CHILEAN JOURNAL OF STATISTICS

Edited by Víctor Leiva and Carolina Marchant

Volume 11 Number 1
April 2020
ISSN: 0718-7912 (print)
ISSN: 0718-7920 (online)

Published by the
Chilean Statistical Society
SOCHE
SOCIEDAD CHILENA DE ESTADÍSTICA

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.

Chilean Journal of Statistics Volume 11 – Number 1 – April 2020

Contents

Chilean Journal of Statistics
Vol. 11, No. 1, April 2020, 53–67

Distribution Theory
Research Paper

A new one-parameter unit-Lindley distribution

Josmar Mazucheli[1], Sudeep R. Bapat[2,*], and André Felipe B. Menezes[1]

[1]Department of Statistics, Universidade Estadual de Maringá, DEs, PR, Brazil
[2]Department 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

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

http://www.soche.cl/chjs

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

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

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

where $0 < x \leq 1$ and $\theta > 0$. Figure (1) shows the PDF of the unit-Lindley distribution for selected values of θ.

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

$$\frac{\mathrm{d}}{\mathrm{d}x} f(x|\theta) = \frac{\theta^2[\theta - 3x]}{[1+\theta]x^5} \exp\left(-\theta\left[\frac{1-x}{x}\right]\right),$$

