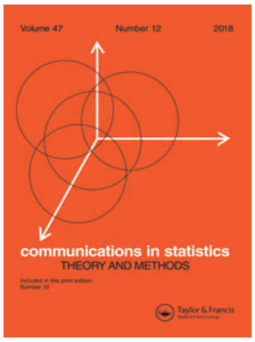

Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Theory and Methods

ISSN: 0361-0926 (Print) 1532-415X (Online) Journal homepage: https://www.tandfonline.com/loi/lsta20

A new skew-bimodal distribution with applications

Altemir da Silva Braga, Gauss M. Cordeiro & Edwin M. M. Ortega

To cite this article: Altemir da Silva Braga, Gauss M. Cordeiro & Edwin M. M. Ortega (2018) A new skew-bimodal distribution with applications, Communications in Statistics - Theory and Methods, 47:12, 2950-2968, DOI: 10.1080/03610926.2017.1343851

To link to this article: https://doi.org/10.1080/03610926.2017.1343851

Published online: 13 Oct 2017.

COMMUNICATIONS IN STATISTICS—THEORY AND METHODS
2018, VOL. 47, NO. 12, 2950–2968
https://doi.org/10.1080/03610926.2017.1343851

A new skew-bimodal distribution with applications

Altemir da Silva Braga[a], Gauss M. Cordeiro[b], and Edwin M. M. Ortega[a]

[a]Departamento de Ciências Exatas, Universidade de São Paulo, Piracicaba, SP, Brazil; [b]Departamento de Estatística, Universidade Federal de Pernambuco, Recife, PE, Brazil

ABSTRACT

The modeling and analysis of experiments is an important aspect of statistical work in a wide variety of scientific and technological fields. We introduce and study the odd log-logistic skew-normal model, which can be interpreted as a generalization of the skew-normal distribution. The new distribution can be used effectively in the analysis of experiments data since it accommodates unimodal, bimodal, symmetric, bimodal and right-skewed, and bimodal and left-skewed density function depending on the parameter values. We illustrate the importance of the new model by means of three real data sets in analysis of experiments.

ARTICLE HISTORY

Received 21 December 2016
Accepted 13 June 2017

KEYWORDS

Log-logistic distribution; maximum-likelihood estimation; mean deviation; skew bimodality; skew distribution.

1. Introduction

The normal distribution is often used in many different areas to model data with symmetric distributions. However, as is widely known, many phenomena cannot always be modeled by the normal distribution, whether by the lack of symmetry or the presence of atypical values. In past decades even when the phenomenon of interest did not present responses for which the assumption of normality was reasonable, the reaction was to try to find some transformation so that the data had at least some semblance of symmetric behavior. The best known of these transformations was the one proposed by Box and Cox (1964).

In recent years, many other distributions have been proposed as alternatives for this type of problem. The elliptical distributions (Fang, Kotz, and Ng 1990) are perhaps the best known among the proposals to preserve the symmetric structure of the Gaussian distribution and allow for heavier or lighter tails than the normal distribution. This optimal property has recently been considered in many studies; see, for example, Cysneiros, Paula, and Galea (2007).

Although this new class of models presents good alternatives to the normal distribution, some of them are not suitable when the distribution of the data or errors of the model are asymmetric (Hill and Dixon 1982). In this context, the skew-elliptical distributions have been used successfully as visible alternatives to model data sets that have asymmetric behavior. The skew-normal (SN) distribution, introduced by Azzalini (1985), is perhaps the pioneer of this new modeling strategy, in which the normal distribution is a special case. Many authors have stressed the importance of using more flexible models in this type of modeling. Among the proposed distributions in this respect, the following stand out: the skew-t (Sahu, Dey,

CONTACT Altemir da Silva Braga altemir@usp.br Departamento de Ciências Exatas. LCE. Avenida Pédua Dias, 11. Piracicaba, São Paulo 13418-900, Brasil.

and Branco 2003), skew-slash (Wang and Genton 2006), skew-slash-t (Punathumparambath 2012), and elliptic-skew (Azzalini and Capitanio 1999), a doubly skewed normal distribution (Arnold, Gómez, and Salinas 2015) and skew-symmetric distributions which include the bimodal ones (Cahoy 2015). However, among these classes, a few include as sub-model the normal distribution.

When the observed data have asymmetry and bimodality, the SN distribution is not appropriate. Therefore, we propose a more general distribution as an alternative, which can model asymmetry and bimodality and include as special cases the normal, SN, and odd log-logistic normal distributions. The new model is called the *odd log-logistic skew-normal* (OLLSN) distribution.

Regression models can be investigated in different forms in analysis of experiments. In this paper, we also propose a regression model based on the OLLSN distribution. The inferential part is carried out using the asymptotic distribution of the maximum-likelihood estimators (MLEs). The performance of these estimators is verified by means of a simulation study

The assessment of the fitted model is an important part of data analysis, particularly in regression models, and residual analysis is a helpful tool to validate the fitted model. For example, examination of the residuals can be used to detect the presence of outlying observations, the absence of components in the systematic part of the model and departures from the error and variance assumptions. We define appropriate residuals to detect influential observations in the OLLSN regression model.

The paper is organized as follows. In Section 2, we define the OLLSN model and obtain its quantile function (qf). In Section 3, we provide structural properties of the OLLSN distribution. In Section 4, we derive explicit expressions for the ordinary and incomplete moments and generating function of the OLLSN distribution. Some inferential tools are discussed in Section 5 and the performance of the MLEs is also investigated by a simulation study. In Section 6, we present the randomized block design model based on the OLLSN distribution. In Section 7, a kind of quantile residual is proposed to assess departures from the underlying OLLSN distribution in linear models to completely randomized design (CRD) and to detect outliers. In Section 8, we prove empirically the potentiality of the new model by means of two real data. Finally, Section 9 ends with some conclusions.

2. The OLLSN model

Statistical distributions are very useful in describing and predicting real world phenomena. Numerous extended distributions have been extensively used over the last decades for modeling data in several areas. Recent developments focus on defining new families that extend well-known distributions and at the same time provide great flexibility in modeling data in practice. Therefore, several classes to generate new distributions by adding one or more parameters have been proposed such as the Marshall and Olkin (1997), beta-G by Eugene, Lee, and Famoye (2002), Kumaraswamy-G (Kw-G) by Cordeiro and de Castro (2011), McDonald normal distribution by Cordeiro et al. (2012), generalized beta-generated distributions by Alexander et al. (2012), Weibull-G by Bourguignon, Rodrigo, and Cordeiro (2014), exponentiated half-logistic by Cordeiro, Morad, and Ortega (2014a), and Lomax generator by Cordeiro et al. (2014b).

Let $g(x)$ and $G(x)$ be the probability density function (pdf) and cumulative distribution function (cdf) of the SN (for $x \in \mathbb{R}$) distribution given by

$$g(z; \lambda) = \frac{2}{\sigma}\phi(z)\,\Phi(\lambda z) \tag{1}$$

and

$$G(z;\lambda)=\Phi(z)-2\,T(z;\lambda)=\Phi_{SN}(z;\lambda) \tag{2}$$

where $z=(x-\mu)/\sigma$, $\phi(\cdot)$ and $\Phi(\cdot)$ are the pdf and cdf of the standard normal distribution, respectively, $\mu\in\mathbb{R}$ is a location parameter, $\sigma>0$ is a dispersion parameter, and $T(z;\lambda)$ is the Owen's function given by (for $z,\lambda\in\mathbb{R}$)

$$T(z;\lambda)=(2\pi)^{-1}\int_0^{\lambda}\frac{\exp\left\{-\frac{1}{2}z^2(1+x^2)\right\}}{1+x^2}dx$$

The parameter λ regulates the skewness and it varies in $\mathbb{R}$. For $\lambda=0$, we have the $N(\mu,\sigma^2)$ density. The function $\Phi_{SN}(\cdot)$ denotes the standard SN cdf.

Based on the transformer odd log-logistic generator (Gleaton and Lynch 2006), we define a new continuous model called the OLLSN distribution, whose cdf follows by integrating the log-logistic density function as

$$F(z;\lambda,\alpha)=\int_0^{\frac{\Phi_{SN}(z;\lambda)}{\bar{\Phi}_{SN}(z;\lambda)}}\frac{\alpha\,t^{\alpha-1}}{(1+t^{\alpha})^2}dt=\frac{\Phi_{SN}^{\alpha}(z;\lambda)}{\Phi_{SN}^{\alpha}(z;\lambda)+[1-\Phi_{SN}(z;\lambda)]^{\alpha}} \tag{3}$$

where $\alpha>0$ is an extra shape parameter.

The pdf corresponding to (3) is given by

$$f(z;\lambda,\alpha)=\frac{2\alpha\phi(z)\,\Phi(\lambda z)\,\Phi_{SN}^{\alpha-1}(z;\lambda)\,[1-\Phi_{SN}(z;\lambda)]^{\alpha-1}}{\sigma\left\{\Phi_{SN}^{\alpha}(z;\lambda)+[1-\Phi_{SN}(z;\lambda)]^{\alpha}\right\}^2} \tag{4}$$

Hereafter, a random variable Z with density function (4) is denoted by $Z\sim$ OLLSN $(0,1,\lambda,\alpha)$. Clearly, the random variable $X=\mu+\sigma Z$ follows the OLLSN$(\mu,\sigma^2,\lambda,\alpha)$ distribution.

