**Aims**
The Chilean Journal of Statistics (ChJS) is an official publication of the Chilean Statistical Society (www.soche.cl). The ChJS takes the place of Revista de la Sociedad Chilena de Estadística, which was published from 1984 to 2000. The ChJS is an international scientific forum strongly committed to gender equality, open access of publications and data, and the new era of information. The ChJS covers a broad range of topics in statistics, data science, data mining, artificial intelligence, and big data, including research, survey and teaching articles, reviews, and material for statistical discussion. In particular, the ChJS considers timely articles organized into the following sections: Theory and methods, computation, simulation, applications and case studies, education and teaching, development, evaluation, review, and validation of statistical software and algorithms, review articles, letters to the editors.

The ChJS editorial board plans to publish one volume per year, with two issues in each volume. On some occasions, certain events or topics may be published in one or more special issues prepared by a guest editor.

**Editors-in-Chief**
Víctor Leiva  
Pontificia Universidad Católica de Valparaíso, Chile

Carolina Marchant  
Universidad Católica del Maule, Chile

**Editors**
Héctor Allende Cid  
Pontificia Universidad Católica de Valparaíso, Chile

José M. Angulo  
Universidad de Granada, Spain

Roberto G. Aykroyd  
University of Leeds, UK

Narayanaswamy Balakrishnan  
McMaster University, Canada

Michelli Barros  
Universidade Federal de Campina Grande, Brazil

Carmen Batanero  
Universidad de Granada, Spain

Ionut Bebu  
The George Washington University, US

Marcelo Bourguignon  
Universidade Federal do Rio Grande do Norte, Brazil

Márcia Branco  
Universidade de São Paulo, Brazil

Oscar Bustos  
Universidad Nacional de Córdoba, Argentina

Luís M. Castro  
Pontificia Universidad Católica de Chile

George Christakos  
San Diego State University, US

Enrico Colosimo  
Università Federale di Minas Gerais, Brazil

Gauss Cordeiro  
Universidade Federal de Pernambuco, Brazil

Francisco Cribari-Neto  
Universidade Federal de Pernambuco, Brazil

Francisco Cysneiros  
Universidade Federal de Pernambuco, Brazil

Mario de Castro  
Universidade de São Paulo, São Carlos, Brazil

José A. Díaz-García  
Universidad Autónoma de Chihuahua, Mexico

Raúl Fierro  
Universidad de Valparaíso, Chile

Isabel Fraga  
Universidade de Concepción, Chile

Manuel Galea  
Pontificia Universidad Católica de Chile

Christian Genest  
McGill University, Canada

Marc G. Genton  
King Abdullah University of Science and Technology, Saudi Arabia

Viviana Giampaoli  
Università Federale di Minas Gerais, Brazil

Patricia Giménez  
Universidad Nacional de Mar del Plata, Argentina

Hector Gómez  
Universidad de Antioquia, Colombia

Daniel Griffith  
University of Texas at Dallas, US

Eduardo Gutiérrez-Peña  
Universidad Nacional Autónoma de Mexico

Nikolai Kolev  
Universidade de São Paulo, Brazil

Eduardo Lalla  
University of Twente, Netherlands

Shuangzhe Liu  
University of Canberra, Australia

Jesús López-Fidalgo  
Universidad de Navarra, Spain

Liliana López-Kleine  
Universidad Nacional de Colombia

Rosangela H. Loschi  
Universidade Federal de Minas Gerais, Brazil

Manuel Mendoza  
Instituto Tecnológico Autónomo de Mexico

Oríetta Nicolis  
Universidad Andrés Bello, Chile

Ana B. Nieto  
Universidad de Salamanca, Spain

Teresa Oliveira  
Universidade Aberta, Portugal

Félix Osorio  
Universidad Técnica Federico Santa María, Chile

Carlos D. Paulino  
Instituto Superior Tecnico, Portugal

Fernando Quintana  
Pontificia Universidad Católica de Chile

Nalini Ravishanker  
University of Connecticut, US

Fabrizio Ruggeri  
Consiglio Nazionale delle Ricerche, Italy

José M. Sarabia  
Universidad de Cantabria, Spain

Helton Saulo  
Universidade de Brasília, Brazil

Pranab K. Sen  
University of North Carolina at Chapel Hill, US

Julio Singer  
Universidade de São Paulo, Brazil

Milan Stehlik  
Johannes Kepler University, Austria

Alejandra Tapia  
Universidad Católica del Maule, Chile

M. Dolores Ugarte  
Universidad Pública de Navarra, Spain

Andrei Volodin  
University of Regina, Canada

**Editorial Assistant**
Mauricio Román  
Chile

**Founding Editor**
Guido del Pino  
Pontificia Universidad Católica de Chile
Contents

Carolina Marchant and Víctor Leiva
Starting a new decade of the Chilean Journal of Statistics in COVID-19 pandemic times with new editors-in-chief 1

Luz Milena Zea Fernandez and Thiago A.N. de Andrade
The erf-G family: new unconditioned and log-linear regression models 3

Thodur Parthasarathy Sripriya, Mamandur Rangaswamy Srinivasan, and Meenakshisundaram Subbiah
Detecting outliers in $I \times J$ tables through the level of susceptibility 25

Adolphus Wagala
A likelihood ratio test for correlated paired multivariate samples 41