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

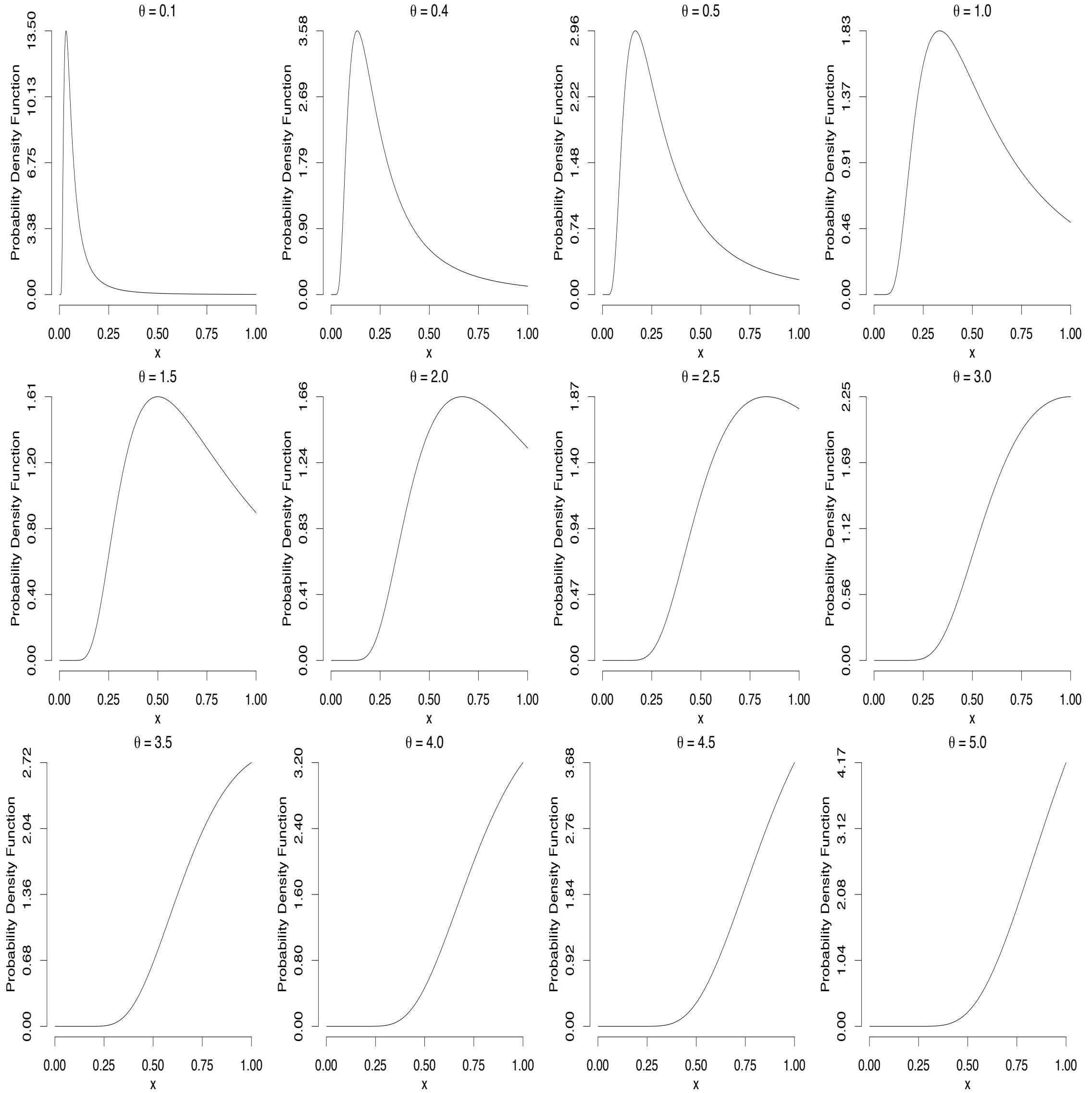

Figure 1. Probability density function of the NUL distribution for selected values of θ.

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

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

This is always < 0 for all x in $(0, 1]$ whenever $\theta > 3/2$. Hence $f(x \mid \theta)$ 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 $\mathrm{d}h(x|\theta)\mathrm{d}x = -[\theta^2(2\theta + 3x)]/[x^3(\theta + x)^2] < 0$ for all $\theta > 0$ the HR function is decreasing in x. Note that $\lim\limits_{x\to 0} h(x|\theta) = \infty$ while $\lim\limits_{x\to 1} h(x \mid \theta) = \theta^2/[1 + \theta]$.

2.4 Moments

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

$$\mu'_k = \mathrm{E}\left(X^k\right) = \int\limits_0^1 k x^{k-1} \left\{1 - \frac{[\theta + x]}{x\,[1+\theta]} \exp\left(-\theta\left[\frac{1-x}{x}\right]\right)\right\} \mathrm{d}x, \quad k = 1, 2, \ldots.$$

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

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

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

where $\mathrm{Ei}(a, z) = \int_1^\infty x^{-a} \exp(-xz)\mathrm{d}x$ is the exponential integral function; see Abramowitz and Stegun (1974).

The k-th incomplete moment about origin is obtained from

$$T_k(t) = \mathrm{E}\left(X^k | X < t\right) = \frac{\theta^2}{[1+\theta]\, F\,(t|\theta)} \int_0^t x^{k-3} \exp\left(-\theta\left[\frac{1-x}{x}\right]\right) \mathrm{d}x, \quad k = 1, 2, \ldots.$$

and for for $k = 1, 2, 3, 4$ we have

$$T_1\,(t) = \tfrac{\theta t}{[\theta+t]}, \qquad T_2\,(t) = \tfrac{\theta^2 \exp\,(\theta) t Ei\left(1, \frac{\theta}{t}\right)}{[\theta+t] \exp\,(\theta[t-1]/t)},$$

$$T_3\,(t) = \tfrac{\theta^2 t \left[t - \theta Ei\left(1, \frac{\theta}{t}\right)\, \exp\,\left(\frac{\theta}{t}\right)\right]}{[\theta+t]}, \qquad T_4\,(t) = \tfrac{\theta^2 t \left[t[t-1] + \theta^2 Ei\left(1, \frac{\theta}{t}\right) \exp\,\left(\frac{\theta}{t}\right)\right]}{2[\theta+t]}.$$

2.5 Mean residual life function

For a nonnegative continuous RV X the mean residual life function is defined as $\mu(t|\theta) = \mathrm{E}(X - t | X > t)$ and is given by

$$\mu(t|\theta) = \frac{1}{S(t|\theta)} \int_t^\infty S(x \mid \theta) \mathrm{d}x.$$

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 θ_1, θ_2 respectively and having PDF's f_X and f_Y. Then the stress-strength reliability measure (Kotz and Pensky, 2003) is given by

$$\begin{aligned} R = P\,(Y < X) &= \int_0^1 f_X(x|\theta_1)\, F_Y(x \mid \theta_2)\, \mathrm{d}x \\ &= \frac{{\theta_1}^2 \left[{\theta_1}^2 \theta_2 + {\theta_1}^2 + \theta_1 + 2\,\theta_1 {\theta_2}^2 + 4\,\theta_1 \theta_2 + 3\,\theta_2 + {\theta_2}^3 + 3\,{\theta_2}^2\right]}{[1+\theta_2]\,[1+\theta_1]\,[\theta_1+\theta_2]^3}. \end{aligned}$$

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

$$Q(p|\theta) = -\frac{\theta}{1 + W[-\exp\,(-(1+\theta))p(1+\theta)]}, \tag{3}$$

