Parametric survival model based on the Lévy distribution

Andrea Valencia-Orozco^{1,a}, José R. Tovar-Cuevas^a

^aEscuela de Estadística, Facultad de Ingeniería, Universidad del Valle, Colombia

Abstract

It is possible that data are not always fitted with sufficient precision by the existing distributions; therefore this article presents a methodology that enables the use of families of asymmetric distributions as alternative probabilistic models for survival analysis, with censorship on the right, different from those usually studied (the Exponential, Gamma, Weibull, and Lognormal distributions). We use a more flexible parametric model in terms of density behavior, assuming that data can be fit by a distribution of stable distribution families considered unconventional in the analyses of survival data that are appropriate when extreme values occur, with small probabilities that should not be ignored. In the methodology, the determination of the analytical expression of the risk function $h(t)$ of the Lévy distribution is included, as it is not usually reported in the literature. A simulation was conducted to evaluate the performance of the candidate distribution when modeling survival times, including the estimation of parameters via the maximum likelihood method, survival function $\hat{S}(t)$ and Kaplan-Meier estimator. The obtained estimates did not exhibit significant changes for different sample sizes and censorship fractions in the sample. To illustrate the usefulness of the proposed methodology, an application with real data, regarding the survival times of patients with colon cancer, was considered.

Keywords: censored data, estimation, Lévy distribution, maximum likelihood, risk function, survival time

1. Introduction

Survival has traditionally been analyzed using statistical methodologies that include nonparametric methods such as Kaplan-Meier (KM) curves, actuarial methods, and some data modeling procedures using analytical probability distributions that are easy to use, such as the Exponential, Weibull, Gamma and Lognormal distributions. However, these distributions may not necessarily reflect the natural behavior of the times observed in a monitoring study, presenting restrictions related to the density form. Datasets are sometimes observed to exhibit important extreme values. An example are patients who survive for periods longer than those expected by the treating physicians, which is mainly caused by the present biological diversity that makes these patients different in terms of their susceptibility to contracting diseases or their response to a performed procedure. Such situations cause statistical methods to provide inaccurate predictions of times, as the likelihood function associated with the probability distribution used for modeling fails to include sufficient information about extreme values, thereby affecting the subsequent estimations (Martínez and Achcar, 2014).

Consequently, the present study proposes modeling the Lévy probability distribution, which is characterized by its tails, which are heavier than a Gaussian distribution; consequently, it may present a greater impulsivity degree in the distribution of times, where some events or group of events still

¹ Corresponding author: Escuela de Estadística, Facultad de Ingeniería, Universidad del Valle, Cali, Colombia.
E-mail: andrea.valencia@correounivalle.edu.co

have an occurrence probability. The model is part of the α -stable family of models. The theory of stable distributions was developed for the first time in the 1920s by Paul Lévy and Aleksander Khinchine (Samorodnitsky and Taqqu, 1994). Subsequently, it has been applied in different areas, such as economics, physics, hydrology, biology and signals processing (Adler *et al.*, 1998). This family of distributions is described by four parameters, α , b , μ , and β , where $\alpha \in (0, 2]$ is the characteristic exponent. This parameter controls the impulsivity degree of the random variable T . $b \in [-1, +1]$ is a parameter that controls the symmetry of the distribution. $\mu \in (-\infty, \infty)$ is a location parameter and $\beta > 0$ a scale parameter. Working with α -stable distributions can be difficult, as the family only has three unique models, for which the density function can be written using analytic functions, including the Lévy model, which has the form $\alpha = 1/2$, $b = 1$, μ, β . This distribution has been applied using classical inference and Bayesian methods in the presence of censored data on the left and the use of covariables (Achcar *et al.*, 2018). However, Yu *et al.* (2010) have also used it in genetic research such as ultrasound imaging modeling. Many of these authors have worked with the Lévy distribution for modeling data variability in situations with heavy tails and asymmetry; however, a literature search revealed no studies that modeled data with survival censorship on the right, whose behavior, in addition to exhibiting an asymmetric shape, reflects the presence of extreme data.

It is possible to use different reparametrizations of the Lévy distribution. According to Nolan (2009) one of the most used is the standard Lévy distribution, with $\alpha = 1/2$, $b = 1$, and $\mu = 0$ of a distribution with only one parameter, in this case β . Generally, the convenience of using a parametric model also depends on its flexibility to reproduce characteristic forms of the risk function as a great utility function in the analysis since it allows specifying an instantaneous death rate or failure in a time interval, given that the individual has survived until the beginning of the interval (Godoy, 2009). However, in the case of the Lévy distribution, the risk function is not usually found explicitly in the literature-usually, only the quotient $f(t)/S(t)$ is reported. In this study, the analytical expression and graphic representation of the risk function are obtained when the Lévy model is assumed for the data likelihood. The function is obtained using some definitions of the survival model, finding different methods to calculate it, and considering that it depends on the Gauss error function (erf), which is related to the normal probability distribution function and can be approximated by the use of certain series. Consequently, depending on availability and simplicity, any of the expressions $h(t)$ reported in this article can be selected.

Using the obtained results, this study compares the proposed distribution with parametric and non-parametric traditional methods (Kaplan-Meier). Additionally, illustrates the methodological proposal using a dataset concerning the survival times of patients with colon cancer (CC) that exhibit a marked asymmetry and censorship presence during the follow-up of individuals.

2. Lévy distribution

The Lévy distribution is a continuous probability distribution, named by the French mathematician Paul Lévy in 1925 during his investigation of the behavior of independent random variables. This distribution belongs to a family of stable distributions that have been of great interest because they allow modeling the asymmetry and arbitrarily larger tails than normal distribution (Buckle, 1995).