Josmar Mazucheli, Sudeep R. Bapat, and André Felipe B. Menezes
A new one-parameter unit-Lindley distribution 53
A new one-parameter unit-Lindley distribution

Josmar Mazucheli¹, Sudeep R. Bapat²,⇤, and André Felipe B. Menezes¹

¹Department of Statistics, Universidade Estadual de Maringá, DEs, PR, Brazil
²Department of Statistics and Applied Probability, University of California, Santa Barbara, USA

(Received: 28 May 2020 · Accepted in final form: 11 December 2019)

Abstract

A large number of useful distributions for data analysis are obtained by transforming different random variables. An example is the one-parameter unit-Lindley distribution, obtained by transforming a random variable which has a Lindley distribution. In this paper, we introduce a new one-parameter unit-Lindley distribution, useful for data analysis in the interval (0,1]. It follows some interesting properties such as having closed form expressions for the moments, belonging to the exponential family. We also analyze a practical application having covariates, by setting up a suitable regression and show that our model fits much better than both unit-Lindley and beta regressions.

Keywords: Maximum likelihood estimation · Proportion data · Regression model · Unit-Lindley distribution · Unit interval.

Mathematics Subject Classification: Primary 60E05 · Secondary 62F10.

1. Introduction

In many practical applications, one encounters data which is spread out in a bounded interval. Moreover this interval happens to be (0,1), where the data would be certain proportions, ratios or standardized scores. Some of the well known distributions having supports in (0,1) are uniform, beta and Kumaraswamy. However all of these contain at least 2 parameters and hence it becomes tedious when it comes to estimation. Further, the beta distribution doesn’t have closed form expressions for the cumulative distribution function (CDF), whereas the Kumaraswamy distribution fails to have a closed form expression for the moments. Some of the only one-parameter distributions in (0,1) are the Topp-Leone distribution (Topp and Leone, 1955) and the newly proposed unit-Lindley distribution by Mazucheli et al. (2019), where the authors have transformed a suitable Lindley distribution. One of the outlook in recent times has been to transform some existing distributions to get more useful distributions having specific properties. A lot of work has been done related to the Lindley distribution in the last few years. Some of the prominent works include the quasi-Lindley distribution by Shanker and Mishra (2013), the log-Lindley distribution by Gómez-Déniz et al. (2014), the power-Lindley distribution or the generalized-Lindley distribution by Nadarajah et al. (2011).

*Corresponding author. Email: bapat@pstat.ucsb.edu
In this paper we propose a new unit-Lindley (NUL) distribution which is a modification to the existing unit-Lindley distribution, by picking a different transformation. This NUL distribution enjoys several interesting properties such as existence of closed form expressions for the moments, the CDF and belonging to the exponential family. Due to its simple formula, one can incorporate a regression setup by involving several covariates in the mean to study their dependence on the response. The advantage of this NUL distribution over the existing unit-Lindley model can be clearly seen through the real-data application which we present in Section 4.

In Section 2 we propose the NUL distribution by providing the density and the distribution functions. We also focus on several interesting properties such as defining the moments, the HR function, the mean residual life function, the quantile function and others. Section 3 involves estimation properties including both method of moments and maximum likelihood (ML) estimators, where we also provide a bias-corrected ML estimators, in addition to a regression modeling. In Section 4, we provide the numerical applications of our work. Extensive simulation analyses are covered by taking a wide range of parameter values. We fit our proposed NUL model to a real-data from finance which involves a ratio of premiums plus uninsured losses and the total assets as the response whereas Section 5 provides brief conclusions.

2. Some Mathematical Results

In the following subsections, we provide a number of key properties of the NUL distribution.

2.1 The NUL distribution

Some probability distributions useful in analyzing data in the unit interval, such as Johnson $S_B$ (Johnson, 1949), Johnson $S'_B$ (Johnson, 1955), unit-Gamma Grassia (1977); Tadikamalla (1981), unit-Logistic (Tadikamalla and Johnson, 1982), log-Lindley (Gómez-Déniz et al., 2014), unit-Inverse-Gaussian (Ghitany et al., 2018), unit-Birnbaum-Saunders (Mazucheli et al., 2018a), unit-Weibull (Mazucheli et al., 2018b) are formulated by transforming specific random variables (RVs). It is important to note that beta and Kumaraswamy (Kumaraswamy, 1980) distributions also can be obtained by transformations.

A unit-Lindley distribution was proposed by Mazucheli et al. (2019) by considering the transformation $X = Y/[1 + Y]$, where $Y \sim \text{Lindley}(\theta)$ (Lindley, 1958). Here we apply the transformation $X = 1/[1 + Y]$, where $Y \sim \text{Lindley}(\theta)$ and propose the distribution of $X$ to be the NUL distribution. One can easily derive its probability density function (PDF) and the CDF say, using the inverse transform method. These expressions are given respectively by

$$f(x|\theta) = \frac{\theta^2}{x^3[1+\theta]} \exp \left(-\theta \left[\frac{1-x}{x}\right]\right),$$  \hspace{1cm} (1)$$

$$F(x|\theta) = \frac{[\theta+x]}{x[1+\theta]} \exp \left(-\theta \left[\frac{1-x}{x}\right]\right),$$  \hspace{1cm} (2)$$

where $0 < x \leq 1$ and $\theta > 0$. Figure (1) shows the PDF of the unit-Lindley distribution for selected values of $\theta$. 
Unlike other distributions such as the unit-Lindley, here we have the possibility of having observations equal to 1 and from (1) the first derivative of $f(x|\theta)$ is

$$\frac{d}{dx} f(x|\theta) = \frac{\theta^2 (\theta - 3x)}{[1 + \theta x^3]} \exp \left(-\theta \left[ \frac{1 - x}{x} \right] \right),$$

which implies that the PDF is unimodal with maximum at $X_{\text{max}} = \theta/3$ for all values of $\theta < 3$ and $X_{\text{max}} = 1$ for $\theta \geq 3$.

