Journal of Data Science **10**(2012), 241-258

A New Long-Term Survival Distribution for Cancer Data

Mari Roman[1], Francisco Louzada[2*], Vicente G. Cancho[2] and José G. Leite[1]
[1] *Universidade Federal de São Carlos and* [2] *Universidade de São Paulo*

Abstract: In this paper we propose a new three-parameters lifetime distribution with decreasing hazard function, the long-term exponential geometric distribution. The new distribution arises on latent competing risks scenarios, where the lifetime associated with a particular risk is not observable, rather we observe only the minimum lifetime value among all risks, and there is presence of long-term survival. The properties of the proposed distribution are discussed, including its probability density function and explicit algebraic formulas for its survival and hazard functions, order statistics, Bonferroni function and the Lorenz curve. The parameter estimation is based on the usual maximum likelihood approach. We compare the new distribution with its particular case, the long-term exponential distribution, as well as with the long-term Weibull distribution on two real datasets, observing its potential and competitiveness in comparison with an usual lifetime distribution.

Key words: Censored data, exponential geometric distribution, latent competing risks, long-term survivals.

1. Introduction

Survival data in presence of competing risks arise in several areas, such as public health, actuarial science, biomedical studies, demography and industrial reliability. In the classical competing risks scenarios the lifetime associated with a particular risk is not observable, rather we observe only the minimum lifetime value among all risks. Simplistically, in reliability, we observe only the minimum component lifetime of a series system. That is, the observable quantities for each component are the minimum lifetime value to failure among all risks, and the cause of failure. Full statistical procedures and extensive literature are available to deal with these problems and interested readers can refer to Cox and Oakes (1984), Crowder *et al.* (1991) and Lawless (2003).

*Corresponding author.

A first difficulty arises if the risks are latent in the sense that there is no information about which factor was responsible for the component failure (or individual death), which can be often observed in field data. We call these latent competing risk data. On many occasions this information is not available or it is impossible that the true cause of the failure is specified by an expert. In reliability, the components can be totally destroyed in the experiment. Further, the true cause of failure can be masked from our view. In modular systems, the need to keep a system running means that a module that contains many components can be replaced without the identification of the exact failing component. Goetghebeur and Ryan (1995) addressed the problem of assessing covariate effects based on a semi-parametric proportional hazards structure for each failure type when the failure type is unknown for some individuals. Reiser *et al.* (1995) considered statistical procedures for analyzing masked data, but their procedure can not be applied when all observations have an unknown cause of failure. Adamidis and Loukas (1998) proposed a compounding distribution, the exponential geometric (EG) distribution, which properly accommodates survival data in presence of latent competing risks. Louzada-Neto (1999) proposed a polyhazard model to deal with lifetime data associated with latent competing risks. Lu and Tsiatis (2001) presented a multiple imputation method for estimating regression coefficients for risk modeling with missing cause of failure. A comparison of two partial likelihood approaches for risk modeling with missing cause of failure is presented in Lu and Tsiatis (2005). Kus (2007) proposed another compounding distribution which properly accommodates survival data in presence of latent competing risks.

A second difficulty arises if a part of the population is not susceptible to the event of interest. For instance, in clinical studies a population can respond favorably to a treatment, being considered cured. Models which consider that part of the population is cured have been widely developed and are usually called long term survival models. Perhaps the most popular type of cure rate model is the mixture model introduced by Boag (1949) and Berkson and Gage (1952). In this model, it is assumed that a certain proportion of the patients, say p, are cured, in the sense that they do not present the event of interest during a long period of time and can be seen as being immune to the cause of failure under study. Later on, Farewell (1977), Farewell (1982), Greenhouse and Wolf (1984), Ghitany and Maller (1992), Ghitany *et al.* (1994), Maller and Zhou (1995), Mackenzie (1996), Chen and Ibrahim (2001), Pons and Lemdani (2003), Perperoglou *et al.* (2007), Cancho *et al.* (2009) and Perdoná and Louzada-Neto (2011) have considered long term mixture modeling.