Stable distributions are described by four parameters (α, b, μ, β) .

- The parameter $\alpha \in [0, 2]$ is the characteristic exponent that controls the degree of impulsivity of the random variable with a stable distribution.
- The parameter $b \in [-1, 1]$ is a parameter that controls the symmetry of the distribution.

- The location parameter $\mu \in (-\infty, \infty)$.
- The scale parameter $\beta \in (0, \infty)$ defines the dispersion or kurtosis of the curve.

In the special case of the Lévy distribution, the stability parameter a is less than 1, therefore, the variance is infinite, the mean of the distribution does not exist and the domain of the density function is between $(0, \infty)$. The Lévy distribution of parameters (μ, β) has the following probability density function (Ahsanullah and Nevzorov, 2014):

$$f(t) = \left(\frac{\beta}{2\pi}\right)^{\frac{1}{2}} \frac{1}{(t-\mu)^{\frac{3}{2}}} \exp\left(-\frac{\beta}{2(t-\mu)}\right); \quad t > \mu, \mu > 0, \beta > 0. \quad (2.1)$$

In addition, it has the following distribution and survival functions:

$$F(t) = \operatorname{erfc}\left(\sqrt{\frac{\beta}{2(t-\mu)}}\right); \quad t > \mu, \mu > 0, \beta > 0, \quad (2.2)$$

$$S(t) = 1 - \operatorname{erfc}\left(\sqrt{\frac{\beta}{2(t-\mu)}}\right); \quad t > \mu, \mu > 0, \beta > 0, \quad (2.3)$$

where the function $\operatorname{erfc}(t) = 1 - \operatorname{erf}(t)$ is the complementary error function and the error function is given by

$$\operatorname{erf}(t) = \frac{2}{\sqrt{\pi}} \int_0^t e^{-z^2} dz. \quad (2.4)$$

It is possible to use different reparametrizations of the Lévy distribution, according to Nolan (2009) one of the most used is the “standard Lévy distribution”, with $\alpha = 1/2$, $b = 1$, and $\mu = 0$ in a distribution of only one parameter, in this case β . The location parameter has the effect of shifting the curve to the right by an amount μ (Figure 1) and changing the domain of the function in the interval $[\mu, \infty)$. The scale parameter has an effect on the kurtosis of the curve (Figure 1), that is, when the scale parameter is small ($\beta < 1$) the curve is more pointed and with a less wide tail (leptokurtic), but as the scale parameter increases, the curve is less pointed and with a wider tail (platykurtic).

3. Determination of the analytical expression of the risk function when it is assumed that the observed data are Lévy distributed

By definition, the risk function for any parametric model associated with a random variable T is designated as

$$h(t) = \frac{f(t)}{S(t)} = -\frac{S'(t)}{S(t)} = -\frac{d}{dt} \log [S(t)]. \quad (3.1)$$

Taking into account the survival function (2.3), the risk function can be expressed as

$$h(t) = -\frac{d}{dt} \log \left[1 - \operatorname{erfc}\left(\sqrt{\frac{\beta}{2(t-\mu)}}\right) \right]; \quad t > 0, \mu > 0, \beta > 0$$

Figure 1: Functions of the survival Lévy model.

then

$$\begin{aligned}
 h(t) &= -\frac{d}{dt} \log \left[\frac{2}{\sqrt{\pi}} \int_0^{\sqrt{\frac{\beta}{2(t-\mu)}}} \exp(-z^2) dz \right] = -\frac{\frac{d}{dt} \left[\frac{2}{\sqrt{\pi}} \int_0^{\sqrt{\frac{\beta}{2(t-\mu)}}} \exp(-z^2) dz \right]}{\frac{2}{\sqrt{\pi}} \int_0^{\sqrt{\frac{\beta}{2(t-\mu)}}} \exp(-z^2) dz} \\
 &= \frac{\frac{1}{\sqrt{\pi}} \frac{1}{\sqrt{\frac{\beta}{2(t-\mu)}}} \left(\frac{\beta}{2(t-\mu)^2} \right) \exp\left(-\frac{\beta}{2(t-\mu)}\right)}{\frac{2}{\sqrt{\pi}} \int_0^{\sqrt{\frac{\beta}{2(t-\mu)}}} \exp(-z^2) dz} = \frac{\left(\frac{\beta}{2\pi}\right)^{\frac{1}{2}} \frac{1}{(t-\mu)^{\frac{3}{2}}} \exp\left(-\frac{\beta}{2(t-\mu)}\right)}{\frac{2}{\sqrt{\pi}} \int_0^{\sqrt{\frac{\beta}{2(t-\mu)}}} \exp(-z^2) dz}. \quad (3.2)
 \end{aligned}$$

The numerator in (3.2) corresponds to the density function of a random variable with a Lévy distribution with parameters μ and β , whereas the denominator contains the survival function (in terms of the error function), for which it is not possible to evaluate the integral directly, as it does not have a closed form using elementary functions. If the integrand is expanded by a Taylor series, the “Taylor series of the error function” is obtained, thereby finding a valid expression for every real

number t and in the whole complex plane (Ayant and Borg, 1974).