![Graphs showing probability density functions for selected values of $\theta$.](figure1.png)

Figure 1. Probability density function of the NUL distribution for selected values of $\theta$. 


2.2 Convexity

Proposition 2.1 The CDF of the NUL is convex for $\theta > 3$.

Proof The second derivative of $F(x|\theta)$ is

$$F''(x|\theta) = \frac{\theta^2[\theta - 3x]}{x^5[1 + \theta]} \exp\left(-\frac{\theta[1 - x]}{x}\right).$$

This implies that for all $x$ in $(0, 1)$, $F''(x|\theta) < 0$ only if $\theta < 0$ therefore it can never be concave and $F''(x|\theta) > 0$ if $\theta > 3$. Hence $F(x|\theta)$ is a convex function of $x$ for $\theta > 3$.

Proposition 2.2 The PDF of the unit-Lindley distribution is log-concave for all $0 < x \leq 1$ if $\theta > 3/2$.

Proof We know that $f(x|\theta)$ is log-concave (log-convex) function of $x$ if for all $x$ in $(0, 1]$ $\frac{d}{dx} \log f(x|\theta)$ is a non-increasing (non-decreasing) function of $x$. Note that

$$\frac{d^2}{dx^2} \log f(x|\theta) = \frac{d}{dx} \frac{f'(x|\theta)}{f(x|\theta)} = \frac{d}{dx} \frac{[\theta - 3x]}{x^2} = -\frac{2[\theta - 3x]}{x^3} - \frac{3}{x^2}.$$

This is always $< 0$ for all $x$ in $(0, 1]$ whenever $\theta > 3/2$. Hence $f(x|\theta)$ is log-concave for all $0 < x \leq 1$, if $\theta > 3/2$.

2.3 Hazard rate function

The hazard rate (HR) function of the unit-Lindley distribution is given by

$$h(x|\theta) = \frac{f(x|\theta)}{1 - F(x|\theta)} = \frac{\theta^2}{[\theta + x]x^2}, \quad 0 < x \leq 1.$$

Since $dh(x|\theta)dx = -[\theta^2(2\theta + 3x)]/[x^3(\theta + x)^2] < 0$ for all $\theta > 0$ the HR function is decreasing in $x$. Note that $\lim_{x \to 0} h(x|\theta) = \infty$ while $\lim_{x \to 1} h(x|\theta) = \theta^2/[1 + \theta]$.

2.4 Moments

The $k$-th moment about origin of the unit-Lindley distribution can be obtained from

$$\mu'_k = E(X^k) = \int_0^1 kx^{k-1} \left(1 - \frac{[\theta + x]}{x[1 + \theta]} \exp\left(-\theta \left[\frac{1-x}{x}\right]\right)\right)dx, \quad k = 1, 2, \ldots.$$

In particular, for $k = 1, 2, 3, 4$ we get

$$\mu'_1 = \frac{\theta}{1+\theta}, \quad \mu'_2 = \frac{\theta^2 \exp(\theta) Ei(1, \theta)}{1+\theta},$$

$$\mu'_3 = \frac{\theta^2[1-\theta \exp(\theta) Ei(1, \theta)]}{1+\theta}, \quad \mu'_4 = \frac{\theta^2[1-\theta+\theta^2 \exp(\theta) Ei(1, \theta)]}{2[1+\theta]},$$

where $Ei(a, z) = \int_{1}^{\infty} x^{-a} \exp(-xz)dx$ is the exponential integral function; see Abramowitz and Stegun (1974).
The \( k \)-th incomplete moment about origin is obtained from
\[
T_k(t) = E(X^k | X < t) = \frac{\theta^2}{[1 + \theta] F(t/\theta)} \int_0^t x^{k-3} \exp \left( -\theta \left[ \frac{1-x}{x} \right] \right) dx, \quad k = 1, 2, \ldots
\]
and for \( k = 1, 2, 3, 4 \) we have
\[
T_1(t) = \frac{\theta t}{[\theta + t]}, \quad T_2(t) = \frac{\theta^2 \exp(\theta t) Ei(1, t)}{[\theta + t] \exp(\theta [t-1]/t)},
\]
\[
T_3(t) = \frac{\theta^2 t [t-\theta Ei(1, \frac{t}{\theta}) \exp(\frac{t}{\theta})]}{[\theta + t]}, \quad T_4(t) = \frac{\theta^2 t [t^2 - \theta Ei(1, \frac{t}{\theta}) \exp(\frac{t}{\theta})]}{2[\theta + t]}.
\]

2.5 MEAN RESIDUAL LIFE FUNCTION

For a nonnegative continuous RV \( X \) the mean residual life function is defined as \( \mu(t|\theta) = E(X - t | X > t) \) and is given by
\[
\mu(t|\theta) = \frac{1}{S(t|\theta)} \int_t^\infty S(x | \theta) dx.
\]
For the NUL distribution, we get
\[
\mu(t|\theta) = \frac{t \{ [(1 + \theta) t - \theta] \delta(t, \theta) - \exp(\theta t) \delta(-t, \theta) \}}{t [\theta + t] \delta \left( \frac{t}{t-1}, \theta \right) - [1 + \theta]},
\]
where \( \delta(t, \theta) = \exp(\theta / t) \).