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

$$\delta\left(X\right) = \int_{x}^{\infty} \left|X - m\right| f(x \mid \theta)\mathrm{d}x = 2\left[mF\left(m\right) - \int_{0}^{m} xf(x \mid \theta)\mathrm{d}x\right], \tag{4}$$

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

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

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

2.9 Exponential family

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

$$f(x|\theta) = \exp\left(Q(\theta)\,T(x \mid \theta) + D(\theta) + S(x \mid \theta)\right).$$

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

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

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, $T(\mathbf{x}) = \sum_{i=1}^{n}\,[1-x_i]/x_i$ is a complete sufficient estimator for θ 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 θ of a NUL distribution. For the ML estimator of θ 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 $\boldsymbol{x} = (x_1, \ldots, x_n)$, the log-likelihood function of θ can be written as

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

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

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

which gives

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

Next

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

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

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

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

where $\xrightarrow{D}$ denotes convergence in distribution and $\mathrm{V}(\widehat{\theta})$ is just the inverse of the expected Fisher information written as $\mathrm{V}(\widehat{\theta}) = \theta^2\,[1+\theta]^2/n\,[\theta^2 + 4\,\theta + 2]$. It is easy to see that for $\psi = g(\theta) = \mathrm{E}\,(X)$ $\widehat{\psi} = \widehat{\mathrm{E}}\,(X) = 1/[1+\widehat{\theta}]$ and $\mathrm{V}(\widehat{\psi}) = \theta^2/n\,[\theta^2 + 4\,\theta + 2]$. Hence, the asymptotic $100\,(1-\alpha)\%$ confidence intervals (CIs) for θ and ψ are given, respectively, by

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

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 θ, given by $\pi(\theta) \propto \sqrt{\mathrm{I}(\theta)}$. However we will not consider it further in this paper.

Cox and Snell (1968) provided a framework to estimate the bias, to $\mathcal{O}(n^{-1})$ 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 $\mathcal{O}(n^{-2})$. Following Cox and Snell (1968) the analytical expression for bias-correction of an scalar $\widehat{\theta}$, given by

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

where

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

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

and

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

Thus, the bias-corrected ML estimator $\widetilde{\theta}$ is

$$\widetilde{\theta} = \widehat{\theta} - \frac{\widehat{\theta}\left[1+\widehat{\theta}\right]\left[\widehat{\theta}^3+6\widehat{\theta}^2+6\widehat{\theta}+2\right]}{n\left[\widehat{\theta}^2+4\widehat{\theta}+2\right]^2},$$

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

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

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

and the corresponding bias-corrected ML estimator $\widetilde{\mu}$ of μ as

$$\widetilde{\mu} = \widehat{\mu} - \frac{2\widehat{\mu}\,[\widehat{\mu}-1]^2}{n\,[\widehat{\mu}^2-2]^2}.$$

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 $\widehat{\theta}_{\mathrm{MME}}$ of θ is given by

$$\widehat{\theta}_{\mathrm{MME}} = \frac{\bar{X}}{1-\bar{X}} = \left[\frac{1}{\bar{X}} - 1\right]^{-1},$$

which is positively biased, that is, $\mathrm{E}(\widehat{\theta}) - \theta > 0$.

Proof Let $\widehat{\theta}_{\mathrm{MME}} = g(\overline{X})$ 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 $\mathrm{E}(g(\overline{X})) > g(\mathrm{E}(\overline{X}))$. Since $g(\mathrm{E}(\overline{X})) = g\left(\theta/[1+\theta]\right) = \theta$ we get $\mathrm{E}(\widehat{\theta}) > \theta$.

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

$$f(y|\mu) = \frac{\mu^2}{[1-\mu]\, y^3} \exp\left(-\frac{\mu\,[1-y]}{y\,[1-\mu]}\right), \tag{5}$$

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

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

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

$$g(\mu_i) = \mathbf{x}_i^\top\, \boldsymbol{\beta}$$

where $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_{(p-1)})^\top$ is a p-dimensional vector of unknown regression coefficients ($p < n$) and $\mathbf{x}_i = (1, x_{i1}, \ldots, x_{i(p-1)})^\top$ denotes the observations on p known covariates. Note that the variance of Y_i is a function of μ_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 R. Some of the most common link functions are:

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

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

$$\ell(\boldsymbol{\beta}) \propto 2\sum_{i=1}^{n}\log(\mu_i) - \sum_{i=1}^{n}\log(1-\mu_i) - \sum_{i=1}^{n}\frac{\mu_i\,[1-y_i]}{y_i\,[1-\mu_i]}, \tag{6}$$

where $\mu_i = g^{-1}(\mathbf{x}_i^\top\, \boldsymbol{\beta})$.