Note that,

$$\begin{aligned} \operatorname{erf}\left(\sqrt{\frac{\beta}{2(t-\mu)}}\right) &= \frac{2}{\sqrt{\pi}} \sum_{n=0}^{\infty} \frac{(-1)^n \left(\sqrt{\frac{\beta}{2(t-\mu)}}\right)^{2n+1}}{n!(2n+1)} \\ &= \frac{2}{\sqrt{\pi}} \left(\left(\frac{\beta}{2(t-\mu)}\right)^{\frac{1}{2}} - \frac{\left(\frac{\beta}{2(t-\mu)}\right)^{\frac{3}{2}}}{3} + \left(\frac{\left(\frac{\beta}{2(t-\mu)}\right)^{\frac{5}{2}}}{10} - \frac{\left(\frac{\beta}{2(t-\mu)}\right)^{\frac{7}{2}}}{42} + \dots \right) \right). \end{aligned} \quad (3.3)$$

To perform an iterative calculation of the abovementioned series, the following alternative formula is used:

$$\operatorname{erf}\left(\sqrt{\frac{\beta}{2(t-\mu)}}\right) = \frac{2}{\sqrt{\pi}} \sum_{n=0}^{\infty} \frac{\sqrt{\frac{\beta}{2(t-\mu)}}}{2n+1} \prod_{i=1}^n \frac{-\left(\frac{\beta}{2(t-\mu)}\right)}{i}.$$

In this manner, the risk function calculated iteratively is determined as

$$\begin{aligned} h(t) &= \frac{\left(\frac{\beta}{2\pi}\right)^{\frac{1}{2}} \frac{1}{(t-\mu)^{\frac{3}{2}}} \exp\left(-\frac{\beta}{2(t-\mu)}\right)}{\frac{2}{\sqrt{\pi}} \sum_{n=0}^{\infty} \frac{\sqrt{\frac{\beta}{2(t-\mu)}}}{2n+1} \prod_{i=1}^n \frac{-\left(\frac{\beta}{2(t-\mu)}\right)}{i}}; \quad t > 0, \mu > 0, \beta > 0. \end{aligned} \quad (3.4)$$

Another manner in which to express the denominator of the risk function is in terms of the standard normal distribution function, where the only difference is its scale and one translation (Ayant and Borg, 1974).

Consequently,

$$\Phi(t) = \frac{1}{2} \left[1 + \operatorname{erf}\left(\frac{t}{\sqrt{2}}\right) \right] \Rightarrow 2\Phi(t) - 1 = \operatorname{erf}\left(\frac{t}{\sqrt{2}}\right), \quad \operatorname{erf}(t) = 2\Phi(t\sqrt{2}) - 1. \quad (3.5)$$

Therefore, considering the expression (3.5), the risk function (3.4) is rewritten as:

$$h(t) = \frac{\left(\frac{\beta}{2\pi}\right)^{\frac{1}{2}} \frac{1}{(t-\mu)^{\frac{3}{2}}} \exp\left(-\frac{\beta}{2(t-\mu)}\right)}{2\Phi\left(\sqrt{\frac{\beta}{2(t-\mu)}}\right) - 1}; \quad t > 0, \mu > 0, \beta > 0. \quad (3.6)$$

Figure 2 shows some graphs related to the risk function (3.6) of the Lévy distribution with different values of the parameter μ , where the scale parameter β has been fixed in each graph. Subsequently, other functions are shown, but under the standard Lévy model, which is characterized by having as a location parameter $\mu = 0$.

4. Estimation via the maximum likelihood method

Let $T_1, T_2, \dots, T_n$ be a sample of independent random variables with censorship drawn from a population with a Lévy distribution and a probability density function (2.1); then, the observed random

Figure 2: Risk functions of the Lévy model.

sample of times has the form $(t_1, \delta_1), (t_2, \delta_2), \dots, (t_n, \delta_n)$. From the n realizations of the (t_i, δ_i) random pair, the log-likelihood function for survival times with censorship on the right is:

$$\begin{aligned}
 l(\mu, \beta \mid t_1, t_2, \dots, t_n) &= \sum_{i=1}^n \delta_i \log \left\{ \left(\frac{\beta}{2\pi} \right)^{\frac{1}{2}} \frac{1}{(t_i - \mu)^{\frac{3}{2}}} \exp \left(-\frac{\beta}{2(t_i - \mu)} \right) \right\} + \sum_{i=1}^n (1 - \delta_i) \log \left\{ 1 - \operatorname{erfc} \left(\sqrt{\frac{\beta}{2(t_i - \mu)}} \right) \right\} \\
 &= \sum_{i=1}^n \delta_i \log \left\{ \frac{\left(\frac{\beta}{2\pi} \right)^{\frac{1}{2}} \frac{1}{(t_i - \mu)^{\frac{3}{2}}} \exp \left(-\frac{\beta}{2(t_i - \mu)} \right)}{1 - \operatorname{erfc} \left(\sqrt{\frac{\beta}{2(t_i - \mu)}} \right)} \right\} + \sum_{i=1}^n \log \left\{ 1 - \operatorname{erfc} \left(\sqrt{\frac{\beta}{2(t_i - \mu)}} \right) \right\}. \quad (4.1)
 \end{aligned}$$

It can be observed in (4.1) that the log-likelihood function has a complex structure that produce derivatives (Annexed) that are difficult to address. Therefore the use of some numerical approximation through iterative methods such as the Newton-Raphson algorithm are suggested to obtain a solution to the problem.

5. Simulation study

A simulation study was performed to evaluate and model different cases that may arise during the analysis, considering the existence of censorship in the estimation process. It was decided to perform $N = 1000$ simulations, taking into account four sample sizes ($n = 60, 100, 500, 1000$) and three percentages of censorship ($C^+ = 10\%, 30\%, 55\%$). The parameters of the censored samples were calculated so that the mentioned censorship fractions C^+ were obtained; using the following expression:

$$C^+ = P(T > C) = \int_{\mu^*}^{\infty} \int_c^{\infty} f_{C,T}(c, t) dt dc = \int_{\mu^*}^{\infty} f_C(c) [1 - F_T(c)] dc,$$

where $t > 0$ and $c > \mu^*$.

