A New Probability Distribution with Applications to Pandemic Data

Isidro Jesús González-Hernández, Rafael Granillo-Macías, Manuel de Jesús Martínez-Téllez, Aidee Domínguez-Islas, Francisca Santana-Robles
Escuela de Superior de Ciudad Sahagún-Ingeniería Industrial
Universidad Autónoma del estado de Hidalgo
Tepeapulco, Hidalgo, México
igonzalez@uaeh.edu.mx, rafaelgm@uaeh.edu.mx, ma448934@uaeh.edu.mx, do338910@uaeh.edu.mx, profe_7739@uaeh.edu.mx

Abstract

In this paper, a new probability distribution called Generalized Continuous Uniform Distribution (GCUD) is presented, which is based on the standard uniform distribution. In this new distribution, a parameter is introduced in the probability density function that is associated with the power of the values of the random variable. The shape properties, the higher order moments, the moment generating function, the failure and survival function, and the model that simulates the GCUD are derived. In addition, this approach allows us to generalize the Generalized Uniform Distribution of Jayakumar and Sankaran (2019), which generates another new distribution that we call $GCUD_{(\mathrm{J-S})}$. To demonstrate the proposed model's potential, we used a set of real data related to the Covid-19 pandemic was obtained, which were adjusted to the new distribution $GCUD_{(\mathrm{J-S})}$. The maximum likelihood method was used to calculate the parameter estimators applying the maxLik package in R language. The results show that the new model is more flexible and appropriate than other models already known in the literature.

Keywords
Probability Density, Maximum Likelihood Estimation, Generalized Uniform Distribution, COVID-19.

1. Introduction
In recent years, various researchers have proposed different generalizations of new distribution functions of continuous random variables to model, more broadly, different behaviors related to survival analysis, for example, the lifespan of a computer or system. Also, these new models have allowed the analysis and study of the failure function (or risk) to describe the reliability of devices subject to use and deterioration. Additionally, these extended distributions provide greater flexibility for modeling various real-life problems (Almuqrin 2023; Mazen Nassar et al. 2018; Torabi et al. 2018).

This research work follows the approach presented in the seminal article by Marshall and Olkin (1997), which was continued by other authors such as Alshangiti et al. (2014); Jayakumar and Sankaran (2016); Jose and Krishna (2011), where they presented the results of the Marshall–Olkin extended uniform distribution, giving different approaches to it to generate a new family of the uniform distribution. The uniform distribution, defined on the interval [0, 1], is closely related to the rest of the distribution functions. From our perspective, we propose a new family of the uniform distribution function, based on a novel presentation of what we call the Powered Mean. Additionally, the work of Jayakumar and Sankaran (2019) was generalized, where the failure and survival functions are relevant.

The document is structured as follows. In the second section, the general conditions of the new Generalized Continuous Uniform Distribution family are defined and discussed, and some interesting properties of the GCUD are also shown. In the third section, the GCUD approach is used to generalize the work of Jayakumar and Sankaran (2019). In the last section, a real data set is used to fit the proposed model and we empirically demonstrate that our model is more appropriate than other competing models. Finally, the conclusions are presented.

1.1 Objectives

- To provide a new probability distribution to analyze COVID-19 data.
- To use the GCUD approach to generalize the Jayakumar and Sankaran (2016) distribution.

2. Method

Based on our research perspective, we present the GCUD as a new family of distribution functions, for a continuous random variable $\boldsymbol{X}$. The respective Probability Density Function (PDF) is defined as follows,

$$f_{X^k}(x) = \begin{cases} \dfrac{x^k}{(b-a)M_k(a,b)}, & a \le x \le b, \ \ k = 0,1,\dots,n, \\ 0, & x < a; x > b. \end{cases} \tag{1}$$

The term $M_k(a,b)$, is defined as an operator called Powered Mean, which is expressed as,

$$M_k(a,b) = \frac{\sum_0^k a^{k-j} b^k}{k+1}. \tag{2}$$

It is easy to show that Equation (1) is a well-defined PDF.