In this paper, we propose a new distribution family conceived inside a latent competing risk scenario with long-term survival, where there is no information about which factor was responsible for the component failure (or individual

death), only the minimum lifetime value among all risks is observed, and a part of the population is not susceptible to the event of interest. Our distribution is fully based on the exponential geometric distribution (Adamidis and Loukas, 1998) on a long term mixture modeling structure. Then, hereafter we shall call it the long term exponential geometric distribution or simplistically the LEG distribution.

The properties of the proposed distribution are discussed, including its probability density function and explicit algebraic formulas for its survival and hazard functions, order statistics, Bonferroni function and the Lorenz curve.

The paper is organized as follows. In Section 2, we describe the genesis for the LEG distribution and present explicit algebraic formulas for its probability density, survival and hazard functions. We also present some proprieties of its hazard function. In Section 3, we derive the kth order statistics, the Bonferroni function and the Lorenz curve. In Section 4 we present the inferential procedure based on maximum likelihood approach. In Section 5 we compare the LEG distribution with its particular case, as well as with the Weibull distribution (an usual lifetime distribution) on two real datasets. Some final comments in Section 6 conclude the paper.

2. Model Formulation

In survival studies, a part of the population may be not susceptible to the event of interest. According to Maller and Zhou (1996), it seems adequate to consider a two components mixture model, in the sense that one component represents the failure or survival time of susceptible individuals to a certain event (in risk individuals - IR), while the other component represents the survival times of the not susceptible individuals to the event (out of risk individuals - OR), allowing infinite survival times. An individual belongs to one group or another with certain probability.

This class of models has been widely used in medicine, especially for data analysis of cancer clinical trials. In general, we observe the time to occurrence of death, or the time until the outbreak of a disease, but in the presence of a significant proportion of cured or immune patients. For instance, consider leukemia, which is a malignant disease of the blood-forming organs. Due to improvements in treatment over the past decades, the leukemia cure rate may reach high proportions (Kersey *et al.*, 1987). For acute lymphoblastic leukemia, which is a common fatal childhood cancer, the cure rate may reach 90% in the near future[1].

Then, the model formulation is described as following. Let Y be a random variable that represents the time until the occurrence of a event of interest, and p

[1] http://www.medicalnewstoday.com/articles/36106.php

be the probability of an individual to belong to the group OR. Considering a population in which exists the possibility of cure, the improper population survival function is given by Maller and Zhou (1996), $S(y) = pS_{OR}(y) + (1-p)S_{IR}(y)$, where $S_{OR}(y)$ and $S_{IR}(y)$ are the survival functions of the individuals OR and IR, respectively. Following Maller and Zhou (1996), the event of interest shall not occur in the group OR, that is, their failure times are infinite, so $S_{OR}(y) = P(Y > y|OR) = 1$, $\forall y \geq 0$. Then, we can rewrite $S(y)$ as,

$$S(y) = p + (1-p)S_{IR}(y). \tag{1}$$

All individuals IR shall present the event of interesting at same time, that is, $\lim_{y\to\infty} S_{IR}(y) = 0$. Consequently, we have $\lim_{y\to\infty} S(y) = p$, and therefore the survival function (not conditional) is improper and its limit corresponds to the individual proportion OR. Also, the event of interest may be caused by an unknown competing cause leading to the so called latent competing risk scenarios (Louzada-Neto, 1999). So, let M denote the unobservable number of causes of the event of interest with probability mass function

$$P(M = m), \tag{2}$$

for $m = 1, 2, \cdots, M$, with M on a infinite range. Let T_m, $m = 1, \cdots, M$, denote the time for the j^{th} cause to produce the event of interest. We assume that, independently but conditional on M, the T_j are independent and identically distributed with survival function $S_0(t)$. And we only observe the random variable given by $Y = \min(T_1, T_2, \cdots, T_M)$. Under this setup, the surviving function for an individual IR is given by

$$S_{IR}(y) = \sum_{m=1}^{\infty} S_0(y)^m P[M = m]. \tag{3}$$

Following Adamidis and Loukas (1998), with M geometrically distributed and T exponentially distributed, we consider $S_{IR}(y)$ define as

$$S_{IR}(y) = \frac{(1-\theta)e^{-\lambda y}}{1-\theta e^{-\lambda y}}, \tag{4}$$

which is the survival function of an EG distributed random variable.