The lower limit μ^* corresponds to the location parameter of the censored samples. The specified functions belong to the same family of distributions and only differ in terms of the values assigned to their parameters, depending on if the data are censored. To solve the system of equations, the algorithm of the “bisection method” of the *NLRoot* library was used in R Core Team (2007) statistical software.

In the particular case of the standard Lévy distribution, it was necessary to approximate the following probability:

$$C^+ = \int_0^\infty \left(\frac{\beta^*}{2\pi}\right)^{\frac{1}{2}} \frac{1}{c^{\frac{3}{2}}} \exp\left(-\frac{\beta^*}{2c}\right) \times \left[1 - \operatorname{erfc}\left(\sqrt{\frac{\beta}{2c}}\right)\right] dc \quad (5.1)$$

considering that $c > 0, \beta > 0$ (uncensored samples) and $\beta^* > 0$ (censored samples).

The random samples $T_i, i = 1, 2, \dots, n$ with the Lévy distribution were generated from standard normal random variables Z_i by applying the following transformation:

$$T_i = \beta \frac{1}{Z_i^2} + \mu$$

therefore $T_i \sim \text{Lévy}(\mu, \beta)$.

The previous step was repeated for the censored random variables C_i based on the parameters μ^* and β^* obtained in equation (5.1). Once the previous random variables T_i and $C_i, i = 1, 2, \dots, n$, were generated, a vector of values was obtained $t_i = \min(T_i, C_i)$. Finally, 12 simulation scenarios were considered. With each generated sample, the maximum likelihood method was applied based on the expression (4.1) to estimate parameters using the libraries *maxLik* and *optimx* in the R Core Team (2007) (Henningsen and Toomet, 2010; Nash and Varadhan, 2011).

A random sample from the 1,000 simulated samples was chosen considering each scenario. With this sample, the survival function, denoted by $S(t)$, was obtained. It was estimated using a “parametric alternative” that uses different families of distributions and a “nonparametric alternative” of the Kaplan-Meier estimator.

Considering the asymptotic normality of the maximum likelihood estimators, the $100(1 - \alpha)\%$ confidence intervals (CIs) of the survival function for the Kaplan-Meier estimator at each fixed time t_i can be calculated as indicated below. These expressions are also include CIs with transformations since it is important to calculate them to symmetrize the distribution of any parameter, stabilize the variance (Ramirez, 2010).

From the parametric approach, the CIs for the survival function $S(t)$ are not fully developed, since the literature only reports intervals when there are survival times Exponential or Weibull (see the works elaborated by Thomas and Grunkemeier (1975), Ramirez (2010)). Consequently, what is proposed is to construct confidence limits for $S(t)$, using the Bootstrap method. This method does not make use of any assumption about the form of the function $S(t)$, on the contrary, it allows each observation to contribute to the construction of the CIs.

The confidence limits of the survival function $S(t)$ of the Lévy model were calculated via the following steps (Akritas, 1986):

- From the sample that was chosen to estimate the survival function, a bootstrap sample was selected $(t_1^*, \delta_1^*; t_2^*, \delta_2^*; \dots; t_n^*, \delta_n^*)$ with replacement.

- Then, the set of artificial data (bootstrap sample) was denoted datos^* ; these data were used to estimate the parameters via the maximum likelihood method.
- Based on the previous estimate, it was obtained $\hat{S}(t_i^*) = \hat{S}(\text{datos}^*)$.
- The previous steps were repeated $B = 3000$ times, independently, until obtaining the set of estimates $\hat{S}(t_i^{*1}), \hat{S}(t_i^{*2}), \dots, \hat{S}(t_i^{*3000})$.
- Then, it was possible to obtain a distribution of the bootstrap estimates for each time t_i .
- For each distribution, the confidence limits were calculated using a method based on the frequency distribution itself, called the “percentile method”. In this method, the confidence limits $100(1 - \alpha)\%$ for a parameter are precisely the two values that contain the central $100(1 - \alpha)\%$ of the 3,000 bootstrap estimates obtained from the original sample. The calculated percentiles were 2.5 and 97.5.
- Based on all the percentiles, the lower and upper confidence limits were drawn, which turned out to be around the respective parametric survival curve, which was accompanied by the function estimated via Kaplan-Meier and some CIs calculated in previous steps.

In addition, a comparison was also made between the survival function $S(t)$ of the standard Lévy distribution, using the true value of the parameter β , and the survival functions $\hat{S}(t)$ of the model for each simulation scenario, using estimates obtained from the parameter, this was done by taking a random sample of the total simulated samples ($N = 1000$) in each scenario. The previous procedure was repeated taking into account the survival function of the Kaplan-Meier estimator and the estimated functions of the standard Lévy model.

To calculate these differences, the following expression was used:

$$\int \|\hat{S}(t) - S(t)\| dt. \quad (5.2)$$

The integral (5.2) was evaluated between the minimum and maximum of the simulated random sample.

5.1. Results

Table 1 indicates that the estimator of the scale parameter β , which characterizes the distribution, turned out to be slightly sensitive to changes in the sample when a large amount of censored information was included. For instance, at 55%, having any size n of those specified, this result was evidenced in the mean square error (MSE) obtained, which in turn exhibited a minimum tendency and standard errors close to zero. The CIs lengths presented similar values under the same sample size and different censorship. The number of times that the true parameter value was included always exceeded the 95% confidence level and the PC were never altered.