2.6 STRESS-STRENGTH RELIABILITY

Let \( X \) and \( Y \) be two independent NUL RVs with parameters \( \theta_1, \theta_2 \) respectively and having PDF’s \( f_X \) and \( f_Y \). Then the stress-strength reliability measure (Kotz and Pensky, 2003) is given by
\[
R = P(Y < X) = \int_0^1 f_X(x|\theta_1) F_Y(x | \theta_2) dx
\]
\[
= \frac{\theta_1^2 \left[ \theta_1^2 \theta_2 + \theta_1^2 + \theta_1 + 2 \theta_1 \theta_2^2 + 4 \theta_1 \theta_2 + 3 \theta_2 + \theta_2^3 + 3 \theta_2^2 \right]}{[1 + \theta_2] [1 + \theta_1] [\theta_1 + \theta_2]}.\]

2.7 QUANTILE FUNCTION

Let \( X \) be a NUL RV with CDF as given in (2). The quantile function, \( Q(p) = F^{-1}(p) \), can then be written as
\[
Q(p|\theta) = -\frac{\theta}{1 + W[-\exp(-(1 + \theta) p(1 + \theta))]}, \quad (3)
\]
such that \( 0 < p < 1 \) and \( W \) is the Lambert \( W \) function which is a multivalued complex function defined as the solution of the equation \( W(z) \exp[W(z)] = z \). For more details on
the Lambert \( W \) function, readers may refer to Corless et al. (1996), Jodrá (2010), Veberić (2012) and references cited therein.

2.8 Mean deviation

As pointed out, for example in Ghitany et al. (2008), the amount of scatter in a population is measured to some extent by the totality of deviations from the mean and the median. These are known as the mean deviation about the mean and the mean deviation about the median and are defined as

\[
\delta(X) = \int_{x}^{\infty} |X - m| f(x \mid \theta) \, dx = 2 \left[ m F(m) - \int_{0}^{m} x f(x \mid \theta) \, dx \right],
\]

with \( m = E(X) \) or \( m = \text{Median}(X) \) respectively. Considering (2) and (1) in (4) we get

\[
\delta(X) = \frac{2m \exp(\theta(m - 1)/m)}{1 + \theta}.
\]

For \( m = E(X) \) we get \( \delta(X) = 2\theta \exp(-1)/(1 + \theta)^2 \). Considering \( m = Q(0.5 \mid \theta) \) we have the expression for the mean deviation about the median, where the expression for \( Q(\cdot \mid \theta) \) is given in (3).

2.9 Exponential family

A distribution belongs to the exponential family (Dobson, 2001) if it is of the form

\[
f(x \mid \theta) = \exp \left( Q(\theta) T(x \mid \theta) + D(\theta) + S(x \mid \theta) \right).
\]

It can be easily seen that the proposed distribution belongs to the exponential family by rewriting the PDF given in (1) as

\[
f(x \mid \theta) = \exp \left( -\frac{\theta(1 - x)}{x} \right) \exp \left( \log \left( \frac{\theta^2}{1 + \theta} \right) \right) \exp \left( \log(x^{-3}) \right),
\]

where \( Q(\theta) = \theta \), \( T(x \mid \theta) = (1 - x)/x \), \( D(\theta) = \log \left( \theta^2/[1 + \theta] \right) \), \( S(x \mid \theta) = \log(x^{-3}) \). Therefore, \( T(x) = \sum_{i=1}^{n} (1 - x_i)/x_i \) is a complete sufficient estimator for \( \theta \) based on a sample of size \( n \) from the proposed distribution. Besides that, since the distribution belongs to an exponential family, a minimum-variance unbiased estimator can be obtained by bias corrected ML estimator.

3. Estimation

In this section, we will derive the method of moments (MME) and ML estimators of parameter \( \theta \) of a NUL distribution. For the ML estimator of \( \theta \) we derive the closed-form expressions for the second order bias-correction. In addition, in this section, we consider regression modeling.
3.1 Maximum likelihood estimation

Let $X_1, \ldots, X_n$ be a random sample from the NUL distribution with PDF, $(1)$. Then, for observed $\mathbf{x} = (x_1, \ldots, x_n)$, the log-likelihood function of $\theta$ can be written as

$$\ell(\theta|\mathbf{x}) \propto 2n \log(\theta) - n \log(1 + \theta) - \theta t(\mathbf{x}).$$

The ML estimate $\hat{\theta}$ of $\theta$ is obtained by solving the following linear equation

$$\frac{d}{d\theta} \ell(\theta|\mathbf{x}) = \frac{2n}{\theta} - \frac{n}{1 + \theta} - t(\mathbf{x}) = 0$$

which gives

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

Next

$$\frac{d^2}{d\theta^2} \ell(\theta|\mathbf{x}) = \frac{n}{(1 + \theta)^2} - \frac{2n}{\theta^2} < 0$$

for all $\theta$, in particular for $\theta = \hat{\theta}$.

