathbb{I}_{[c]}(y_i) - \sum_{\substack{i=1 \\ y_i \in (0,1)}}^{n} \frac{y_i}{1-y_i} + \right.$$

$$\left. \left[\left(\sum_{\substack{i=1 \\ y_i \in (0,1)}}^{n} \frac{y_i}{1-y_i}\right)^2 + \left\{n - \sum_{i=1}^{n}\mathbb{I}_{[c]}(y_i)\right\}^2 - 6\sum_{\substack{i=1 \\ y_i \in (0,1)}}^{n} \frac{y_i}{1-y_i}\left(n - \sum_{i=1}^{n}\mathbb{I}_{[c]}(y_i)\right)\right]^{\frac{1}{2}}\right)$$

The Fisher Information matrix for the Zero and One inflated Unit Lindley law is

$$K(v) = \begin{bmatrix} k_{\alpha\alpha} & k_{\alpha p} & k_{\alpha\theta} \\ k_{p\alpha} & k_{pp} & k_{p\theta} \\ k_{\theta\alpha} & k_{\theta p} & k_{\theta\theta} \end{bmatrix}$$

where

$$k_{\alpha\alpha} = \frac{n}{\alpha(1-\alpha)}$$

$$k_{\alpha p} = k_{\alpha\theta} = k_{p\alpha} = k_{p\theta} = k_{\theta\alpha} = k_{\theta p} = 0$$

$$k_{pp} = \frac{n\alpha}{p(1-p)}$$

$$k_{\theta\theta} = \frac{n(1-\alpha)\left(\theta^2 + 4\theta + 2\right)}{\theta^2(1+\theta)^2}$$

Suppose $\hat{v} = (\hat{\alpha}, \hat{p}, \hat{\theta})$. In large samples, $\hat{v}$ is asymptotically normally distributed, i.e., $\hat{v} \xrightarrow{D} N_3\left(v, k(v)^{-1}\right)$ where $k(v)$ is the Fisher Information Matrix. Let $\delta \in (0, 0.5)$. Then $(1-\delta) \times 100\%$ asymptotic confidence intervals for α, p and θ are given respectively by

$$\hat{\alpha} \pm Z_{1-\frac{\alpha}{2}} SE(\hat{\alpha}), \ \hat{p} \pm Z_{1-\frac{\alpha}{2}} SE(\hat{p}) \text{ and } \hat{\theta} \pm Z_{1-\frac{\alpha}{2}} SE(\hat{\theta}),$$

where the symbols have their usual meaning.

4.3. Conditional Mean Estimation. The conditional mean of $Y \sim ULZI(\alpha, \theta)$ given $y \in (0, 1)$ is obtained as:

$$
\begin{aligned}
E(Y|y \in (0,1)) &= \int_0^1 y.iuL_c(y; \alpha, \theta) dy \\
&= \frac{1}{1+\theta},
\end{aligned}
$$

which does not depend on α. Consequently, the conditional mean estimate (CME) of θ is $\left\{\frac{1}{E(Y|y \in (0,1))} - 1\right\}$. The CME of θ for $Y \sim ULZOI(y; \alpha, p, \theta)$ is also obtained likewise.

5. ASSESSMENT OF ESTIMATORS: SIMULATION STUDY

In this section, a Monte Carlo simulation study is conducted for the purpose of evaluation and comparison of the finite-sample behavior of the maximum likelihood estimates, bias-corrected estimates obtained by using the Cox-Snell Methodology and the conditional mean estimates of the parameters of both the *ULZI* and *ULZOI* distributions.

Samples of size $n = 25, 30, 50, 100, 500$ and 1000 are generated and 1000 Monte Carlo replications is considered. The five sets of parameters of *ULZI* distribution used for this numerical exercise are $\alpha = 0.2, E(Y|y \in (0,1)) = 0.1$; $\alpha = 0.2, E(Y|y \in (0,1)) = 0.4$; $\alpha = 0.2, E(Y|y \in (0,1)) = 0.7$; $\alpha = 0.5, E(Y|y \in (0,1)) = 0.1$ and $\alpha = 0.5, E(Y|y \in (0,1)) = 0.4$, which corresponds to θ values $7.0, 1.0, 0.14, 4.0$ and 0.25 respectively. The four sets of parameters of ULZOI distribution used are $\alpha = 0.3, E(Y|y \in (0,1)) = 0.4, p = 0.3$; $\alpha = 0.3, E(Y|y \in (0,1)) = 0.6, p = 0.5$; $\alpha = 0.5, E(Y|y \in (0,1)) = 0.4, p = 0.3$ and $\alpha = 0.5, E(Y|y \in (0,1)) = 0.6, p = 0.5$, which corresponds θ values $1.26, 0.56, 1.0$ and 0.43 respectively. To simulate n observations from $ULZI(\alpha, \theta)$ distribution, the following algorithm was implemented:

Algorithm to generate from $ULZI(\alpha,\theta)$.