According to the values obtained when taking the differences between the survival functions, it was observed that, under any simulation scenario, the two curves approximated inadequately when the sample size increased, especially when it was $n = 500$ and $n = 1000$, respectively. The percentage of censorship did not influence the differences of these functions (Table 1). However, it is observed that as the percentage of censorship increases, the difference is greater between the survival curves of the standard Lévy model and the function obtained from the Kaplan-Meier estimator (Table 1).

Table 1: Estimation of the β parameter with its different performance measures, standard Lévy distribution

$\beta = 0.05$									
Cens	β^*	n	$E(\hat{\beta})$	$SE(\hat{\beta})$	$MSE(\hat{\beta})$	$ALI_{\hat{\beta}}$	$PC_{\hat{\beta}}$	$\int \ \hat{S}(t) - S(t)\ dt$	$\int \ \hat{S}(t) - S(t)_{KM}\ dt$
10%	1.993	60	0.05200	0.00949	0.00009	0.03721	0.943	0.03146	1.5623
		100	0.05121	0.00724	0.00006	0.02839	0.948	0.02389	1.4631
		500	0.05015	0.00317	0.00001	0.01243	0.949	0.00278	1.3650
		1000	0.05003	0.00224	0.00000	0.00877	0.950	0.00168	1.2694
30%	0.193	60	0.05129	0.00939	0.00008	0.03681	0.961	0.00515	1.6872
		100	0.05087	0.00721	0.00005	0.02828	0.958	0.02296	1.5817
		500	0.05020	0.00318	0.00001	0.01247	0.948	0.00652	1.4809
		1000	0.05003	0.00224	0.00000	0.00879	0.951	0.00118	1.3870
55%	0.036	60	0.05112	0.00939	0.00009	0.03866	0.940	0.03413	2.6871
		100	0.05107	0.00763	0.00005	0.02989	0.941	0.03246	2.3874
		500	0.05023	0.00335	0.00001	0.01312	0.946	0.00296	2.2277
		1000	0.05007	0.00236	0.00000	0.00924	0.946	0.00207	2.2821

β^* = parameter of random samples censored; SE = Standard error; MSE = mean square error; ALI = average length of the confidence interval; PC = probability of coverage of the confidence interval; $\hat{S}(t)$ = estimated $S(t)$; $S(t)$ = survival function of the standard Lévy model; $S(t)_{KM}$ = Kaplan-Meier survival function.

Figure 3 shows that the chances of survival are high at the beginning of the time interval. This behavior was repeated in all simulation scenarios. Under the standard Lévy distribution, it is possible to calculate the survival probabilities for censored times. The lengths of the CIs for the Kaplan-Meier estimator increased according to the increase in censorship and moderate sample sizes ($n = 60$, $n = 100$). With respect to the survival function of the model, the observed CIs had very similar amplitudes, and in all cases, these amplitudes were lower than CIs obtained using the Kaplan-Meier estimator.

6. Colon cancer data

To illustrate the proposed method, a set of data provided by a Health Services Provider Institution in Colombia was used. The data file contains basic information regarding individuals treated during oncology consultations in the period between October 2003 and January 2016. The information includes sociodemographic information, care type, and procedures performed (chemotherapy, biopsy, and radiotherapy) on patients from different regions of Colombia diagnosed with CC. The information includes 88 patients (women and men) who entered the study gradually as they underwent consultation. The event of interest was “death associated with a failure attributable to CC”, which occurred for 79 of the individuals. The remaining individuals experienced diverse situations that classified them as *alive* and the event of interest did not occur because the patient left the study or experienced another cause other than the one of interest.

It is worth mentioning that the times included in the analysis correspond to information about patients treated with different procedures, considering their condition, in addition to patients with different stages of CC (stages II, III, and IV) since the available information was quite restricted. To illustrate the methodological proposal, it was decided via consultation with experts to combine the information about all patients to obtain preliminary results regarding the overall survival of the disease.

6.1. Exploratory data analysis

Figure 4 presents the frequency distribution and the boxplot of the time variable. A larger observation than the rest of the data is shown, reflecting great asymmetry and therefore a histogram with a long

Figure 3: Survival functions of the standard model Lévy and the Kaplan-Meier estimator, taking into account the censorship and the size of the sample.

tail to the right. The shortest survival time found among the patients was 0.24 months, however, the longest-surviving patient reached 147.5 months (4425 days). The censorship percentage exhibited by the dataset was 10.2%.

6.2. Estimation of parameters and model selection

Table 2 reports the results of the estimations via the maximum likelihood method of candidate distributions. Based on the Akaike information criterion (AIC), Bayesian information criterion (BIC)

Figure 4: Distribution of survival times in months of patients with colon cancer. (a) Histogram of survival times; (b) Box-plot of survival times.

Table 2: Estimation of the parameters for the candidate distributions

	Distribution	Parameter	Estimation	SE	CI (95%)	$-2 \times \log [L(\cdot)]$	AIC	BIC	AICc
Common distribution	Exponential	β	0.122	0.013	(0.095, 0.149)	490.2	492.2	494.7	492.3
		α	0.877	0.117	(0.646, 1.108)	489.2	493.2	498.2	493.4
	Gamma	β	0.106	0.020	(0.067, 0.145)				
		α	0.853	0.063	(0.728, 0.978)	485.1	489.1	494.0	489.2
	Weibull	β	0.133	0.018	(0.105, 0.182)				
		α	1.440	0.128	(1.188, 1.691)	472.3	476.3	481.3	476.5
Unconventional distribution	Lognormal	μ	1.440	0.128	(1.188, 1.691)				
		σ	1.176	0.093	(0.992, 1.359)				
	Lévy	μ	-0.111	0.108	(-0.324, 0.101)	161.1	326.3	331.3	324.3
		β	2.365	0.487	(1.415, 3.316)				
	Standard Lévy	$\mu = 0$							
		β	1.994	0.301	(1.403, 2.584)	323.8	325.8	328.2	325.8

SE = Standard error; CI = confidence interval; AIC = Akaike information criterion; BIC = Bayesian information criterion; AICc = corrected AIC.

and Akaike information criterion corrected (AICc), it was found that from the candidate distributions, those that presented a better fit were the standard Lévy distribution (325.8, 328.2, 325.8) and the Lognormal distribution (476.3, 481.3, 476.5). When the two parameters of the model $\text{Lévy}(\mu, \beta)$ were considered unknown, it was found that the AIC was less than the $\text{Lévy}(\mu = 0, \beta)$; however, for the former, its location parameter μ was not significant.