Since $d^2\ell(\theta|\mathbf{x})/d\theta^2$ is data-independent, we have that $nE[d^2 \log f(X|\theta)/d\theta^2] = d^2\ell(\theta|\mathbf{x})/d\theta^2$. Thus, the expected Fisher information is $I(\hat{\theta}) = 2n/\theta^2 - n/[1 + \theta]^2$. From the large sample theory (Lehmann and Casella (1998, pp. 461-463)), the asymptotic distribution of ML estimator $\hat{\theta}$ of $\theta$ is such that

$$\sqrt{n} (\hat{\theta} - \theta) \xrightarrow{D} N\left(0, V(\hat{\theta})\right),$$

where $\xrightarrow{D}$ denotes convergence in distribution and $V(\hat{\theta})$ is just the inverse of the expected Fisher information written as $V(\hat{\theta}) = \theta^2 [1 + \theta]^2/n [\theta^2 + 4 \theta + 2]$. It is easy to see that for $\psi = g(\theta) = E(X) \hat{\psi} = E(X) = 1/[1 + \theta]$ and $V(\hat{\psi}) = \theta^2/n [\theta^2 + 4 \theta + 2]$.

Hence, the asymptotic 100 $(1 - \alpha)\%$ confidence intervals (CIs) for $\theta$ and $\psi$ are given, respectively, by

$$\hat{\theta} \pm z_{\alpha/2} \sqrt{\frac{\hat{\theta}^2 [1 + \theta]^2}{n [\theta^2 + 4 \theta + 2]}} \quad \text{and} \quad \frac{1}{1 + \theta} \pm \frac{z_{\alpha/2}}{1 + \theta} \sqrt{\frac{\hat{\psi}^2}{n [\theta^2 + 4 \theta + 2]}},$$

where $z_{\alpha/2}$ is the upper $\alpha/2$ quantile of the standard normal distribution.

It is important to note that for a Bayesian setup, we can use the Jeffreys invariant prior for $\theta$, given by $\pi(\theta) \propto \sqrt{I(\theta)}$. However we will not consider it further in this paper.

Cox and Snell (1968) provided a framework to estimate the bias, to $O(n^{-1})$ for the ML estimates of parameters of regular densities. Hence, subtracting the estimated bias from the original ML estimator produces a bias-corrected estimator (BCE) that is unbiased to $O(n^{-2})$. Following Cox and Snell (1968) the analytical expression for bias-correction of an scalar $\hat{\theta}$, given by

$$B\left(\hat{\theta}\right) = (\kappa^{11})^2 \left[ 0.5 \kappa_{111} + \kappa_{111,1} \right] + O(n^{-2}),$$
where

\[ \kappa_{11} = \mathbb{E} \left[ \frac{d^2}{d\theta^2} \ell(\theta | x) \right]^{-1} = \frac{\theta^2 (1 + \theta)^2}{n (\theta^2 + 4 \theta + 2)}, \]

\[ \kappa_{11,1} = \mathbb{E} \left[ - \frac{d^2}{d\theta^2} \ell(\theta | x) \times \frac{d}{d\theta} \ell(\theta | x) \right] = 0, \]

and

\[ \kappa_{111} = \mathbb{E} \left[ - \frac{d^3}{d\theta^3} \ell(\theta | x) \right] = \frac{2 n (\theta^3 + 6 \theta^2 + 6 \theta + 2)}{\theta^3 (1 + \theta)^3}. \]

Thus, the bias-corrected ML estimator \( \tilde{\theta} \) is

\[ \tilde{\theta} = \hat{\theta} - \frac{\hat{\theta} \left[ \hat{\theta} + \hat{\theta}^2 \right] \left[ \hat{\theta}^3 + 6 \hat{\theta}^2 + 6 \hat{\theta} + 2 \right]}{n \left[ \hat{\theta}^2 + 4 \hat{\theta} + 2 \right]^2}, \]

where the right hand side is \( \mathcal{B} \left( \hat{\theta} \right) \).

Re-parameterizing (1) in terms of the mean \( \mu = \theta / \left[ 1 + \theta \right] \), the ML of \( \mu \) is obtained as

\[ \hat{\mu} = \frac{1}{2n} \left[ 3 n + t(x) - \sqrt{t(x)^2 + 6 n t(x) + n^2} \right], \]

and the corresponding bias-corrected ML estimator \( \tilde{\mu} \) of \( \mu \) as

\[ \tilde{\mu} = \hat{\mu} - \frac{2 \hat{\mu} \left[ \hat{\mu} - 1 \right]^2}{n \left[ \hat{\mu}^2 - 2 \right]^2}. \]

### 3.2 Method of moment estimation

Let \( X_1, \ldots, X_n \) be a random sample from the unit-Lindley distribution with PDF (1). Then, the MME \( \hat{\theta}_{\text{MME}} \) of \( \theta \) is given by

\[ \hat{\theta}_{\text{MME}} = \frac{\bar{X}}{1 - \bar{X}} = \left[ \frac{1}{\bar{X}} - 1 \right]^{-1}, \]

which is positively biased, that is, \( \mathbb{E}(\hat{\theta}) - \theta > 0 \).