The Cumulative Distribution Function (CDF) corresponding to Equation (1) is given by,

$$F_{X^k}(x) = P(X < x) = \begin{cases} 0, & x < a, \\ \dfrac{x^{k+1} - a^{k+1}}{b^{k+1} - a^{k+1}}, & a \le x < b, \\ 1, & x \ge b, \end{cases} \tag{3}$$

where $k = 0, 1, \dots, n$.

Figure 1 shows the form of $f_{X^k}(x)$ for different values of k. It can be seen in the figure that the FDP deviates to the right each time the value of parameter k increases. The shape properties of the graph are of great importance because it allows professionals and researchers to determine whether any of these distributions fit the data set that is being used to analyze or solve a specific problem.

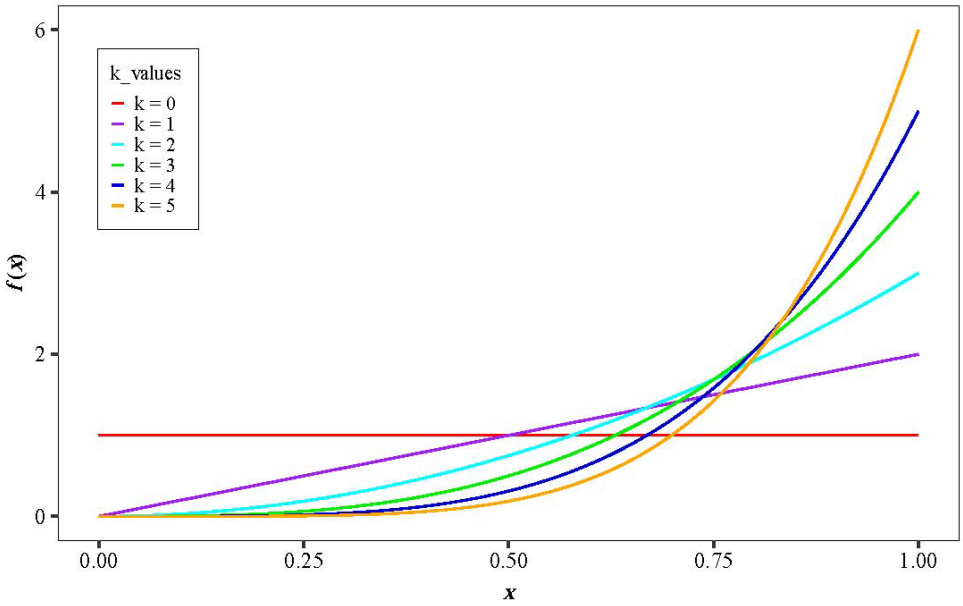

Figure 1. $f_{X^k}(x)$ for $a = 0$, $b = 1$, and $k = 0, 1, 2, 3, 4, 5$.

2.1 General properties of the GCUD

In this section, some general properties of the GCUD are studied to show the flexibility of this new family of distributions, which will allow the development of a generalization of the model proposed by Jayakumar and Sankaran (2016), this generalization will be presented in section three. From Equations (1) and (3), the Failure Function and the Survival Function can be obtained, respectively, as shown below:

$$h_{X^k}(x) = \frac{(k+1)x^k}{1-x^{k+1}}, \quad \text{for } a = 0,\ b = 1 \text{ and } k = 0,1,\dots,n. \tag{4}$$

In the rest of the work, we will use the most usual notation for the Survival Function $S_{X^k}(x) = \bar{F}_{X^k}(x)$. Where:

$$S_{X^k}(x) = 1 - F_{X^k}(x) = 1 - x^{k+1}, \quad \text{for } a = 0, b = 1 \text{ and } k = 0,1,\dots,n. \tag{5}$$

Survival analysis is a subject that has acquired great importance for researchers in various areas, in such a way that both the failure function and the survival function deal with a non-negative continuous or discrete random variable X, which is related to data that analyze or study lifetimes.