The estimation of parameters makes it possible to compare the empirical frequency distribution and theoretical distribution functions. The empirical frequency distribution was constructed using the Kaplan-Meier method due to the presence of censored data. It was observed in Figure 5 that the exponential and gamma distributions fitted the dataset the least well. In addition, the standard Lévy distribution curve exhibited a different behavior compared to the other distributions. This is because this distribution is more flexible and allows modeling extreme data (147.5 months) located in the right tail of the curve.

6.3. Estimation of survival functions

With the purpose of comparing the two curves of the $S(t)$ function, confidence limits were added; these provide information about the accuracy of the survival estimation for patients with CC. Under

Figure 5: Adjustment of the candidate parametric models. (a) Histogram versus probability distributions; (b) Adjustment of the proposed distributions; (c) Empirical distribution function versus the estimated distribution function.

this condition and based on the standard Lévy distribution, the confidence limits under this model exhibited shorter lengths, especially at the end; however, this behavior was similar to the Kaplan-Meier estimator when its limits were drawn using the transformation $\log(-\log(\hat{S}(t)))$ (Figure 6).

6.4. Estimation of the risk function

One of the main functions of the survival model is the risk function. With this, the manner in which the instantaneous death rate changes in patients with CC over time. Figure 7 shows that that from the distributions with the best fit, the risk function $\hat{h}(t)$ of each one tends to behave in the same manner at the beginning of the time interval. The risk is pronounced in the case of the standard Lévy distribution with greater intensity for short times and an immediate disposition to death is found as a failure associated with CC from the first moment that the patient's life expectancy begins to be counted.

7. Discussion and conclusion

Usually, in a survival analysis, datasets have observations associated with variables whose natural behavior is asymmetric, and some sample units can exhibit large values in comparison with others. Additionally, it is possible to have censored data on the right that makes the scenario quite complex at the time of the estimation process. To analyze this type of data, the Lévy distribution could be used

Figure 6: Survival functions of the times with censorship, coming from patients with colon cancer, under the parametric and nonparametric (Kaplan-Meier) approach. (a) Standard Lévy distribution; (b) Lognormal distribution.

Figure 7: Risk function for survival times from patients with colon cancer. (a) Standard Lévy distribution; (b) Lognormal distribution.

as a good alternative to common parametric models available in the literature: Exponential, Gamma, Weibull and Lognormal.

In addition, if the assumptions of the parametric models exist, a more robust analysis can be performed in comparison with the non-parametric methods (Kaplan-Meier curves and actuarial methods), since taking into account some assumptions and the selection of a distribution of hypothetical probability for survival times, statistical inference is more accurate, which makes standard deviations smaller when there are no such assumptions (Hashemian *et al.*, 2013).

According to a preliminary analysis regarding the performance of the new distribution, assuming that the location parameter μ is known, it can be concluded from the simulation results that the estimation performance of the parameter β is not affected by the size of the sample since in situations in which the size is relatively small (approximately 60), the estimator tends to behave apparently well in relation to the true value, even when it is subjected to censorship fractions greater than 30%. Significant effects are observed when constructing the confidence limits of the Kaplan-Meier estimator, especially when the censorship is equal to 55%, as with the three interval types (linear or without

transformation, $\log$, $\log(-\log)$), its amplitudes exceed the calculated CI for the $S(t)$ function of the standard Lévy model using the bootstrap method. Solanas and Olivera (1992), mention that the CIs obtained via the bootstrapping method are narrower, enabling informative CIs to be obtained. The previous scenario is highlighted in situations in which the sample is less than $n = 100$, thus generating an increase in the degree of uncertainty regarding the true value of the survival probability.

Usually in the literature, the confidence limits of the survival function are reported for times of interest as the median or some arbitrary time, which differs from the proposal that focuses on building CIs via the percentile method for functions $S(t)$ of the parametric models, applying the bootstrapping method to data with censorship since the expressions that represent them are unknown. However, other authors have performed different procedures, mainly for central tendency measures of the Nelson-Aalen and Kaplan-Meier estimators, based on alternative bootstrapping methods, such as the bootstrap-t parametric, bootstrap-t parametric transformation, root of the ratio log-likelihood parametric bootstrap, percentile parametric bootstrap and bias-corrected parametric bootstrap methods. For instance, Emerson (1982) provided a method to construct nonparametric CIs for the median of the survival distribution using the sign test under different censorship percentages and using the exponential and Weibull distributions to model the distribution of survival times. Reid (1981) proposed two methods to find CIs for the median time using bootstrapping methods. Tomas and Grunkemeier (1975) calculated CIs for survival probabilities with four different sample sizes and censorship percentages using the Weibull distribution to fit the distribution of survival times. Other authors who performed studies related to CIs for median survival time include Jeng and Meeker (2000), Barber and Jennison (1999), Borgan and Liestol (1990), Ole *et al.* (1987), Jennison and Turnbull (1985), Slud *et al.* (1984), Anderson *et al.* (1982), and Brookmeyer and Crowley (1982).

