Model Assisted Statistics and Applications 17 (2022) 199–210
DOI 10.3233/MAS-221354
IOS Press

The Dual-Dagum family of distributions: Properties, regression and applications to COVID-19 data

Elisângela Candeias Biazatti[a,b,*], Gauss Moutinho Cordeiro[a] and Maria do Carmo Soares de Lima[a]
[a]*Department of Statistics, Federal University of Pernambuco, Recife, PE, Brazil*
[b]*Department of Mathematical and Statistics, Federal University of Rondônia, Ji-Paraná, RO, Brazil*

Abstract. A new Dual-Dagum-G (DDa-G) family is defined as a good competitor to the Beta-G and Kumaraswamy-G generators, which are widely applied in several areas. Some of its mathematical properties are addressed. We obtain the maximum likelihood estimates, and some simulations prove the consistency of the estimates. The flexibility of this family is shown through a COVID-19 data set. We propose a new regression based on a special distribution of the DDa-G family, and provide a sensitivity analysis by using data from 1,951 COVID-19 patients collected in Curitiba, Brazil.

Keywords: Beta-G family, COVID-19, Dagum distribution, Kumaraswamy-G family, regression, simulations, sensitivity analysis

1. Introduction

Over the last few decades, many generators have been studied in the distribution theory literature. Two generators that stand out are the Beta-G (B-G) (Eugene et al., 2002) and Kumaraswamy-G (Kw-G) (Cordeiro & de Castro, 2011) classes.

Regarding the B-G family, we can say that, although it contains the incomplete and complete beta functions, its flexibility in terms of adjustment to real data is widespread. Several authors introduced new distributions in this family in different contexts: cancer recurrence (Paranaíba et al., 2011), waiting times before service of 100 bank customers (Abd El-Bar & Ragab, 2015), test on the endurance of deep groove ball bearings (Abu-Zinadah & Bakoban, 2017), survival times of 33 patients suffering from acute Myelogeneous Leukaemia (Mead et al., 2017), among others. More than one-hundred different published distributions in this class can be found to date.

The second family stands out because of the simplicity of its density function, which does not include complicated functions. Further, its suitability for the most diverse types of data sets is widely discussed in the literature. We can cite, for example, the work that originated this family and used data from adult numbers of T. confusum cultured at 29°C (Eugene et al., 2002). Since then, many sub-models in this family have been proposed for various types of data sets: fatigue (Cordeiro et al., 2010), body mass index (Mameli, 2015), component lifetime (Cordeiro et al., 2016), among others.

We know through an analysis of these works that the fits of both classes to real data have a better performance compared to other known classes. We can note that the data sets studied in the aforementioned works are of different types. However, many authors end up repeating the same data sets used in previous works by other authors.

*Corresponding author: Elisângela Candeias Biazatti, Department of Mathematical and Statistics, Federal University of Rondônia, 76900-726, R. Rio Amazonas, 351, Ji-Paraná, RO, Brazil. E-mail: elisangela.biazatti@unir.br.

ISSN 1574-1699/$35.00

In this sense, we define a new class from the Dagum distribution (Dagum, 1975) and use data bases never published before. The data bases in question concern a very current topic: COVID-19. We understand the importance of studies on this pandemic that impacted the world, and then use COVID-19 data from two cities in Brazil.

The remainder of the paper is organized as follows. Section 2 defines the new family. In Section 3, we present some of its generated distributions. The main properties of the new family are reported in Section 4. Estimation including the case of censoring is addressed in Section 5. A simulation study is done in Section 6. In Section 7, we construct the *Log-Dual-Dagum-Weibull* regression, and estimate the parameters. Two applications to real data are reported in Section 8, including a regression application and a sensitivity analysis. Conclusions end the paper in Section 9.

2. The new family

The new generator is defined based on the survival function of the Dagum distribution (Dagum, 1977). Kleiber and Kotz (2003) and Kleiber (2008) analyzed characteristics and properties of this distribution. The Dagum distribution presents forms of the increasing, decreasing, bathtub and inverted bathtub risk function (Domma, 2002). This behavior has aroused the interest of several authors to study it in survival analysis (Domma et al., 2011a, b). In this sense, we propose the *Dual-Dagum-G* (DDa-G) family.

Let W be a Dagum random variable with two positive shape parameters a and b. The cumulative distribution function (cdf) of the DDa-G family (for $x \in \mathbb{R}$) is

$$\begin{aligned} F(x;a,b,\boldsymbol{\theta}) &= \mathrm{P}(\mathrm{X} \leqslant \mathrm{x}) = \mathrm{P}(\mathrm{W} \geqslant -\log[\mathrm{G}(\mathrm{x};\boldsymbol{\theta})]) \\ &= 1 - \{1 + (-\log[G(x;\boldsymbol{\theta})])^{-a}\}^{-b}, \end{aligned} \tag{1}$$

where $G(x) = G(x;\boldsymbol{\theta})$ is the baseline cdf, and $\boldsymbol{\theta}$ is its parameter vector.

Henceforth, Eq. (1) refers to the random variable $X \sim$ DDa-G(a,b).

The probability density function (pdf) of X has the form

$$f(x;a,b,\boldsymbol{\theta}) = \frac{abg(x;\boldsymbol{\theta})(-\log[G(x;\boldsymbol{\theta})])^{-a-1}\{1 + (-\log[G(x;\boldsymbol{\theta})])^{-a}\}^{-b-1}}{G(x;\boldsymbol{\theta})}, \tag{2}$$

where $g(x) = \partial G(x)/\partial x$.

Equations (1) and (2) do not involve complicated mathematical functions, which is an advantage of this family when compared, for example, with the Beta generator.