The ML estimates $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ 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

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

where $\widehat{L}$ 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

$$r_i = -\log\left(1 - \widehat{F}(y_i)\right), \quad i = 1, \ldots, n,$$

where $\widehat{F}$ 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 θ 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, `rlindley` function in 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 θ 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.

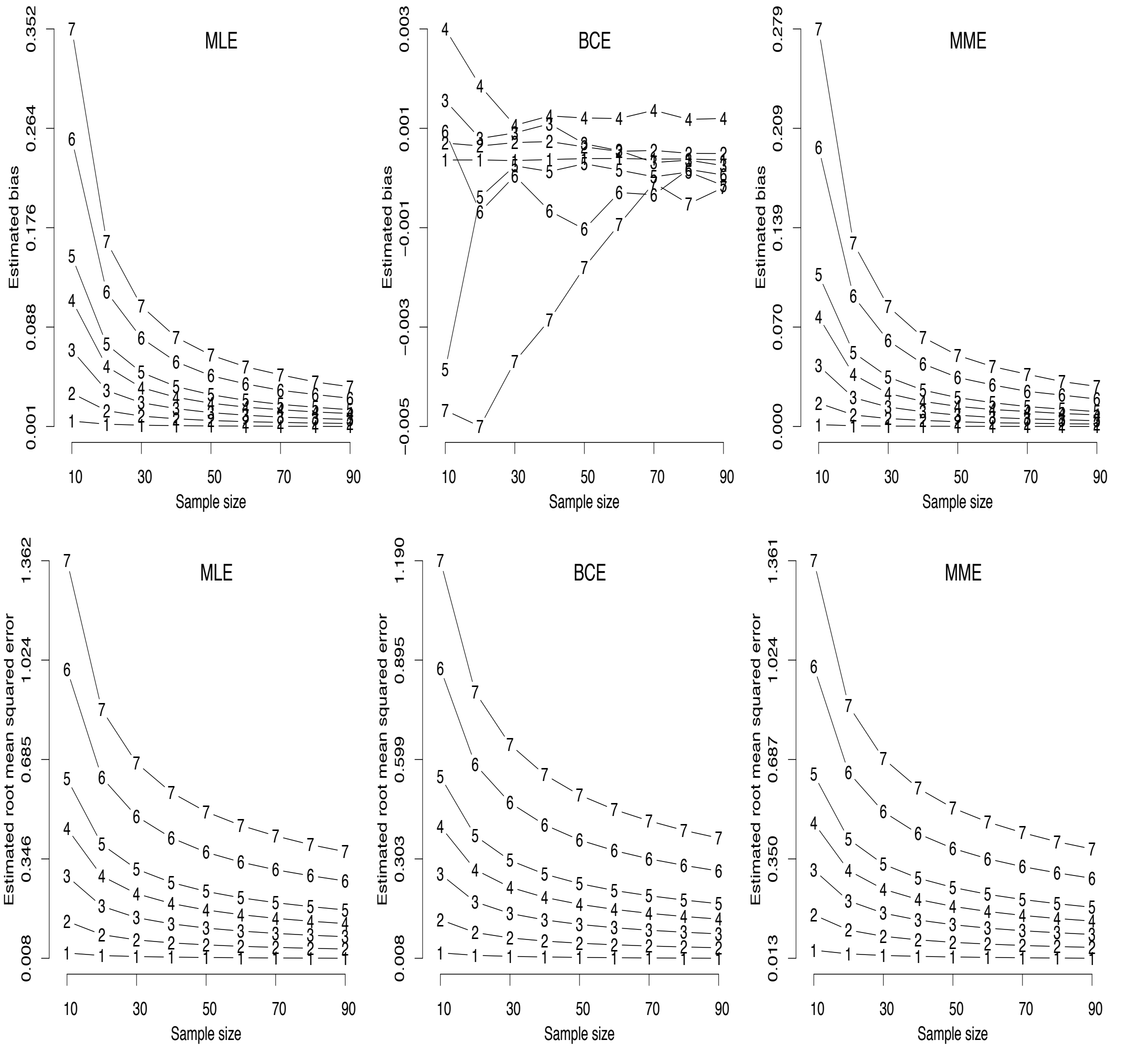

Figure 2. Upper Panel: Estimated bias. Lower Panel: Estimated root mean squared error. (1: $\theta = 0.1$, 2: $\theta = 0.5$, 3: $\theta = 1.0$, 4: $\theta = 1.5$, 5: $\theta = 2.0$, 6: $\theta = 3.0$ and 7: $\theta = 4.0$).