It is important to remember is that this study includes the analytical expression of the risk function of the Lévy model along with its graphical representation, which is similar to the density function $f(t)$ of the model, according to the behavior they exhibit. This behavior is mainly characterized by the intensity that the initial values exhibit in the time interval, considering the values of the scale parameter β , which determines the dispersion in the data, and the location parameter μ that shifts the risk function to the right as it increases. The previous behavior is observed once the function $h(t)$ is deduced from the mathematical expression $h(t) = -(d/dt) \log [S(t)]$, from which a special or nonelementary function is obtained. This function is called the error function (also known as the Gauss error function) is especially used in the probability field, statistics and differential equations. Similarly, other methods to calculate $h(t)$ are shown, expressing the error function in terms of other better-known functions, as the standard normal probability distribution function.

As a special application and motivation for the present study, a set of medical data that contains a few people with very large values of the response over time that they survive the event was considered. From the obtained results, it was detected that the standard distributions are not good candidates to adapt to this type of data. However, after the data were fitted using the Lévy distribution with the location parameter $\mu = 0$ and β unknown, called the standard Lévy, it was concluded that this type of model can be very effective at fitting data. This is because the distribution is more flexible, which in turn allows modeling extreme observations, which are usually located in the right tail of the curve. The assigned probability is different from zero, which admits that some events or group of events still have a nonzero probability of occurring. This condition does not occur with traditional models, which exhibit a relatively fast approach to zero in their density curve, suggesting that all available information was not included when the corresponding fitting was performed. This result implies that if the frequency distribution of the observed data exhibits a marked asymmetry and extreme values, the analyst must necessarily consider the Lévy distribution as a candidate among other commonly

used distributions in the inference procedure, as the predictions of survival times may be inaccurate or barely optimistic or could be based on methodologies unsuitable for the analyzed times, ignoring possible influencing factors. This paper also extends the study of Achcar *et al.* (2013) by considering censored information of survival times and maximum likelihood estimation. The Lévy distribution is a special case of the α -stable family considered by Achcar *et al.* (2013); however, this paper provides a new approach to analyze such data. In contrast to the proposed model, new probability distributions have been introduced for analysis in recent years by many researchers, such as in the studies of Cancho *et al.* (2012), Roman *et al.* (2012), Cordeiro *et al.* (2011), and De Pascoa *et al.* (2011).

Among the limitations of the study, we refer to the low number of patients included. The cancer staging could also be a good alternative for a specific analysis of patients, since a global survival analysis for colon cancer was carried out based on available information. For future research work, other types of asymmetrical distributions with characteristics can be considered flexible functions of density and risk, which favor modeling, when faced with sets of data that present extreme values of importance, such as the Gamma-Kumaraswamy Generalized distribution, which contain a large number of special sub-models that are useful to compare in terms of goodness of fit with the proposed study distribution.

In this sense, the use of these models can provide great benefits for medical research from new procedures of parametric survival analysis, where some units of the sample can have very large values compared to others. This type of methodology can be used for the analysis of times in the area of health as well as in the area of industrial quality control with elements that have a longer duration than the rest of the data.

Acknowledgement

The authors express their sincere thanks to Dr. Jose William Martinez for authorizing the use of cancer data to illustrate the concepts discussed in this article.

Appendix: Annexed

The derivatives of the log-likelihood in relation to the parameters μ and β are expressed by equations (A.1) and (A.2). In the case of the standard Lévy model, equation (A.2) would be used only for the inference process, evaluating $\mu = 0$.

$$\begin{aligned} \frac{\partial l(\mu, \beta \mid t_1, t_2, \dots, t_n)}{\partial \mu} &= \sum_{i=1}^n \delta_i \left\{ -\frac{3}{2(t_i - \mu)} + \frac{\beta}{2(t_i - \mu)^2} \right\} + \sum_{i=1}^n \delta_i \left\{ \frac{\frac{\beta}{\sqrt{\pi}(t_i - \mu)^2} \exp \left[-\left(\frac{\beta}{2(t_i - \mu)} \right) \right]}{1 - \operatorname{erfc} \left(\sqrt{\frac{\beta}{2(t_i - \mu)}} \right)} \right\} \\ &\quad - \sum_{i=1}^n \frac{\frac{\beta}{\sqrt{\pi}(t_i - \mu)^2} \exp \left[-\left(\frac{\beta}{2(t_i - \mu)} \right) \right]}{1 - \operatorname{erfc} \left(\sqrt{\frac{\beta}{2(t_i - \mu)}} \right)} = 0, \end{aligned} \quad (\text{A.1})$$

$$\begin{aligned} \frac{\partial l(\mu, \beta \mid t_1, t_2, \dots, t_n)}{\partial \beta} &= \sum_{i=1}^n \delta_i \left\{ \frac{1}{2\beta} - \frac{1}{2(t_i - \mu)} \right\} + \sum_{i=1}^n \delta_i \left\{ \frac{\frac{1}{\sqrt{\pi}(t_i - \mu)} \exp \left[-\left(\frac{\beta}{2(t_i - \mu)} \right) \right]}{1 - \operatorname{erfc} \left(\sqrt{\frac{\beta}{2(t_i - \mu)}} \right)} \right\} \\ &\quad + \sum_{i=1}^n \frac{\frac{1}{\sqrt{\pi}(t_i - \mu)} \exp \left[-\left(\frac{\beta}{2(t_i - \mu)} \right) \right]}{1 - \operatorname{erfc} \left(\sqrt{\frac{\beta}{2(t_i - \mu)}} \right)} = 0. \end{aligned} \quad (\text{A.2})$$