On the other hand, a function of great importance in the calculation of moments of higher order is the generating function of moments. We will see that, once again, we can obtain a compact expression of it in terms of the Powered Mean. The PDF moment generating function is given by,

$$\emptyset_x(t) = \sum_{l=0}^{\infty} \frac{t^l}{l!} \frac{M_{k+l}}{M_k}. \tag{6}$$

Next, we show the computation of higher order moments for the GCUD that we are proposing, which allows us to determine the mean, variance, skewness, and kurtosis of the new model. The moment r^{th} of the PDF is given by,

$$\mu_k^r = E_k[x^r] = \frac{M_{k+r}(a,b)}{M_k(a,b)}. \tag{7}$$

For the case $a = 0$ y $b = 1$ is expressed as,

$$\mu_k^r = \frac{k+1}{k+r+1}. \tag{8}$$

For the skewness and kurtosis coefficients, we have:

$$\gamma 3k = E\left[\left(\frac{x-\mu}{\sigma}\right)^3\right] = \frac{1}{\sigma^3}[E(x^3) - 3\mu E(x^2) + 2\mu^3] \tag{9}$$

$$\gamma 4k = \left[\left(\frac{x-\mu}{\sigma}\right)^4\right] = \frac{1}{\sigma^4}[E(x^4) - 4\mu E(x^3) + 6\mu^2 E(x^2) - 3\mu^4] \tag{10}$$

To demonstrate the flexibility of the properties GCUD, Table 1 shows the corresponding calculations for mean (μ_k), variance (σ_k^2), skewness (γ_{3k}) y kurtosis (γ_{4k}) for $a = 0, b = 1$ and different values of k. The data in the table indicates that GCUD has a negative bias for values of $k \geq 1$. Besides, the GCUD It is a leptokurtic family.

Table 1. Mean, variance and the coefficients of skewness and kurtosis for the GCUD

(a, b, k)	Mean	Variance	Skewness	Kurtosis
$(0, 1, 0)$	0.5	0.08333	0	-1.2
$(0, 1, 1)$	0.66667	0.05556	-0.56569	-0.6
$(0, 1, 2)$	0.75	0.03750	-0.86066	0.095
$(0, 1, 3)$	0.8	0.02667	-1.04978	0.696
$(0, 1, 4)$	0.83333	0.01984	-1.18322	1.2
$(0, 1, 5)$	0.85714	0.01531	-1.28300	1.62

3. Generalization of the Jayakumar and Sankaran (2019) distribution using GCUD

After showing the characteristics of the family GCUD, we will see below how this new approach provides greater versatility in the modeling of specific statistical applications and in data analysis, which has allowed us to generalize the results obtained by Jayakumar and Sankaran (2019). From the perspective of these authors, they introduce what they call the Generalized Uniform Distribution, where the parameters (α, θ) are considered.

$$\bar{G}(x, \alpha, \theta) = \frac{\alpha^{\theta}}{1-\alpha^{\theta}}\left[[F(x) + \alpha\bar{F}(x)]^{-\theta} - 1\right], \quad \text{for } \theta > 0, \alpha > 0 \text{ and } x \epsilon R. \tag{11}$$

The corresponding FDA is,

$$G(x, \alpha, \theta) = \frac{1 - \alpha^{\theta}[x(1-\alpha) + \alpha]^{-\theta}}{1-\alpha^{\theta}}, \tag{12}$$

and the PDF is given by,

$$g(x, \alpha, \theta) = \frac{(1-\alpha)\theta\alpha^{\theta}}{(1-\alpha^{\theta})[x(1-\alpha) + \alpha]^{\theta+1}}. \tag{13}$$

From the approach of GCUD, now it is introduced $F_{X^k}(x) = x^{k+1}$, y $S_{X^k}(x) = \bar{F}_{X^k}(x) = 1 - F_{X^k}(x) = 1 - x^{k+1}, 0 < x < 1$, where we obtain a new family of distributions with three parameters (α, θ, k), which will be defined as $GCUD_{(\text{J-S})}$. Where the FDA is,

$$G_{(J-S)}(x, \alpha, \theta, k) = \frac{1 - \alpha^{\theta}[x^k(1-\alpha) + \alpha]^{-\theta}}{1-\alpha^{\theta}}. \tag{14}$$