The hazard rate function (hrf) of X is

$$h(x;a,b,\boldsymbol{\theta}) = \frac{abg(x;\boldsymbol{\theta})(-\log[G(x;\boldsymbol{\theta})])^{-a-1}\{1 + (-\log[G(x;\boldsymbol{\theta})])^{-a}\}^{-1}}{G(x;\boldsymbol{\theta})}.$$

3. Special DDa-G distributions

3.1. Dual-Dagum-Weibull (DDa-W)

The DDa-W density (for $x > 0$) is defined from Eq. (2) and the Weibull cdf with scale $\lambda > 0$ and shape $k > 0$, namely

$$f_{\text{DDa-W}}(x) = \frac{ab\left\{\frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}\exp\left[-(\frac{x}{\lambda})^{k}\right]\right\}\left\{1 + \left[-\log\left(1 - \exp\left[-\left(\frac{x}{\lambda}\right)^{k}\right]\right)\right]^{-a}\right\}^{-b-1}}{\left[-\log\left(1 - \exp\left[-\left(\frac{x}{\lambda}\right)^{k}\right]\right)\right]^{a+1}\left\{1 - \exp\left[-\left(\frac{x}{\lambda}\right)^{k}\right]\right\}}, \tag{3}$$

where all parameters are positive. For $k = 1$, we obtain the Dual-Dagum-Exponential distribution.

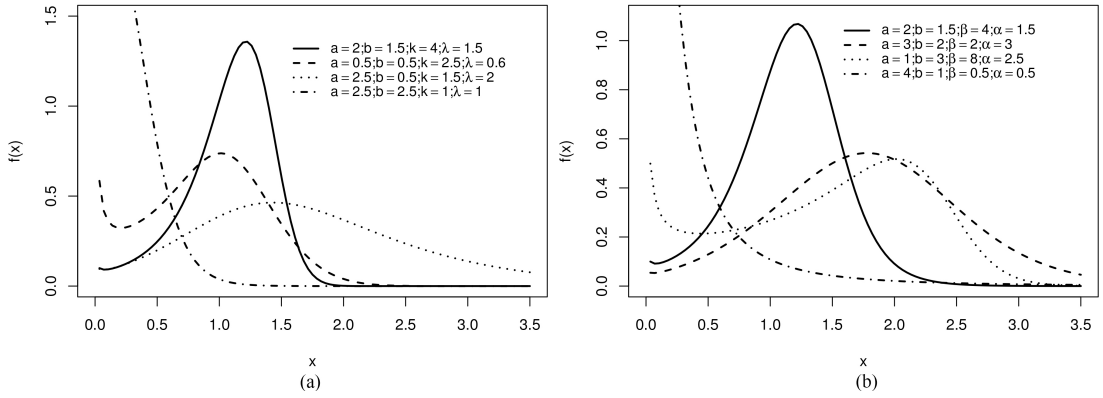

Fig. 1. Shapes of the (a) DDa-W(a, b, k, λ) and (b) DDa-LL(a, b, β, α) densities.

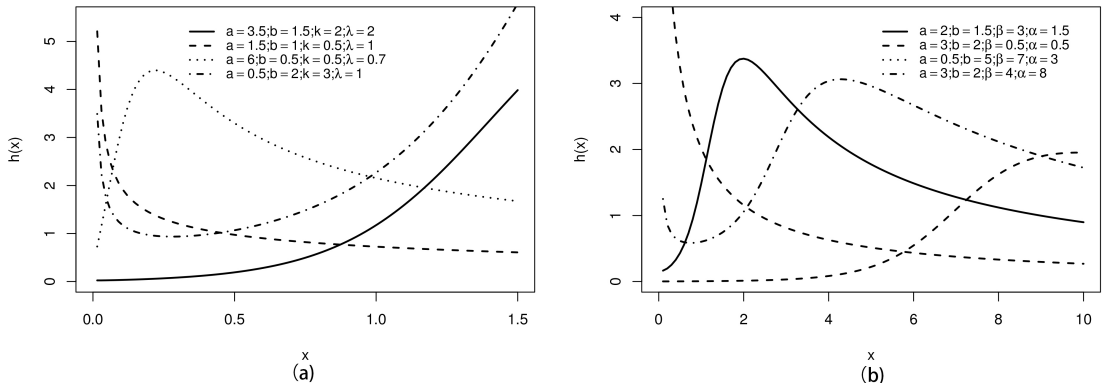

Fig. 2. Shapes of the (a) DDa-W(a, b, k, λ) and (b) DDa-LL(a, b, β, α) hazard rates.

3.2. Dual-Dagum-log-logistic (DDa-LL)

The cdf of the log-logistic (LL) distribution is (for $x, \alpha, \beta > 0$)

$$G(x) = 1 - \left[1 + \left(\frac{x}{\alpha}\right)^{\beta}\right]^{-1}.$$

Inserting this expression and its derivative in Eq. (2) leads to the DDa-LL density (for $x > 0$)

$$f_{\text{DDa-LL}}(x) = \frac{ab\beta x^{\beta-1}[1 + (\frac{x}{\alpha})]^{-2}\{1 + [-\log(1 - [1 + (\frac{x}{\alpha})^{\beta}]^{-1})]^{-a}\}^{-b-1}}{\alpha^{\beta}\{1 - [1 + (\frac{x}{\alpha})^{\beta}]^{-1}\}[-\log(1 - [1 + (\frac{x}{\alpha})^{\beta}]^{-1})]^{a+1}}.$$

Figures 1 and 2 display shapes of the pdf and hrf of the previous generated models, which show their flexibility in fitting data with different shapes. For example, the Weibull pdf presents only decreasing and unimodal shapes, whereas the DDaW pdf has an extra shape: decreasing-increasing-decreasing.

4. Some properties

4.1. Linear representation

For any real p, the power series holds (Ward, 1934)

$$[-\log(1 - t)]^{p} = \sum_{i=0}^{\infty} \rho_i(p) t^{i+p}, \quad |t| < 1, \tag{4}$$

where

$$\rho_0(p) = 1 \quad \text{and} \quad \rho_i(p) = p\psi_{i-1}(i+p-1), \quad i \geqslant 1,$$

and $\psi_0(p) = 1/2, \psi_1(p) = (2+3p)/24$, etc., are the Stirling polynomials.