The density function (4) allows greater flexibility of its tails and can be widely applied in many areas of engineering and biology. The normal distribution is a special case of (4) when $\lambda=0$ and $\alpha=1$. If $\lambda\neq 0$ and $\alpha=1$, we obtain the SN distribution and, for $\lambda=0$ and $\alpha\neq 1$, it reduces to the OLLN distribution (Braga et al. 2016), respectively.

We can write

$$\alpha=\frac{\log\left\{\Phi_{SN}^{\alpha}(z;\lambda)/[1-\Phi_{SN}(z;\lambda)]^{\alpha}\right\}}{\log\left\{\Phi_{SN}(z;\lambda)/[1-\Phi_{SN}(z;\lambda)]\right\}}$$

and then the parameter α represents the quotient of the log odds ratio for the generated and baseline distributions.

Equation (3) has tractable properties specially for simulations, since the qf of Z has a simple form. Let $F(z;\lambda,\alpha)=u$ and $\Phi_{SN}^{-1}(z;\lambda)$ be the inverse function of $\Phi_{SN}(z;\lambda)$. We have

$$Q(u)=Q_{SN}\,[h(u,\alpha);\lambda] \tag{5}$$

where $u\sim U(0,1)$ and $Q_{SN}[h(u,\alpha);\lambda]=\Phi_{SN}^{-1}[h(u,\alpha);\lambda]$ is the qf of the SN distribution at

$$h(u,\alpha)=u^{\frac{1}{\alpha}}\left[u^{\frac{1}{\alpha}}+(1-u)^{\frac{1}{\alpha}}\right]^{-1}$$

It is not possible to study the behavior of the parameters of the OLLSN distribution by taking derivatives. We can verify skew-bimodality of the new distribution in the plots of Figures 1 and 2 by combining some values of λ, α and μ. Figure 1 reveals different types of bimodality, whereas Figure 2a and 2b reveal the different types of asymmetrical bimodality.

Let $f(z;0,\alpha)$ be the density of the OLLN$(0,1,\alpha)$ distribution. By using $\Phi(-z)=1-\Phi(z)$, we can verify that $f(-z;0,\alpha)=f(z;0,\alpha)$. So, we prove that the OLLSN distribution

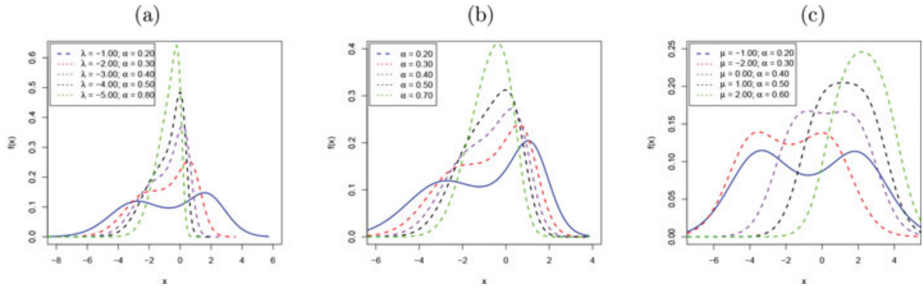

Figure 1. Plots of the OLLSN density function for some parameter values. (a) For different values of the λ and α with $\mu = 0$ and $\sigma = 1$. (b) For different values of the α with $\lambda = -1.8$, $\mu = 0$ and $\sigma = 1$. (c) For different values of the μ and α with $\lambda = 0.3$ and $\sigma = 1$.

is symmetric about 0 for $\lambda = 0$, and then the parameters σ and α characterize the kurtosis and skew-bimodality of this distribution.

3. Useful expansions

First, we define the exponentiated-SN ("Exp-SN" for short) distribution with power parameter $c > 0$ by raising the baseline SN cdf $G(z; \lambda)$ to a power parameter c, say $W_c \sim \text{Exp}^c(SN)$. Then, the cdf and pdf of W_c are given by

$$H_c(z) = \Phi_{SN}\,(z; \lambda)^c \qquad \text{and} \qquad h_c(z) = 2\,c\,\sigma^{-2}\,\phi\,(z)\;\Phi\,(\lambda z)\;\Phi_{SN}\,(z; \lambda)^{c-1}$$

respectively. In a general context, the properties of the exponentiated-G (Exp-G) distributions have been studied by several authors for some baseline G models, see Nadarajah and Kotz (2006).

First, we obtain an expansion for $F(z; \lambda, \alpha)$ using a power series for $\Phi_{SN}^{\alpha}(z)$ (for $\alpha > 0$)

$$\Phi_{SN}^{\alpha}\,(z; \lambda) = \sum_{k=0}^{\infty} a_k\,\Phi_{SN}\,(z; \lambda)^k \tag{6}$$

where

$$a_k = a_k(\alpha) = \sum_{j=k}^{\infty} (-1)^{k+j} \binom{\alpha}{j} \binom{j}{k}$$

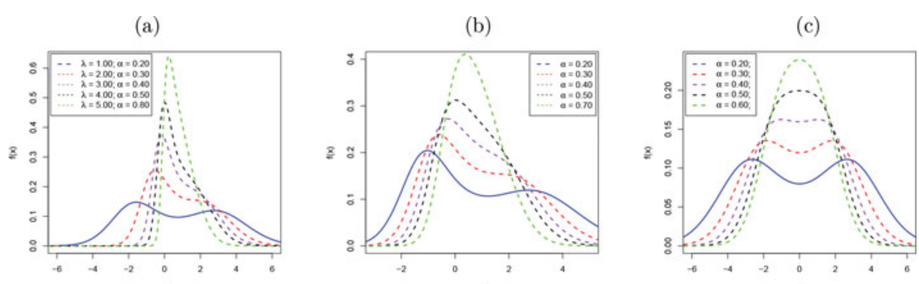

Figure 2. Plots of the OLLSN density function for some parameter values. (a) For different values of the α and λ with $\mu = 0$ and $\sigma = 1$. (b) For different values of the α and $\lambda = 1.8$, $\mu = 0$ and $\sigma = 1$. (c) For different values of the α with $\lambda = 0$, $\mu = 0$ and $\sigma = 1$.

For any real $\alpha > 0$, we consider the generalized binomial expansion

$$[1 - \Phi_{SN}(z; \lambda)]^{\alpha} = \sum_{k=0}^{\infty} (-1)^k \binom{\alpha}{k} \Phi_{SN}(z; \lambda)^k \tag{7}$$

Inserting (6) and (7) in Equation (3), we obtain

$$F(z; \lambda, \alpha) = \frac{\sum_{k=0}^{\infty} a_k \, \Phi_{SN}(z; \lambda)^k}{\sum_{k=0}^{\infty} b_k \, \Phi_{SN}(z; \lambda)^k}$$

where $b_k = a_k + (-1)^k \binom{\alpha}{k}$ for $k \geq 0$.

The ratio of the two power series in the last equation can be expressed as

$$F(z; \lambda, \alpha) = \sum_{k=0}^{\infty} c_k \, \Phi_{SN}(z; \lambda)^k \tag{8}$$

where $c_0 = a_0/b_0$ and the coefficients c_k's (for $k \geq 0$) are determined from the recurrence equation

$$c_k = b_0^{-1} \left(a_k - \sum_{r=1}^{k} b_r \, c_{k-r} \right)$$

The pdf of $Z \sim \text{OLLSN}(0, 1, \lambda, \alpha)$ follows by differentiating (8) as

$$f(z; \lambda, \alpha) = \sum_{k=0}^{\infty} c_{k+1} \, h_{k+1}(z; \lambda) \tag{9}$$

where $h_{k+1}(z; \lambda) = 2(k+1)\sigma^{-2} \phi(z) \, \Phi(\lambda z) \, \Phi_{SN}^{k}(z; \lambda)$ is the Exp-SN density function with power parameter $k + 1$.

Equation (9) reveals that the OLLSN density function is a linear combination of the Exp-SN densities. Thus, some of its structural properties such as the ordinary and incomplete moments and generating function can be determined from well-established properties of the Exp-SN distribution. This equation is the main result of this section.

Let Y_{k+1} have the Exp-SN density function $h_{k+1}(x)$ with power parameter $k + 1$. In this section, we prove that the mathematical properties of Y_{k+1} can be determined from those of the normal distribution.