Step 1. n random numbers from $U(0,1)$ were generated, say $U_i; i = 1,2,\ldots n$

Step 2. If $U_i < \alpha$, then we assign $y_i = 0$.

Step 3. If $U_i \geq \alpha$, then we draw a random number from the Lindley(θ) distribution, say x_i and assign $y_i = \dfrac{x_i}{1+x_i}$.

Observations are simulated from the $ULZOI(\alpha,\theta,p)$ distribution using the following algorithm:

Algorithm to generate from $ULZOI(\alpha,\theta,p)$.

Step 1. n random numbers from $U(0,1)$ were generated, say $U_i; i = 1,2,\ldots n$

Step 2. If $U_i \leq \alpha p$, then we assign $y_i = 0$.

If $U_i \leq \alpha$, then we assign $y_i = 1$.

Step 3. Otherwise, we draw a random number from the Lindley(θ) distribution, say x_i and assign $y_i = \dfrac{x_i}{1+x_i}$.

The performance evaluation of the estimates was done based on the estimated bias and Root Mean Square Error (RMSE).Table 1 and Table 2 present the simulation results for the ULZI distribution and ULZOI distribution respectively.

		$\alpha = 0.2, \theta = 7$								
	n	Mean			Bias			RMSE		
		MLE	BCMLE	CME	MLE	BCMLE	CME	MLE	BCMLE	CME
θ	25	7.298	6.988	7.755	0.892	-0.011	0.755	0.667	0.590	0.663
	50	7.165	7.007	7.130	0.165	0.007	0.130	0.110	0.108	0.112
	100	7.093	7.015	7.075	0.093	0.015	0.075	0.057	0.057	0.058
	200	7.045	7.006	7.034	0.045	0.006	0.034	0.028	0.028	0.029
	500	7.029	7.013	7.024	0.028	0.013	0.024	0.011	0.011	0.011
	1000	7.023	7.016	7.018	0.024	0.016	0.018	0.006	0.006	0.006
	n	Mean			Bias			RMSE		
		MLE			MLE			MLE		
α	25	0.198			-0.001			0.012		
	50	0.202			0.003			0.006		
	100	0.201			0.001			0.003		
	200	0.201			0.001			0.001		
	500	0.200			0.000			0.000		
	1000	0.199			0.000			0.000		
		$\alpha = 0.5, \theta = 0.25$								
	n	Mean			Bias			RMSE		
		MLE	BCMLE	CME	MLE	BCMLE	CME	MLE	BCMLE	CME
θ	25	0.262	0.250	0.254	0.012	0.000	0.004	0.009	0.009	0.012
	50	0.256	0.251	0.253	0.006	0.001	0.003	0.004	0.004	0.006
	100	0.252	0.250	0.250	0.002	-0.000	0.000	0.002	0.003	0.003
	200	0.252	0.251	0.251	0.002	0.001	0.001	0.001	0.001	0.001
	500	0.250	0.250	0.250	0.000	0.000	0.000	0.000	0.000	0.000
	1000	0.250	0.250	0.250	0.000	0.000	0.000	0.000	0.000	0.000
	n	Mean			Bias			RMSE		
		MLE			MLE			MLE		
α	25	0.501			0.003			0.039		
	50	0.498			-0.002			0.008		
	100	0.499			-0.001			0.004		
	200	0.499			-0.001			0.002		
	500	0.500			0.000			0.001		
	1000	0.499			-0.000			0.000		

Table 1: Simulation results for ULZI distribution for different set of values of θ and α

	n	Mean			Bias			RMSE		
		$\alpha = 0.3, p = 0.5, \theta = 0.56$
		MLE	BCMLE	CME	MLE	BCMLE	CME	MLE	BCMLE	CME
θ	50	0.565	0.556	0.563	0.005	-0.004	0.003	0.008	0.008	0.010
	100	0.563	0.559	0.562	0.003	-0.001	0.002	0.004	0.004	0.005
	200	0.562	0.559	0.561	0.002	-0.001	0.001	0.002	0.002	0.002
	500	0.562	0.561	0.561	0.002	0.001	0.001	0.001	0.001	0.001

	n	Mean	Bias	RMSE
		MLE	MLE	MLE
α	50	0.296	-0.003	0.007
	100	0.301	0.001	0.004
	200	0.300	-0.000	0.002
	500	0.300	0.000	0.001

	n	Mean	Bias	RMSE
		MLE	MLE	MLE
p	50	0.506	0.006	0.014
	100	0.502	0.002	0.007
	200	0.501	0.001	0.004
	500	0.498	-0.002	0.002

	n	Mean			Bias			RMSE		
		$\alpha = 0.5, p = 0.3, \theta = 1$
		MLE	BCMLE	CME	MLE	BCMLE	CME	MLE	BCMLE	CME
θ	50	1.025	0.999	1.015	0.025	-0.001	0.015	0.017	0.017	0.015
	100	1.009	0.996	1.003	0.009	-0.004	0.003	0.008	0.008	0.010
	200	1.005	0.999	1.003	0.005	-0.001	0.003	0.004	0.004	0.005
	500	1.003	1.001	1.003	0.003	0.001	0.003	0.002	0.002	0.002

	n	Mean	Bias	RMSE
		MLE	MLE	MLE
α	50	0.503	0.003	0.008
	100	0.501	0.001	0.004
	200	0.500	0.000	0.002
	500	0.501	0.001	0.001

	n	Mean	Bias	RMSE
		MLE	MLE	MLE
p	50	0.298	-0.002	0.011
	100	0.299	-0.001	0.005
	200	0.300	-0.000	0.003
	500	0.301	0.002	0.001