By using Eq. (4) in Eq. (1) gives

$$(-\log[G(x)])^{-a} = \sum_{i=0}^{\infty} \rho_i(a)[1-G(x)]^{i+a}.$$

By expanding the binomial term,

$$(-\log[G(x)])^{-a} = \sum_{j=0}^{\infty} \delta_j(a) G(x)^j,$$

where $\delta_j(a) = (-1)^{j+1} \sum_{i=0}^{\infty} \rho_i(a) \binom{i+a}{j}$ (for $j \geqslant 0$), and then we obtain

$$F(x;a,b) = 1 - \left[\sum_{j=0}^{\infty} \lambda_j(a) G(x)^j\right]^{-b},$$

where $\lambda_0(a) = 1 + \delta_0(a)$ and $\lambda_j(a) = \delta_j(a)$ for $j \geqslant 1$.

Next, we use a theorem of Henrici (1993) for a power series raised to any real power different from zero

$$F(x;a,b) = 1 - \sum_{j=0}^{\infty} \nu_j(a,b) G(x)^j, \tag{5}$$

where the coefficients are determined recursively from $\nu_0(a,b) = \lambda_0(a)^{-b}$ and (for $j \geqslant 1$)

$$\nu_j(a,b) = \frac{1}{j\lambda_0(a)} \sum_{m=0}^{j-1} [m(b-1) - jb]\lambda_{j-m}(a)\nu_m(a,b).$$

Formulas for other functions may be found in Hairer et al. (1993).

A random variable T_c following the exponentiated-G (exp-G) distribution and power $c > 0$, say $T_c \sim$ exp-G(c), has cdf $\Pi_c(x) = G(x)^c$, and pdf $\pi_c(x) = cG(x)^{c-1}g(x)$. Many exponentiated distributions were given in Table 1 of Tahir and Nadarajah (2015).

By differentiating Eq. (5) and using the concept of exp-G distribution, we can write

$$f(x;a,b) = \sum_{j=0}^{\infty} \omega_{j+1}(a,b)\pi_{j+1}(x), \tag{6}$$

where $\omega_{j+1}(a,b) = -\nu_{j+1}(a,b)$.

Equation (6) is the linear representation for the DDa-G family density in terms of exp-G densities. So, it can provide some mathematical properties for sub-models of the new family from exp-G properties.

4.2. Quantile function

Let $Q_G(u) = G^{-1}(u)$ be the quantile function (qf) of G (for $0 < u < 1$). Inverting $F(x) = u$ in Eq. (1), the qf of X becomes

$$x = Q(u) = F^{-1}(u) = Q_G\left(\mathrm{e}^{-[-1+(1-\mathrm{u})^{-1/\mathrm{b}}]^{-1/\mathrm{a}}}\right). \tag{7}$$

Equation (7) reveals that the qf of the proposed family is a function of the baseline qf.

The skewness and kurtosis of X can be calculated from the first four ordinary moments (see Section 4.3). Alternatively, approximations for the skewness (S) and kurtosis (K) can be obtained from Eq. (7) and the formulae

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

Table 1
Simulation results for the MLEs

Setup	Sample size	Parameter	Average	Bias	MSE
Setup 1	$n = 50$	a	6.08333	0.08333	0.57738
		b	0.50948	0.00948	0.00564
		λ	0.70188	0.00188	0.00185
	$n = 100$	a	6.03727	0.03727	0.26878
		b	0.50437	0.00437	0.00268
		λ	0.70065	0.00065	0.00094
	$n = 150$	a	6.03482	0.03482	0.17810
		b	0.50325	0.00325	0.00173
		λ	0.70019	0.00019	0.00062
Setup 2	$n = 50$	a	5.57856	0.07856	0.47019
		b	0.76402	0.01402	0.01268
		λ	0.60096	0.00096	0.00126
	$n = 100$	a	5.53366	0.03366	0.21645
		b	0.75664	0.00664	0.00602
		λ	0.60048	0.00048	0.00063
	$n = 150$	a	5.53103	0.03103	0.14320
		b	0.75506	0.00505	0.00387
		λ	0.60006	0.00006	0.00042
Setup 3	$n = 50$	a	6.59325	0.09325	0.65217
		b	0.81513	0.01513	0.01443
		λ	0.75093	0.00093	0.00139
	$n = 100$	a	6.53986	0.03986	0.30061
		b	0.80719	0.00719	0.00684
		λ	0.75030	0.00030	0.00069
	$n = 150$	a	6.53657	0.03657	0.19869
		b	0.80534	0.00534	0.00441
		λ	0.75002	0.00002	0.00046

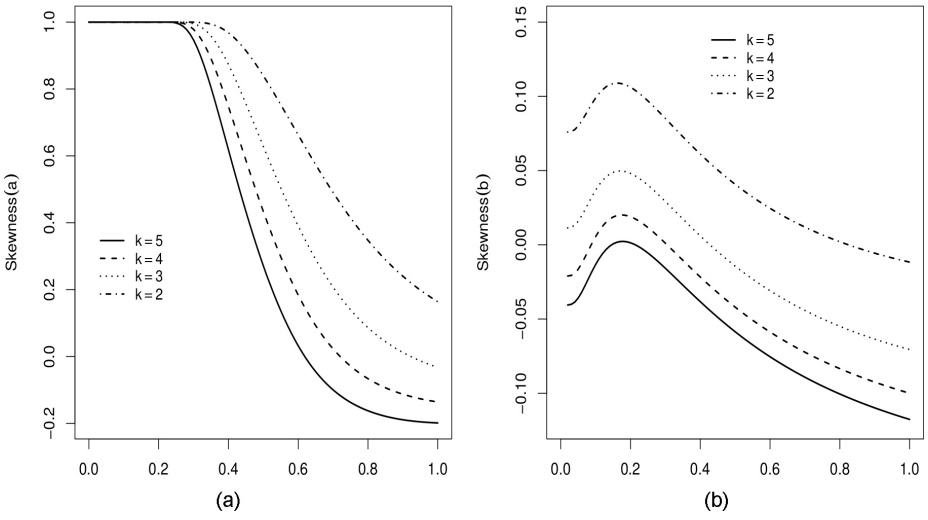

Fig. 3. Bowley's skewness of the DDa-W distribution. (a) as function of a ($b = \lambda = 2$) and (b) as function of b ($a = \lambda = 2$).

and

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

reported by Kenney and Keeping (1961) and Moors (1988), respectively.

Figures 3 and 4 display the skewness and kurtosis of the DDa-W distribution as functions of both a and b.