4.2 Empirical illustration

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.

- 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

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

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.

	NUL		UL		Beta	
Parameter	MLE	95% CI	MLE	95% CI	MLE	95% CI
β_0	4.3789	(2.6395, 6.1183)	3.0506	(0.8132, 5.2879)	1.8880	(-0.4096, 4.1855)
β_1	-0.0050	(-0.0273, 0.0173)	-0.0592	(-0.0830, -0.0354)	-0.0121	(-0.0394, 0.0151)
β_2	-0.0112	(-0.3780, 0.3556)	1.8972	(1.2649, 2.5295)	0.1780	(-0.2763, 0.6322)
β_3	-0.8943	(-1.0709, -0.7176)	-0.6606	(-0.8889, -0.4322)	-0.5115	(-0.7524, -0.2705)
β_4	1.7145	(1.0244, 2.4046)	4.5081	(2.8651, 6.1511)	1.2362	(0.3359, 2.1366)
β_5	-0.0538	(-0.1878, 0.0801)	0.0885	(-0.1143, 0.2912)	-0.0122	(-0.1836, 0.1593)
β_6	0.0012	(-0.0317, 0.0340)	-0.0846	(-0.1415, -0.0277)	-0.0037	(-0.0455, 0.0380)
ϕ	-	-	-	-	6.3305	(4.1300, 8.5311)

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.

Criteria	NUL	UL	Beta
AIC (%)†	-224.9780 (68.26%)	-77.3946 (16.17%)	-159.4460 (15.57%)
AICC (%)	-223.2549 (68.44%)	-75.6715 (16.23%)	-157.1960 (15.33%)
BIC (%)	-208.9447 (69.16%)	-61.3614 (16.42%)	-141.1223 (14.42%)

†: % 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

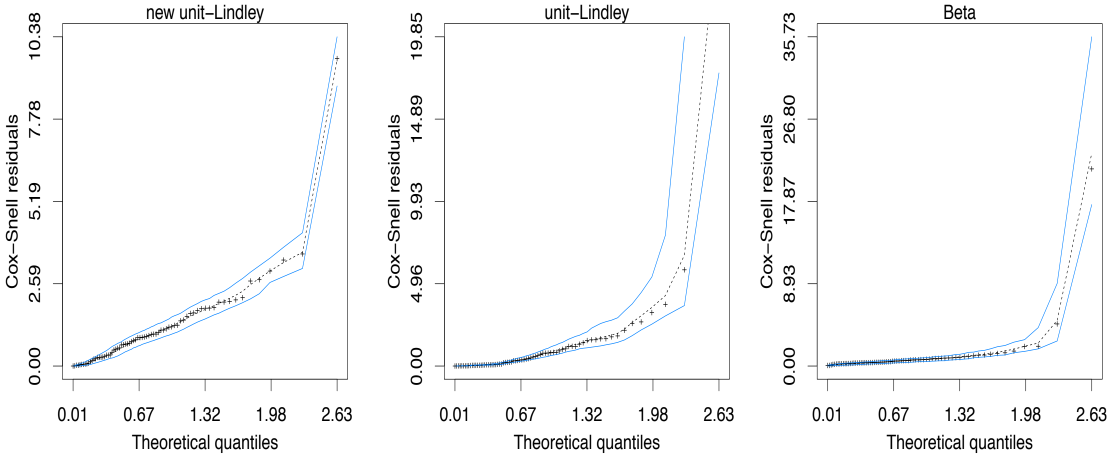

Figure 3. The half-normal plot with simulated envelope for the Cox-Snell residuals.

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

Abramowitz, M. and Stegun, I.A., 1974. Handbook of Mathematical Functions with Formulas, graphs, and Mathematical Tables. National Bureau of Standards Applied Mathematics Series. Dover Pub- lications, Incorporated, New York.

Akaike, H., 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control, 19, 716–723.

Atkinson, A.C., 1981. Two graphical displays for outlying and influential observations in regression. Biometrika, 68, 13–20.

Cavanaugh, J.E., 1997. Unifying the derivations for the Akaike and corrected Akaike information criteria. Statistics and Probability Letters, 33, 201–208.