Further, let W be a $SN(\lambda)$ random variable. Next, we give a power series for the cdf of W. First, for $0 < \lambda < 1$, the cdf of W can be expressed as (Castellares et al. 2012)

$$\Phi_{SN}(z; \lambda) = \sum_{r=0}^{\infty} p_r \, z^r \tag{10}$$

where the coefficients $p_r = p_r(\lambda)$ are functions of λ given by

$$p_0 = \frac{1}{2} - \frac{1}{\pi} \arctan(\lambda)$$

$$p_{2r} = \frac{(1+\lambda^2)^r}{\lambda^{2r-1} \pi \, 2^r} \sum_{k=r}^{\infty} \frac{(-1)^{k+1} \binom{2k+1}{2r} \lambda^{2k} \Gamma\left((k-r)+1\right)}{(2k+1)k!}$$

and

$$p_{2r+1} = \frac{(1+\lambda^2)^{r+1/2}}{\lambda^{2r}\pi\, 2^{r+1/2}} \sum_{k=r}^{\infty} \frac{(-1)^k \binom{2k+1}{2r+1} \lambda^{2k} \Gamma\left((k-r)+1/2\right)}{(2k+1)k!}$$

Second, for $\lambda \geq 1$, the cdf of W can be expressed as

$$\Phi_{SN}(z;\lambda) = 2\Phi(z)\Phi(\lambda z) - \Phi_{SN}(\lambda z;\lambda^{-1}) \tag{11}$$

where $\Phi_{SN}(z;\lambda^{-1})$ can be expanded as in (10).

The standard normal cdf $\Phi(z)$ can be expressed as a power series $\Phi(z) = \sum_{i=0}^{\infty} e_i z^i$, $|z| < \infty$, where $e_0 = (1+\sqrt{2/\pi})^{-1}/2$, $e_{2i+1} = \frac{(-1)^i}{\sqrt{2\pi}(2i+1)2^i i!}$ for $i = 0, 1, 2\ldots$ and $e_{2i} = 0$ for $i = 1, 2, \ldots$

Then, for $\lambda \geq 1$, we can write using (10) and (11)

$$\Phi_{SN}(z;\lambda) = 2\sum_{i,j=0}^{\infty} \lambda^j e_i e_j z^{i+j} - \sum_{r=0}^{\infty} \lambda^r q_r z^r \tag{12}$$

where $q_r = q_r(\lambda) = p_r(\lambda^{-1})$.

We define the set $I_r = \{(i,j), i, j = 0, 1, 2, \ldots; i+j = r\}$ for $r \geq 0$. Then, we can rewrite (12) as

$$\Phi_{SN}(z;\lambda) = \sum_{r=0}^{\infty} w_r z^r \tag{13}$$

where $w_r = 2\lambda^j e_i e_j - \lambda^r q_r$ for any $(i,j) \in I_r$, $r \geq 0$.

Third, if $\lambda < 0$, we can use the relation $\Phi_{SN}(z;-\lambda) = 2\Phi(z) - \Phi_{SN}(z;\lambda)$ and apply the previous expansions to obtain $\Phi_{SN}(z;\lambda) = \sum_{r=0}^{\infty} g_r^{\star} z^r$, where $g_r^{\star} = 2e_r - g_r$ for $r \geq 0$.

Equations (10) and (13) have the same form except for the coefficients. The results below hold for all $\lambda > 0$ and we will work with (10) and define $\Phi_{SN}(z;\lambda) = \sum_{r=0}^{\infty} g_r z^r$, where $g_r = p_r$ when $0 < \lambda < 1$ and $g_r = w_r$ when $\lambda > 1$. If $\lambda < 0$, we will have only to change g_r by $g_r^{\star}$.

We use throughout a result of Gradshteyn and Ryzhik (2000, Section 0.314) for a power series raised to a positive integer k

$$\Phi_{SN}(z;\lambda)^k = \left(\sum_{r=0}^{\infty} g_r z^r\right)^k = \sum_{r=0}^{\infty} d_{k,r} z^r \tag{14}$$

where the coefficients $d_{k,r}$ (for $r = 1, 2, \ldots$) are easily determined from the recurrence equation

$$d_{k,r} = (rg_0)^{-1} \sum_{m=1}^{r} [m(k+1) - r]\, g_m\, d_{k,r-m} \tag{15}$$

and $d_{k,0} = g_0^k$. The coefficient $d_{k,r}$ can be obtained from $d_{k,0}, \ldots, d_{k,r-1}$ and then from the quantities $g_0, \ldots, g_r$.

By using (10) and (14) and the power series for $\Phi(z)$ given before, the density of Y_{k+1} can be rewritten as

$$h_{k+1}(z;\lambda) = 2(k+1)\sigma^{-2}\phi(z) \sum_{i,r=0}^{\infty} e_i \lambda^i d_{k,r} z^{i+r} \tag{16}$$

Combining (9) and (16), the OLLSN density can be reduced to

$$f(z;\lambda,\alpha)=\phi(z)\sum_{i,r=0}^{\infty} m_{i,r}\, z^{i+r} \tag{17}$$

where $m_{i,r}=2\sigma^{-2}\,\lambda^{i}\,e_i\,\sum_{k=0}^{\infty}(k+1)\,c_{k+1}\,d_{k,r}$.

Equation (17) reveals that some mathematical properties of the OLLSN distribution can be determined from those of the standard normal distribution. This equation is the main result of this section.

4. Mathematical properties

In this section, we provide some mathematical properties of $Z\sim \text{OLLSN}(0,1,\lambda,\alpha)$ based on Equation (17). The properties of $X\sim \text{OLLSN}(\mu,\sigma^2,\lambda,\alpha)$ can follow from those of Z by simple linear transformation. The formulae derived throughout the paper can be easily handled in software platforms such as **Maple**, **Mathematica,** and **Matlab** because of their ability to deal with analytic expressions of formidable size and complexity.

4.1. Moments

The nth ordinary moment of Z is given by

$$E(Z^n)=\sum_{i,r=0}^{\infty} m_{i,r}\,\delta_{n+i+r} \tag{18}$$

where $\delta_{n+i+r}=0$ if $n+i+r$ is odd and $\delta_{n+i+r}=\sigma^{n+i+r}\,(n+i+r)!!$ if $n+i+r$ is even, where $p!!=p(p-2)\ldots 3$ denotes the double factorial.

The central moments, cumulants, skewness, and kurtosis of Z are determined from (18) using well-known results.

4.2. Incomplete moments

The nth incomplete moment of Z is given by $m_n(y)=\int_0^y x^n\, f(x)dx$. Based on the linear representation (17) and the monotone convergence theorem, we obtain

$$m_n(y)=\frac{1}{\sqrt{2\pi}}\sum_{i,r=0}^{\infty} m_{i,r}\int_{-\infty}^{y} A(i+r,y) \tag{19}$$

where $A(k,y)=\int_{-\infty}^{y} z^k\,\mathrm{e}^{-z^2/2}\,dz$ (for $k\geq 0$) depends if $y<0$ and $y>0$.

For determining the integral $A(k,y)$ we use some known results on special functions. First,

$$G(k)=\int_0^{\infty} x^k\,\mathrm{e}^{-\frac{x^2}{2}}\,dx=2^{(k-1)/2}\,\Gamma\left(\frac{k+1}{2}\right)$$

Second, we consider the confluent hypergeometric function ${}_1F_1(a;b;z)=\sum_{k=0}^{\infty}\frac{(a)_k}{(b)_k}\frac{z^k}{k!}$, and Whittaker function (Whittaker and Watson 1990, pp. 339–351) defined by

$$M_{k,m}(x)=\mathrm{e}^{-x/2}\,x^{m+1/2}\;{}_1F_1\left(\frac{1}{2}+m-k,1+2m;x\right)$$

All these functions can be easily computed using Mathematica. See http://mathworld.wolfram.com/ConfluentHypergeometricFunctionoftheFirstKind.html.

We can prove that $A(k, y) = (-1)^k G(k) + (-1)^{k+1} H(k, y)$ for $y < 0$ and $A(k, y) = (-1)^k G(k) + H(k, y)$ for $y > 0$, where $H(k, y) = \int_0^y x^k \, e^{-x^2/2} \, dx$ is given by

$$H(k, y) = \frac{2^{k/4+1/4} \, y^{k/2+1/2} \, e^{-y^2/4}}{(k/2 + 1/2)(k + 3)} \, M_{k/4+1/4, k/4+3/4}(y^2/2) + \frac{2^{k/4+1/4} \, y^{k/2-3/2} \, e^{-y^2/4}}{k/2 + 1/2} \, M_{k/4+5/4, k/4+3/4}(y^2/2)$$

Based on these results, we can obtain $m_n(y)$ from (19).

An important application of the first incomplete moment, say $m_1(y)$, is related to the Bonferroni and Lorenz curves of Z defined by $B(\pi) = m_1(q)/(\pi \mu_1')$ and $L(\pi) = m_1(q)/\mu_1'$, respectively, where $q = Q_Z(\pi)$ is obtained from the qf (5) for a given probability π.

A second application of $m_1(y)$ refers to the mean residual life, which represents the expected additional life length for a unit which is alive at age t, and the mean inactivity time, which represents the waiting time elapsed since the failure of an item on condition that this failure had occurred in $(0, t)$, given by $\nu_1(t) = [1 - m_1(t)]/[1 - F(t)] - t$ and $\xi(t) = t - m_1(t)/F(t)$, respectively.

A third one is related to the mean deviations of Z about the mean μ_1' and about the median M given by $\delta_1 = 2[\mu_1' \, F(\mu_1') - M_1(\mu_1')]$ and $\delta_2 = \mu_1' - 2m_1(M)$, where $\mu_1' = E(Z)$ and $M = Q(1/2)$ is the median obtained from (5).

