Chilean Journal of Statistics
Volume 15, Number 1 (June 2024), Pages 27–43
DOI: 10.32372/chjs.15-01-02

Distribution Theory and Statistical Modeling
Research Article

The Weibull flexible generalized family of bimodal distributions: Properties, simulation, regression models, and applications

Gauss M. Cordeiro[1], Elisângela C. Biazatti[1,2,*], Edwin M.M. Ortega[3], Maria do Carmo S. de Lima[1], and Luís H. de Santana[4]

[1]Department of Statistics, Federal University of Pernambuco, Recife, Brazil
[2]Department of Mathematics and Statistics, Federal University of Rondônia, Ji-Paraná, Brazil
[3]Department of Exact Sciences, University of São Paulo, Piracicaba, Brazil
[4]Technology Department, State University of Maringá, Umuarama, Brazil

(Received: 03 September 2023 · Accepted in final form: 05 December 2023)

Abstract

In the last two decades, several classes of distributions have been introduced to provide great flexibility in modeling real data. We define a new Weibull flexible-G family and construct a regression model based on this distribution. Maximum likelihood methods are adopted. Three real applications reveal that the new models can perform better fits than other competing ones.

Keywords: Maximum-likelihood method · Monte Carlo simulation · New flexible family · T-X family · Weibull-G class.

Mathematics Subject Classification: Primary 62E10 · Secondary 62F10.

1. Introduction, literature review, and motivating example

In this section, we provide an introduction, bibliographical review, and a motivating example from breast cancer to justify the development of the proposed methodology.

1.1 Introduction

In the theory of new distributions, the search for models that fit well various data sets led to several types of research in this field. We can mention two recent researches: unit-Gompertz distribution (Sindhu et al., 2023) and the Inflated log-Lindley distribution (Bhattacharjee and Chakraborty, 2023). However, although many models in the literature present some flexibility in terms of adjustment to different types of databases, the addition of many parameters sometimes brought identifiability problems, and estimation, among others. Some approaches include constructing new continuous distributions by enlarging other common continuous ones (Alzaatreh et al., 2013; Tahir et al., 2022).

*Corresponding author. Email: elisangela.biazatti@unir.br (E.C. Biazatti)

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

http://www.soche.cl/chjs

Combining the Weibull-G (W-G) (Bourguignon et al., 2014) and the flexible generalized (F-G) (Tahir et al., 2022) classes, we define a competitive family to the exponentiated-G (Exp-G) (Lehmann, 1953), Marshall-Olkin-G (MO-G) (Marshall and Olkin, 1997) and gamma-G (Zografos and Balakrishnan, 2009) classes.

Some recent distributions that exploit this class are: Weibull Birnbaum-Saunders (Benkhelifa, 2021), Weibull-Power Lomax (Hussain et al., 2020), and Weibull inverse Lomax —WIL— (Hassan and Mohamed (2019) and Jamilu et al. (2019)).

The great advantage of the F-G family is that it involves adding just one parameter. This eliminates, for example, the possible difficulty in estimating parameters. Additionally, the mathematical structure of the family is very simple and does not involve complex mathematical functions. The choice of the W-G family follows its simplicity and applicability in several areas. Additionally, as the novelty of the aticle, we will see that the combination of the two families can generate distributions that have bimodality as a characteristic.

1.2 Motivating example from breast cancer

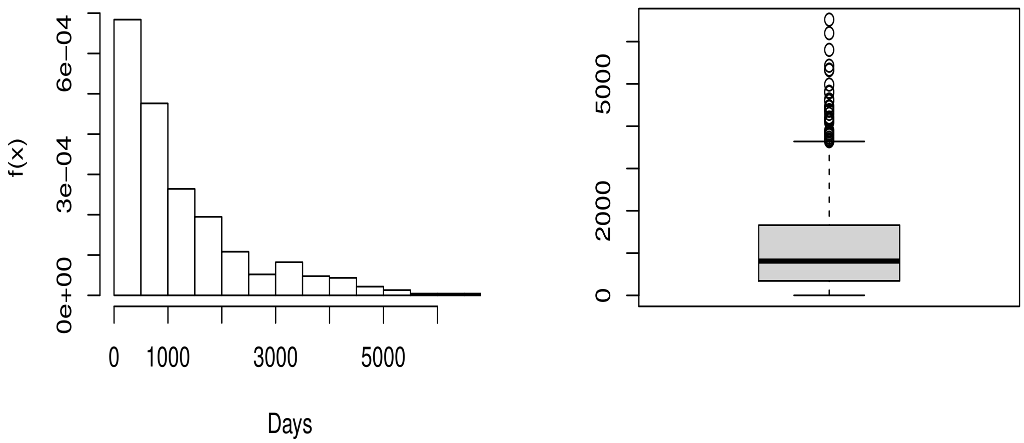
Cancer arises from the transformation of normal cells into tumor cells in a multi-stage process that generally progresses from a pre-cancerous lesion to a malignant tumor (WHO, 2022). This characteristic comprises studies that may use time as a variable of interest, for example, that elapsed between the diagnosis of the disease and death. Survival time is a non-negative random variable, and in many studies, it is necessary to consider probability models to describe the lifetimes of cancer patients. It is common in practical situations for survival time to be related to covariates that explain its variability. Further, a very common characteristic of survival data is the presence of censored observations, that is, for some individuals under study, the exact time of interest is not known, but only that it occurs to the right or left of a certain value.

The motivation for our study came from the fact that breast cancer is one of the most commonly occurring cancers in Brazil, ranked among the most prevalent cancer types among women. The data set studied here was obtained from the database available on the website of the National Cancer Institute (INCA, https://www.inca.gov.br/BasePopIncidencias/Home.action accessed on 20 May 2024), and refers to the mortality caused by breast cancer in the city of João Pessoa, Paraíba (Brazil) as measured by the death index. The sample is composed of 230 women diagnosed with breast cancer between 1999 and 2016. The data on these women refer to personal characteristics, date of diagnosis, follow-up period, and date of death (or end of monitoring). Data are collected on variables: x_i-time (days) of follow-up of women undergoing treatment for breast cancer until death and v_{i1}-morphology of neoplasms (0=adenocarcinoma, 1=malignant neoplasm). Most cases of breast cancer originate in glandular tissues, which can lead to adenocarcinoma. The disease can cause symptoms such as the presence of nodules; breast deformity; nipple retraction; redness; nipple discharge; pain and swelling in the armpits. A malignant neoplasm is one in which cells grow uncontrollably and spread to adjacent organs and tissues, a process called metastasis.

Table 1 reports basic statistics for the breast cancer data set related to minimum and maximum values, mean, median, standard deviation (SD) as well as, skewness and kurtosis. Figure 1 shows the histogram and box-plot for the data under analysis. A right skewness is noted in the histogram and confirmed by the skewness result reported in Table 1. The box-plot in Figure 1(b) detects some atypical values.

Table 1. Descriptive summary for breast cancer data.