Table 2: Simulation results for ULZOI distribution for different set of values of θ, α and p

Table 1 shows that for the ULZI distribution, the bias corrected estimate of both θ achieves substantial bias reduction over the conditional mean estimate and maximum likelihood estimate whereas the RMSE of MLE and BCMLE are smaller than those of CME. Both the bias and RMSE decreases with an increase in n. For moderately large and large sample sizes, the bias of α is seen to be negative and the RMSE of α also decreases with an increase in n.

From Table 2, it is evident that for the ULZOI distribution, for small and moderately large sample sizes, the bias of BCMLE of θ is negative and the RMSE of both MLE and BCMLE coincide and are less than that of the CME. The RMSE of the estimates of θ, α and p decrease with an increase in n. Further, the bias of p is seen to be negative for moderately large and large sample sizes. Table 3 and 4 display the confidence intervals and empirical coverage probabilities at 90% and 95% for the sample sizes $n = 100, 200, 500$ and 1000 for the ULZI and ULZOI distribution respectively.

		$\alpha = 0.2, \theta = 7$							
	%	Confidence interval				Coverage probability			
		n=50	n=100	n=500	n=1000	n=50	n=100	n=500	n=1000
θ	95%	(5.047,8.952)	(5.619,8.381)	(6.383,7.617)	(6.563,7.436)	94.4	94	95.2	94.5
	90%	(5.361,8.638)	(5.841,8.159)	(6.481,7.518)	(6.634,7.366)	90.8	89.6	91.6	88.7
α	95%	(0.089,0.311)	(0.122,0.278)	(0.165,0.235)	(0.175,0.225)	95.1	93.7	95.2	94.4
	90%	(0.107,0.293)	(0.134,0.266)	(0.171,0.229)	(0.179,0.221)	88.5	88.9	91.6	88.9
		$\alpha = 0.5, \theta = 0.25$							
	%	Confidence interval				Coverage probability			
		n=50	n=100	n=500	n=1000	n=50	n=100	n=500	n=1000
θ	95%	(0.207,0.293)	(0.219,0.281)	(0.236,0.264)	(0.240,0.260)	76.9	77.8	76.9	79.6
	90%	(0.214,0.286)	(0.224,0.276)	(0.238,0.262)	(0.242,0.258)	69.7	67.9	69.6	71.4
α	95%	(0.361,0.639)	(0.402,0.598)	(0.456,0.543)	(0.469,0.531)	92.8	95	92.9	94.6
	90%	(0.384,0.616)	(0.417,0.582)	(0.463,0.538)	(0.474,0.526)	86.5	92	88.7	89.8

Table 3: Confidence intervals and coverage probabilities for ULZI distribution for different set of values of θ and α

	%	Confidence interval				Coverage probability			
		$\alpha = 0.3, p = 0.5, \theta = 0.56$							
		n=50	n=100	n=500	n=1000	n=50	n=100	n=500	n=1000
θ	95%	(0.485,0.635)	(0.507,0.613)	(0.562,0.594)	(0.536,0.584)	86.9	88.8	87.9	86.2
	90%	(0.497,0.623)	(0.515,0.605)	(0.532,0.588)	(0.540,0.580)	79.8	80.8	81.7	79.3
α	95%	(0.210,0.390)	(0.236,0.364)	(0.260,0.340)	(0.272,0.328)	93.2	95.1	95.6	94.2
	90%	(0.224,0.375)	(0.247,0.353)	(0.266,0.334)	(0.276,0.324)	89	90	89.5	88
p	95%	(0.321,0.679)	(0.373,0.627)	(0.420,0.580)	(0.443,0.557)	95.1	94.4	94.4	95.1
	90%	(0.344,0.650)	(0.394,0.606)	(0.432,0.567)	(0.453,0.547)	89.9	88.4	88.7	89.7
		$\alpha = 0.5, p = 0.3, \theta = 1$							
	%	Confidence interval				Coverage probability			
		n=50	n=100	n=500	n=1000	n=50	n=100	n=500	n=1000
θ	95%	(0.846,1.154)	(0.891,1.109)	(0.931,1.069)	(0.951,1.049)	85.4	85.4	85.4	83.7
	90%	(0.871,1.129)	(0.909,1.091)	(0.942,1.058)	(0.959,1.041)	79.7	78.9	77.1	75.6
α	95%	(0.402,0.598)	(0.431,0.569)	(0.456,0.544)	(0.469,0.531)	95	93.2	95.1	93.8
	90%	(0.418,0.582)	(0.442,0.558)	(0.463,0.538)	(0.474,0.526)	91.5	88.7	90	90.9
p	95%	(0.173,0.427)	(0.210,0.390)	(0.243,0.357)	(0.260,0.340)	95.1	95.5	93.9	94.7
	90%	(0.193,0.407)	(0.225,0.375)	(0.252,0.348)	(0.266,0.334)	90.6	90.1	89	90.1