4.3. *Moments based on quantiles*

We provide a further insight of the effects of the parameters α and λ on the skewness and kurtosis of Z by considering these measures based on quantiles. The shortcomings of the classical kurtosis measure are well-known. There are many heavy-tailed distributions for which this measure is infinite, so it becomes uninformative precisely when it needs to be.

The Bowley's skewness is based on quartiles

$$B = \frac{Q(3/4) - 2Q(1/2) + Q(1/4)}{Q(3/4) - Q(1/4)}$$

whereas the Moors' kurtosis is based on octiles

$$M = \frac{Q(7/8) - Q(5/8) - Q(3/8) + Q(1/8)}{Q(6/8) - Q(2/8)}$$

These measures are less sensitive to outliers and they exist even for distributions without moments. In Figure 3 and 4, we plot B and M respectively, for different values of α and λ. These plots indicate how these measures can be sensitive to the two shape parameters α and λ.

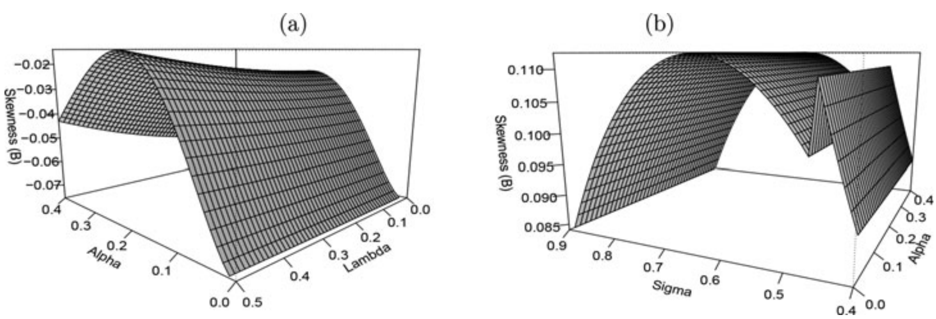

Figure 3. Bowley's skewness for the OLLSN distribution. Plots (a) as functions of $\lambda \in [0, 0.5]$ with $\alpha \in [0, 0.4]$ and (b) as functions of $\lambda \in [0, 0.9]$ with $\alpha \in [0, 0.4]$.

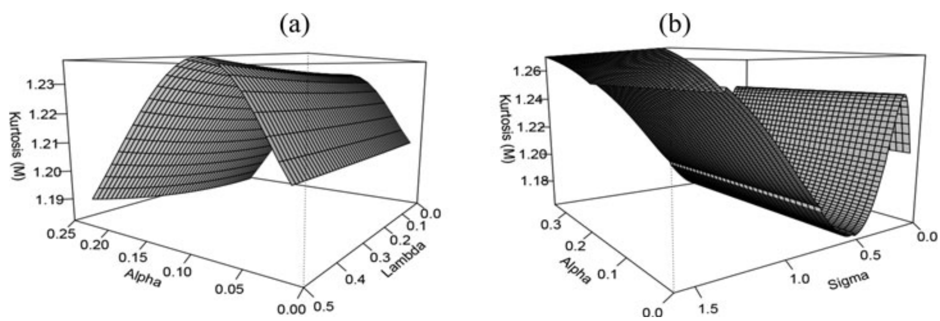

Figure 4. Moors' kurtosis for the OLLSN distribution. Plots (a) as functions of $\lambda \in [0, 0.5]$ with $\alpha \in [0, 0.25]$ and (b) as functions of $\lambda \in [0, 1.6]$ with $\alpha \in [0, 0.35]$.

4.4. Generating function

In this section, we calculate the moment generating function (mgf) of the OLLSN distribution. First, we require the formula, which is a special case of the result by Prudnikov, Brychkov, and Marichev (1986, equation 2.3.15.8),

$$J(n,t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} z^n \exp\left(-tz - \frac{z^2}{2}\right) dz = (-1)^n \frac{\partial^n}{\partial t^n}(e^{t^2/2}) \tag{20}$$

Theorem. *Let Z have the OLLSN distribution. The mgf of Z is given by*

$$M_Z(-t) = \int_{-\infty}^{\infty} e^{-tz} f(z) dz = \sum_{i,r=0}^{\infty} (-1)^{i+r} m_{i,r} J(i+r,t)$$

where $J(i+r,t)$ is given by (20).

Proof. We can write using (17)

$$M_Z(-t) = (2\pi)^{-\frac{1}{2}} \sum_{i,r=0}^{\infty} m_{i,r} \int_{-\infty}^{\infty} z^{i+r} \exp\left(-\frac{z^2}{2} - tz\right) dz.$$

By interchanging the sums and the integral and using (20), the theorem is proved. Clearly, the mgf of $Y = \mu + \sigma Z$ is given by $M_Y(-t) = e^{\mu t} M_Z(-\sigma t)$ □

5. Maximum-likelihood estimation

Let $x_1, \ldots, x_n$ be a random sample of size n from the OLLSN$(\mu, \sigma, \lambda, \alpha)$ distribution. In this section, we determine the MLEs of the model parameters from complete samples only. The log-likelihood function for the vector of parameters $\boldsymbol{\theta} = (\mu, \sigma, \lambda, \alpha)^T$ is given by

$$l(\boldsymbol{\theta}) = n \log\left(\frac{2\alpha}{\sigma}\right) + \sum_{i=1}^{n} \log[\phi(z_i)] + \sum_{i=1}^{n} \log[\Phi(\lambda\, z_i)] + (\alpha - 1) \sum_{i=1}^{n} \log[\Phi_{SN}(z_i; \lambda)]$$
$$+ (\alpha - 1) \sum_{i=1}^{n} \log[1 - \Phi_{SN}(z_i; \lambda)] - 2 \sum_{i=1}^{n} \log\left\{\Phi_{SN}^{\alpha}(z_i; \lambda) + [1 - \Phi(z_i; \lambda)]^{\alpha}\right\} \tag{21}$$

where $z_i = (x_i - \mu)/\sigma$.

The components of the score vector $U(\boldsymbol{\theta})$ can be requested from the authors. The MLE $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ can be determined from the non linear likelihood equations $U_\mu(\boldsymbol{\theta}) = 0, U_\sigma(\boldsymbol{\theta}) = 0, U_\lambda(\boldsymbol{\theta}) = 0$ and $U_\alpha(\boldsymbol{\theta}) = 0$. These equations cannot be solved analytically and statistical software can be utilized to solve them numerically. We can utilize iterative techniques such as a Newton–Raphson type algorithm to obtain $\widehat{\boldsymbol{\theta}}$. We employ the Optim script in the **R** software. We choose as initial values for μ and σ their MLEs $\hat{\mu}$ and $\hat{\sigma}$ under the special normal model.

The 4×4 total observed information matrix is given by $J(\boldsymbol{\theta})$, whose elements can be evaluated numerically. The asymptotic distribution of $(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is $N_4(0, K(\boldsymbol{\theta})^{-1})$ under standard regularity conditions, where $K(\boldsymbol{\theta}) = E[J(\boldsymbol{\theta})]$ is the expected information matrix. Based on the approximate multivariate normal $N_4(0, J(\widehat{\boldsymbol{\theta}})^{-1})$ distribution, where $J(\widehat{\boldsymbol{\theta}})^{-1}$ is the observed information matrix evaluated at $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$, we can construct approximate confidence intervals for the model parameters.

The likelihood ratio (LR) statistic can be used to compare the OLLN distribution with some of its special models. We can compute the maximum values of the unrestricted and restricted log-likelihoods to obtain LR statistics for testing some of its sub-models.

5.1. Simulation study

We perform a simulation study in order to verify some properties of the MLE $\widehat{\boldsymbol{\theta}}$. The data are simulated from the OLLSN$(\mu, \sigma^2, \lambda, \alpha)$ distribution. We use the qf given in Equation (5). We consider the following values for the parameters:

- **Scenario 1** $\mu = 0.00, \quad \sigma = 1.00, \quad \lambda = 1.00, \quad \alpha = 0.20$
- **Scenario 2** $\mu = 0.00, \quad \sigma = 1.00, \quad \lambda = 1.00, \quad \alpha = 0.20$

For each scenario, the observations are generated by taking $n = 100, 150,$ and 300. The results are obtained from 1,000 Monte Carlo simulations performed using the **R** software with the **Optim** function. We determine the average estimates (AEs), biases, and mean-squared errors (MSEs). For each generated sample, the parameters are estimated by maximum likelihood. The results are reported in Table 1.

Table 1 indicates that the MSEs and the biases of the estimates of μ, σ, λ, and α decay toward zero when the sample size increases, as expected under standard asymptotic theory. The AEs of the parameters tend to be closer to the true parameter values when n increases. This fact supports the asymptotic normal distribution as an adequate approximation to the distribution of the MLEs in finite sample. The normal approximation can be oftentimes

Table 1. The AEs and MSEs based on 1,000 simulations of the OLLSN distribution.