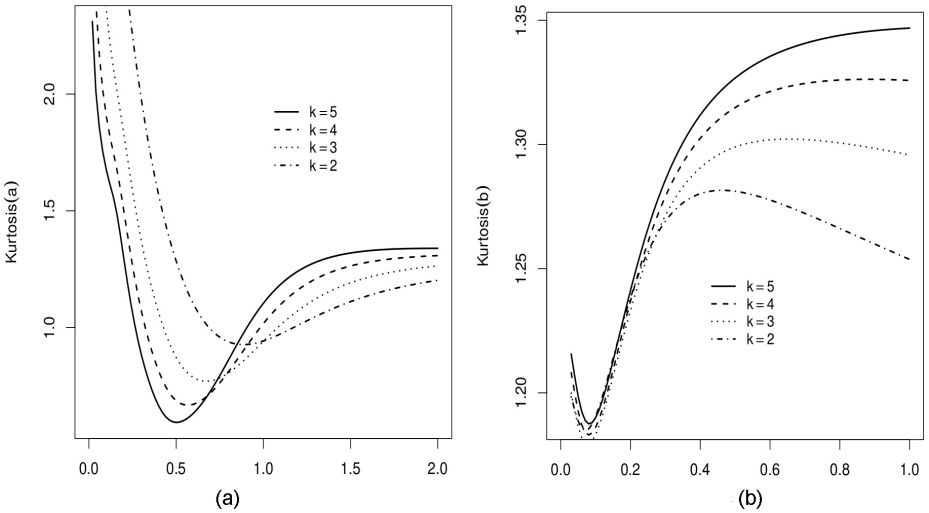

Fig. 4. Moors' kurtosis of the DDa-W distribution. (a) as function of a ($b = \lambda = 2$) and (b) as function of b ($a = \lambda = 2$).

4.3. Moments

From now on, let $T_j \sim$ exp-G$(j+1)$. The nth ordinary moment of X comes from Eq. (6)

$$E(X^n) = \sum_{j=0}^{\infty} \omega_{j+1}(a,b) E(T_{j+1}^n).$$

Moments for several exp-G distributions reported by Nadarajah and Kotz (2006) give $E(X^n)$.

4.4. Generating functions

The generating function (gf) $M(t)$ of X follows from Eq. (6) as

$$M(t) = \sum_{j=0}^{\infty} \omega_{j+1}(a,b) M_{j+1}(t),$$

where $M_{j+1}(t)$ is the gf of T_{j+1}.

5. Estimation

The estimation of the unknown parameters of the DDa-G distribution is performed by the maximum likelihood method. Let $x_1, \ldots, x_n$ be a sample from Eq. (2). Let $\boldsymbol{\theta}$ be the $q \times 1$ parameter vector in $G(\cdot)$. The log-likelihood function $\log L = \log L(a, b, \boldsymbol{\theta})$ has the form

$$\log L = n \log a + n \log b + \sum_{i=1}^{n} \log g(x_i) - (a+1) \sum_{i=1}^{n} \log\{-\log[G(x_i)]\}$$
$$- (b+1) \sum_{i=1}^{n} \log\{1 + (-\log[G(x_i)])^{-a}\} - \sum_{i=1}^{n} \log[G(x_i)].$$

The `R` software has the AdequacyModel computational library (Marinho et al., 2019) as a good alternative for maximizing $\log L(a, b, \boldsymbol{\theta})$, obtain the maximum likelihood estimates (MLEs), their standard errors, and some statistical measures to evaluate the adequacy of a fitted distribution.

6. Simulation study

We adopt the exponential (E) baseline (with the expected value λ) for the simulations to assess the accuracy of the MLEs in the DDa-E(a, b, λ) model. In order to generate observations from the DDa-E distribution, we adopt the inversion method following the steps:

1. Generate $U \sim U(n, 0, 1)$;
2. Return $X = F^{-1}(U)$, where $X \sim$ DDa-E(a, b, λ)

We consider 2,000 Monte Carlo replications and the BFGS algorithm in the `R` software for maximizing the log-likelihood, obtain the MLEs and their averages, biases and mean square errors (MSEs). The simulation process is carried out as below:

1. Simulate DDa-E observations for fixed $n \in \{50, 100, 150\}$ by means of the previous scheme.
2. Three scenarios considered are: $a = 6$, $b = 0.5$ and $\lambda = 0.7$ (Setup 1); $a = 5.5$, $b = 0.75$ and $\lambda = 0.6$ (Setup 2); and $a = 6.5$, $b = 0.8$ and $\lambda = 0.75$ (Setup 3). The values of the true parameters are chosen arbitrary.
3. We calculate the MLEs from each generated data set, and obtain the averages, biases and MSEs.

Table 1 reports these findings. The average estimates converge to the true parameter values and the biases decrease when n increases (for all scenarios).

7. Log-Dual-Dagum-Weibull (LDDa-W) regression

If X follows density Eq. (3), then $Y = \log(X)$ has the LDDa-W distribution, which reparameterized in terms of $k = \sigma^{-1}$ and $\lambda = \mathrm{e}^{\mu}$, has the form (for $y \in \mathbb{R}$)

$$f_Y(y) = \frac{ab\exp[(\frac{y-\mu}{\sigma}) - \exp(\frac{y-\mu}{\sigma})]\{1 + [-\log(1 - \exp[-\exp(\frac{y-\mu}{\sigma})])]^{-a}\}^{-b-1}}{\sigma[-\log(1 - \exp[-\exp(\frac{y-\mu}{\sigma})])]^{a+1}\{1 - \exp[-\exp(\frac{y-\mu}{\sigma})]\}}, \tag{8}$$

where $\mu \in \mathbb{R}$, $\sigma > 0$, $a > 0$ and $b > 0$.