Minimum	Maximum	Mean	Median	SD	Skewness	Kutorsis
1	6520	1206.37	811	1222.86	1.58	5.38



(a) (b)

Figure 1. Histogram (a) and box-plot (b) for breast cancer data.

Based on this exploratory analysis, we postulate the Weibull flexible generalized (WF-G) family as a suitable alternative to model these data. The remainder of this article is organized as follows. In Section 2, the WF-G family of distributions as well as some special models and basic properties are introduced. Maximum likelihood (ML) estimation, a simulation study, and the WF-G regression model are addressed in Section 3. Three real applications are reported in Section 4. Conclusions end the article in Section 5.

2. New family, special models and properties

In this section, we introduce the WF-G family of distributions as well as some special models and basic properties.

2.1 The Weibull flexible generalized distribution

The cumulative distribution function (CDF) of the F-G family is given by

$$H_G(x) = 1 - \overline{G}(x)^{G(x)}, \quad x \in \mathrm{R}, \tag{2.1}$$

where $G(x)$ is a parent CDF with parameter vector $\boldsymbol{\xi}$ and $\overline{G}(x) = 1 - G(x)$.

The CDF of the W-G class (Bourguignon et al., 2014) with unity scale and shape parameter $b > 0$ is stated

$$W_G(x) = 1 - \exp\left(-\left[\frac{G(x)}{\overline{G}(x)}\right]^b\right). \tag{2.2}$$

The CDF of the WF-G family is defined by inserting Equation (2.1) in Equation (2.2) and formulated as

$$F_G(x) = W_{H_G}(x) = 1 - \exp\left(-\left[\overline{G}(x)^{-G(x)} - 1\right]^b\right). \tag{2.3}$$

By differentiating Equation (2.3), the WF-G probability density function (PDF) is

$$f_G(x) = b\,g(x)\,\overline{G}(x)^{-G(x)}\left[\frac{G(x)}{\overline{G}(x)} - \log\overline{G}(x)\right]\left[\overline{G}(x)^{-G(x)} - 1\right]^{b-1}\exp\left(-\left[\overline{G}(x)^{-G(x)} - 1\right]^{b}\right), \tag{2.4}$$

where $g(x) = \mathrm{d}G(x)/\mathrm{d}x$.

2.2 Four special models

The Weibull flexible-Weibull (WFW) PDF (for $x > 0$) follows easily from Equation (2.4) and the Weibull CDF with shape α and scale λ, where all parameters are positive constants. For $\alpha = 1$, it gives the Weibull flexible-exponential (WFE) PDF.

The beta-prime (BP) CDF (for $x, \alpha, \lambda > 0$) is $G(x) = I_z(\alpha, \lambda)$, where I_z is the regularized incomplete beta function, and $z = z(x) = x/(1 + x)$. The Weibull flexible-beta prime (WFBP) PDF is determined by inserting this CDF and its derivative in Equation (2.4).

The log-logistic (LL) CDF (for $x, \lambda, \alpha > 0$) with shape α and scale λ. The Weibull flexible-log-logistic (WFLL) PDF can be obtained from Equation (2.4).

The gamma CDF (for $x > 0$) with scale $\lambda > 0$ and shape $\alpha > 0$ is equal to the incomplete gamma function ratio $G(x) = \gamma_1(\alpha, x/\lambda)$. The Weibull flexible-gamma (WFGa) density is determined from Equation (2.4).

The plots of the densities above in Figure 2 reveal that they are very flexible including bimodality for the WFW(b, α, λ) and WFGa(b, α, λ) models, when $\boldsymbol{\theta} = (0.15, 10, 4)$ and $\boldsymbol{\theta} = (0.09, 8, 3)$ for WFW, and $\boldsymbol{\theta} = (0.09, 72, 24)$ and $\boldsymbol{\theta} = (0.08, 62, 40)$ for WFGa, where $\boldsymbol{\theta} = (b, \alpha, \lambda)$. For the WFW PDF curves in Figure 2 when $\boldsymbol{\theta} = (2, 2, 5)$, we note an unimodal PDF, which has a value of approximately 0.6 when $x = 5$. In addition, some hazard rates in Figure 3 have bathtub, increasing, decreasing, crescent-descending-crescent and unimodal shapes. For the same parameter values in Figure 3, the failure rate increases over time. For $\boldsymbol{\theta} = (0.05, 5, 2)$, the PDF decreases up to $x = 2$, and from that point on there is an increase in the number of failures, where the function is approximately 0.28 when $x = 3.5$. On the other hand, for $\boldsymbol{\theta} = (0.15, 10, 4)$, we note bimodality in the PDF, and the HRF in Figure 3 shows an increase. However, when $\boldsymbol{\theta} = (0.09, 8, 3)$, the bimodality is seen in the PDF curve, and the HRF in Figure 3 has a high rate of failures in the initial times, and as the time progresses, failures begin to increase.

2.3 Properties

The quantile function (QF) of X, say $Q(u) = F_G^{-1}(u)$, follows by inverting Equation (2.3) under the steps below:

1. Generate $u \sim \mathrm{U}(0, 1)$;
2. Set $z = 1 + [-\log(1 - u)]^{1/b}$;
3. Solve numerically for x a non-linear equation $\left[\overline{G}(x)^{-G(x)}\right] = z$, for example, for a Newton-Raphson algorithm.

Another alternative for generating random variables from a PDF f, when it is not possible to find an expression for the QF F_G^{-1}, is the acceptance-rejection method. The algorithm below generate these random variables:

1. Generate y from $g(y)$, where $g(y)$ is an alternative PDF such that observations can be easily generated (inverse transform method, for example);
2. Generate $u \sim \mathrm{U}(0, 1)$;
3. If $u \leq f(y)/c(g(y))$, then $x = y$, where $c = \max[f(y)/g(y)]$. Otherwise, discard y as an observation from X and return to Step 1.