Scenario 1								
$n = 100$			$n = 150$			$n = 300$		
$\boldsymbol{\theta}$	AE	MSE	$\boldsymbol{\theta}$	AE	MSE	$\boldsymbol{\theta}$	AE	MSE
μ	−0.0119	0.2033	μ	−0.0085	0.0383	μ	−0.0053	0.0293
σ	1.0680	0.5407	σ	1.0270	0.0408	σ	1.0154	0.0291
λ	1.0165	0.0390	λ	0.9931	0.0146	λ	1.0221	0.0241
α	0.2449	0.7054	α	0.2132	0.6436	α	0.2081	0.6994
Scenario 2								
$n = 20$			$n = 30$			$n = 50$		
$\boldsymbol{\theta}$	AE	MSE	$\boldsymbol{\theta}$	AE	MSE	$\boldsymbol{\theta}$	AE	MSE
μ	−0.0166	0.1254	μ	−0.0194	0.2192	μ	0.0171	0.1313
σ	1.0578	0.2261	σ	1.1682	1.5601	σ	1.0662	0.1897
λ	1.0232	0.0507	λ	1.0089	0.0222	λ	1.0087	0.0315
α	0.2311	0.7279	α	0.2832	0.6763	α	0.2346	0.6854

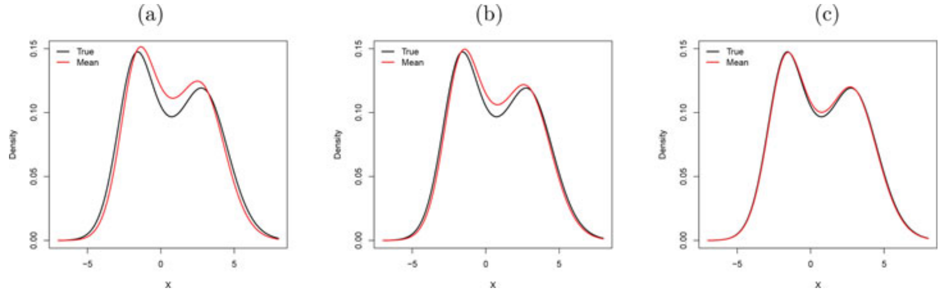

Figure 5. Density functions of the OLLSN distribution at the true parameter values and at the average estimates for $\mu = 0$, $\sigma = 1$, $\lambda = 1$ and $\alpha = 0.20$. (a) $n = 1020$; (b) $n = 150$; (c) $n = 300$.

improved by making bias adjustments to the MLEs. In Figure 5, we display plots of the true densities and the density functions evaluated at the AEs given in Table 1 for selected parameter values and sample sizes. These plots are in agreement with the standard asymptotic theory for the MLEs.

6. Completely randomized design model

In this section, we use the OLLSN model in experimental design. It is often necessary to plan an experiment so that the treatments are not systematically subject to any bias (defect), thus invalidating the conclusions. The proper design of experiments allows simultaneously investigating various treatments in a single experiment without invalidating the model assumptions. One of these assumptions is normality. However, when the data present asymmetry, kurtosis or bimodality, the normal model will not be suitable. For this reason, we consider the OLLSN distribution for the random errors. The CRD model is characterized by not imposing any restriction on randomization of the treatments. In this case, all the treatments have the same chance of occupying any experimental unit or plot. This design is used when all the conditions for the performance of the experiment are controlled and the experimental units are considered homogeneous. The statistical model associated with experiments with one factor is given by

$$Y_{ij} = m + \tau_i + \epsilon_{ij} \tag{22}$$

where Y_{ij} represents the observed value of the group that received treatment i, m is the overall mean effect, τ_i is the effect of treatment i applied to the treated group, and $\epsilon_{ij} \sim \text{OLLSN}(0, \sigma, \lambda, \alpha)$ is the effect of the uncontrolled factors in the experimental group, for $i = 1, \dots, I$ and $j = 1, \dots, J$, where I denotes the number of treatments and J the number of repetitions.

Let $y_{11}, \dots, y_{IJ}$ be a sample of size n from the OLLSN distribution. The log-likelihood function for the vector of parameters $\boldsymbol{\theta} = (m, \boldsymbol{\tau}^T, \sigma, \lambda, \alpha)^T$, where $\boldsymbol{\tau} = (\tau_1, \dots, \tau_I)^T$, is given by

$$l(\boldsymbol{\theta}) = n \log\left(\frac{2\,\alpha}{\sigma}\right) + \sum_{i=1}^{I}\sum_{j=1}^{J} \log[\phi(z_{ij})] + (\alpha - 1)\sum_{i=1}^{I}\sum_{j=1}^{J} \log\{\Phi_{SN}(z_{ij};\lambda)[1 - \Phi_{SN}(z_{ij};\lambda)]\}$$
$$+ \sum_{i=1}^{I}\sum_{j=1}^{J} \log[\Phi(\lambda z_{ij})] - 2\sum_{i=1}^{I}\sum_{j=1}^{J} \log\{\Phi_{SN}^{\alpha}(z_{ij};\lambda) + [1 - \Phi_{SN}(z_{ij};\lambda)]^{\alpha}\} \tag{23}$$

where $z_{ij} = (y_{ij} - m - \tau_i)/\sigma$.

The MLE $\widehat{\boldsymbol{\theta}}$ of the model parameters can be obtained by maximizing the log-likelihood (23) using the Optim function in **R**. In addition, we obtain the parameter estimates and their standard errors by means of the "L-BFGS-B" or "Nelder-Mead" methods. We can provide upon request the data and the **R** script to perform the calculations. The components of the score vector $U(\boldsymbol{\theta})$ can also be requested from the authors.

The asymptotic distribution of $\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}$ is multivariate normal $N_{I+4}(0, K(\boldsymbol{\theta})^{-1})$, where $K(\boldsymbol{\theta})$ is the $(I+4) \times (I+4)$ expected information matrix. The asymptotic covariance matrix $K(\boldsymbol{\theta})^{-1}$ of $\widehat{\boldsymbol{\theta}}$ can be approximated by the inverse of the observed information matrix $-J(\boldsymbol{\theta})$, i.e., we can use $-J(\boldsymbol{\theta})^{-1}$ to obtain an approximation for the large-sample covariance matrix of the MLEs. The elements of this matrix can be determined by simple double differentiation of $l(\boldsymbol{\theta})$ with respect to the model parameters and them evaluated numerically. Then, an asymptotic confidence interval with significance level γ for each parameter θ_r is given by

$$IC_r = \left(\hat{\theta}_r - z_{\gamma/2}\sqrt{-\widehat{J}^{r,r}}, \hat{\theta}_r + z_{\gamma/2}\sqrt{-\widehat{J}^{r,r}} \right)$$

where $-\widehat{J}^{r,r}$ is the rth diagonal of $-J(\widehat{\boldsymbol{\theta}})^{-1}$ estimated at $\hat{\boldsymbol{\theta}}$, for $r = 1, \ldots, I+4$, and $z_{\gamma/2}$ is the quantile $1 - \gamma/2$ of the standard normal distribution

7. Residual analysis

When attempting to adjust a model to a dataset, the validation of the fit should be analyzed by a specific statistic with the purpose of measuring the goodness of fit. Once the model is chosen and fitted, the analysis of the residuals is an efficient way to check the model adequacy. The residuals also aim other purposes, such as to detect the presence of aberrant points (outliers), identify the relevance of an additional factor omitted from the model and verify if there are indications of serious deviance from the distribution considered for the random error. Further, since the residuals are used to identify discrepancies between the fitted model and the dataset, it is convenient to define residuals that take into account the contribution of each observation to the goodness-of-fit measure used.

In summary, the residuals allow measuring the model fit for each observation and enable studying whether the differences between the observed and fitted values are due to chance or to a systematic behavior that can be modeled. The quantile residual proposed by Dunn and Smyth (1996) is as a measure of the discrepancy between y_{ij} and $\hat{\mu}_{ij}$ given by

$$\widehat{qr}_{ij} = \Phi^{-1} \left\{ \frac{\Phi_{SN}^{\hat{\alpha}}\left(\frac{y_{ij}-\hat{m}-\hat{\tau}_i}{\hat{\sigma}}; \hat{\lambda}\right)}{\Phi_{SN}^{\hat{\alpha}}\left(\frac{y_{ij}-\hat{m}-\hat{\tau}_i}{\hat{\sigma}}; \hat{\lambda}\right) + \left[1 - \Phi_{SN}\left(\frac{y_{ij}-\hat{m}-\hat{\tau}_i}{\hat{\sigma}}; \hat{\lambda}\right)\right]^{\hat{\alpha}}} \right\} \tag{24}$$

where $\Phi(\cdot)^{-1}$ is the inverse cumulative standard normal distribution.