Table 4: Confidence intervals and coverage probabilities for ULZOI distribution for different set of values of θ, α and p

It can be seen from Table 3 that the accuracy of the empirical confidence intervals increases with an increase in the sample size for the single inflation case for both the parameters. The coverage probabilities of both α and θ are also seen to be above 85% for all sample sizes when the value of θ is high, i.e. when $\theta = 7$, whereas the coverage probability of θ is seen to be low (less then 80%) when true value of θ is low, i.e. $\theta = 0.25$.

6. APPLICATIONS

In this section we consider real life data arising from High School Leaving Examination results of the State of Manipur, in India for the year 2020 ([3], [2] and [4]). In both the cases, the variable of interest is the proportion of students who have passed the exam and it takes values in the interval $[0, 1]$. The data have been collected from the website E-pao, an e-platform for Manipuris and it consists of 523 observations pertaining to private schools, 305 observations

pertaining to government schools and 67 observations pertaining to aided schools. We compare our model with the famed inflated beta distribution [12].

Table 5 displays the descriptive statistics of the data sets on proportion of students studying in government schools, private schools and aided schools of Manipur, India who have passed the H.S.L.C. Examination, 2020:

Data set	Minimum	Maximum	Mean	1st quartile	Median	3rd quartile
Aided schools	0	1	0.4219	0.0238	0.325	0.7456
Government schools	0	1	0.3632	0.0476	0.3143	0.6154
Private schools	0	1	0.6705	0.4706	0.752	0.9189
All schools combined	0	1	0.5471	0.25	0.6	0.8621

Table 5: Descriptive statistics

We see from Table 5 that 75% of the data points in the data set pertaining to aided schools, government schools, private schools and all schools combined exceed 0.0238, 0.0476, 0.4706 and 0.5471 respectively. Further, the mean exceeds the median for aided and government schools. Thus, it is clear that the data sets on proportion of students studying in aided and government schools are right skewed. Similarly, the data sets on proportion of students studying in private schools and all schools combined are left skewed.

The Zero and zero-and-one inflated unit Lindley distributions and Zero and zero-and-one inflated Beta distributions are fitted to each of the four data sets, where the estimation of parameters have been done using the maximum likelihood method. The Kolmogorov-Smirnov (KS) test statistic is used to compare the fit of each of the distributions to the data sets. Table 6 displays the maximum likelihood estimates and standard errors of the parameters of the Zero Inflated Unit Lindley distribution and Zero Inflated Beta distribution which have been fitted to the four data sets. It also shows the Kolmogorov-Smirnov test statistic values for each case.

Data set	Distribution	MLE & SE of the parameters	K-S test statistic value
Aided schools	ULZI	MLE: $\alpha = 0.2787, \theta = 0.5273$	
		SE: $\alpha = 0.0547, \theta = 0.0459$	0.2357
	ZIB	MLE: $\alpha = 0.2787, \mu = 0.5143, \phi = 2.5075$	
		SE: $\alpha = 0.0547, \mu = 0.0389, \phi = 0.4595$	0.3809
Government schools	ULZI	MLE: $\alpha = 0.2438, \theta = 0.7617$	
		SE: $\alpha = 0.0255, \theta = 0.0328$	0.2411
	ZIB	MLE: $\alpha = 0.2438, \mu = 0.4207, \phi = 2.5035$	
		SE: $\alpha = 0.0255, \mu = 0.0174, \phi = 0.2096$	0.3452
Private schools	ULZI	MLE: $\alpha = 0.0440, \theta = 0.2288$	
		SE: $\alpha = 0.0094, \theta = 0.0065$	0.3148
	ZIB	MLE: $\alpha = 0.0440, \mu = 0.6599, \phi = 2.6921$	
		SE: $\alpha = 0.0094, \mu = 0.0112, \phi = 0.1591$	0.6466
All schools combined	ULZI	MLE: $\alpha = 0.1303, \theta = 0.3011$	
		SE: $\alpha = 0.0117, \theta = 0.0067$	0.3141
	ZIB	MLE: $\alpha = 0.1303, \mu = 0.5772, \phi = 2.1898$	
		SE: $\alpha = 0.0117, \mu = 0.0099, \phi = 0.0988$	0.4616

Table 6: ML estimates and standard errors of the parameters of and Kolmogorov-Smirnov test statistic values for ULZI and ZIB distributions fitted to the four data sets