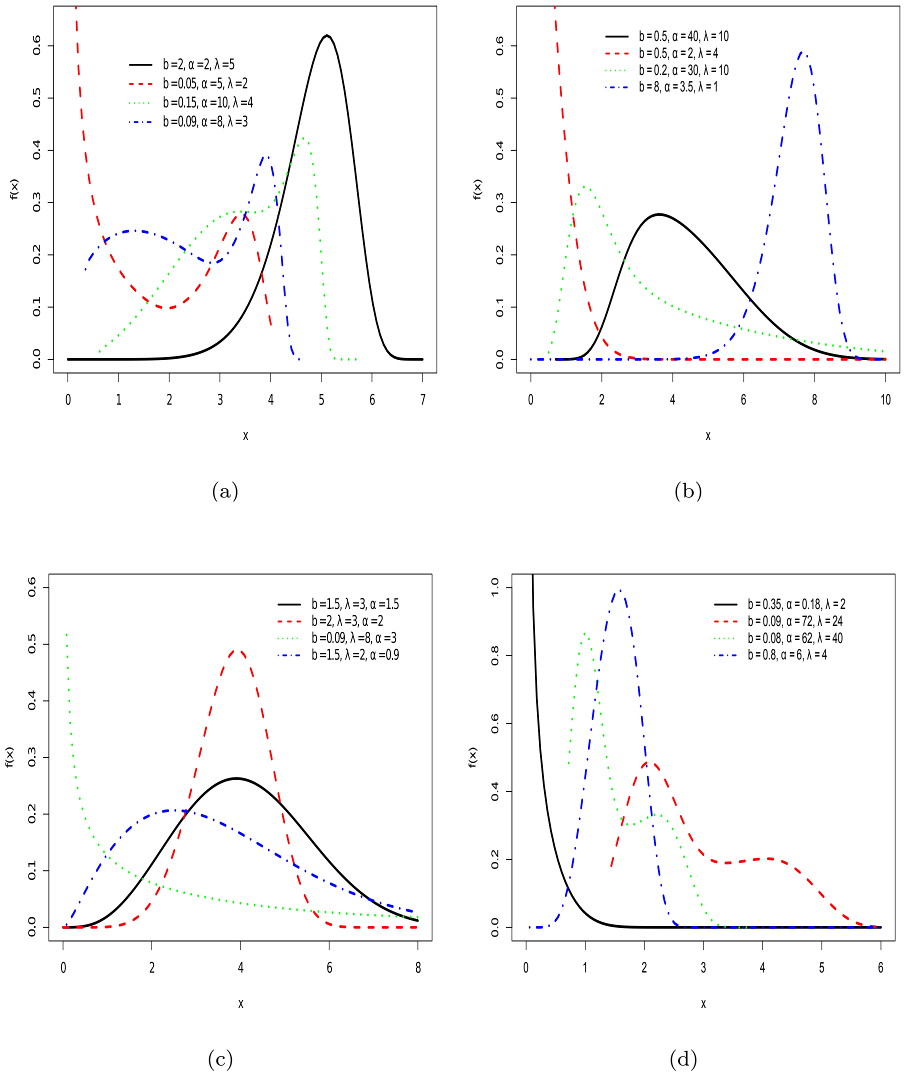


Figure 2. PDF shapes for some special cases of the WF-G family: (a) WFW, (b) WFBP, (c) WFLL and (d) WFGa.

Hereafter, let $T_s \sim$ exp-G(s) be the exponentiated-G (exp-G) random variable with power $s > 0$, whose PDF is $\pi_s(x) = s\, G(x)^{s-1}\, g(x)$. Many exp-G properties have been studied by several authors (Tahir and Nadarajah, 2015).

The linear representation for the WF-G PDF holds (Appendix A)

$$f_G(x) = \sum_{i=0,j=1}^{\infty} \rho_{i,j}\, \pi_{i+2jb}(x), \tag{2.5}$$

where $\rho_{i,j} = (-1)^{j+1}\, \gamma_{i,j}/j!$, and $\gamma_{i,j}$ is given there. Equation (2.5) allows obtaining WF-G properties from known properties of more than forty exp-G models.

The corresponding moments are formulated from Equation (2.5) as

$$\mu'_n = E(X^n) = \sum_{i=0,j=1}^{\infty} \rho_{i,j}\, E(T_{i+2jb}^n) = \sum_{i=0,j=1}^{\infty} (i+2jb)\, \rho_{i,j}\, \delta_{n,i+2jb}\,,$$

where $\delta_{n,d} = \int_0^1 Q(u)^n\, u^{d-1}\, \mathrm{d}u$. The nth incomplete moment of X is determined in a similar manner.

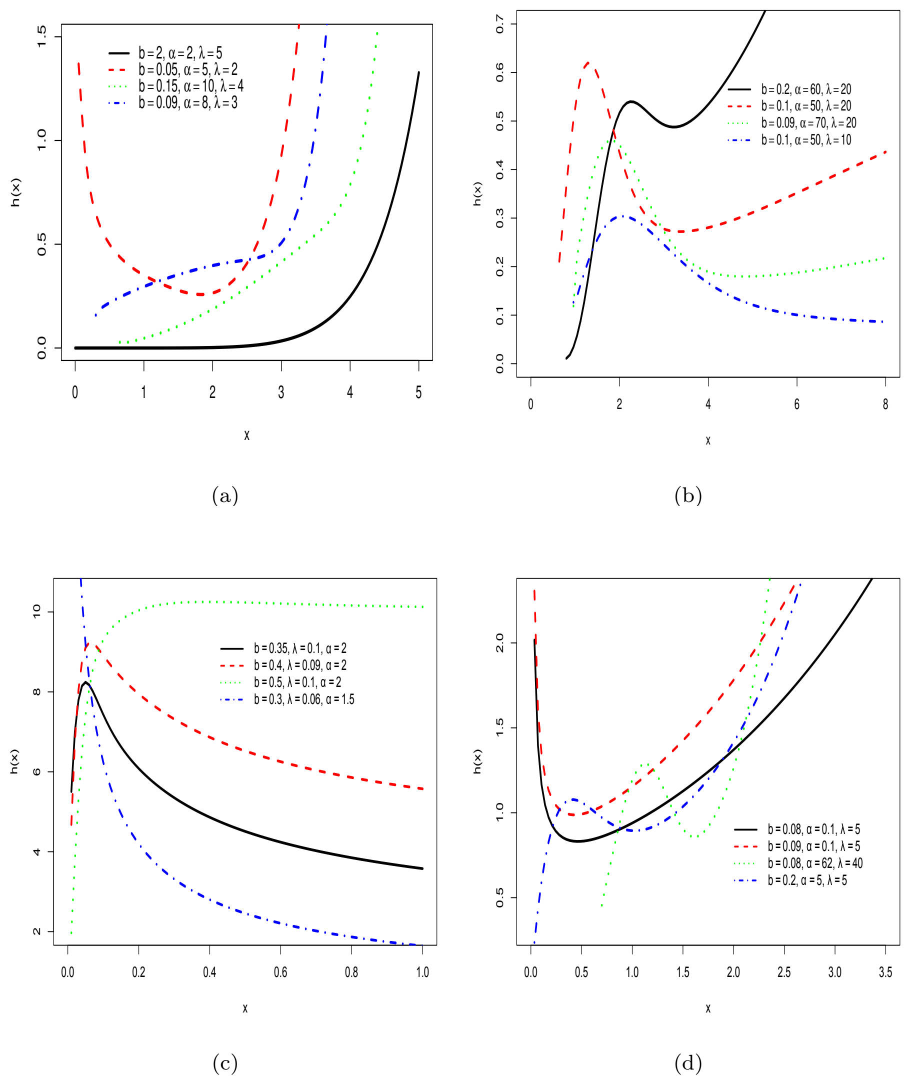


Figure 3. HRF for four special models of the WF-G family: (a) WFW, (b) WFBP, (c) WFLL and (d) WFGa.