The survival function corresponding to Eq. (8) is

$$S(y) = \left\{1 + \left[-\log\left(1 - \exp\left[-\exp\left(\frac{y-\mu}{\sigma}\right)\right]\right)\right]^{-a}\right\}^{-b}.$$

The pdf of the standardized random variable $Z = (Y - \mu)/\sigma$ (for $z \in \mathbb{R}$) is

$$\pi(z; a, b) = \frac{ab\exp[(z) - \exp(z)]\{1 + [-\log(1 - \exp[-\exp(z)])]^{-a}\}^{-b-1}}{[-\log(1 - \exp[-\exp(z)])]^{a+1}\{1 - \exp[-\exp(z)]\}}. \tag{9}$$

The lifetimes x_i are affected by known explanatory variables $\mathbf{v}_i = (v_{i1}, \ldots, v_{ip})^\top$ in many applications. Consider a sample $(y_1, \mathbf{v}_1), \ldots, (y_n, \mathbf{v}_n)$ of independent observations, where $y_i = \min\{\log(x_i), \log(c_i)\}$, and the log-lifetime $\log(x_i)$ and the log-censoring $\log(c_i)$ are assumed independent (under non-informative censoring).

The LDDa-W regression for the response variable y_i is defined by

$$y_i = \mathbf{v}_i^\top \gamma + \sigma z_i, i = 1, \ldots, n, \tag{10}$$

where z_i is the random error with density Eq. (9), $\gamma = (\gamma_1, \ldots, \gamma_p)^\top$, $\sigma > 0$ is a scale, and $a > 0$ and $b > 0$ are shape parameters. The location vector $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)^\top$, where $\mu_i = \mathbf{v}_i^\top \gamma$, is $\boldsymbol{\mu} = \mathbf{V}\gamma$, where $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_n)^\top$ is a known model matrix. The log-DDa-exponential (LDDa-E) regression is given by Eq. (10) with $\sigma = 1$.

Let F and C be the sets representing the observed lifetimes and censoring times, respectively. The total log-likelihood function for $\boldsymbol{\theta} = (a, b, \sigma, \gamma^\top)^\top$ can be determined from Eqs (9) and (10) as

$$l(\boldsymbol{\theta}) = q[\log(a) + \log(b) - \log(\sigma)] + \sum_{i \in F}\left[\left(\frac{y_i - \mathbf{v}_i^\top\gamma}{\sigma}\right) - \exp\left(\frac{y_i - \mathbf{v}_i^\top\gamma}{\sigma}\right)\right]$$

$$- (b+1)\sum_{i \in F}\log\left\{1 + \left[-\log\left(1 - \exp\left[-\exp\left(\frac{y_i - \mathbf{v}_i^\top\gamma}{\sigma}\right)\right]\right)\right]^{-a}\right\}$$

$$
\begin{aligned}
&- (a+1)\sum_{i\in F}\log\left[-\log\left(1-\exp\left[-\exp\left(\frac{y_i-\mathbf{v}_i^{\top}\gamma}{\sigma}\right)\right]\right)\right]\\
&-\sum_{i\in F}\log\left\{1-\exp\left[-\exp\left(\frac{y_i-\mathbf{v}_i^{\top}\gamma}{\sigma}\right)\right]\right\}\\
&- b(n-q)\sum_{i\in C}\log\left\{1+\left[-\log\left(1-\exp\left[-\exp(\frac{y_i-\mathbf{v}_i^{\top}\gamma}{\sigma})\right]\right)\right]^{-a}\right\}, \qquad (11)
\end{aligned}
$$

where q is the observed number of failures. The MLE $\widehat{\boldsymbol{\theta}}$ of $\boldsymbol{\theta}$ can be found by maximizing Eq. (11).

8. Applications

The Weibull and Birnbaum-Saunders distributions are taken as baselines to prove the flexibility of the new family. The data sets were obtained from the open data portal of the Federal Government linked to the Ministry of Health and comprise events from 2020–2021 (accessed on August 23, 2021). The data portal is available at https://dados.gov.br/dataset/bd-srag-2020.

All computations are done in `R` using `GenSA`, `MASS` and `AdequacyModel` libraries, and the `goodness.fit()` function with the "`SANN`" method. The initial parameter values to maximize the log-likelihood were obtained through a heuristic method by using the `MASS` package in the `R` language.

The new distributions are compared with well-known models belonging to the Kw-G and B-G classes using the statistics: Cramér-von Mises (W^*), Anderson-Darling (A^*), Kolmogorov- Smirnov (KS), and p-values of the KS test. We present below the alternative densities for the applications.

– The Kumaraswamy-Weibull (Kw-W) density (Cordeiro et al., 2010) (for $x>0$)

$$
\begin{aligned}
f(x;a,b,k,\lambda) = ab\frac{k}{\lambda}\left(\frac{x}{\lambda}\right)^{k-1}\exp\left\{-\left(\frac{x}{\lambda}\right)^{k}\right\}\left\{1-\exp\left[-\left(\frac{x}{\lambda}\right)^{k}\right]\right\}^{a-1}\\
\times\left\{1-\left(1-\exp\left[-\left(\frac{x}{\lambda}\right)^{k}\right]\right)^{a}\right\}^{b-1}, \quad a,b,k,\lambda>0.
\end{aligned}
$$

– The Beta Weibull (B-W) density (Lee et al., 2007), and explored by Cordeiro et al. (2013) (for $x>0$)

$$
f(x;a,b,k,\lambda) = \frac{k}{\lambda B(a,b)}\left(\frac{x}{\lambda}\right)^{k-1}\exp\left[-b\left(\frac{x}{\lambda}\right)^{k}\right]\left\{1-\exp\left[-\left(\frac{x}{\lambda}\right)^{k}\right]\right\}^{a-1},
$$

where $a,b,k,\lambda>0$ and $B(a,b)=\int_0^1 x^{a-1}(1-x)^{b-1}dx$ is the beta function.