Figure 3, 4, 5 and 6 show the plot of the observed distribution function and the distribution function of ULZI and ZIB distributions fitted to the four data sets. By visual inspection of the figures 3 to 6, it is clear that the proposed Zero inflated Unit Lindley distribution is a better fit to all the four data sets in comparison to the popular Inflated Beta distribution. From table 6, we observe that for each of the four data sets, the value of the K-S test statistic for the ULZI distribution is less than that of the ZIB distribution. Thus, we conclude that the proposed zero inflated unit Lindley distribution is able to model each of the four data sets better than the zero inflated Beta distribution.
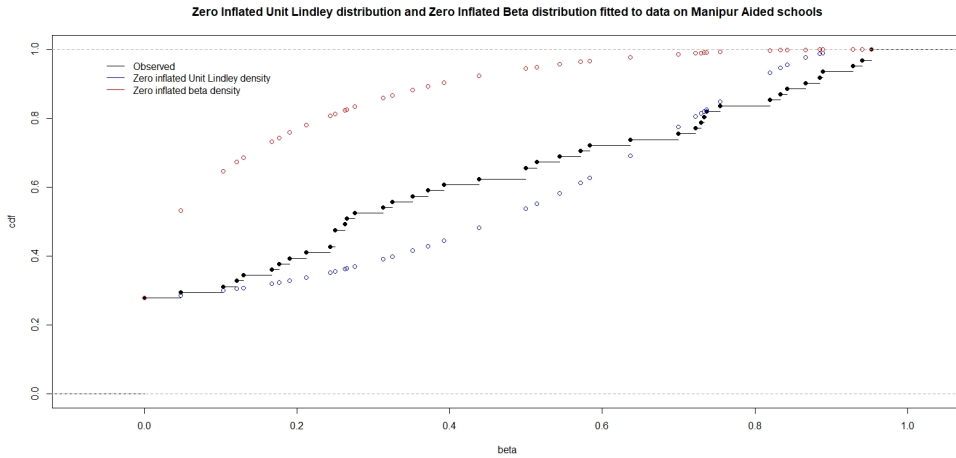
FIGURE 3. Observed d.f and c.d.f. plot of the fitted ULZI and ZIB distributions for proportion of students passing the H.S.L.C. exam 2020, Manipur from aided schools
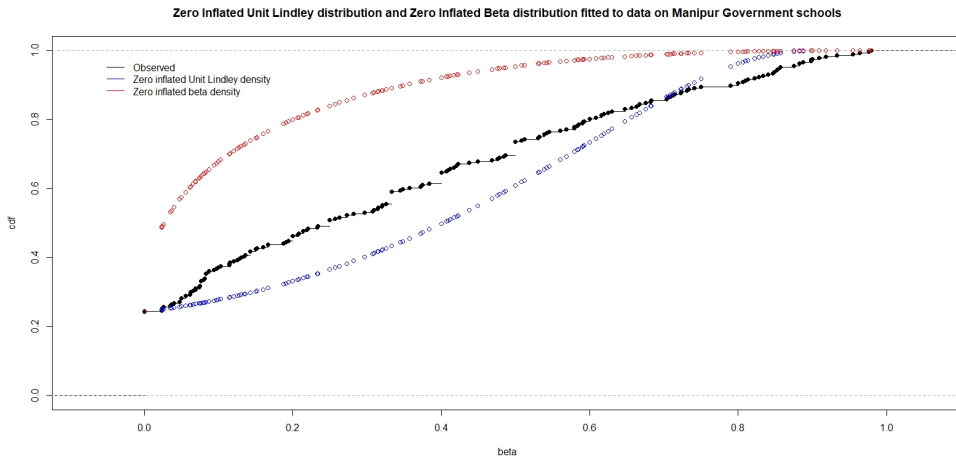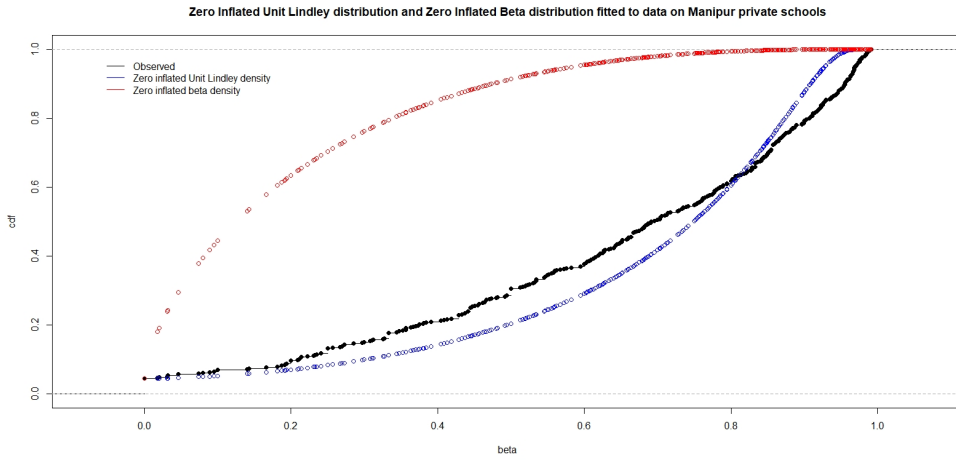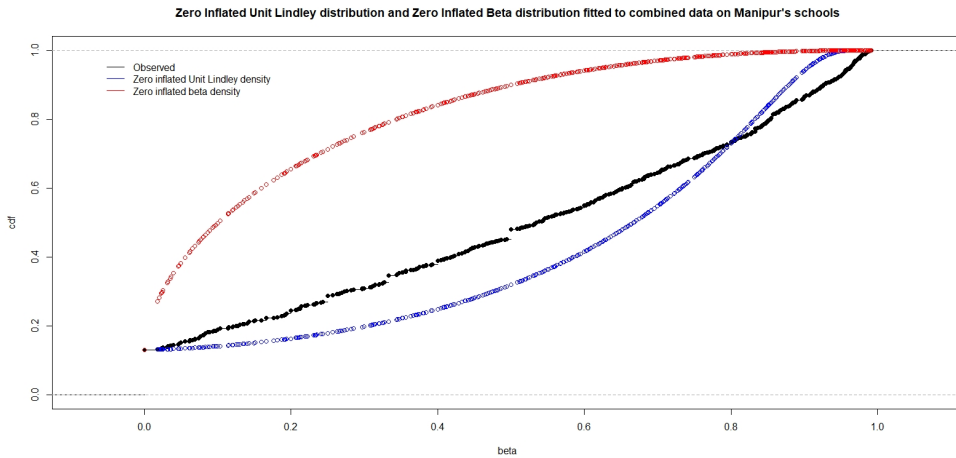