Further, the moment generating function $M_X(t) = E(\exp(t\,X))$ of X has the form

$$M_X(t) = \sum_{i=0,j=1}^{\infty} \rho_{i,j}\, M_{T_{i+2jb}}(t) = \sum_{i=0,j=1}^{\infty} (i+2jb)\,\rho_{i,j}\,\tau_{t,i+2jb}\,,$$

where $M_{T_{i+2jb}}(t)$ is the moment generating function of T_{i+2jb}, and $\tau_{t,p} = \int_0^1 \exp(tQ(u))\,u^{p-1}du$ can be integrated numerically from the baseline QF.

3. Estimation, simulations, and associated regression model

In this section, we estimate the parameters of the WF-G distribution, provide a simulation study, and introduce the WF-G regression model.

3.1 ML estimation

Let $x_1, \ldots, x_n$ be a set of observations from Equation (2.4). The log-likelihood function for $\boldsymbol{\theta} = (b, \boldsymbol{\xi})^\top$ has the form

$$l = l(\boldsymbol{\theta}) = n\,\log(b) + \sum_{i=1}^{n} \log g(x_i) - \sum_{i=1}^{n} \Big[G(x_i)\,\log \overline{G}(x_i)\Big] + \sum_{i=1}^{n} \log\left[\frac{G(x_i)}{\overline{G}(x_i)} - \log \overline{G}(x_i)\right]$$
$$+(b-1)\sum_{i=1}^{n} \log\left[\overline{G}(x_i)^{-G(x_i)} - 1\right] - \sum_{i=1}^{n}\left[\overline{G}(x_i)^{-G(x_i)} - 1\right]^{b}. \quad (3.6)$$

The ML estimates, say $\widehat{\boldsymbol{\theta}}$, can be found by maximizing Equation (3.6) via the SAS (`PROC NLMIXED`) and the `R` software (`AdequacyModel` library (Marinho et al., 2019), `maxLik` function in `maxLik` library (Henningsen and Toomet, 2011) and `optim` function).

3.2 Monte Carlo simulations

We use the acceptance-rejection method to generate WFW random numbers as follows: (i) generate y from $g(y) = (\alpha/\lambda)(y/\lambda)^{\alpha-1}\exp(-(y/\lambda)^{\alpha})$; (ii) obtain $u \sim \mathrm{U}(0,1)$; (iii) if $u \leq f(y)/c(g(y))$, then $x = y$, where f has WFW PDF, and $c = \max[f(y)/g(y)]$. Otherwise, discard y and return to Step (i). The Monte Carlo simulation is carried out for sample sizes $n = 80, 100$, and 150, repeated 5,000 times and computed the average estimates (AEs), their biases (Bias), and mean squared errors (MSEs) for the five scenarios below:

- Scenario 1: Decreasing PDF $X \sim \mathrm{WFW}(b = 0.35, \alpha = 0.18, \lambda = 2)$;
- Scenario 2: Decreasing-increasing-decreasing PDF $X \sim \mathrm{WFW}(b = 0.05, \alpha = 2, \lambda = 0.9)$;
- Scenario 3: Unimodal PDF $X \sim \mathrm{WFW}(b = 1.2, \alpha = 1.5, \lambda = 0.5)$;
- Scenario 4: Bimodal PDF $X \sim \mathrm{WFW}(b = 0.19, \alpha = 3.5, \lambda = 2)$;
- Scenario 5: Bathtub PDF $X \sim \mathrm{WFW}(b = 0.009, \alpha = 7, \lambda = 0.5)$.

We utilize the `R` software with the `maxLik` function from the `maxLik` library and BFGS method. Appendix B provides PDF plots for each scenario. The findings of Table 2 indicate that the estimates are consistent since the AEs tend to the true parameters and the biases and MSEs decay when n increases. The simulation study is carried out on a computer with a 64-bit Windows operating system, 7th generation Intel(R) Core(TM) i5 processor, 2.50GHz and 4 Gb of RAM. We use the `R` software version 4.0.3, Studio Interface. The simulation times in minutes for each scenario are reported in Table 3. We can note that Scenarios 2 and 5 required more execution time.

Table 3. Run-times (in minutes) for each scenario.

	Scenario				
n	1	2	3	4	5
80	3.26	8.64	3.51	4.33	9.81
100	3.56	13.17	4.18	5.75	11.90
150	4.39	22.04	5.42	8.12	21.74

3.3 The WF-G regression model

Based on the WF-G PDF given in Equation (2.4), where λ_i and α are the baseline parameters, we construct a regression model linking the parameter λ_i to the explanatory vector

Table 2. Simulations from the WFW distribution.

Scenario 1 ($b = 0.35$, $\alpha = 0.18$, $\lambda = 2$)									
	$n = 80$			$n = 100$			$n = 150$		
$\boldsymbol{\theta}$	AE	Bias	MSE	AE	Bias	MSE	AE	Bias	MSE
b	0.4479	0.0979	0.0966	0.4243	0.0743	0.0652	0.3982	0.0482	0.0338
α	0.1843	0.0043	0.0045	0.1835	0.0035	0.0035	0.1829	0.0029	0.0025
λ	2.5013	0.5013	1.9641	2.4869	0.4869	1.7858	2.4483	0.4483	1.5266
Scenario 2 ($b = 0.05$, $\alpha = 2$, $\lambda = 0.9$)									
	$n = 80$			$n = 100$			$n = 150$		
$\boldsymbol{\theta}$	AE	Bias	MSE	AE	Bias	MSE	AE	Bias	MSE
b	0.1694	0.1194	0.0226	0.1632	0.1132	0.0191	0.1517	0.1017	0.0124
α	2.5301	0.5301	0.6382	2.4958	0.4959	0.5252	2.4496	0.4496	0.3730
λ	0.9914	0.0914	0.0207	0.9894	0.0894	0.0179	0.9878	0.0878	0.0146
Scenario 3 ($b = 1.2$, $\alpha = 1.5$, $\lambda = 0.5$)									
	$n = 80$			$n = 100$			$n = 150$		
$\boldsymbol{\theta}$	AE	Bias	MSE	AE	Bias	MSE	AE	Bias	MSE
b	0.9724	-0.2277	0.2590	0.9865	-0.2135	0.2441	1.0263	-0.1737	0.2063
α	2.3646	0.8646	2.0440	2.2951	0.7951	1.8311	2.1281	0.6281	1.3508
λ	0.5058	0.0058	0.0004	0.5054	0.0054	0.0003	0.5044	0.0044	0.0003
Scenario 4 ($b = 0.19$, $\alpha = 3.5$, $\lambda = 2$)									
	$n = 80$			$n = 100$			$n = 150$		
$\boldsymbol{\theta}$	AE	Bias	MSE	AE	Bias	MSE	AE	Bias	MSE
b	0.2708	0.0808	0.0222	0.2581	0.0681	0.0132	0.2435	0.0535	0.0068
α	3.2450	-0.2549	0.2762	3.2714	-0.2286	0.2139	3.3053	-0.1947	0.1531
λ	1.9817	-0.0183	0.0027	1.9861	-0.0139	0.0018	1.9909	-0.0090	0.0009
Scenario 5 ($b = 0.009$, $\alpha = 7$, $\lambda = 0.5$)									
	$n = 80$			$n = 100$			$n = 150$		
$\boldsymbol{\theta}$	AE	Bias	MSE	AE	Bias	MSE	AE	Bias	MSE
b	0.2133	0.2043	0.0523	0.1985	0.1895	0.0396	0.1817	0.1727	0.0312
α	6.7272	-0.2728	0.6368	6.7973	-0.2027	0.3799	6.8906	-0.1094	0.1357
λ	0.4989	-0.0011	0.0002	0.4989	-0.0010	0.0002	0.4997	-0.0003	9.3e-05

$\boldsymbol{v}_i^\top = (v_{i1}, \ldots, v_{ip})$ as

$$\lambda_i = \exp(\boldsymbol{v}_i^\top \boldsymbol{\eta}) \qquad i = 1, \ldots, n, \tag{3.7}$$

where $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_p)^\top$. The WF-G regression model in Equation (3.7) can be adopted for many different types of data.