References

- Achcar JA, Barros EA, Tovar JR, and Mazucheli J (2018). Use of Lévy distribution to analyze longitudinal data with asymmetric distribution and presence of left censored data, *Communications for Statistical Applications and Methods*, **25**, 43–60.
- Achcar JA, Lopes S, Mazucheli J, and Linhares R (2013). A Bayesian approach for stable distributions: some computational aspects, *Open Journal of Statistics*, **3**, 268–277.
- Adler R, Feldman R, and Taqqu M (1998). *A Practical Guide to Heavy Tails: Statistical Techniques and Applications*, Springer Science & Business Media, United States of America.
- Ahsanullah M and Nevzorov VB (2014). Some inferences on the Lévy distribution, *Journal of Statistical Theory and Applications*, **13**, 205–211.
- Akritas M (1986). Bootstrapping the Kaplan-Meier estimator, *American Statistical Association*, **81**, 1032–1038.
- Anderson J, Bernstein L, and Pike M (1982). Confidence intervals for probabilities of survival and quantiles in life-table analysis, *Biometrics*, **38**, 407–416.
- Ayant Y and Borg M (1974). *Funciones Especiales* (pp. 1–242), Madrid, Alhambra.
- Barber S and Jennison C (1999). Symmetric tests confidence intervals for survival probabilities and quantiles of censored data, *Biometrics*, **55**, 430–436.
- Borgan O and Liestol K (1990). A note on confidence intervals and bands for the survival function based on transformations, *Scandinavian Journal of Statistics*, **17**, 35–41.
- Brookmeyer R and Crowley J (1982). A Confidence interval for the median survival time, *Biometrics*, **38**, 29–41.
- Buckle D (1995). Bayesian inference for stable distributions, *Journal of the American Statistical Association*, **90**, 605–613.
- Cancho VG, Louzada F, and Barriga G (2012). The geometric Birnbaum-Saunders regression model with cure rate, *Journal of Statistical Planning and Inference*, **142**, 993–1000.
- Cordeiro GM, Ortega E, and Silva GO (2011). The exponentiated generalized Gamma distribution with application to lifetime data, *Journal of Statistical Computation and Simulation*, **81**, 827–842.
- De Pascoa MA, Ortega EM, and Cordeiro GM (2011). The Kumaraswamy generalized Gamma distribution with application in survival analysis, *Statistical Methodology*, **8**, 411–433.
- Emerson J (1982). Nonparametric Confidence intervals for the median in the presence of right censoring, *Biometrics*, **81**, 17–27.
- Godoy AM (2009). *Introduccion al analisis de supervivencia en R*, Master's thesis, Universidad Nacional Autonoma de Mexico.
- Hashemian AH, Beiranvand B, Rezaei M, and Reissi D (2013). A comparison between Cox regression and parametric methods in analyzing kidney transplant survival, *World Applied Sciences Journal*, **26**, 502–507.
- Henningsen A and Toomet O (2010). A package for maximum likelihood estimation in R, *Computational Statistics*, **3**, 443–458.
- Jeng SL and Meeker W (2000). Approximate confidence interval procedures for type I censored data, *Technometrics*, **42**, 135–148.
- Jennison C and Turnbull B (1985). Repeated confidence intervals for the median, *Biometrika*, **70**, 619–625.
- Martínez E and Achcar J (2014). Bayesian bivariate generalized Lindley model for survival data with a cure fraction, *Computer Methods and Programs in Biomedicine*, **117**, 145–157.
- Nash JC and Varadhan R (2011). Unifying optimization algorithms to aid software system users:

- optimx for R, *Journal of Statistical Software*, **9**, 1–14.
- Nolan JP (2009). *Stable Distributions: Models for Heavy-Tailed Data*, University of California, United States of America.
- Ole B, Borgan O, and Liestol K (1987). Confidence intervals and confidence bands for the cumulative hazard rate function and their small sample properties, *Scandinavian Journal of Statistics*, **14**, 221–233.
- Ramirez J (2010). *Comparacion de intervalos de confianza para la funcion de supervivencia, en un tiempo de interes, con censura a derecha* (Master's thesis), Universidad Nacional de Colombia Sede Medellin.
- R Core Team (2007). *R: A Language and environment for statistical computing* Vienna Austria: R Foundation for Statistical Computing.
- Reid N (1981). Estimating the median survival time, *Biometrika*, **68**, 601–608.
- Roman M, Louzada F, Cancho VG, and Leite JG (2012). A new long-term survival distribution for cancer data, *Journal of Data Science*, **10**, 241–258.
- Samorodnitsky G and Taqqu M (1994). *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*, Chapman and Hall/CRC, London.
- Slud E, Byar D, and Green S (1984). A Comparison of reflected versus test-based confidence intervals for the median survival time, *Biometrics*, **40**, 587–600.
- Solanas A and Olivera V (1992). Bootstrap: fundamentos e introduccion a sus aplicaciones, *Anuario de Psicologia*, **55**, 143–154.
- Thomas D and Grunkemeier G (1975). Confidence interval estimation of survival probabilities for censored data, *Journal of American Statistical Association*, **70**, 865–871.
- Yu Z, Anh V, Wang Y, Mao D, and Wanliss J (2010). Modeling and simulation of the horizontal component of the geomagnetic field by fractional stochastic differential equations in conjunction with empirical mode decomposition, *Journal of Geophysical Research: Space Physics*, **115**(A10).

Received February 8, 2019; Revised July 12, 2019; Accepted August 27, 2019