In turn, the corresponding PDF is given by,

$$g_{(J-S)}(x, \alpha, \theta, k) = \frac{\alpha^{\theta}k\theta(1-\alpha)\alpha^{k-1}}{1 - \alpha^{\theta}[x(1-\alpha) + \alpha]^{\theta+1}}. \tag{15}$$

The survival function is expressed as,

$$\bar{G}_{(J-S)}(x, \alpha, \theta, k) = \frac{\alpha^{\theta}}{1-\alpha^{\theta}}\left[[x^k(1-\alpha) + \alpha]^{-\theta} - 1\right]. \tag{16}$$

For $\theta > 1, 0 < \alpha < 1$ and $k = 0, 1, 2, 3, \ldots, n$.

Figure 2 shows the PDF of the $GCUD_{(\text{J-S})}$, for a value of $k = 3, \theta = 5$ and considering different values of α. As can be seen in the figure, this new distribution provides several major advantages by fitting a variety of different data types

related to lifetime. In addition, it can be viewed as a suitable model for fitting data which may not be properly fitted by other common distributions and can also be used in a variety of problems in different areas such as financial, industrial reliability and survival analysis.

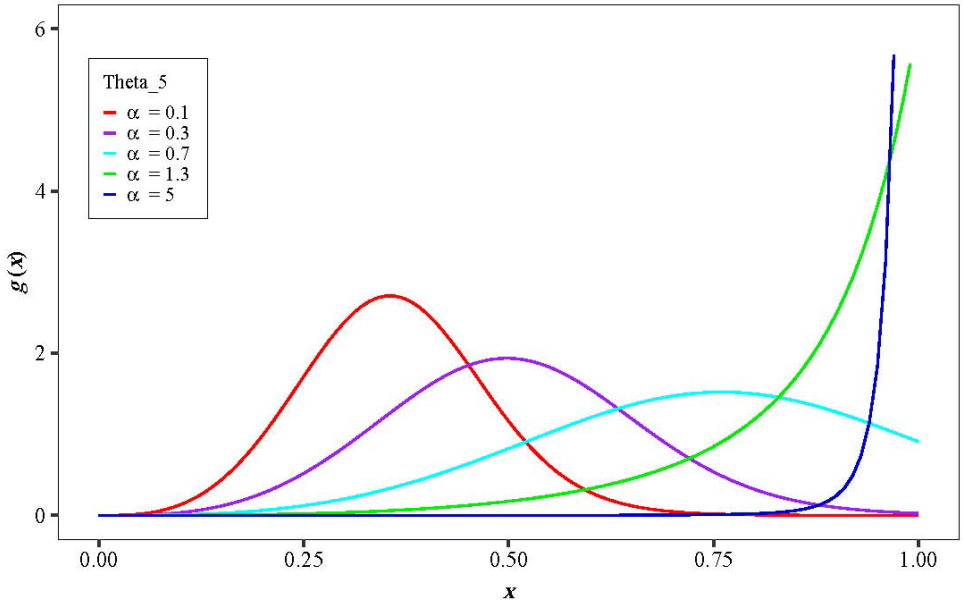

Figure 2. $g_{(\mathrm{J-S})}(x)$ for $k = 3, \theta = 5$, and $\alpha = 0.1, 0.3, 0.7, 1.3, 5$.

The importance of our generalization lies in the possibility of creating a wide range of different failure functions (or hazard function), which can be applied to various survival analyzes or reliability studies in areas such as medicine, engineering, economics, and other disciplines. In our case, considering the $GCUD_{(\mathrm{J-S})}$, where, $\bar{F}_{X^k}(x) = 1 - x^{k+1}$, we obtain a new family of failure functions given in terms of the parameter k,

$$h_{(J-S)}(x;\ \alpha, \theta, k) = \frac{\theta(1\ -\ \alpha)\ kx^{k-1}}{(\alpha + (1-\alpha)x^k)[1 - [\alpha + (1-\alpha)x^k]^\theta]}. \tag{17}$$