Considering independent observations $(x_1, \boldsymbol{v}_1), \ldots, (x_n, \boldsymbol{v}_n)$, the total log-likelihood function for $\boldsymbol{\theta} = (b, \alpha, \boldsymbol{\eta}^\top)^\top$ is given by

$$
\begin{aligned}
l(\boldsymbol{\theta}) = {} & n\log(b) + \sum_{i=1}^{n} \log(g(x_i; \lambda_i, \alpha)) - \sum_{i=1}^{n} G(x_i; \lambda_i, \alpha)\log(1 - G(x_i; \lambda_i, \alpha)) \\
& + \sum_{i=1}^{n} \log\left(\frac{G(x_i; \lambda_i, \alpha)}{[1 - G(x_i; \lambda_i, \alpha)]} - \log[1 - G(x_i; \lambda_i, \alpha)]\right) \\
& + (b-1)\sum_{i=1}^{n} \left\{[1 - G(x_i; \lambda_i, \alpha)]^{-G(x_i;\lambda_i,\alpha)} - 1\right\} \\
& - \sum_{i=1}^{n} \left\{[1 - G(x_i; \lambda_i, \alpha)]^{-G(x_i;\lambda_i,\alpha)} - 1\right\}^{b}. \qquad (3.8)
\end{aligned}
$$

By maximizing Equation (3.8) through the `NLMixed` subroutine in SAS, it is found the ML estimate $\widehat{\boldsymbol{\theta}}$.

4. Real data illustrations

In this section, the applicability of the WF-G family is proved in three real applications, where the BP, Ga, LL, and W are the baseline models. We compare the WF-G family with the Kumaraswamy-G (Kw-G) (Cordeiro and de Castro, 2011), beta-G (B-G) (Eugene et al., 2002), and their own baselines. Other competitors are the Exp-G (Lehmann, 1953), MO-G (Marshall and Olkin, 1997) and gamma-G (G-G) (Zografos and Balakrishnan, 2009).

4.1 Description of data sets

The real data sets used correspond: First, we consider the deaths of women of childbearing age (10 to 49 years old) in the five Brazilian regions from the Mortality Information System from 2016 to 2020 at http://tabnet.datasus.gov.br/cgi/tabcgi.exe?sim/cnv/mat10uf.def (accessed on 20 May 2024). According to the place of residence of the deceased, the original values are divided by 100. This device is used because our proposal is a continuous distribution (named as "Data set I"). In the second set, the variable refers to the time until the maintenance time from 14 October 2002 to 16 May 2005 in one pressure reduction and measurement stations (ERPM-A) (Sivini, 2006) (named as "Data set II"). The third data set was described in Section 1.2.

4.2 Software

All the computations are done using the `R` software. We calculate some descriptive statistics for the data sets, ML estimates, their standard errors (SEs), and some adequacy statistics to compare the fitted models. As far as the estimation methods used in the applications are concerned, we alternate between the BFGS and CG methods of the `AdequacyModel` package reported in the `goodness.fit()` function in `R` software. In addition to these, the function also supports the SANN, Nelder-Mead, and PSO (default) methods. The BFGS and CG are chosen because they produce the best results when compared to the others. The application codes are at the link https://github.com/elisangelacbiazatti/WF-G (accessed on 20 May 2024).

4.3 Analyses of data sets I and II

Table 4 reports a descriptive summary for Data set I and Data set II. Table 5 provides the ML estimates and their SEs. The Cramér-von Mises (W^*), Anderson-Darling (A^*) statistics, the Akaike information criterion (AIC), corrected Akaike information criterion (CAIC), and Bayesian information criterion (BIC) for choosing of the best model are given in Table 6. The WFBP model is better than the other distributions for Data set I and the WFGa model is superior to the others for Data set II. Some estimated PDFs are plotted in Figure 4, thus revealing the superiority of the WFBP distribution for Data set I and the superiority of the WFGa distribution for Data set II. Figure 5 shows the empirical and the estimated CDFs of these models, which confirms that the WFBP and WFGa distributions are the best ones for Data set I and Data set II, respectively. For the Data set I, the generalized likelihood ratio (GLR) test (Vuong, 1989) shows that the proposed distribution is more adequate than the EBP (GLR=26.08), MOBP (GLR=28.97), GBP (GLR=40.83), KwBP (GLR=122.38), and BBP (GLR=6.90) models at a 5% significance level. By comparing the WFBP and BP distributions leads to LR=58.26 (p-value = 2.28e-14). So, the WFBP model is the best one to describe the current data. And for the Data set II GLR test, the WFGa model outperforms the EGa (GLR=15.08), MOGa (GLR=5.68), GGa (GLR=25.83), KwGa (GLR=7.36), and BGa (GLR=9.56) distributions at a significance level of 5%. A comparison of the WFGa distribution with its Ga model gives LR=11.11 (p-value=0.00086).

Table 4. Descriptive statistics.

Data set I						
Min	Max	Mean	Median	SD	Skewness	Kurtosis
48.73	313.33	133.41	87.62	90.23	0.69	1.94
Data set II						
Min	Max	Mean	Median	SD	Skewness	Kurtosis
1.17	8	2.92	1.59	2.26	0.93	2.39

Table 5. ML estimates and SEs in parentheses.