Considering the EG survival function (4) and the definition given in (1), the improper survival function (5) of a LEG distributed nonnegative random variable, Y, denoting the lifetime of a component in some population is given by,

$$S(y) = \frac{p + (1-\theta-p)e^{-\lambda y}}{1-\theta e^{-\lambda y}}, \tag{5}$$

where, $y > 0$, $\lambda > 0$, $\alpha > 0$, $0 < \theta < 1$, and $0 < p < 1$.

Its pdf is directly obtained by considering $f(y) = -dS(y)/dy$, that is, it is given by

$$f(y) = \frac{\lambda e^{-\lambda y}(1 - \theta - p + p\theta)}{(1 - \theta e^{-\lambda y})^2}, \tag{6}$$

where, λ is scale parameter, θ is shape parameter and p is the long-term parameter. Figure 1 shows the LEG pdf and survival function for some values of the vector $\phi = (\lambda, \theta)$, with $p = 0, 0.25, 0.50$ and $\theta = 0.001, 0.1, 0.5, 0.75, 0.99$. Without loss of generality, we fixed $\lambda = 1$.

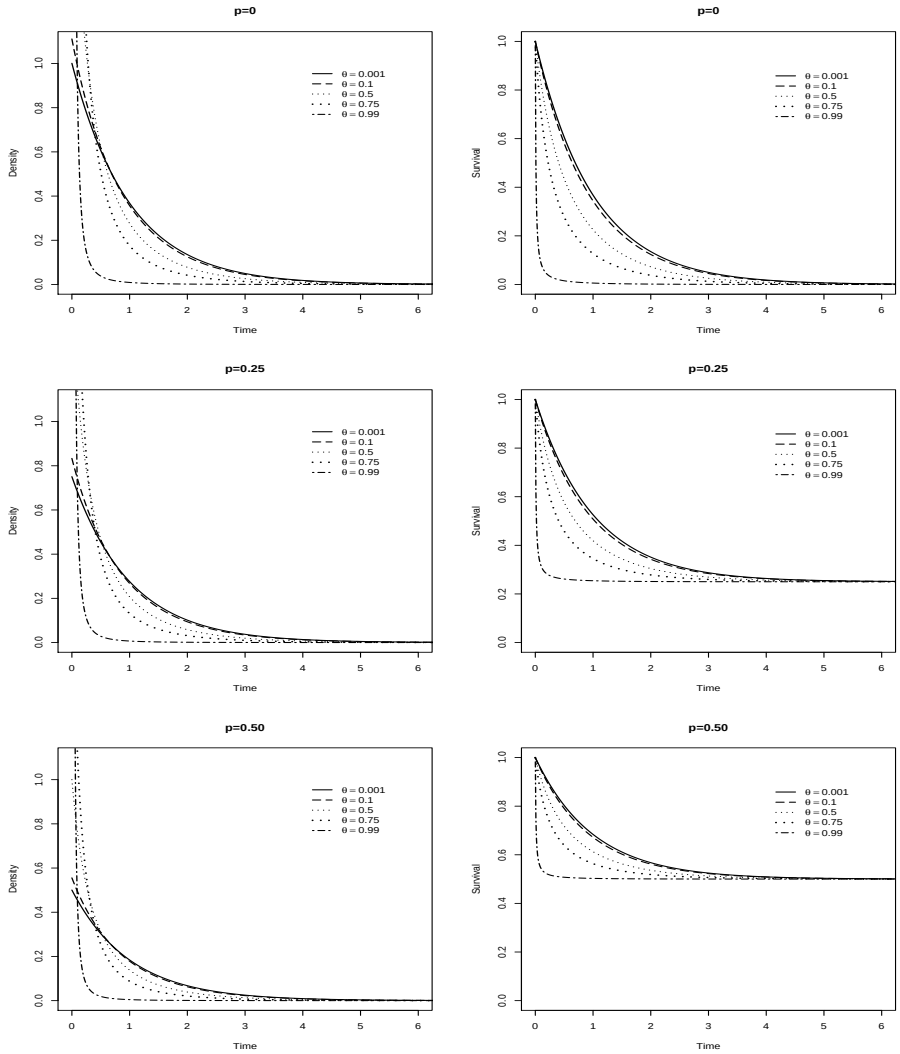

Figure 1: LEG pdf and survival function for selected values of parameters and fixed $\lambda = 1$

The quantile function, obtained by inverting the distribution function, $F(y) = 1 - S(y)$, defined in $[0, 1]$, is given by

$$Q(u) = \frac{1}{\lambda} \ln \left(\frac{1 - p - u\theta}{1 - p - u} \right), \tag{7}$$

where u has the uniform distribution $U(0, 1)$. Then, we may simulate a variable LEG distributed random variable by considering the inverse transformation of the cumulated function given in (7).

From (7), the median is given by

$$\widetilde{Y} = Q(0.5) = \frac{1}{\lambda} \ln \left(\frac{1 - p - 0.5\theta}{1 - p - 0.5} \right). \tag{8}$$

From (6) and (5) it is easy to verify that the hazard function is given

$$h(y) = \frac{(1 - \theta - p + p\theta)\lambda e^{-\lambda y}}{(1 - \theta e^{-\lambda y})(p + (1 - \theta - p)e^{-\lambda y})}, \tag{9}$$