**Proof** Let \( \hat{\theta}_{\text{MME}} = g(\bar{X}) \) and \( g(t) = t / [1 - t] \) for \( t > 0 \). Since \( g''(t) = -2 / [t - 1]^3 > 0 \) for all \( t < 1 \), \( g(t) \) is strictly convex. Thus, by Jensen’s inequality, we have \( \mathbb{E}(g(\bar{X})) > g(\mathbb{E}(\bar{X})). \) Since \( g(\mathbb{E}(\bar{X})) = g(\theta / [1 + \theta]) = \theta \) we get \( \mathbb{E}(\hat{\theta}) > \theta \).

### 3.3 Regression analysis

We will now present a real data analysis in order to showcase the applicability of the proposed distribution. Since the NUL distribution has a closed form expression for the
mean we are able to introduce a new regression model for bounded response variable. The re-parametrized PDF of the NUL distribution is given by

\[
f(y | \mu) = \frac{\mu^2}{[1 - \mu] y^3} \exp \left( -\frac{\mu [1 - y]}{y [1 - \mu]} \right),
\]

where \( 0 < y \leq 1 \) and \( 0 < \mu \leq 1 \). Under this parametrization the mean and variance of NUL distribution are given by

\[
E(Y) = \mu \quad \text{and} \quad \text{Var}(Y) = \frac{\mu^2}{1 - \mu} \left[ \text{Ei} \left( 1, \frac{\mu}{1 - \mu} \right) \exp \left( \frac{\mu}{1 - \mu} \right) + \mu - 1 \right].
\]

Let \( Y_1, \ldots, Y_n \) be \( n \) independent RVs, where \( Y_i \sim \text{NUL}(\mu_i), i = 1, \ldots, n \) with PDF, given by (5). The NUL regression model is defined assuming that the mean of \( Y_i \) can be written as

\[
g(\mu_i) = x_i^T \beta
\]

where \( \beta = (\beta_0, \ldots, \beta_{(p-1)})^T \) is a \( p \)-dimensional vector of unknown regression coefficients \( (p < n) \) and \( x_i = (1, x_{i1}, \ldots, x_{i(p-1)})^T \) denotes the observations on \( p \) known covariates. Note that the variance of \( Y_i \) is a function of \( \mu_i \) and, as a consequence of the covariate values, which implies that non-constant response variances are naturally accommodated into the model.

We shall assume that the mean link function \( g \) is a strictly monotonic and twice differentiable function that maps \((0, 1)\) into \( \mathbb{R} \). Some of the most common link functions are:

(i) logit: \( g(\mu) = \log(\mu / (1 - \mu)) \);
(ii) probit: \( g(\mu) = \Phi^{-1}(\mu) \), where \( \Phi^{-1} \) is the standard normal quantile function;
(iii) complementary log-log: \( g(\mu) = \log[\log(1 - \mu)] \).

Inferences about the regression coefficients \( \beta \) can be performed under the likelihood paradigm (Lehmann and Casella, 1998). The log-likelihood function based on a sample of \( n \) independent observations is

\[
\ell(\beta) \propto 2 \sum_{i=1}^{n} \log(\mu_i) - \sum_{i=1}^{n} \log(1 - \mu_i) - \sum_{i=1}^{n} \frac{\mu_i [1 - y_i]}{y_i [1 - \mu_i]},
\]

where \( \mu_i = g^{-1}(x_i^T \beta) \).

The ML estimates \( \hat{\beta} \) of \( \beta \) are obtained by maximizing the log-likelihood function defined in (6) using standard optimization methods, such as Newton-Raphson or quasi-Newton. In this paper, the ML estimate were obtained by the quasi-Newton method available in the SAS/NLMIXED procedure (https://www.sas.com/).

For comparison purpose, we also considered the beta and unit-Lindley regression models. The PDF of the alternative regression models are:

- Beta regression (Cepeda-Cuervo, 2001; Ferrari and Cribari-Neto, 2004):

\[
f(y | \mu, \phi) = \frac{\Gamma(\phi)}{\Gamma(\mu \phi) \Gamma([1 - \mu] \phi)} y^{\mu \phi - 1} [1 - y]^{[1 - \mu] \phi - 1}, \quad 0 < y < 1
\]

where \( 0 < \mu < 1 \) denotes the mean and \( \phi > 0 \) is a precision parameter.
• Unit-Lindley regression (Mazucheli et al., 2019):

\[ f(y|\mu) = \frac{(1-\mu)^2}{\mu[1-y]^3} \exp\left(-\frac{y(1-\mu)}{\mu[1-y]}\right), \quad 0 < y < 1 \]

where \(0 < \mu < 1\) denotes the mean.

To discriminate and choose the best among the proposed models, the Akaike (AIC) (Akaike, 1974), Schwarz (BIC) (Schwarz, 1978) and corrected Akaike (AICC) (Cavanaugh, 1997) information criteria were used. These measures are defined as follows

\[
\text{AIC} = 2p - 2 \log \hat{L}, \quad \text{BIC} = \log(n) p - 2 \log \hat{L}, \quad \text{AICC} = \frac{2n[p+1]}{n-p-2} - 2 \log \hat{L}
\]

where \(\hat{L}\) is the likelihood evaluated at the ML estimates, \(p\) is the number of parameters in the model and \(n\) the number of observations. The decision rule, in all these criteria, is favorable to the model with the lowest value (Held and Sabanés Bové, 2014). To quantify the uncertainty associated with these criteria, the non-parametric Bootstrap approach was used to decide on the final model. We considered 10,000 independent runs and calculated the percentage of times each model was selected.