Now, consider the estimation of unknown parameters using the maximum verisimilitude method (Eghwerido et al. 2023; Koleoso 2023; Okasha and Kayid, 2016; Torabi et al. 2018). For a sample of the random variable $(x_1, x_2, \ldots, x_n)$, starting from equation (14), in which an additional parameter k was introduced, which has been worked on in the generalizations proposed throughout this work. The corresponding maximum likelihood function is given by,

$$\log L = n \log\left[\frac{\alpha^\theta \theta k(1-\alpha)}{1-\alpha^\theta}\right] - (\theta+1)\sum_{i=1}^{n} \log\left(x_i^k(1-\alpha)+\alpha\right) + (k-1)\sum_{i=1}^{n}\log(x_i) \tag{18}$$

The maxLik package of the R programming language was used to calculate the parameter estimators.

4. Application to real data

In this section we present the practical utility of $GCUD_{(\mathrm{J-S})}$ through the analysis of a real data set to show the potential of the new family of distributions. The data set is related to the global health problem currently being experienced by the pandemic caused by a new strain of coronavirus (COVID-19), which has infected more than 675 million people worldwide and has caused the death of more than 6.87 million people as of February 28, 2023. The data correspond to people who died from COVID-19 and also had diabetes. The time from symptoms to death of the person was analyzed. Data refer to Mexico; This information was obtained from the Ministry of Health of the Government of Mexico (https://www.gob.mx/salud/documentos/datos-abierto-152127?idiom=es). The data corresponds to March 18 (first death from COVID and person with diabetes) to April 17, 2020. Up to that date, a total of 427 data were obtained.

It should be noted that the information from the referred source is available in days (dates). However, it was necessary, for compatible calculation purposes, to divide each data by the longest life time of the infected people (30.6 days), in order to obtain values of the study variable, in the interval $0 < x < 1$, since this is a requirement of our model.

The distribution fit $GCUD_{(\mathrm{J-S})}$ it is compared to the following distributions that are used to analyze lifetimes, such as: Weibull distribution, exponentiated Weibull distribution (EW), New Marshall–Olkin Weibull distribution ($NMOW$) of Cui et al. (2020), and Generalized Marshall-Olkin Exponential distribution ($GMOE$) from Garcia et al. (2020).

Table 2 presents the calculations obtained from the five distributions for the values of the estimators of each distribution, as well as the log-verismilitude ($-log\ L$), Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC). According to Jayakumar and Sankaran (2016), $AIC = -2logL + 2k$ y $BIC = -2logL + klog(n)$. L is the verisimilitude function evaluated on the estimates of maximum verisimilitude, k is the number of parameters and n is the sample (data set). In addition, the Crammer-von Mises (W*), Anderson-Darling (A*), and Kolmogorov-Smirnov (KS) statistics and their corresponding p-value are calculated to test goodness of fit. It can be noted that in Table 2, the KS statistic of the distribution $GCUD_{(\mathrm{J-S})}$ is the smallest compared to the other distributions, and therefore the value corresponding to the p-value is the highest, showing that this new distribution produces the best fit for the Covid-19 data set.

Table 2. Parameter estimates and goodness-of-fit statistics for COVID-19 data

Model	**MLEs**	**-logL**	**AIC**	**BIC**	**W***	**A***	**A-S**	**P - Value**
Weibull	$\hat{\lambda} = 1.9003$ $\hat{\beta} = 0.3517$	179.15	363.43	371.5	0.074	0.508	0.032	0.765
EW	$\hat{\alpha} = 1.3997$ $\hat{\lambda} = 3.3573$ $\hat{\beta} = 1.5951$	180.88	367.76	379.93	0.041	0.302	0.023	0.976
NMOW	$\hat{\theta} = 2.0100$ $\hat{\lambda} = 6.2445$ $\hat{\beta} = 2.1766$	180.47	366.94	379.11	0.048	0.365	0.024	0.958
GMOE	$\hat{\lambda} = 7.6100$ $\hat{\beta} = 8.8881$ $\hat{\delta} = 6.2150$	182.50	371.00	383.17	0.017	0.114	0.024	0.963
$GCUD_{(\mathrm{J-S})}$	$\hat{\theta} = 4.0627$ $\hat{\alpha} = 0.2696$ $\hat{k} = 2.1334$	182.96	371.928	384.00	0.040	0.322	0.019	0.997

Also, it can be seen in Figure 3 that our model presents excellent flexibility, and it can be considered that the model is competitive with other widely accepted and used distributions such as the Weibull distribution or the Weibull Exponential, among others.