FIGURE 4. Observed d.f and c.d.f. plot of the fitted ULZI and ZIB distributions for proportion of students passing the H.S.L.C. exam 2020, Manipur from government schools

Further, the zero-and-one inflated unit Lindley distribution (ULZOI) and zero-and-one inflated Beta distribution (ZOIB) are fitted to each of the four data sets, with the estimation of parameters being carried out using the maximum likelihood method and the goodness of fit being tested using the Kolmogorov-Smirnov statistic. Table 7 exhibits the maximum likelihood estimates and standard errors of the parameters of the ULZOI and ZOIB distributions which

FIGURE 5. Observed d.f and c.d.f. plot of the fitted ULZI and ZIB distributions for proportion of students passing the H.S.L.C. exam 2020, Manipur from private schools

FIGURE 6. Observed d.f and c.d.f. plot of the fitted ULZI and ZIB distributions for proportion of students passing the H.S.L.C. exam 2020, Manipur from all schools combined

have been fitted to the four data sets, along with the Kolmogorov-Smirnov test statistic values for each case.

Data set	Distribution	MLE(SE) of the parameters	K-S test statistic value
Aided schools	ULZOI	MLE: $\alpha = 0.2022, \theta = 0.3011, p = 0.4088$	
		SE: $\alpha = 0.0132, \theta = 0.0080, p = 0.0361$	0.2881
	ZOIB	$\alpha = 0.2022, \mu = 0.5772, \phi = 2.1898, p = 0.4088$	
		SE: $\alpha = 0.0132, \mu = 0.0099, \phi = 0.0988, p = 0.0361$	0.3967
Government schools	ULZOI	MLE: $\alpha = 0.2984, \theta = 0.7617, p = 0.2416$	
		SE: $\alpha = 0.0262, \theta = 0.0387, p = 0.0490$	0.2237
	ZOIB	$\alpha = 0.2984, \mu = 0.4207, \phi = 2.5035, p = 0.2416$	
		SE: $\alpha = 0.0262, \mu = 0.0174, \phi = 0.2096, p = 0.0490$	0.2885
Private schools	ULZOI	MLE: $\alpha = 0.1281, \theta = 0.2288, p = 0.6866$	
		SE: $\alpha = 0.0143, \theta = 0.0076, p = 0.0581$	0.2870
	ZOIB	MLE: $\alpha = 0.1405, \mu = 0.6599, \phi = 2.6922, p = 0.6866$	
		SE: $\alpha = 0.0143, \mu = 0.0112, \phi = 0.1593, p = 0.0581$	0.5706
All schools combined	ULZOI	MLE: $\alpha = 0.2022, \theta = 0.3011, p = 0.4088$	
		SE: $\alpha = 0.0132, \theta = 0.0080, p = 0.0361$	0.2881
	ZOIB	MLE: $\alpha = 0.2022, \mu = 0.5772, \phi = 2.1898, p = 0.4088$	
		SE: $\alpha 0.0132, \mu = 0.0099, \phi = 0.0988, p = 0.0361$	0.3967

Table 7: ML estimates and standard errors of the parameters of and Kolmogorov-Smirnov test statistic values for ULZOI and ZOIB distributions fitted to the four data sets

Table 7 clearly shows that for each of the four data sets, the K-S test statistic value for the ULZOI distribution is smaller than that for the ZOIB distribution and so, the zero-and-one inflated unit Lindley distribution is able to model these data sets better than the zero-inflated Beta distribution. Figure 7, 8, 9 and 10 show the plot of the observed distribution function and the distribution function of ULZI and ZIB distributions fitted to the four data sets.