Atkinson (1985) suggested the construction of envelopes to enable better interpretation of the probability normal plot of the residuals. These envelopes are simulated confidence bands that contain the residuals, such that if the model is well-fitted, the majority of points will be within these bands and randomly distributed. The construction of the confidence bands follows the steps:

- Fit the proposed model and calculate the residuals $\widehat{qr}_{ij}$'s;
- Simulate k samples of the response variable using the fitted model;

- Fit the model to each sample and calculate the residuals $\widehat{qr}_{ij}$, $i = 1, \ldots, I$ and $j = 1, \ldots, J$;
- Arrange each sample of IJ residuals in rising order to obtain $\widehat{qr}_{(ij)k}$ for $k = 1, \ldots, K$;
- For each ij, obtain the mean, minimum and maximum $\widehat{qr}_{(ij)k}$, namely

$$\widehat{qr}_{(ij)M} = \sum_{k=1}^{K} \frac{\widehat{qr}_{(ij)k}}{K}, \quad \widehat{qr}_{(ij)B} = \min\{\widehat{qr}_{(ij)k} : 1 \leq k \leq K\}$$
$$\text{and} \qquad \widehat{qr}_{(ij)H} = \max\{\widehat{qr}_{(ij)k} : 1 \leq k \leq K\}$$

- Include the means, minimum and maximum together with the values of $\widehat{qr}_{ij}$ against the expected percentiles of the standard normal distribution.

The minimum and maximum values of $\widehat{qr}_{ij}$ form the envelope. If the model under study is correct, the observed values should be within the lower and upper bands and distributed randomly.

8. Applications

In this section, we provide two applications to show the flexibility of the OLLSN distribution in relation to the OLLN, SN, and normal models.

8.1. Application 1: Temperature and production of soybeans data

- The first dataset refers to the temperature variable (°C) obtained from daily readings for the period from January 1 to December 31, 2011 in the city of Piracicaba. The data were supplied by the Department of Biosystems Engineering of the Luiz de Queiroz Higher School of Agriculture (ESALQ) in the University of São Paulo (USP). The temperatures were measured at the Piracicaba Meteorological Station, located at latitude 22°42′30″S, longitude 47°38′30″W, and altitude of 546 meters.
- The second dataset refers to the production of soybeans in the municipality of Lucas do Rio Verde in the period from 1990 to 2012. This town is one of the 15 leading soybean producers in the state of Mato Grosso. The data are the crop yields in kilograms of beans per hectare (Kg/ha) obtained from the Brazilian Institute of Geography and Statistics (IBGE).

In order to compare the distributions, we consider some goodness-of-fit measures including the Akaike information criterion (AIC), consistent Akaike information criterion (CAIC) and Bayesian information criterion (BIC). The MLEs of the parameters μ, σ, λ and α are obtained using the Optim script in the **R** software. These estimates and their standard errors (SEs) are given in Table 2. These results indicate that the OLLSN distribution has the lowest AIC, CAIC and BIC values among all fitted models, and so it could be chosen as the best model in both applications.

In addition to comparing the models, we consider LR statistics and formal tests. First, the OLLSN model includes some sub-models thus allowing their evaluation relative to each other and to a more general model. The values of the LR statistics for testing some sub-models of the OLLSN distribution are given in Table 3. The figures in this table reveal the superiority of the OLLSN model in relation to the others to fit both datasets.

Table 2. MLEs, SEs (in parentheses) and information criteria for temperature and production of soybeans data.

Temperature	μ	σ	λ	α	AIC	CAIC	BIC
OLLSN	24.020 (0.1221)	1.466 (0.1326)	− 2.530 (0.1861)	0.228	1748.3	1748.5	1763.9
	μ	σ	α		AIC	CAIC	BIC
OLLN	21.9071 (0.1221)	0.8915 (0.1326)	0.1861 (0.1861)		1790.4	1790.5	1802.1
	μ	σ	λ				
SN	26.3603 (0.1332)	4.9948 (0.2139)	− 9.7087 (2.4430)		1764.7	1764.8	1776.4
Normal	22.3271 (0.1542)	2.9463 (0.1090)			1828.7	1828.8	1836.5
Soybean	μ	σ	λ	α	AIC	CAIC	BIC
OLLSN	3134.400 (0.002)	160.919 (0.001)	− 2.328 (0.002)	0.155 (0.027)	345.5	350.7	350.0
	μ	σ	α		AIC	CAIC	BIC
OLLN	2823.607 (15.331)	115.850 (14.811)	0.1522 (0.0419)		348.3	351.8	351.7
	μ	σ	λ				
SN	2863.500 (15428.161)	437.520 (522.636)	− 0.00006 (35.993)		351.0	354.5	354.4
Normal	2863.478 (91.231)	437.529 (64.509)			349.0	351.2	351.2

In order to assess if the model is appropriate, the estimated pdfs and cdfs of the fitted distributions are displayed in Figures 6 and 7. Based on these plots, we can conclude that the OLLSN model yields the best fits to both datasets.

8.2. Application 2: Completely randomized design model—Soybean data

The experiment was conducted at Block 3, of the Geraldo Schultz Research Center, located in the municipality of Iracemápolis, São Paulo state, with average altitude of 570 m (longitude 47° 30' 10.81" W and latitude 22° 38' 49.14" S). The climate in the region is classified as Cwa according to the Köppen classification (tropical highlands, with rain mainly in the summer and dry winters). The soil was classified as Dystrophic Red Latosol according to the Brazilian

Table 3. LR tests.

Temperature	Hypotheses	LR statistic	*p*-value
OLLSN vs normal	$H_0: \alpha = 1,\ \lambda = 0$ vs $H_1: H_0$ is false	84.3	< 0.001
OLLSN vs OLLN	$H_0: \alpha \neq 1,\ \lambda = 0$ vs $H_1: H_0$ is false	43.9	< 0.001
OLLSN vs SN	$H_0: \alpha = 1,\ \lambda \neq 0$ vs $H_1: H_0$ is false	18.3	< 0.001
Soybean	Hypotheses	LR statistic	*p*-value
OLLSN vs normal	$H_0: \alpha = 1,\ \lambda = 0$ vs $H_1: H_0$ is false	7.5	0.024
OLLSN vs OLLN	$H_0: \alpha \neq 1,\ \lambda = 0$ vs $H_1: H_0$ is false	4.7	0.006
OLLSN vs SN	$H_0: \alpha = 1,\ \lambda \neq 0$ vs $H_1: H_0$ is false	7.4	0.023

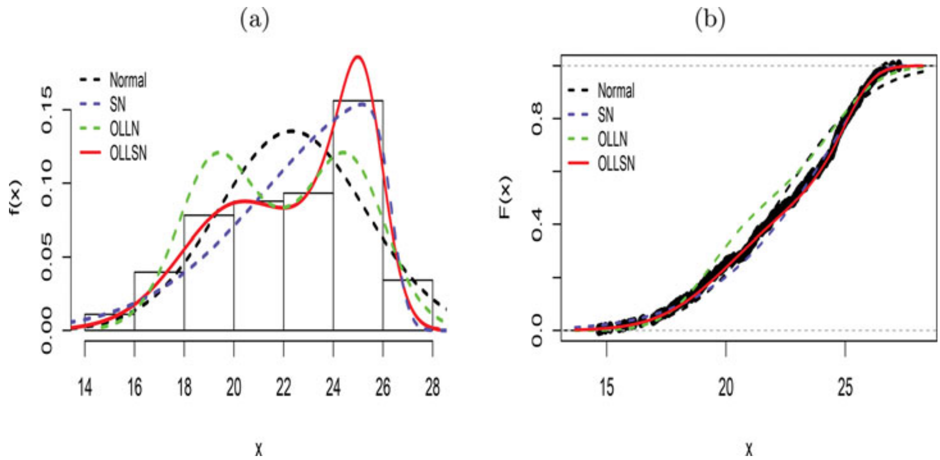

Figure 6. (a) Estimated densities of the OLLSN, OLLN, SN, and normal models for the temperature data. (b) Empirical cdf and estimated cdfs of the OLLSN, OLLN, SN, and normal models for the temperature data.

Soil Classification System (Rhodic Hapludox according to the Soil Taxonomy). The objective of the experiment was to assess the level of nitrogen and protein in soybeans (Glycine max L. Merryl) under the effects of different treatments composed of the elements boron (B) and sulfur (S). The higher the concentration in the soil, the greater the uptake of these nutrients by the plants should be. The experimental design was completely randomized with 4 repetitions and 7 treatments. Each plot was composed of 6 rows with length of 7 m. The useful portion of the plot was composed of two rows with length of 5 m. The study was carried out by the company Produquimica, in Iracemápolis, São Paulo state, Brazil, during the 2014-2015 growing season.