for which the initial value is finite and given by $\lim_{Y \to 0} h(y)$, that is,

$$h(0) = \lambda(1 - \theta - p - p\theta)/(1 - \theta)^2.$$

The hazard function (9) is decreasing. According with a theorem proposed by Glaser (1980), let $g(y) = 1/h(y) = (1 - F(y))/f(y)$ be the general reciprocal hazard rate, for which the derivate is given by $h'(y) = h(y)\vartheta(y) - 1$, where $\vartheta(y) = -f'(y)/f(y)$, and the shape of $h(y)$ depends on the behavior of ϑ'. If $\vartheta'(y) > 0$ ($\vartheta'(y) < 0$), for all $y > 0$, implies an increasing (decreasing) hazard rate. If $\vartheta'(y)$ change the sign, with $\vartheta'(y) = 0$ for some $y_0 > 0$, and $\vartheta'(y) < 0$ for $y < y_0$ and $\vartheta'(y) > 0$ for $y > y_0$, we have an increasing hazard hate if $\lim_{y \to 0} f(y) = 0$ and a $\cup$-shaped hazard rate if $\lim_{y \to 0} f(y) = \infty$. Similarly, we obtain decreasing and $\cap$-shaped hazard rate if the inequalities in the preceding conditions are reserved. Then, considering the LEG distribution given in (6), $\vartheta'(y) = 2\lambda^2\theta e^{-\lambda y}/\left(1 - \theta e^{-\lambda y}\right) > 0$ for all $t > 0$, implying in a decreasing hazard function. Figure 2 corroborates the above results by showing some of the possible shapes of the hazard function for fixed $\lambda = 1$, and some combinations of p and θ values.

3. Some Properties of the LEG Distribution

Moments of order statistics play an important role in quality control testing and reliability, where a practitioner needs to predict the failure of future items based on the times of a few early failures. These predictors are often based on moments of order statistics. Considering the LEG distribution, the kth order statistic is given as follows.