To assess the adequacy of the regression models we used the Cox-Snell residuals and examined the half-normal plot with simulated envelope (Atkinson, 1981). The Cox-Snell residuals are defined as

\[ r_i = -\log \left(1 - \hat{F}(y_i)\right), \quad i = 1, \ldots, n, \]

where \(\hat{F}\) is the estimated CDF. A notable property of the Cox-Snell residuals is that if the regression model fits the data well, \(r_i\)’s follow a standard exponential distribution.

4. Numerical results

In this section, we conduct a Monte Carlo simulation in order to evaluate and compare the finite-sample behavior of the ML estimators, its bias-corrected counterpart obtained by the Cox-Snell methodology (BCE) and the MME of the parameter \(\theta\) of the NUL distribution. In addition, in this section, an empirical illustration is conducted.

4.1 Simulation study

We have generated samples ranging from 10 to 90 with a gap of 10 and \(\theta = 0.1, 0.5, 1.0, 1.5, 2.0, 3.0\) and 4.0. To simulate observations from the proposed distribution we generated \(Y\) from Lindley distribution (see, \texttt{rlindley} function in \texttt{LindleyR} library) and then used the transformation \(X = 1/[1+Y]\). The simulation experiment was repeated \(M = 20,000\) times. The performance evaluation was done based on the estimated bias and estimated root mean squared error (RMSE).

Figure 2 shows that ML estimates and MME of \(\theta\) are positively biased, while the BCE estimator achieve substantial bias reduction, especially for small and moderate sample sizes. It is also observed that the RMSE decreases as \(n\) increases, as expected. Additionally, the RMSE of the corrected estimates are smaller than those of the uncorrected estimates.
The real data set considered is presented by Schmit and Roth (1990), and corresponds to the 73 responses to a questionnaire sent to 374 risk managers of large North American organizations. The objective of Schmit and Roth (1990) was to evaluate the cost effectiveness with the management philosophy of controlling the company’s exposure to various property losses and accidents, taking into account company characteristics such as size and type of industry.

The response variable $y$ (Firm cost) is the firm-specific ratio of premiums plus uninsured losses divided by total assets. The covariates associated with this response variable are:

- $X_1$ (Assume): firm-specific ratio of the summation of per occurrence retention levels, as measured by the corporate risk manager.
- $X_2$ (Cap): 1 if the firm uses a captive and 0 otherwise.
- $X_3$ (Sizelog): log of the firm’s total asset value.
- $X_4$ (Indcost): industry average of premiums plus uninsured losses divided by total assets, as measured by the 1985 Cost of Risk Survey (a measure of risk).
- $X_5$ (Central): importance of local manager in choosing local retention levels, as measured by the corporate risk manager.

4.2 Empirical illustration

Figure 2. Upper Panel: Estimated bias. Lower Panel: Estimated root mean squared error. (1: $\theta = 0.1$, 2: $\theta = 0.5$, 3: $\theta = 1.0$, 4: $\theta = 1.5$, 5: $\theta = 2.0$, 6: $\theta = 3.0$ and 7: $\theta = 4.0$).
• $X_6$ (Soph): importance of analytical tools in making risk management decisions, as measured by the corporate risk manager.

We assume that the regression structure for the mean is given by

$$\text{logit}(\mu_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6}, \quad i = 1, \ldots, 73,$$

where $x_{ij}$ are the values of the covariate $X_j$.

The point estimates and the 95% confidence intervals for the parameters of the three regression models are given in Table 1. It is observed that the NUL and beta regression models have the same significant covariates to explain the response variable, which are the Sizelog and Indcost variables.

### Table 1. The ML estimates and the 95% confidence intervals.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>NUL MLE</th>
<th>95% CI</th>
<th>UL MLE</th>
<th>95% CI</th>
<th>Beta MLE</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\beta_0$</td>
<td>4.3789</td>
<td>(2.6395, 6.1183)</td>
<td>3.0506</td>
<td>(0.8132, 5.2879)</td>
<td>1.8880</td>
<td>(-0.4096, 4.1855)</td>
</tr>
<tr>
<td>$\beta_1$</td>
<td>-0.0050</td>
<td>(-0.0273, 0.0173)</td>
<td>-0.0592</td>
<td>(-0.0830, -0.0354)</td>
<td>-0.0121</td>
<td>(-0.0394, 0.0151)</td>
</tr>
<tr>
<td>$\beta_2$</td>
<td>-0.0112</td>
<td>(-0.3780, 0.3556)</td>
<td>1.8972</td>
<td>(1.2649, 2.5295)</td>
<td>0.1780</td>
<td>(-0.2763, 0.6322)</td>
</tr>
<tr>
<td>$\beta_3$</td>
<td>-0.8943</td>
<td>(-1.0709, -0.7176)</td>
<td>-0.6606</td>
<td>(-0.8889, -0.4322)</td>
<td>-0.5115</td>
<td>(-0.7524, -0.2705)</td>
</tr>
<tr>
<td>$\beta_4$</td>
<td>1.7145</td>
<td>(1.0244, 2.4046)</td>
<td>4.5081</td>
<td>(2.8651, 6.1511)</td>
<td>1.2362</td>
<td>(0.3359, 2.1366)</td>
</tr>
<tr>
<td>$\beta_5$</td>
<td>-0.0538</td>
<td>(-0.1878, 0.0801)</td>
<td>0.0885</td>
<td>(-0.1143, 0.2912)</td>
<td>-0.0122</td>
<td>(-0.1836, 0.1593)</td>
</tr>
<tr>
<td>$\beta_6$</td>
<td>0.0012</td>
<td>(-0.0317, 0.0340)</td>
<td>-0.0846</td>
<td>(-0.1415, -0.0277)</td>
<td>-0.0037</td>
<td>(-0.0455, 0.0380)</td>
</tr>
<tr>
<td>$\phi$</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>-</td>
<td>6.3305</td>
<td>(4.1300, 8.5311)</td>
</tr>
</tbody>
</table>