- **Response variable** ($\mathbf{Y}_{ij}$): Concentration of nitrogen and protein the soybeans; $i = 1, \ldots, 7$ and $j = 1, \ldots, 4$,
- **Treatments**:
  - **Treat 1** - Controle;
  - **Treat 2** - Sulfurgran (elemental S);

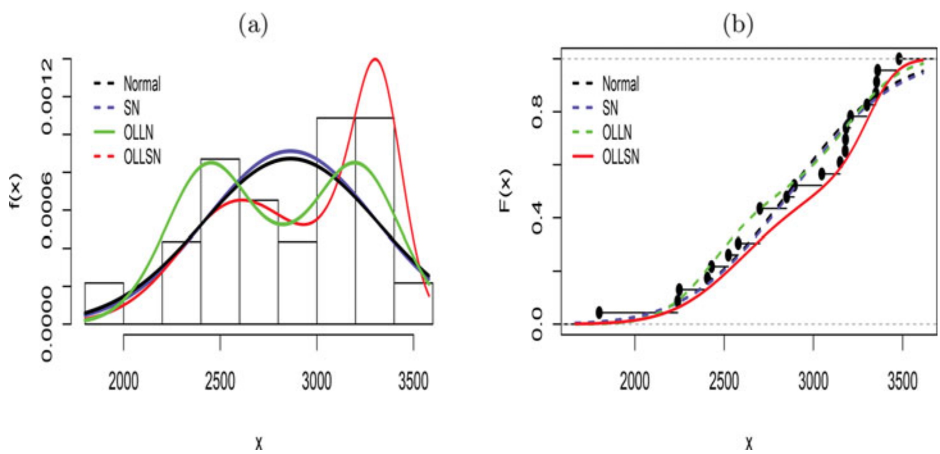

Figure 7. (a) Estimated densities of the OLLSN, OLLN, SN, and normal models for the production of soybeans data. (b) Empirical cdf and estimated cdfs of the OLLSN, OLLN, SN, and normal models for the production of soybeans data.

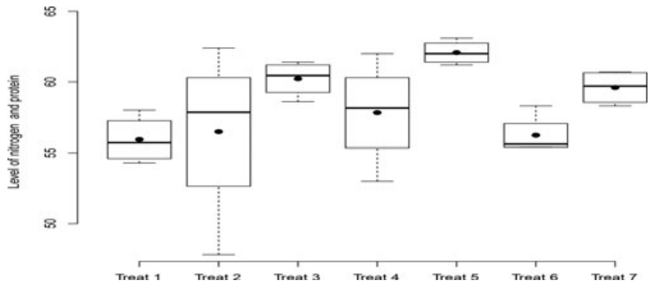

Figure 8. Boxplot for each treatment where • is the average.

- **Treat 3** - Sulfugran + Borosol (elemental S + boric acid);
- **Treat 4** - Sufurgran + ActiveBor (elemental S + sodium octoborate);
- **Treat 5** - Sulfurgran + Ulexite (elemental S + ulexite);
- **Treat 6** - Sulfurgran + Produbor (elemental S + partially acidified ulexite);
- **Treat 7** - Sulfurgran B-MAX (elemental S + ulexite in the same pellet).

According to the general model given in Equation (22), the OLLSN regression model considering the dummy variables is given by

$$Y_{ij} = m + \sum_{k=2}^{I} \beta_k D_k + \epsilon_{ij}, \quad k = 2, \ldots, 7 \tag{25}$$

where m is the effect of the treatment of reference for comparison, $D_2, D_3, D_4, D_5, D_6, D_7$ are the variables *dummies* for the treatment levels, the coefficients $\beta_2 = \tau_2 - \tau_1, \beta_3 = \tau_3 - \tau_1, \ldots,$ $\beta_7 = \tau_7 - \tau_1$ are the effects of the treatment differences, τ_i is the effect of treatment i and ϵ_{ij} $\sim$ OLLSN$(0, \sigma, \lambda, \alpha)$ is the effect of the uncontrolled factors in the experimental group, for $i = 1, \ldots, 7$ and $j = 1, \ldots, 4$.

Figure 8 shows the average differences between the treatments. Then, using model (25), the mathematical expectations for each of the differences are given by

- $E(Y_{ij}) = m$ if $D_2 = \cdots = D_7 = 0$ (represents the effect of the τ_1)
- $E(Y_{ij}) = m + \beta_2$ if $D_2 = 1, D_3 = D_4, \ldots, = D_7 = 0$ $(\beta_2 = \tau_2 - \tau_1)$
- $E(Y_{ij}) = m + \beta_3$ if $D_3 = 1, D_2 = D_4, \ldots, = D_7 = 0$ $(\beta_3 = \tau_3 - \tau_1)$
- $E(Y_{ij}) = m + \beta_4$ if $D_4 = 1, D_2 = D_3, \ldots, = D_7 = 0$ $(\beta_4 = \tau_4 - \tau_1)$
- $E(Y_{ij}) = m + \beta_5$ if $D_5 = 1, D_2 = \cdots = D_7 = 0$ $(\beta_5 = \tau_5 - \tau_1)$
- $E(Y_{ij}) = m + \beta_6$ if $D_6 = 1, D_2 = \cdots = D_7 = 0$ $(\beta_6 = \tau_6 - \tau_1)$
- $E(Y_{ij}) = m + \beta_7$ if $D_7 = 1, D_2 = \cdots = D_6 = 0$ $(\beta_7 = \tau_7 - \tau_1)$

Table 4 lists the MLEs and SEs of the estimates of the parameters for the OLLSN, S and normal regression models fitted to these data using the Optim script in the R software.

By fitting model (22), a restriction on the solution is imposed, i.e., only the effect of $\tau_1 = 0$ is considered. Thus, the estimates of the parameters of the treatments (τ_2, τ_3, τ_4, τ_5, τ_6, and τ_7) represent mean differences in relation to treatment τ_1. This means that the interpretations should be carried out in relation to the treatment that is carried out under the restriction τ_1. Further, for the three fitted models (normal, SN, and OLLSN), the estimates of the parameters are coherent with the results displayed in Figure 8, since the average effects of all the estimated differences present positive values, as can be verified in Table 4.

The values of the AIC, CAIC, and BIC statistics to compare the OLLSN, SN, and normal regression models are listed in Table 4. Note that the OLLSN regression model outperforms the SN and normal models irrespective of the criteria and then the proposed regression model can be used effectively in the analysis of these data. A comparison of the OLLSN regression

Table 4. MLEs, SEs, and information criteria for soybean data.

Normal			SN			OLLSN		
$\boldsymbol{\theta}$	MLE	SE	$\boldsymbol{\theta}$	MLE	SE	$\boldsymbol{\theta}$	MLE	SE
m	55.925	1.272	m	55.904	8.283	m	52.318	5.531
β_2	0.550	1.799	β_2	0.549	1.799	β_2	2.062	1.501
β_3	4.300	1.799	β_3	4.299	1.799	β_3	4.356	1.320
β_4	1.899	1.799	β_4	1.899	1.799	β_4	2.306	1.485
β_5	6.150	1.799	β_5	6.149	1.799	β_5	6.166	1.293
β_6	0.300	1.799	β_6	0.300	1.799	β_6	0.221	1.326
β_7	3.675	1.799	β_7	3.674	1.799	β_7	3.714	1.323
σ	2.544	0.340	σ	2.544	0.346	σ	3.090	1.553
λ			λ	0.010	4.031	λ	0.501	0.196
α			α			α	2.516	1.557
AIC	CAIC	BIC	AIC	CAIC	BIC	AIC	CAIC	BIC
147.764	160.705	158.422	149.763	166.263	161.753	139.210	160.010	152.532

Table 5. LR tests for soybean data.

Models	Hypotheses	LR statistic	p-value
OLLSN vs normal	$H_0 : \lambda = 0$ and $\alpha = 1$ vs $H_1 : H_0$ é false	12.5532	<0.001
OLLSN vs SN	$H_0 : \lambda \neq 0$ and $\alpha = 1$ vs $H_1 : H_0$ is false	12.8172	<0.001

model with some of its sub-models using LR statistics is performed in Table 5. The figures in this table, specially the p-values, indicate that the OLLSN regression model yields better fits to these data than the other sub-models.

In relation to the estimates of the standard errors, the normal distribution presents equal values (1.799), indicating homogeneity within the treatments, something not verified in Figure 8. The results of the SN model are very close, while the OLLSN distribution provides a better model to the variance within the treatments. For example, the treatment (τ_2) is the one that presents the largest variance and also the highest standard error (1.501), while τ_5 presents the smallest variance and also the lowest standard error (1.293). Hence, the OLLSN distribution gives smaller estimates of the standard errors than the normal and SN models.

Then, by using the OLLSN distribution to explain the concentration of nitrogen and protein in soybeans, it can be stated that the Treatment 6 (Sulfurgran + Produbor) presents the smallest average difference (0.221), which is not statistically significant, while Treatment 5 (Sulfurgran + ulexite) presents the largest average difference (6.166), which is statistically significant. In practical terms, the higher the concentration of nitrogen and protein implies better

Table 6. Results of the comparison of the seven treatments for soybean data.