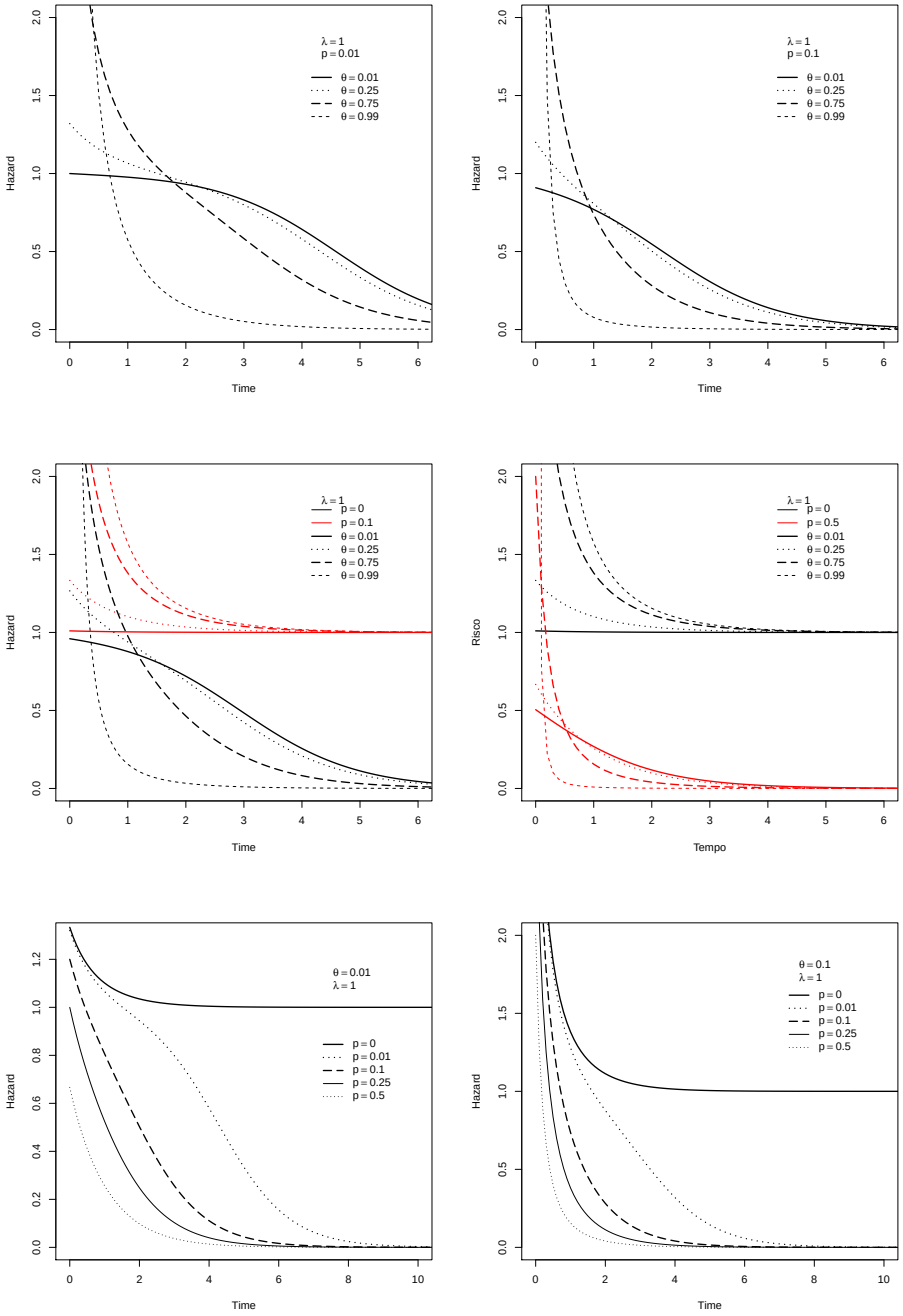

Figure 2: Hazard rate function of LEG distribution for selected values of the parameters. Fixed $\lambda = 1$