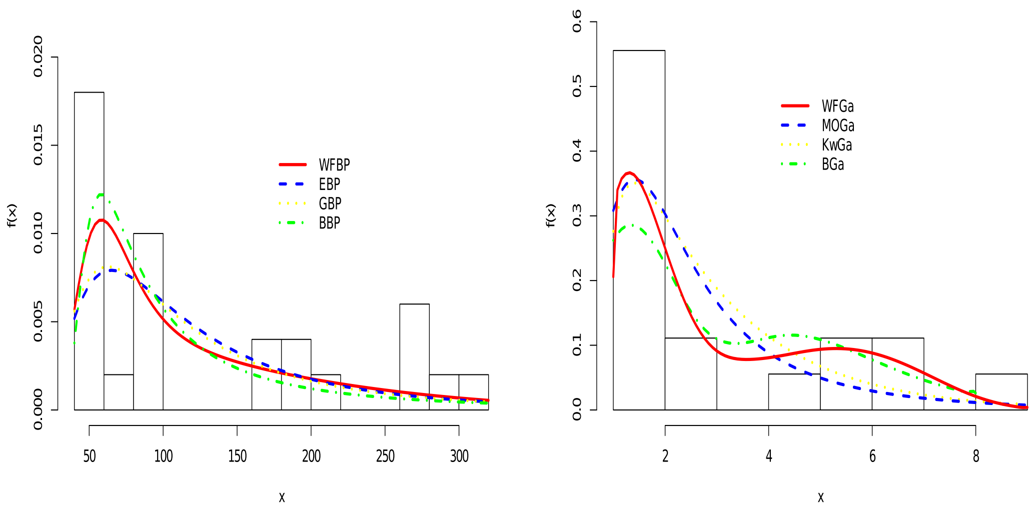
Distribution	Data set I				Distribution	Data set II			
WFBP(b,α,λ)	-	0.1906	1000.0000	9.7030	WFGa(b,α,λ)	-	0.0532	37.1449	12.2864
	-	(0.0280)	(5.0e-05)	(5.0e-05)		-	(0.0088)	(0.0335)	(0.0349)
EBP(a,α,λ)	3.2695	-	99.0545	2.3089	EGa(a,α,λ)	100.0000	-	0.0303	0.4235
	(0.0013)	-	(2.0e-05)	(2.0e-06)		(0.2384)	-	(0.0086)	(0.0512)
MOBP(a,α,λ)	0.5874	-	252.4765	2.4319	MOGa(a,α,λ)	0.1408	-	2.9453	0.5664
	(0.4659)	-	(0.0079)	(0.0079)		(0.0018)	-	(0.0045)	(0.0886)
GBP(a,α,λ)	8.9287	-	43.2847	4.7581	GGa(a,α,λ)	1.4578	-	1.4867	0.7043
	(1.0e-05)	-	(8.2e-07)	(0.3503)		(0.5412)	-	(0.2323)	(0.2362)
KwBP(a,b,α,λ)	1.0057	0.0337	1.0021	6.3045	KwGa(a,b,α,λ)	12.3413	0.1219	2.1156	4.5919
	(0.1304)	(0.0067)	(0.0043)	(0.0045)		(8.7e-06)	(0.0212)	(2.0e-05)	(8.8e-08)
BBP(a,b,α,λ)	59.4215	0.2331	115.2104	6.7879	BGa(a,b,α,λ)	0.0536	0.0714	53.8206	17.1392
	(0.0168)	(0.0001)	(0.0008)	(0.0010)		(0.0149)	(0.0218)	(0.0025)	(0.0025)
BP(α,λ)	-	-	6.7986	0.3682	Ga(α,λ)	-	-	2.0799	0.7118
	-	-	(2.5797)	(0.0837)		-	-	(0.6455)	(0.2496)

4.4 Motivating example (continuation)

We consider the WF-G regression model to explain breast cancer data in Brazil. Table 7 reports the ML estimates for some fitted regression models to these data via the `NLMixed` procedure in SAS. These findings indicate that the new regression models (WFW, WFLL, and WFGa) have the lowest AIC, CAIC, and BIC values.

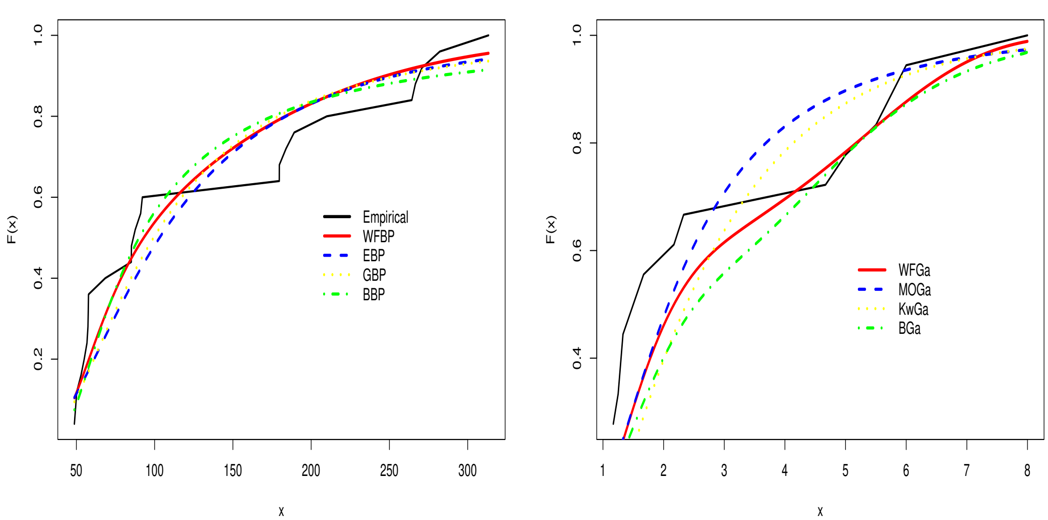
Table 6. Statistics of the fitted models.

	Data set I					Data set II					
Distribution	W^*	A^*	AIC	CAIC	BIC	Distribution	W^*	A^*	AIC	CAIC	BIC
WFBP	0.1512	0.9662	281.1157	282.2585	284.7723	WFGa	0.1820	1.0994	64.9046	66.6189	67.5757
EBP	0.1974	1.2602	288.3671	289.5100	292.0238	EGa	0.2722	1.5873	74.9510	76.6653	77.6222
MOBP	0.2254	1.3785	310.3452	311.4880	314.0018	MOGa	0.2511	1.5089	73.7879	75.5023	76.4591
GBP	0.1932	1.2301	289.7740	290.9169	293.4306	GGa	0.2855	1.6332	75.9348	77.6491	78.6059
KwBP	0.2184	1.3585	369.0378	371.0378	373.9133	KwGa	0.2342	1.4053	72.7931	75.8700	76.3546
BBP	0.1795	1.1772	289.1914	291.1914	294.0669	BGa	0.2267	1.3138	70.4991	73.5761	74.0606
BP	0.2107	1.3106	337.4414	337.9869	339.8792	Ga	0.2863	1.6363	74.0150	74.8150	75.7958



(a) (b)

Figure 4. Estimated PDFs for some models in Data set I (a) and Data set II (b).