Hypotheses	Estimates	Lwr	Upr	Hypotheses	Estimates	Lwr	Upr
$H_0 : \tau_2 - \tau_1 = 0$	2.048*	0.075	4.020	$H_0 : \tau_4 - \tau_3 = 0$	−2.060*	− 3.964	− 0.156
$H_0 : \tau_3 - \tau_1 = 0$	4.362*	2.627	6.097	$H_0 : \tau_5 - \tau_3 = 0$	1.802*	0.163	3.440
$H_0 : \tau_4 - \tau_1 = 0$	2.309*	0.359	4.260	$H_0 : \tau_6 - \tau_3 = 0$	−4.135*	− 5.822	− 2.449
$H_0 : \tau_5 - \tau_1 = 0$	6.161*	4.462	7.860	$H_0 : \tau_7 - \tau_3 = 0$	-0.652^{ns}	− 2.332	1.028
$H_0 : \tau_6 - \tau_1 = 0$	0.227^{ns}	− 1.515	1.971	$H_0 : \tau_5 - \tau_4 = 0$	3.856*	1.984	5.729
$H_0 : \tau_7 - \tau_1 = 0$	3.709*	1.970	5.448	$H_0 : \tau_6 - \tau_4 = 0$	−2.081*	− 3.991	− 0.170
$H_0 : \tau_3 - \tau_2 = 0$	2.305*	0.378	4.231	$H_0 : \tau_7 - \tau_4 = 0$	$1.401^{\mathbf{ns}}$	− 0.505	3.307
$H_0 : \tau_4 - \tau_2 = 0$	0.246^{ns}	− 1.857	2.349	$H_0 : \tau_6 - \tau_5 = 0$	−5.937*	− 7.585	− 4.289
$H_0 : \tau_5 - \tau_2 = 0$	4.108*	2.211	6.004	$H_0 : \tau_7 - \tau_5 = 0$	−2.453*	− 4.096	− 0.811
$H_0 : \tau_6 - \tau_2 = 0$	-1.829^{ns}	− 3.762	0.104	$H_0 : \tau_7 - \tau_6 = 0$	3.489*	1.800	5.179
$H_0 : \tau_7 - \tau_2 = 0$	1.654*	− 0.275	3.584				

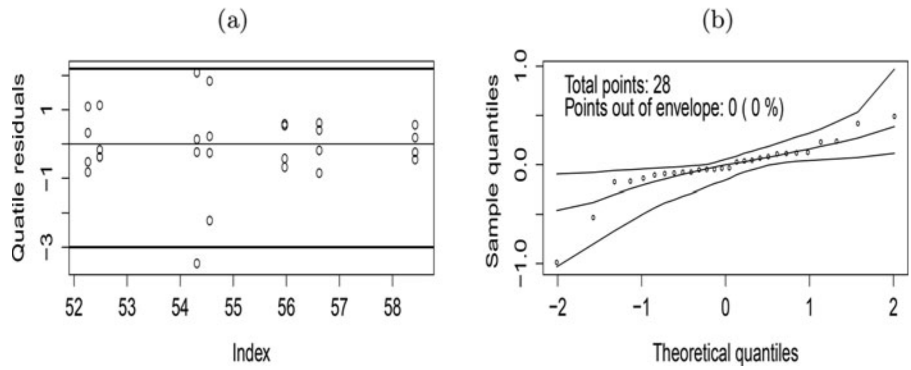

Figure 9. (a) Index plot of the quantile residual for the soybean data. (b) Normal probability plot with envelope for the quantile residual from the fitted OLLSN regression model fitted to soybean data.

quality of soybeans. Table 6 shows all the comparisons between the treatments. These comparisons are carried out by means of confidence intervals at 5% significance. The asterisks (*) indicate the existence of a statistically significant difference between average differences estimated for the treatments, at the 5% level, while (*ns*) indicates no significant difference.

8.3. Residual analysis

To detect possible outlying observations by fitting the OLLSN regression models to the soybean data, Figure 9(a) provides the index plot of $\widehat{qr}_{ij}$. By analyzing the quantile residual plot, one observation appears as a possible outlier, thus indicating that the model is well-fitted.

To detect possible departures from the assumptions of distribution errors for model (25) as well as outlying observations, we display in Figure 9(b) the normal probability plot for the quantile residual with the generated envelope (Atkinson 1985). We note that the plot in Figure 9(b) indicates that the OLLSN regression model for soybean data does not seem unsuitable to fit the data. Also, no observation appears as a possible outlier.

9. Concluding remarks

We introduce a four-parameter continuous model, called the OLLSN distribution, which extends the normal, SN and OLLN (Braga et al. 2016) distributions. The proposed distribution is more versatile than the SN and OLLN distributions, since it can be adjusted to bimodal data. We provide a mathematical treatment of the new distribution including expansions for the density function, moments, generating and qfs. The model parameters are estimated by the method of maximum likelihood. Furthermore, based on the OLLSN distribution, we propose an extended regression model for completely randomized design. This extended regression model is very flexible and can be used in many practical situations. Two applications of the new models to real data are given to prove empirically that they can provide consistently better fits than other special models. Our formulas related to the new distribution and the extended regression model are manageable, and with the use of modern computer resources, the proposed models may be useful alternatives to applied statisticians.

Acknowledgments

We are very grateful to a referee and an associate editor for helpful comments that considerably improved the paper. We gratefully acknowledge financial support from CAPES and CNPq.

References

Alexander, C., G. M. Cordeiro, E. M. M. Ortega, and J. M. Sarabia. 2012. Generalized beta-generated distributions. *Computational Statistics and Data Analysis* 56:1880–89.

Arnold, B. C., H. W. Gómez, and H. S. Salinas. 2015. A doubly skewed normal distribution. *Statistics* 49:842–58.

Atkinson, A. C. 1985. *Plots, transformations and regression: An introduction to graphical methods of diagnostics regression analysis*. 2nd ed. Oxford: Clarendon Press.

Azzalini, A. 1985. A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics* 12:171–8.

Azzalini, A., and A. Capitanio. 1999. Statistical application of the multivariate skew normal distribution. *Journal of the Royal Statistical Society* 61:579–602.

Braga, A. S., G. M. Cordeiro, E. M. M. Ortega, and J. N. da Cruz. 2016. The odd log-logistic normal distribution: Theory and applications in analysis os experiments. *Journal of Statistical Theory and Practice* 10:311–35.

Bourguignon, M., B. S. Rodrigo, and G. M. Cordeiro. 2014. The Weibull-G family of probability distributions. *Journal of Data Science* 12:53–68.

Box, G., and D. R. Cox. 1964. An analysis of transformations. *Journal of the Royal Society* 26:211–52.

Cahoy, D. O. 2015. Some skew-symmetric distributions which include the bimodal ones. *Communications in Statistics—Theory and Methods* 44:554–63.

Castellares, F., G. M. Cordeiro, M. C. Santos, and L. C. Montenegro. 2012. A power series expansion for the skew normal distribution function. *Journal of Applied Probability and Statistics* 7:1–17.

Cordeiro, G. M., R. J. Cintra, L. C. Rêgo, and E. M. M. Ortega. 2012. The McDonald Normal Distribution. *Pakistan Journal of Statistics and Operation Research* 8:301–29.

Cordeiro, G. M., A. Morad, and E. M. M. Ortega. 2014a. The exponentiated half-logistic family of distributions: Properties and applications. *Journal of Probability and Statistics* 1:1–21.

Cordeiro, G. M., E. M. Ortega, B. V. Popovic, and R. R. Pescim. 2014b. The Lomax generator of distributions: Properties, minification process and regression model. *Applied Mathematics and Computation* 247:465–86.

Cordeiro, G. M., and M. de Castro. 2011. A New family of generalized distribution. *Journal of Statistical Computation and Simulation* 81:883–93.

Cysneiros, F. J. A., G. A. Paula, and M. Galea. 2007. Heteroscedastic symmetrical linear models. *Statistics and Probability Letters* 77:1084–90.

Dunn, P. K., and G. K. Smyth. 1996. Randomized quantile residuals. *Journal Computational and Graphical Statistics* 5:236–44.

Eugene, N., C. Lee, and F. Famoye. 2002. Beta-normal distribution and its applications. *Communications in Statistics—Theory and Methods* 31:497–512.

Fang, K., S. Kotz, and K. Ng. 1990. *Symmetric multivariate and related distribution*. London: Chapman & Hall.

Gleaton, J. U., and J. D. Lynch. 2006. Properties of generalized log-logistic families of lifetime distributions. *Journal of Probability and Statistical Science* 4:51–64.

Gradshteyn, I. S., and I. M. Ryzhik. 2000. *Table of integrals, series, and products*. 6th ed. A. Jeffrey and D. Zwillinge, eds. San Diego: Academic Press.

Hill, M., and W. J. Dixon. 1982. Robustness in real life: A study of clinical laboratory data. *Biometrics* 38:377–96.

Marshall, A. W., and I. Olkin. 1997. A new method for adding a parameter to a family of distributions with application to the exponential and Weibull families. *Biometrika* 84:641–52.

Nadarajah, S., and S. Kotz. 2006. The exponentiated type distribution. *Acta Applicandae Mathematicae* 92:97–111.

Prudnikov, A. P., Y. A. Brychkov, and O. I. Marichev. 1986. *Integrals and series: Elementary functions*. New York: Gordon and Breach Science Publishers.

Punathumparambath, B. 2012. The multivariate skew-slash t and skew-slash Cauchy distributions. *Model Assisted Statistics and Applications* 7:33–40.

Sahu, S. K., D. K. Dey, and M. D. Branco. 2003. A new class of multivariate skew distributions with applications to Bayesian regression models. *The Canadian Journal of Statistic* 31:129–50.

Wang, J., and M. G. Genton. 2006. The multivariate skew-slash distribution. *Journal os Statistical Planning and Inference* 136:209–20.

Whittaker, E. T., and G. N. Watson. 1990. *A course in modern analysis*. Cambridge: University Press.