Table 2 gives the values of the likelihood-based statistics and one can see that the NUL regression model provides the best fit, since it has the lowest values of AIC, AICC and BIC. It is also observed that the NUL was selected approximately 68% of the times as opposed to the UL and beta models.

### Table 2. The likelihood-based statistics of fit.

<table>
<thead>
<tr>
<th>Criteria</th>
<th>NUL</th>
<th>UL</th>
<th>Beta</th>
</tr>
</thead>
<tbody>
<tr>
<td>AIC (%)</td>
<td>-224.9780 (68.26%)</td>
<td>-77.3946 (16.17%)</td>
<td>-159.4460 (15.57%)</td>
</tr>
<tr>
<td>AICC (%)</td>
<td>-223.2549 (68.44%)</td>
<td>-75.6715 (16.23%)</td>
<td>-157.1960 (15.33%)</td>
</tr>
<tr>
<td>BIC (%)</td>
<td>-208.9447 (69.16%)</td>
<td>-61.3614 (16.42%)</td>
<td>-141.1223 (14.42%)</td>
</tr>
</tbody>
</table>

†: % of times out of 10,000 non-parametric Bootstrap runs that the model is selected.

In Figure 3 we present the half-normal plots for the Cox-Snell residuals with simulated envelopes. It is observed for the NUL regression model that all points lie inside the envelopes, suggesting that there is no serious violation of the model assumptions. We can conclude that NUL regression model provides a good fit to these data and therefore can be used for inference purposes.

From the inference results of NUL model (see Table 1) it is observed that the mean of Firm cost is negatively related to the log of the firm’s total asset value (Sizelog). In contrast, the measure of risk (Indcost) has a positive impact on the mean response.

5. Concluding remarks

The ideas in this paper stem from a recent work which proposed a unit-Lindley distribution by transforming a Lindley random variable appropriately. We applied a slightly
different transformation, yet again to a Lindley random variable and introduced a new one-parameter unit-Lindley distribution which is capable of describing data which is limited to the interval (0,1]. Several mathematical properties of the new distribution are presented in detail and parameter estimation is discussed considering the methods of maximum likelihood and moments. We also derived an analytical expression for the bias corrected maximum likelihood estimator. Using a simple re-parametrization of the new distribution we introduced a newer regression model to describe data in a bounded interval. An application of the proposed model to a real dataset from finance shows a better and more parsimonious fit than the classical beta regression model. As such we envisage that the new model attracts the attention of practitioners across all relevant fields of science.

A few related ideas for future work could be to provide a Fisher scoring algorithm for parameter estimation, and to check if this algorithm is equivalent to an iteratively re weighted least squares, as the model belongs to the exponential family.

ACKNOWLEDGMENTS

The authors would like to sincerely thank the editors-in-chief of ChJS and the anonymous referees for their valuable and constructive comments which greatly improved an earlier manuscript.

REFERENCES


INFORMATION FOR AUTHORS

The editorial board of the Chilean Journal of Statistics (ChJS) is seeking papers, which will be refereed. We encourage the authors to submit a PDF file of the manuscript in a free format to Editors of the ChJS (E-mail: chilean.journal.of.statistics@gmail.com). Submitted manuscripts must be written in English and contain the name and affiliation of each author followed by a leading abstract and keywords. The authors must include a “cover letter” presenting their manuscript and mentioning: “We confirm that this manuscript has been read and approved by all named authors. In addition, we declare that the manuscript is original and it is not being published or submitted for publication elsewhere”.

PREPARATION OF ACCEPTED MANUSCRIPTS

Manuscripts accepted in the ChJS must be prepared in Latex using the ChJS format. The Latex template and ChJS class files for preparation of accepted manuscripts are available at http://chjs.mat.utfsm.cl/files/ChJS.zip. Such as its submitted version, manuscripts accepted in the ChJS must be written in English and contain the name and affiliation of each author, followed by a leading abstract and keywords, but now mathematics subject classification (primary and secondary) are required. AMS classification is available at http://www.ams.org/mathscinet/msc/. Sections must be numbered 1, 2, etc., where Section 1 is the introduction part. References must be collected at the end of the manuscript in alphabetical order as in the following examples:


References in the text must be given by the author’s name and year of publication, e.g., Gelfand and Smith (1990). In the case of more than two authors, the citation must be written as Tsay et al. (2000).

COPYRIGHT

Authors who publish their articles in the ChJS automatically transfer their copyright to the Chilean Statistical Society. This enables full copyright protection and wide dissemination of the articles and the journal in any format. The ChJS grants permission to use figures, tables and brief extracts from its collection of articles in scientific and educational works, in which case the source that provides these issues (Chilean Journal of Statistics) must be clearly acknowledged.