– The Beta-Birnbaum-Saunders (B-BS) density (Cordeiro & Lemonte, 2011) (for $x>0$ and $a,b,\alpha,\beta>0$)

$$
\begin{aligned}
f(x;a,b,\alpha,\beta) = \frac{x^{-3/2}(x+\beta)}{2\alpha\sqrt{2\pi\beta}}\frac{\exp(\alpha^{-2})}{B(a,b)}\exp\left\{-\frac{1}{2\alpha^2}\left(\frac{x}{\beta}+\frac{\beta}{x}\right)\right\}\\
\times\Phi\left[\frac{1}{\alpha}\left(\sqrt{\frac{x}{\beta}}-\sqrt{\frac{\beta}{x}}\right)\right]^{a-1}\left\{1-\Phi\left[\frac{1}{\alpha}\left(\sqrt{\frac{x}{\beta}}-\sqrt{\frac{\beta}{x}}\right)\right]\right\}^{b-1}.
\end{aligned}
$$

– The Kumaraswamy-Birnbaum-Saunders (Kw-BS) density (Saulo et al., 2012) (for $x>0$ and $a,b,\alpha,\beta>0$)

$$
\begin{aligned}
f(x;a,b,\alpha,\beta) = \frac{abx^{-3/2}(x+\beta)}{2\alpha\sqrt{2\pi\beta}}\exp(\alpha^{-2})\exp\left\{-\frac{1}{2\alpha^2}\left(\frac{x}{\beta}+\frac{\beta}{x}\right)\right\}\\
\times\Phi\left[\frac{1}{\alpha}\left(\sqrt{\frac{x}{\beta}}-\sqrt{\frac{\beta}{x}}\right)\right]^{a-1}\left\{1-\Phi\left[\frac{1}{\alpha}\left(\sqrt{\frac{x}{\beta}}-\sqrt{\frac{\beta}{x}}\right)\right]^{a}\right\}^{b-1}.
\end{aligned}
$$

In the following, we calculate descriptive statistics, MLEs, their standard errors (SEs) and adequacy statistics to compare the fitted distributions to the data sets.

Table 2
Parameter estimation results for COVID-19 times in Recife, and adequacy measures

Distribution	MLEs and SEs				W^*	A^*	KS	p-value
DDa-W	a	b	k	λ	**0.14252**	**0.98512**	**0.04054**	**0.3121**
	6.72231	0.39752	0.52249	37.76017				
	(0.01294)	(0.01831)	(0.01294)	(< 0.0001)				
DDa-BS	a	b	α	β	0.21022	1.38258	0.04364	0.2331
	3.30213	0.42399	1.33094	13.4665				
	(0.40970)	(0.03065)	(0.12385)	(2.967e-06)				
Kw-W	a	b	k	λ	0.37456	2.23561	0.05660	0.05389
	19.77064	1.84816	0.32830	0.59472				
	(2.27358)	(0.77297)	(0.04448)	(0.16338)				
Kw-BS	a	b	α	β	0.56382	3.35188	0.06824	0.01047
	4.67480	40.85901	5.42802	89.55090				
	(0.16191)	(3.29676)	(0.12325)	(0.01177)				
BW	a	b	k	λ	0.42545	2.48047	0.06282	0.02334
	77.06372	47.75535	0.13930	19.46675				
	(4.713e-03)	(1.404e-04)	(4.095e-03)	(0.02961)				
Beta-BS	a	b	α	β	0.78468	4.42605	0.07809	0.00205
	6.41931	10.69270	37.27032	34.29943				
	(0.19522)	(0.40651)	(0.04304)	(0.03022)				
Weibull			k	λ	1.8997	10.20654	0.11559	5.702e-07
			1.18881	22.24782				
			(0.03583)	(0.83527)				
BS			α	β	0.69621	4.13159	0.07788	0.00214
			0.94854	14.34992				
			(0.02826)	(0.51174)				

8.1. COVID-19 data in Recife

The first application represents the times (in days) of 564 COVID-19 patients from the date of entry in the Intensive Care Unit (ICU) until cure in Recife (State of Pernambuco). In this context, the cure characterizes the evolution of the case as hospital discharge. Discharge from hospital can only mean that the patient no longer needs hospitalization.

The descriptive statistics for the time until cure for COVID-19 data in Recife include: mean = 20.85, SD = 20.58, skewness = 2.94, kurtosis = 17.53, and minimum and maximum values 1 and 206, respectively. For these data, the total time under test (TTT) plot (Aaset, 1987) indicated an inverted bathtub form for the hrf (not shown here).

The values of the statistics W^*, A^*, KS, and the p-values of the KS test, are reported in Table 2. The DDa-W model is better than the Kw-G and B-G classes for both Birnbaum-Saunders and Weibull baselines as shown in Table 2. So, the proposed family can provide better fits than the well-known B-G and Kw-G classes.

The Vuong test (Vuong, 1989) also reveals that the DDa-W distribution is better than the DDa-BS ($LR = 2.47$), Kw-W ($LR = 16.39$) and B-W ($LR = 14.76$) distributions for a level of significance of 5%.

Figure 5 displays the histogram of the data, where x represents the time and the fitted DDa-W density and some other densities. The plots confirm that the DDa-W distribution provides the best fit to the current data. Further, the Q-Q plots of the best four fitted distributions (not shown here) indicate that the DDa-W distribution provides a superior fit to the current data.

8.2. Regression modeling applied to COVID-19 data in Curitiba (Brazil)

The study comprises the time (in days) elapsed from the date of hospitalization until death by the coronavirus, of 1,951 patients in Curitiba-PR, with all observations failing, that is, censored times were not considered in the study, with occurrences of death in 2020 and 2021.

The explanatory variables are (for $i = 1, \ldots, 1951$):

- v_{i1}: sex (1 = masculine, 2 = feminine);
- v_{i2}: age (in years).

Table 3
Estimation results from some fitted regressions to the COVID-19 data in Curitiba, and the adequacy measures

	LDDa-W		Kw-Gu		LBW		LW	
Parameter	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
γ_0	5.635*	0.419	3.496*	0.311	3.378*	0.508	4.090*	0.100
γ_1	−0.122*	0.035	−0.147*	0.037	−0.068**	0.039	−0.193*	0.035
γ_2	−0.009*	0.001	−0.013*	0.001	−0.012*	0.001	−0.013*	0.001
σ	2.355	0.349	1.419	0.186	1.973	0.608	0.758	0.012
a	5.876	0.709	3.076	0.795	6.065	3.688	–	–
b	1.677	0.348	1.813	0.234	3.492	1.124	–	–
AIC	**4880.25**		4882.50		4895.15		4990.87	
CAIC	**4880.29**		4882.54		4895.19		4990.89	
BIC	**4913.70**		4915.95		4928.61		5013.18	

*p-value < 0.0001; **p-value < 0.1.