Proposition 3.1 Let $Y_1, Y_2, \cdots, Y_n$ be iid random variables such that Y_k follows $LEG\ (p, \lambda, \theta)$ for $k = 1, 2, \cdots, n$. The pdf of the kth order statistic, say $Y_{k:n}$, is given by (for $y > 0$)

$$f_{k:n}(y) = g_{k:n}(y)(1-p)^k \left(\frac{p(1-e^{-\lambda y})}{e^{-\lambda y}} + 1 \right),$$

where $g_{k:n}$ is order statistic of EG with parameters λ and θ.

Proof 3.1 We now derive an explicit expression for the density of the ith order statistic $Y_{k:n}$, say $f_{k:n}(y)$, in a random sample of size n from the LEG distribution. It is well known that

$$f_{k:n}(y) = \frac{1}{B(k, n-k+1)} f(y)(F(y))^{k-1}(S(y))^{n-k}, \tag{10}$$

where, $B(k, n-k+1) = \dfrac{(n-k)!(k-1)!}{n!}$.

Using the definition, we have,

$$\begin{aligned}
f_{k:n}(y) &= \frac{1}{B(k, n-k+1)} f(y)(F(y))^{k-1}(S(y))^{n-k} \\
&= \frac{\lambda e^{-\lambda y}}{B(k, n-k+1)} \frac{(1-\theta-p+p\theta)}{(1-\theta e^{-\lambda y})^2} \left(\frac{(p-1)e^{-\lambda y} + 1 - p}{1-\theta e^{-\lambda y}} \right)^{k-1} \\
&\quad \times \left(1 - \frac{(p-1)e^{-\lambda y} + 1 - p}{1-\theta e^{-\lambda y}} \right)^{n-k} \\
&= \frac{\lambda e^{-\lambda y}(1-\theta-p+p\theta)}{B(k, n-k-1)\,(1-\theta e^{-\lambda y})^{n+1}} \left((p-1)e^{-\lambda y} + 1 - p \right)^{k-1} \\
&\quad \times \left(p(1-e^{-\lambda y}) + e^{-\lambda y}(1-\theta) \right)^{n-k} \\
&= \frac{\lambda(1-\theta)e^{-\lambda y}}{B(k, n-k+1)\,(1-\theta e^{-\lambda y})^{n+1}} \left(1 - e^{-\lambda y} \right)^{k-1} (1-p)^k \\
&\quad \times \left(p(1-e^{-\lambda y}) + e^{-\lambda y}(1-\theta) \right)^{n-k} \\
&= \frac{\lambda(1-\theta)e^{-\lambda y} \left((1-\theta)e^{-\lambda y} \right)^{n-k} \left(1 - e^{-\lambda y} \right)^{k-1}}{B(k, n-k+1)\,(1-\theta e^{-\lambda y})^{n+1}} (1-p)^k \left(\frac{p(1-e^{-\lambda y})}{e^{-\lambda y}(1-\theta)} + 1 \right)^{n-k}.
\end{aligned}$$

But, as given in the Appendix A, the order statistic of a EG distribution is given by,

$$g_{k:n}(y) = \frac{1}{B(k, n-k+1)} \frac{\lambda(1-\theta)e^{-\lambda y}}{(1-\theta e^{-\lambda y})^{n+1}} \left(1 - e^{-\lambda y} \right)^{k-1} \left((1-\theta)e^{-\lambda y} \right)^{n-k}. \tag{11}$$

Then,

$$f_{k:n}(y) = g_{k:n}(y)(1-p)^k \left(\frac{p(1-e^{-\lambda y})}{e^{-\lambda y}(1-\theta)} + 1 \right)^{n-k}, \tag{12}$$

concluding the proof. □

The Bonferroni and Lorenz curves and the Gini index have many applications not only in economics to study income and poverty, but also in other fields such as reliability, medicine and insurance.