(a) (b)

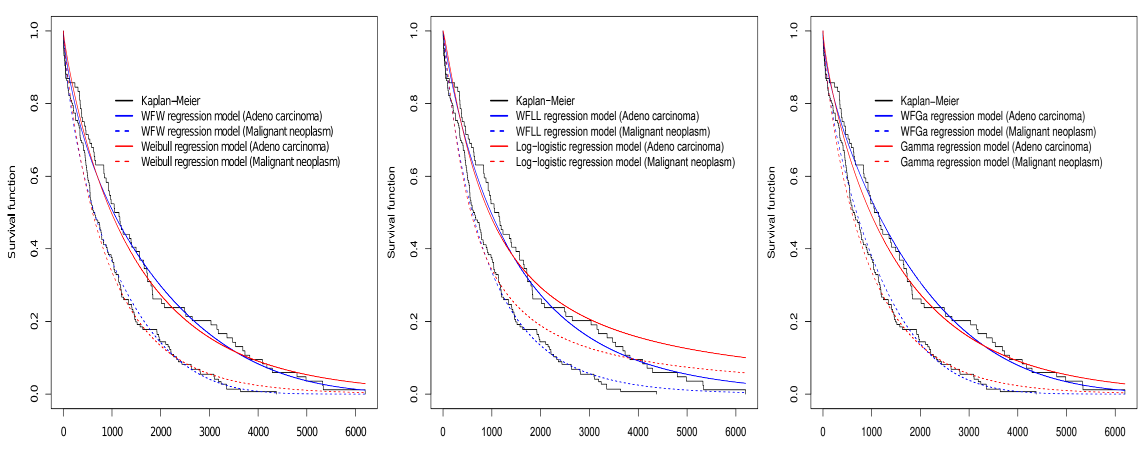
Figure 5. Empirical and estimated CDFs for some models in Data set I (a) and Data set II (b).

For the new regression models, v_1 is significant at 5% and then there is a significant difference between patients with adenocarcinoma and malignant neoplasm concerning the length of follow-up of women undergoing treatment for breast cancer until death.

Table 7. Estimation results.

Model	b	α	η_0	η_1	AIC	CAIC	BIC
WFW	0.4418	0.8966	7.3187	-0.4834	3690.7	3690.7	3704.4
	(0.1792)	(0.3299)	(0.1248)	(0.1216)			
			[<0.0001]	[<0.0001]			
Weibull		0.8873	7.3071	-0.4966	3700.6	3700.7	3710.9
		(0.0481)	(0.1250)	(0.1543)			
			[<0.0001]	[<0.0001]			
WFLL	13.5966	0.04874	-5.7269	-0.4974	3703.6	3703.8	3717.4
	(5.5600)	(0.0195)	(5.2241)	(0.1554)			
			[<0.0001]	[<0.0001]			
LL		1.1635	6.8481	-0.5005	3761.2	3761.4	3771.6
		(0.0660)	(0.1588)	(0.1992)			
			[<0.0001]	[<0.0001]			
WFGa	0.9474	0.3016	9.0515	-0.4860	3686.2	3686.4	3700.0
	(0.7088)	(0.2869)	(1.7313)	(0.1190)			
			[<0.0001]	[<0.0001]			
Gamma		0.7838	7.6012	-0.4959	3695.9	3696.0	3706.2
		(0.0629)	(0.1471)	(0.1547)			
			[<0.0001]	[<0.0001]			

The empirical and estimated survival functions of the WFW, WFLL, WFGa, Weibull, LL, and gamma models in Figure 6 reveal that the first three models provide good fits to the current data. The plots in Figure 6 show that patients with adenocarcinoma have a longer survival time than patients with malignant neoplasm.



(a) (b) (c)

Figure 6. Estimated survival functions from the fitted regressions and the empirical survival.

5. Conclusions

We proposed the Weibull flexible generalized family and provided some of its mathematical properties. The parameters are estimated by the maximum likelihood, and the consistency of the estimators was accessed from a simulation study. We proved that the new models outperformed other well-known ones using three applications. The fact there is not an analytical solution for the quantile function of the Weibull Flexible generalized family constitutes a limitation, as this factor impacts, for example, the generation of random samples from special cases of this family, and also quantile measures that could assist in the calculation of skewness and kurtosis.

Appendix A

For $0 \leq x < 1$, we obtain the convergent power series stated as

$$t(x) = (1-x)^{-x} - 1 = x^2 \sum_{i=0}^{\infty} \alpha_i \, x^i, \tag{5.9}$$

where $\alpha_0 = 1$, $\alpha_1 = 1/2$, $\alpha_2 = 5/6$, $\alpha_3 = 3/4$ and so on. From a power series raised to a real power (Apostol, 1974, p. 239) and Equation (5.9), we obtain

$$\left[\overline{G}(x)^{-G(x)} - 1\right]^b = \left(G(x)^2 \sum_{i=0}^{\infty} \alpha_i \, G(x)^i\right)^b = G(x)^{2b} \sum_{i=0}^{\infty} \beta_i(b) \, G(x)^i \,,$$

where

$$\beta_i = \beta_i(b) = \begin{cases} 1, & i = 0\,, \\ \dfrac{1}{i} \displaystyle\sum_{k=0}^{i-1} [\, i\, b - k(b+1)\,] \, \alpha_k \, \beta_{i-k}, & i > 0. \end{cases}$$

The expansion is formulated as $\exp(-x) = 1 + \sum_{j=1}^{\infty} (-1)^j \, x^j/j!$, for $x \in \mathrm{R}$. Hence, the WF-G family CDF stated in Equation (2.3) can be expressed as a convergent power series

$$F_G(x) = 1 - \exp\left(-\left[\overline{G}(x)^{-G(x)} - 1\right]^b\right) = \sum_{i=0, j=1}^{\infty} \frac{(-1)^{j+1}\, \gamma_{i,j}}{j!} \, G(x)^{i+2jb} \,,$$

where

$$\gamma_{i,j} = \gamma_{i,j}(b) = \begin{cases} 1, & i = 0, \\ \dfrac{1}{i} \displaystyle\sum_{k=0}^{i-1} [\, i\, j - k\,(j+1)\,] \gamma_{k,j} \, \beta_{i-k}, & i > 0. \end{cases}$$

The linear representation of the WF-G PDF follows after simple differentiation of $F_G(x)$.