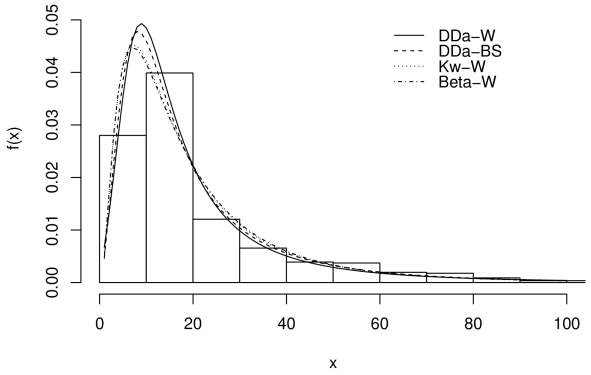

Fig. 5. Estimated DDa-W, DDa-BS, Kw-W and Beta-W densities.

The computational part is developed in `R` using `survival` library, with `optim()` function, and the "`SANN`" method.

We adopt the Akaike information criterion (AIC), corrected Akaike information criterion (CAIC), and Bayesian information criterion (BIC) to choose the appropriate model. We compare the fits of the LDDa-W Eq. (8) with the log-Kumaraswamy-Weibull or Kumaraswamy Gumbel (Kw-Gu) (Cordeiro et al., 2012), log-beta Weibull (LBW) and log-Weibull (LW) models. The densities for the alternative regressions are reported below:

– The LW (or Gumbel) density function

$$f(y;\mu,\sigma) = \frac{1}{\sigma}\exp\left\{\left(\frac{y-\mu}{\sigma}\right) - \exp\left(\frac{y-\mu}{\sigma}\right)\right\},$$

where $y \in \mathbb{R}$, $\sigma > 0$ and $\mu \in \mathbb{R}$; see Johnson et al. (1995).

– The LBW density function

$$f(y;a,b,\mu,\sigma) = \frac{1}{\sigma B(a,b)}\exp\left\{\left(\frac{y-\mu}{\sigma}\right) - b\exp\left(\frac{y-\mu}{\sigma}\right)\right\}\left\{1-\exp\left[-\exp\left(\frac{y-\mu}{\sigma}\right)\right]\right\}^{a-1},$$

where $y \in \mathbb{R}$, $\sigma > 0$ and $\mu \in \mathbb{R}$. For more details, see Ortega et al. (2013).

– The Kw-Gu density function

$$f(y;a,b,\mu,\sigma) = \frac{ab}{\sigma}\exp\left\{\left(\frac{y-\mu}{\sigma}\right) - \exp\left(\frac{y-\mu}{\sigma}\right)\right\}\left\{1-\exp\left[-\exp\left(\frac{y-\mu}{\sigma}\right)\right]\right\}^{a-1}$$
$$\times\left[1-\left(1-\exp\left[-\exp\left(\frac{y-\mu}{\sigma}\right)\right]\right)^{a}\right]^{b-1},$$

where $y \in \mathbb{R}$, $\sigma > 0$ and $\mu \in \mathbb{R}$.

The failure rate function is useful to aid in model identification more suitable for the variable time. In this context, the TTT plot (not shown here) for the data under study shows an increasing appearance for the most part, but due to

its final behavior, it indicates an inverting bathtub risk function. The descriptive statistics for the time until death for COVID-19 data in Curitiba include: mean = 16.86, SD = 15.54, skewness = 3.75, kurtosis = 29.77, and minimum and maximum values of 1 and 205, respectively.

Next, we provide results from the fit of the regression

$$y_i = \gamma_0 + \gamma_1 v_{i1} + \gamma_2 v_{i2} + \sigma z_i, \tag{12}$$

where $z_1, \ldots, z_{1951}$ are independent random variables with density function Eq. (9).

Table 3 provides some findings from the regressions fitted to the current data. They indicate that LDDa-W model provides the best fit to the data. Further, all covariates (v_{i1}: sex and v_{i2}: age) are significant at 1%, and then the categories of these explanatory variables are statistically different.

Thus, the time to death decreases when the age increases. Regarding the patient's gender, male patients present smaller time until death than female patients, since the estimate of its coefficient is negative.

After the LDDa-W regression estimation, the plots of the empirical and estimated survival functions support the model adequacy to these data.

Also, as part of the analysis, it is important to verify if there are observations influencing the model's adjustment. A sensitivity analysis was carried out to investigate this fact using the Cook's distance and will be presented below.

8.2.1. Sensitivity analysis and influential observations

Under the Generalized Cook Distance (Cook, 1977), the observations #349, #826 and #897 are the ones that stand out the most, thus indicating that they can be possible influential observations.

The observation #349 refers to a female individual, aged 84 years and with a time of hospitalization until death of 172 days. The observation #826 is identified as a 78-year-old male and has a time to death of 6 days. And the observation #897 refers to the male individual aged 61 years old, whose hospitalization time until death was only 3 days. The observations #349, #826 and #897 represent individuals with peculiar behaviors, but do not show signs of error in data collection or transcription, and therefore must be kept in the database. The final model is given in Eq. (12).

The impact of possible influential observations detected should be analyzed in order to assess the estimates and sensitivity of the model. This analysis considers new estimates for the model parameters from sub-samples referring to the withdrawal of these observations individually and in groups.

It is considered that the changes in the estimated values for the parameters are not very expressive and there is significance of the explanatory variables when considering the level of 10%. In addition, there was no change in the sign of the coefficient of the explanatory variables, so the inclusion or exclusion of the identified observations does not presuppose changes in the interpretation of the results.

9. Conclusions

One of the main objectives of distribution theory is to define a family of models to better explain lifetime phenomenon in several areas of knowledge. We proposed the Dual-Dagum-G ("DDa-G") family, which can generalize all classical continuous distributions. Its parameters are estimated by maximum likelihood, and a simulation study showed the consistency of the estimators. We showed the flexibility of the new family by means of two real COVID-19 data sets. We proved that the new Log-Dual-Dagum-Weibull (LDDaW) regression outperformed regressions based on well-known Kumaraswamy-G and Beta-G generators. After verifying the good fit of the new regression, a sensitivity analysis was performed, where it was possible to verify the occurrence of influential observations. As future work, it could be interesting to investigate other methods of sensitivity analysis, such as the local influence, if the results obtained through the Cook Distance prevail and still carry out a residual analysis for the new regression model.