Proposition 3.2 The Bonferroni curve of the distribution function $F(y)$ of a LEG distribution is given by

$$B_F[F(y)] = \frac{\lambda(1-\theta-p+p\theta)(1-\theta e^{-\lambda y})}{\ln\left(\frac{1-p-0.5\theta}{1-p-0.5}\right)(1-p)(1-e^{-\lambda y})} \left[\frac{1}{\lambda\theta} \ln\left(\frac{\theta e^{-\lambda y}-1}{\theta-1}\right) - \frac{y}{e^{\lambda y}-\theta} \right].$$

Proof 3.2 The Bonferroni curve $B_F[F(x)]$ is given by

$$B_F[F(y)] = \frac{1}{\mu F(y)} \int_0^y x f(x) dx,$$

where $\mu = \widetilde{Y}$ given in (8), $F(y) = 1 - S(y)$ and $f(x)$ given in (6).

From the relationship between the Bonferroni curve, $B_F[F(y)]$ and the mean residual lifetime given by Theorem 2.1 of Pundir *et al.* (2005), the Bonferroni curve equation is obtained as

$$\begin{aligned} B_F[F(y)] &= \frac{\lambda(1-\theta e^{-\lambda y})}{\ln\frac{1-p-0.5\theta}{1-p-0.5}\,(1-p-(1-p)e^{-\lambda y})} \int_0^y \frac{x\lambda(1-\theta-p+p\theta)e^{-\lambda x}}{(1-\theta e^{-\lambda x})^2} \\ &= \frac{\lambda(1-\theta e^{-\lambda y})(1-\theta-p+p\theta)}{\ln\frac{1-p-0.5\theta}{1-p-0.5}\,(1-p-(1-p)e^{-\lambda y})} \int_0^y \frac{x\lambda e^{-\lambda x}}{(1-\theta e^{-\lambda x})^2}. \end{aligned}$$

The proof is completed by solving the above integral. □

The Lorenz curve of $F(y)$ that follows a LEG distribution can be obtained by considering the expression $L_F[F(y)] = B_F[F(y)]F(y)$.

Proposition 3.3 The scaled total time and cumulative total time for a model LEG are given respectively by,

$$S_F[F(y)] = \frac{p\lambda y}{\ln\left(\frac{1-p-0.5\theta}{1-p-0.5}\right)} + \frac{1-\theta-p+p\theta}{\theta\ln\left(\frac{1-p-0.5\theta}{1-p-0.5}\right)} \ln\left(\frac{\theta e^{-\lambda y}-1}{\theta-1}\right),$$

and

$$C_F = \frac{p\lambda(1-\theta-p+p\theta)}{\ln\left(\frac{1-p-0.5\theta}{1-p-0.5}\right)}\left(\theta\lambda\ln\left(\frac{\theta e^{-\lambda}-1}{\theta-1}\right)-\frac{1}{e^{\lambda}-\theta}\right)$$
$$+\frac{(1-\theta-p+p\theta)}{\theta\ln\left(\frac{1-p-0.5\theta}{1-p-0.5}\right)}\left(\frac{\ln\left(\frac{\theta e^{-\lambda}-1}{\theta-1}\right)}{\theta(\theta e^{-\lambda}-1)}\right).$$

Proof 3.3 The scaled total time and cumulative total time transform of a distribution function $F(y)$ (Pundir *et al.*, 2005) are defined by

$$S_F[F(y)] = \frac{1}{\mu}\int_0^y \bar{F}(x)dx \tag{13}$$

and

$$C_F = \int_0^1 S_F[F(y)]f(y)dy, \tag{14}$$

respectively. If $F(y)$ is the LEG distribution function specified as $1-S(y)$, where $S(y)$ is given in (5) and $\mu = \widetilde{Y}$ given in (8) then using (13) and (14) the proof is concluded. □

From (14), the Gini index can be obtained from the relationship $G = 1 - C_F$, where C_F is the cumulative total time given in the Proposition 3.1. Figure 3 presents the Bonferroni and Scaled total time plots according to different θ and p parameters. The parameter λ is assumed to be equal to 1.