It is evident from figures 7 to 10 that the zero-and-one inflated unit Lindley distribution is a better fit than the zero-and-one inflated Beta distribution to each of the four data sets. The same is confirmed by comparing the Kolmogorov-Smirnov test statistic values also. This is an obvious result as all the four data sets contain both zeroes and ones.
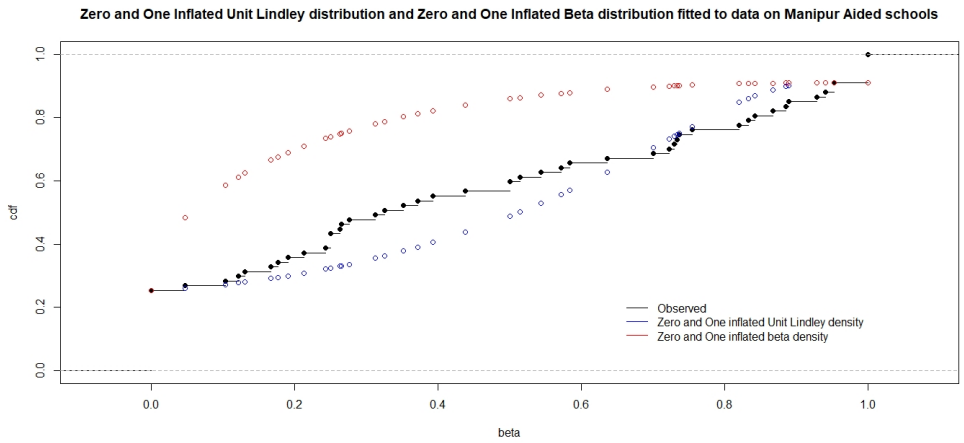
FIGURE 7. Observed d.f and c.d.f. plot of the fitted ULZOI and ZOIB distributions for proportion of students passing the H.S.L.C. exam 2020, Manipur from aided schools
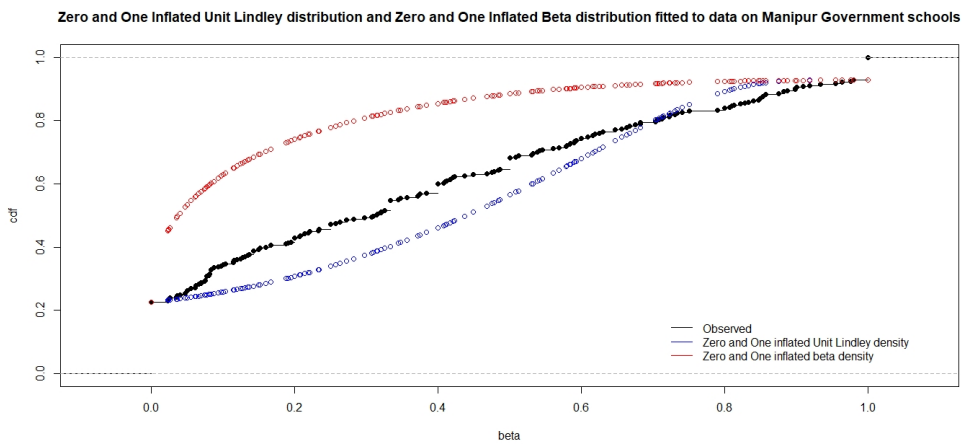