Appendix B

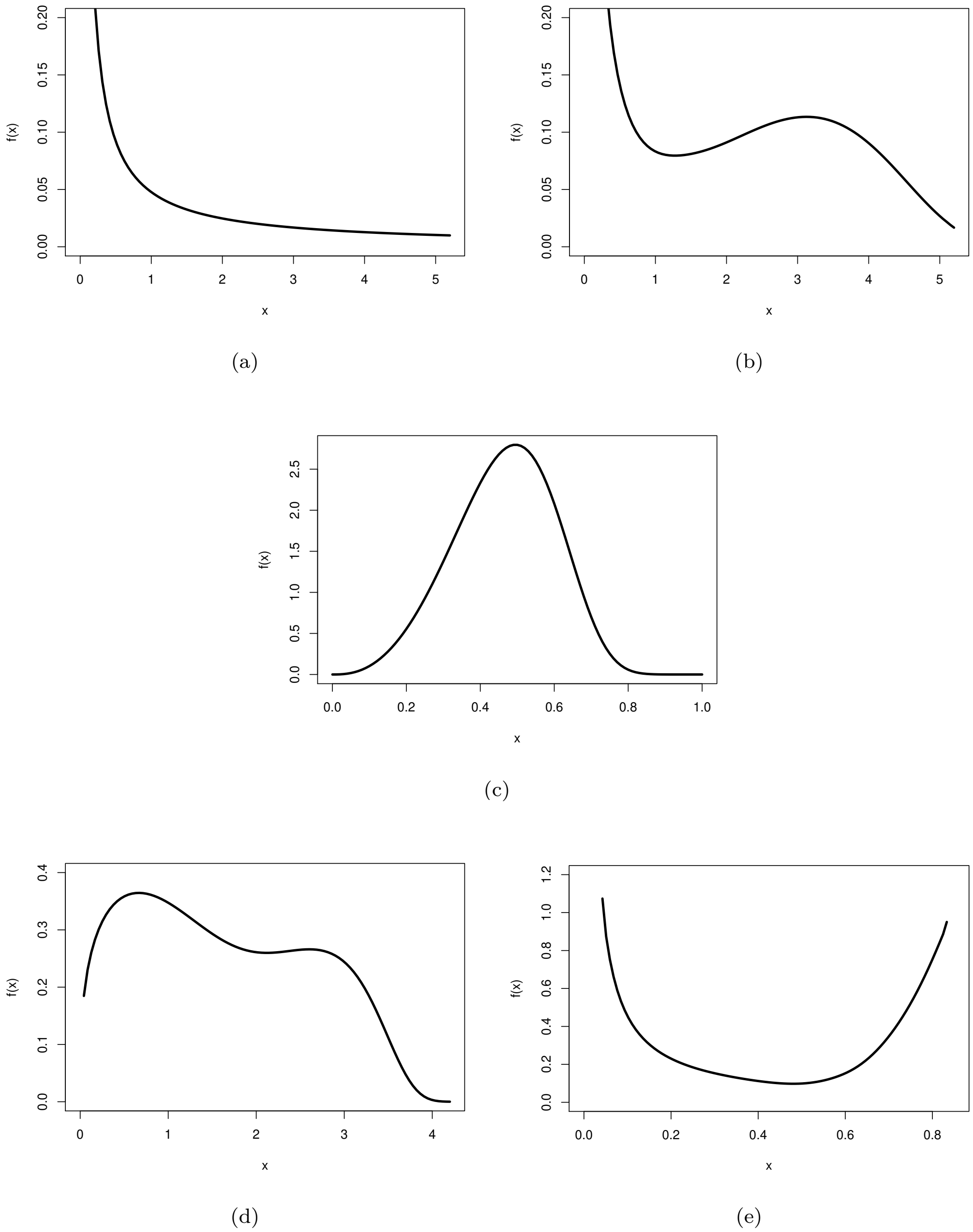


Figure 7. Estimated PDF for the scenarios 1(a), 2(b), 3(c), 4(d) and 5(e).

Author contributions Conceptualization: G.M.C., E.C.B., E.M.M.O., M.C.S.L., L.H.S.; data curation: G.M.C., E.C.B., E.M.M.O., M.C.S.L.; formal analysis: G.M.C., E.C.B., E.M.M.O., M.C.S.L.; investigation: G.M.C., E.C.B., E.M.M.O.; methodology: G.M.C., E.C.B., E.M.M.O., M.C.S.L.; software: E.C.B., E.M.M.O., M.C.S.L.; supervision: G.M.C.; validation: G.M.C., E.C.B., E.M.M.O., M.C.S.L., L.H.S.; visualization: G.M.C., E.C.B., E.M.M.O., M.C.S.L., L.H.S.; writing—original draft: G.M.C., E.C.B., E.M.M.O., M.C.S.L., L.H.S.; writing—review and editing: G.M.C., E.C.B., E.M.M.O., M.C.S.L., L.H.S.; All authors read and agreed to the published version of the article.

Funding This research received no external funding.

Conflicts of interest The authors declare no conflict of interest.

References

Alzaatreh, A., Lee, C., and Famoye, F., 2013. A new method for generating families of continuous distributions. Metron, 71, 63–79.

Apostol, T.M., 1974. Mathematical Analysis, Reading, Mass. Addison-Wesley, London, Canada.

Bhattacharjee, S., and Chakraborty, S., 2023. Inflated log-Lindley distribution for modeling continuous data bounded in unit interval with possible mass at boundaries. Chilean Journal of Statistics, 14, 123–142.

Benkhelifa, L., 2021. The Weibull Birnbaum-Saunders distribution and its applications. Statistics, Optimization and Information Computing, 9, 61–81.

Bourguignon, M., Silva, R.B., and Cordeiro, G.M., 2014. The Weibull-G family of probability distributions. Journal of Data Science, 12, 53–68.

Cordeiro, G.M. and de Castro, M., 2011. A new family of generalized distributions. Journal of Statistical Computation and Simulation, 81, 883–898.

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

Hassan, A.S. and Mohamed, R.E., 2019. Weibull inverse Lomax distribution. Pakistan Journal of Statistics and Operation Research, 15, 587–603.

Henningsen, A. and Toomet, O., 2011. maxLik: A package for maximum likelihood estimation in R. Computational Statistics, 26, 443–458.

Hussain, N.A., Doguwa, S.I.S., and Yahaya, A., 2020. The Weibull-power Lomax distribution: properties and application. Communication in Physical Sciences, 6, 869–881.

Jamilu, Y.F, Sani, I.D, and Audu, I., 2019. The Weibull-inverse Lomax (WIL) distribution with application on bladder cancer. Biometrics and Biostatistics International Journal, 8, 195–202.

Lehmann, E.L., 1953. The power of rank tests. The Annals of Mathematical Statistics, 24, 23–43.

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

Marshall, A.W. and Olkin, I., 1997. A new method for adding a parameter to a family of distributions with application to the exponential and Weibull families. Biometrika, 84, 641–652.

Sindhu, T.N., Shafiq, A., Mazucheli, J., Ozel, G., and Alves, B., 2023. Some additional facts about the unit-Gompertz distribution. Chilean Journal of Statistics, 14, 143–171.

Sivini, P.G.L., 2006. Desenvolvimento de banco de dados de confiabilidade: uma aplicação em estações redutoras de pressão de gás natural. Dissertation submitted to the Federal University of Pernambuco, Recife, Pernambuco, Brazil.

Tahir, M.H., Hussain, M.A., and Cordeiro, G.M., 2022. A new flexible generalized family for constructing many families of distributions. Journal of Applied Statistics, 49, 1615–1635.

Tahir, M.H., and Nadarajah, S., 2015. Parameter induction in continuous univariate distributions: Well-established G families. Annals of the Brazilian Academy of Sciences, 87, 539–568.

Vuong, Q.H., 1989. Likelihood ratio tests for model selection and non-nested hypotheses. Econometrica, 57, 307–333.

World Health Organization. 2022. Cancer. https://www.who.int/en/news-room/fact-sheets/detail/cancer (accessed on 09 April 2024).

Zografos, K. and Balakrishnan, N., 2009. On families of beta-and generalized gamma-generated distributions and associated inference. Statistical Methodology, 6, 344–362.