Cepeda-Cuervo, E., 2001. Variability modeling in generalized linear models. Ph.D. thesis, Mathematics Institute, Universidade Federal do Rio de Janeiro.

Corless, R.M., Gonnet, G.H., Hare, D.E.G., Jeffrey, D.J., and Knuth, D.E., 1996. On the Lambert W function. Advances in Computational Mathematics, 5, 329–359.

Cox, D.R. and Snell, E.J., 1968. A general definition of residuals. Journal of the Royal Statistical Society, Series B, 30, 248–275.

Dobson, A.J., 2001. An Introduction to Generalized Linear Models, Second Edition. Chapman and Hall/CRC.

Ferrari, S. and Cribari-Neto, F., 2004. Beta regression for modeling rates and proportions. Journal of Applied Statistics, 31, 799–815.

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

Ghitany, M.E., Mazucheli, J., Menezes, A.F.B., and Alqallaf, F., 2018. The unit-inverse Gaussian distribution: A new alternative to two-parameter distributions on the unit interval. Communications in Statistics: Theory and Methods, 48, 1–19.

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

Grassia, A., 1977. On a family of distributions with argument between 0 and 1 obtained by trans- formation of the Gamma distribution and derived compound distributions. Australian Journal of Statistics, 19, 108–114.

Held, L. and Sabanés Bové, D., 2014. Applied Statistical Inference-Likelihood and Bayes. Springer, New York.

Jodrá, P., 2010. Computer generation of random variables with Lindley or Poisson-Lindley distribution via the Lambert W function. Mathematics and Computers in Simulation, 81, 851–859.

Johnson, N.L., 1949. Systems of frequency curves generated by methods of translation. Biometrika 36, 149–176.

Johnson, N.L., 1955. Systems of frequency curves derived from the first law of Laplace. TEST, 5, 283–291.

Kotz, S. and Pensky, M., 2003. The Stress-Strength Model and its Generalizations: Theory and Applications. World Scientific, Singapore.

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

Lehmann, E.J. and Casella, G., 1998. Theory of Point Estimation. Springer, Berlin.

Lindley, D.V., 1958. Fiducial distributions and Bayes′theorem. Journal of the Royal Statistical Society B, 20, 102–107.

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

Mazucheli, J., Menezes, A.F.B., and Dey, S., 2018a. The unit-Birnbaum-Saunders distribution with applications. Chilean Journal of Statistics, 1, 47–57.

Mazucheli, J., Menezes, A.F.B., and Ghitany, M.E., 2018b. The unit-Weibull distribution and associated inference. Journal of Applied Probability and Statistics, 13, 1–22.

Nadarajah, S., Bakouch, H.S., Tahmasbi, R., 2011. A generalized Lindley distribution. Sankhya B, 73, 331–359.

Schmit, J.T. and Roth, K., 1990. Cost effectiveness of risk management practices. Journal of Risk and Insurance, 57, 455–470.

Schwarz, G., 1978. Estimating the dimension of a model. The Annals of Statistics, 6, 461–464.

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

Tadikamalla, P.R., 1981. On a family of distributions obtained by the transformation of the Gamma distribution. Journal of Statistical Computation and Simulation, 13, 209–214.

Tadikamalla, P.R. and Johnson, N.L., 1982. Systems of frequency curves generated by transformations of Logistic variables. Biometrika, 69, 461–465.

Topp, C.W. and Leone, F.C., 1955. A family of J-Shaped frequency functions. Journal of the American Statistical Association, 50, 209–219.

Veberić, D., 2012. Lambert W function for applications in physics. Computer Physics Communications, 183, 2622–2628.

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:

Arellano-Valle, R., 1994. Elliptical Distributions: Properties, Inference and Applications in Regression Models. Unpublished Ph.D. Thesis. Department of Statistics, University of São Paulo, Brazil.

Cook, R.D., 1997. Local influence. In Kotz, S., Read, C.B., and Banks, D.L. (Eds.), Encyclopedia of Statistical Sciences, Vol. 1., Wiley, New York, pp. 380-385.

Rukhin, A.L., 2009. Identities for negative moments of quadratic forms in normal variables. Statistics and Probability Letters, 79, 1004-1007.

Stein, M.L., 1999. Statistical Interpolation of Spatial Data: Some Theory for Kriging. Springer, New York.

Tsay, R.S., Peña, D., and Pankratz, A.E., 2000. Outliers in multivariate time series. Biometrika, 87, 789-804.

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.