FIGURE 8. Observed d.f and c.d.f. plot of the fitted ULZOI and ZOIB distributions for proportion of students passing the H.S.L.C. exam 2020, Manipur from government schools
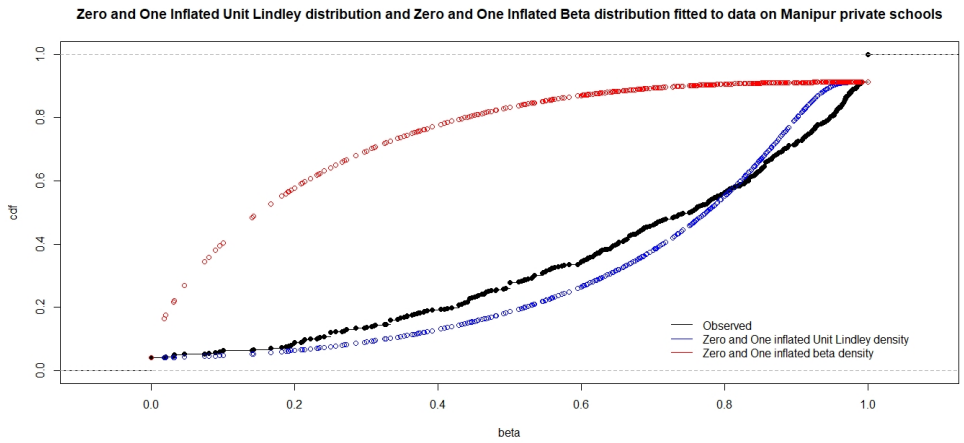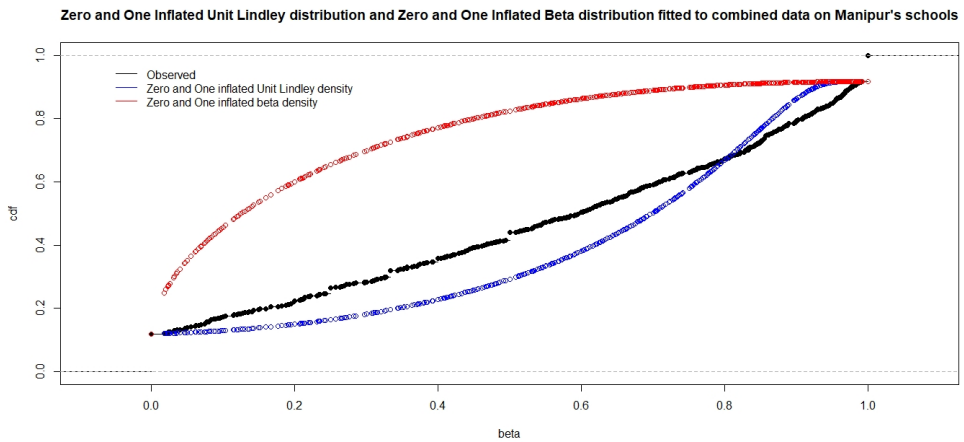
FIGURE 9. Observed d.f and c.d.f. plot of the fitted ULZOI and ZOIB distributions for proportion of students passing the H.S.L.C. exam 2020, Manipur from private schools

FIGURE 10. Observed d.f and c.d.f. plot of the fitted ULZOI and ZOIB distributions for proportion of students passing the H.S.L.C. exam 2020, Manipur from all schools combined

7. CONCLUDING REMARKS

Applied statisticians may often encounter data sets where the variable of interest assumes values in the interval $(0, 1)$, for instance, proportions, ratios, etc. However, the variable may also display the phenomenon of inflation, i.e., it may take the values zero and/or one with positive probabilities. In other words, the interest may lie in modeling variables assuming values in the intervals $[0, 1)$, $(0, 1]$ or $[0, 1]$. Inflated Beta distribution is the leading probability model existing in the current literature to model such variables. In the present work, the zero-or-one inflated version and zero-and-one inflated version of the unit Lindley distribution is discussed and its various important distributional aspects are studied. The finite sample behaviour of the point estimates and interval estimates are studied with the help of Monte Carlo simulation studies. Real life data modelling is performed on data sets of pass proportion of students to show the advantage of modelling them using the proposed distributions over the existing inflated versions of Beta distribution. Our findings confirmed the usefulness of the inflated version of unit Lindley over the considered data.

CONFLICT OF INTERESTS

The author(s) declare that there is no conflict of interests.

REFERENCES

[1] J. Aitchinson, On the distribution of a positive random variable having a discrete probability mass at the origin. J. Amer. Stat. Assoc. 50(1955), 901-908.

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

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

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

[5] B. Burch, J. Egbert, Zero-inflated beta distribution applied to word frequency and lexical dispersion in corpus linguistics. J. Appl. Stat. 47(2)(2020), 337-353.

[6] E. M. Hashimoto, E. M. Ortega, G. M. Cordeiro, V. G. Cancho, C. Klauberg, Zero-spiked regression models generated by gamma random variables with application in the resin oil production. J. Stat. Comput. Simul. 89(1)(2019), 52-70.

[7] N. Johnson, S. Kotz, N. Balakrishnan, Continuous univariate distributions. John Wiley and Sons, New York, 1995.

[8] P. Kumaraswamy, A generalized probability density function for double-bounded random processes. J. Hydrol. 46(1-2)(1980), 79-88.

[9] L. Liu, Y. C. T. Shih, R. L. Strawderman, D. Zhang, B. A. Johnson, H.Chai. Statistical analysis of zero-inflated nonnegative continuous data: A review. Stat. Sci. 34(2)(2019), 253-279.

[10] J. Mazucheli, A.F.B. Menezes, S. Chakraborty, On the one parameter unit-Lindley distribution and its associated regression model for proportion data. J. Appl. Stat. 46(4)(2019), 700-714.

[11] F. Cribari-Neto, J. Santos, Inflated Kumaraswamy distributions, An. Acad. Bras. Ciênc. 91 (2019) e20180955.

[12] R. Ospina, S.L.P. Ferrari, Inflated beta distributions. Stat. Papers. 51 (2010), 111-126.

[13] V. Tomazella, G. H. Pereira, J. S. Nobre, M. Santos-Neto, Zero-adjusted reparameterized Birnbaum–Saunders regression model. Stat. Probab. Lett. 149 (2019), 142-145.