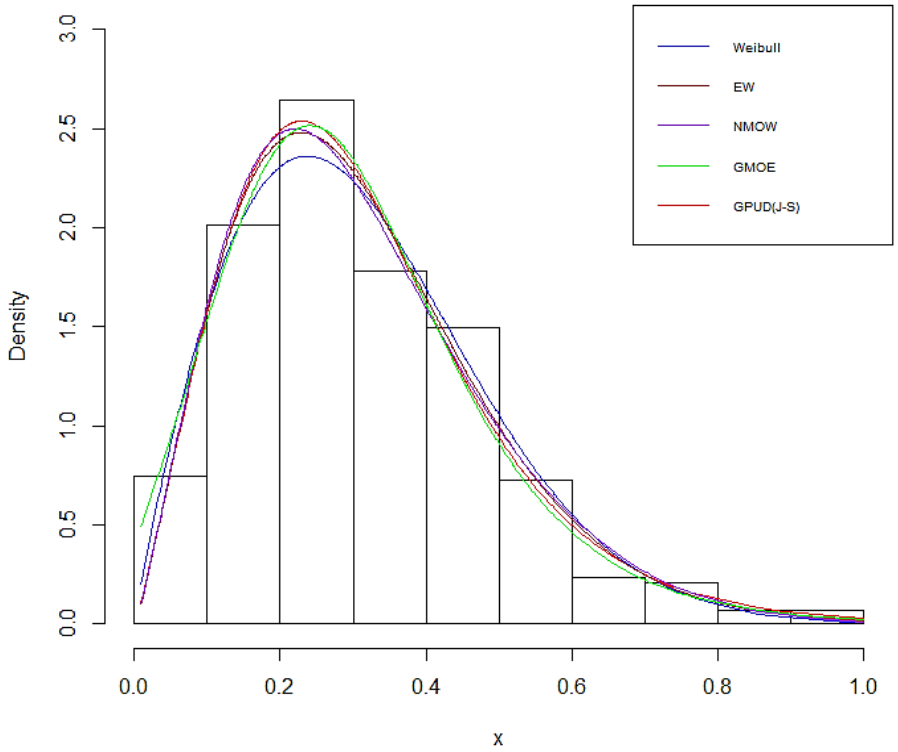

Figure 3. FDP fit for W, EW, NMOW, GMOE and $GCUD_{(J-S)}$ for COVID-19 data.

5. Conclusion

This article introduced a new family of the standard uniform distribution with three parameters, called Generalized Continuous Uniform Distribution (GCUD). The method used in this proposal incorporates a parameter $\boldsymbol{k}$ to the power of the values of the continuous random variable, which favors a greater diversity of the probability density and failure functions. Also, some properties are derived from the new distribution. On the other hand, this approach allowed us to generalize the model presented by Jayakumar and Sankaran (2016), which allowed us to generate a new family of distributions, called $\boldsymbol{GCUD_{(J-S)}}$, which presents excellent flexibility in the cumulative distribution function due to the presence of the parameter $\boldsymbol{k}$. To demonstrate the above, a set of real data related to Covid-19 was adjusted, the maxLik package in R-language was used to find the parameter estimators. The results obtained show that the $\boldsymbol{GCUD_{(J-S)}}$. It can be considered as a valid alternative to known distributions, such as the Weibull, Exponential Weibull, New Marshall-Olkin Weibull distributions, among others, with the advantage that it provides the flexibility of working with the parameter in the values of the variable. random.

References

Almuqrin, M. A., A new flexible distribution with applications to engineering data, *Alexandria Engineering Journal*, vol. 69, pp. 371-382, 2023.

Alshangiti, A. M., Kayid, M. and Alarfaj, B., A new family of Marshall-Olkin extended distributions, *Journal of Computational and Applied Mathematics*, vol. 271, 2014.