Note: Computer codes in `R` language can be provided to readers free upon request.

References

Aarset, M.V. (1987). How to identify a bathtub hazard rate. *IEEE Transactions on Reliability*, *36*, 106-108.

Abd El-Bar, A.M.T., & Ragab, I.E. (2015). The beta generalized inverted exponential distribution. *International Information Institute*, *18*, 421-430.

Abu-Zinadah, H.H., & Bakoban, R.A. (2017). The Beta Generalized Inverted Exponential Distribution with real data applications. *Revstat – Statistical Journal*, *15*, 65-88.

Cook, R.D. (1977). Detection of influential observation in linear regression. *Technometrics*, *19*, 15-18.

Cordeiro, G.M., & de Castro, M. (2011). A new family of generalized distributions. *Journal of Statistical Computation and Simulation*, *81*, 883-898.

Cordeiro, G.M., & Lemonte, A.J. (2011). The β-Birnbaum-Saunders distribution: An improved distribution for fatigue life modeling. *Computational Statistics & Data Analysis*, *55*, 1445-1461.

Cordeiro, G.M., Nadarajah, S., & Ortega, E.M.M. (2012). The kumaraswamy gumbel distribution. *Statistical Methods & Applications*, *21*, 139-168.

Cordeiro, G.M., Nadarajah, S., & Ortega, E.M.M. (2013). General results for the beta Weibull distribution. *Journal of Statistical Computation and Simulation*, *83*, 1082-1114.

Cordeiro, G.M., Ortega, E.M.M., & Nadarajah, S. (2010). The Kumaraswamy Weibull distribution with application to failure data. *Journal of the Franklin Institute*, *347*, 1399-1429.

Cordeiro, G.M., Saboor, A., Khan, M.N., Ozel, G., & Pascoa, M.A.R. (2016). The kumaraswamy exponential-weibull distribution: Theory and applications. *Hacettepe Journal of Mathematics and Statistics*, *45*, 1203-1229.

Dagum, C. (1975). A model of income distribution and the conditions of existence of moments of finite order. *Bulletin of the International Statistical Institute*, *46*, 199-205.

Dagum, C. (1977). A new model of personal income-distribution-specification and estimation. *Economie Appliquáe*, *30*, 413-437.

Domma, F. (2002). L'andamento della hazard function nel modello di Dagum a tre parametri. *Quaderni di Statistica*, *4*, 1-12.

Domma, F., Giordano, S., & Zenga, M. (2011a). Maximum likelihood estimation in Dagum distribution with censored samples. *Journal of Applied Statistics*, *38*, 2971-2985.

Domma, F., Latorre, G., & Zenga, M. (2011b). Reliablity studies of Dagum distribution. *Working Paper*, *206*, 1-17.

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

Hairer, E., Norsett, S., & Wanner, G. (1993). Solving Ordinary Differential Equations I. *Springer*, *8*, 528.

Henrici, P. (1993). Applied and Computational Complex Analysis, Volume 3: Discrete Fourier analysis-Cauchy integrals-Construction of Conformal Maps-Univalent Functions. *John Wiley & Sons Inc.*, *3*, 656.

Johnson, N.L., Kotz, S., & Balakrishnan, N. (1995). Continuous univariate distributions. *John Wiley & Sons*, *2*, 732.

Kenney, J.F., & Keeping, E.S. (1961). Mathematics of statistics. *D. Van Nostrand Company*, *1*, 429.

Kleiber, C. (2008). A guide to the Dagum distributions. *Modeling Income Distributions and Lorenz Curves – Springer*, *5*, 97-117.

Kleiber, C., & Kotz, S. (2003). Statistical size distributions in economics and actuarial sciences. *John Wiley & Sons*, *470*, 352.

Lee, C., Famoye, F., & Olumolade, O. (2007). Beta-weibull distribution: Some properties and applications to censored data. *Journal of Modern Applied Statistical Methods*, *6*, 173-186.

Mameli, V. (2015). The Kumaraswamy skew-normal distribution. *Statistics & Probability Letters*, *104*, 75-81.

Marinho, P.R.D., Silva, R.B., Bourguignon, M., Cordeiro, G.M., & Nadarajah, S. (2019). AdequacyModel: An R package for probability distributions and general purpose optimization. *PLoS ONE*, *14*, e0221487.

Mead, M., Afify, A.Z., Hamedani, G., & Ghosh, I. (2017). The beta exponential fréchet distribution with applications. *Austrian Journal of Statistics*, *46*, 41-63.

Moors, J. (1988). A quantile alternative for kurtosis. *Journal of the Royal Statistical Society. Series D (The Statistician)*, *37*, 25-32.

Nadarajah, S., & Kotz, S. (2006). The exponentiated type distributions. *Acta Applicandae Mathematica*, *92*, 97-111.

Ortega, E.M., Cordeiro, G.M., & Kattan, M.W. (2013). The log-beta weibull regression model with application to predict recurrence of prostate cancer. *Statistical Papers*, *54*, 113-132.

Paranaíba, P.F., Ortega, E.M.M., Cordeiro, G.M., & Pescim, R.R. (2011). The Beta Burr XII distribution with application to lifetime data. *Computational Statistics & Data Analysis*, *55*, 1118-1136.

Saulo, H., Leão, J., & Bourguignon, M. (2012). The kumaraswamy birnbaum-saunders distribution. *Journal of Statistical Theory and Practice*, *6*, 745-759.

Tahir, M.H., & Nadarajah, S. (2015). Parameter induction in continuous univariate distributions: Well-established G families. *Anais da Academia Brasileira de Ciências*, *87*, 539-568.

Vuong, Q.H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, *57*, 307-333.

Ward, M. (1934). The Representation of Stirling's Numbers and Stirling's Polynomials as Sums of Factorials. *American Journal of Mathematics*, *56*, 87-95.