Cui, W., Yan, Z. and Peng, X., A New Marshall Olkin Weibull Distribution, *Engineering Letters*, vol. 28, no. 1, 2020.

Eghwerido, J. T., Oguntunde, P. E. and Agu, F. I., The Alpha Power Marshall-Olkin-G Distribution: Properties, and Applications, *Sankhya*, vol. 85, pp. 172–197, 2023.

García, V., Martel-Escobar, M. and Vázquez-Polo, F. J., Generalising Exponential Distributions Using an Extended Marshall-Olkin Procedure, *Symmetry*, vol. 12, no. 3, 2020.

Jayakumar, K. y Sankaran, K. K., On a generalisation of uniform distribution and its properties, *Statistica*, vol. 76, no. 1, 2016.

Jose, K. K. and Krishna, E., Marshall-Olkin extended uniform distribution, *ProbStat Forum*, vol. 4, 2011.

Koleoso, P. O., The properties of odd Lomax-Dagum distribution and its application, Scientific African, vol. 19, e01555, 2023.

Marshall, A. W. and Olkin, I., A new method for adding a parameter to a family of distributions with application to the exponential and Weibull families, *Biometrika*, vol. 84, no. 3, 1997.

Nassar, M., Afify, A. Z., Dey, S. and Kumar, D., A new extension of Weibull distribution: Properties and different methods of estimation, *Journal of Computational and Applied Mathematics*, vol. 336, 2018.

Okasha, H. M. and Kayid, M., A new family of Marshall-Olkin extended generalized linear exponential distribution, *Journal of Computational and Applied Mathematics*, vol. 296, 2016.

Torabi, H., Bagheri, F. L. and Mahmoudi, E., Estimation of parameters for the Marshall–Olkin generalized exponential distribution based on complete data, *Mathematics and Computers in Simulation*, vol. 146, 2018.

Biographies

Isidro J. González-Hernández obtained a Bachelor's Degree in Industrial Engineering and a Master's Degree in Industrial Engineering from the Autonomous University of the State of Hidalgo. He also obtained a PhD in Strategic Planning and Technology Management from the Popular Autonomous University of the State of Puebla. He is currently a research professor in the academic area of industrial engineering at the Autonomous University of the State of Hidalgo. His lines of research are focused on the design, modeling and optimization of the supply chain and logistics; as well as statistical modeling. He is a member of the National System of Researchers of CONACyT in Mexico. In addition, he has developed consulting projects in companies such as: Quaker State, HBPO, RG Refresco Embotelladora de México, Automotriz Serva, Barcel, Alpura and Bimbo.

Rafael Granillo-Macías is a member of the National System of Researchers (SNI) graduated in Industrial Engineering from the Autonomous University of the State of Hidalgo (UAEH), Bachelor of Business Administration from the Tecmilenio University, Master of Science with a specialty in Industrial Engineering from the Instituto Tecnológico y de Estudios Superiores de Monterrey (ITESM) Campus Estado de México and PhD in Logistics and Supply Chain Management from the Popular Autonomous University of the State of Puebla. In the professional field, he has collaborated in the supply and logistics area at Coca Cola FEMSA. He has international certifications such as Certified Supply Chain Professional by the American Production and Inventory Control Society (APICS) and Certificate Associated Project Management Professional by the Project Management Institute (PMI).

Manuel de Jesús Martínez-Téllez is currently studying for a degree in industrial engineering at the Autonomous University of the State of Hidalgo.

Aidee Domínguez-Islas is currently studying for a degree in industrial engineering at the Autonomous University of the State of Hidalgo.

Francisca Santana-Robles is a full-time researcher in the Industrial Engineering Degree, belonging to the Escuela Superior de Ciudad Sahagún-UAEH. She is an Industrial Engineer from the Technological Institute of Pachuca, Master of Science in Industrial Engineering from the Autonomous University of the State of Hidalgo and PhD of Science in Industrial Engineering from the same university. Her areas of interest are: modeling and simulation of the supply chain, agri-food value chain, modeling and optimization of logistics systems and simulation with colored Petri nets.