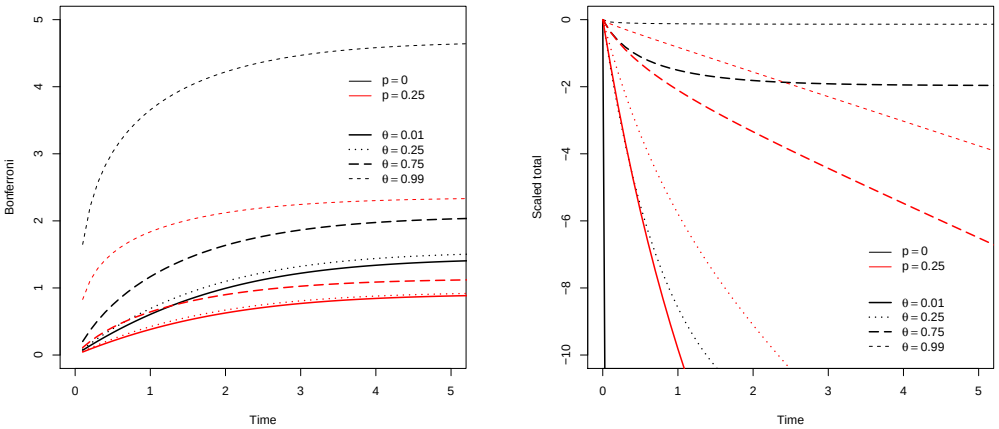

Figure 3: Bonferroni and scaled total plots to LEG distribution. Fixed, $\lambda = 1$

4. Inference

In this section we described the inferential and hypothesis testing procedures, which are based on the usual maximum likelihood approach as well as in the asymptotic large sample theory.

Let us consider the situation when the failure time Y in Section 2 is not completely observed and is subject to right censoring. Let C_i denote the censoring time. In a sample of size n, we then observe $Z_i = \min\{Y_i, C_i\}$ and $\delta_i = \mathrm{I}(Y_i \leq C_i)$, where $Y = \min(t_1, T_2, \cdots, T_M)$, $\delta_i = 1$ if Z_i is a failure time and $\delta_i = 0$ if it is right censored, for $i = 1, \cdots, n$. The likelihood of $\boldsymbol{\psi} = (\theta, \lambda, p)$ under non-informative censoring is given by Klein and Moeschberger (2003)

$$L(\boldsymbol{\psi}; \boldsymbol{D}) \propto \prod_{i=1}^{n} f(z_i; \boldsymbol{\psi})^{\delta_i} \, S(z_i; \boldsymbol{\psi})^{1-\delta_i}, \tag{15}$$

where $\boldsymbol{D} = (\boldsymbol{z}, \boldsymbol{\delta})$, $\boldsymbol{z} = (z_1, \cdots, z_n)^\top$, and $\boldsymbol{\delta} = (\delta_1, \cdots, \delta_n)^\top$, whereas $f(\cdot; \boldsymbol{\psi})$ and $S(\cdot; \boldsymbol{\psi})$ are given in (6) and (5), respectively, the log-likelihood for LEG distribution is given by

$$\begin{aligned}
\ell(\boldsymbol{\psi}; \boldsymbol{D}) \propto & \sum_{i=1}^{n} \ln(\lambda(1 - \theta - p - p\theta) - \lambda \sum_{i=1}^{n} \delta_i z_i \\
& + \sum_{i=1}^{n} (1 - \delta_i) \ln\left(p + (1 - \theta - p)e^{-\lambda z_i}\right) + \sum_{i=1}^{n} (1 - \delta_i) \ln\left(p + (1 - \theta - p)e^{-\lambda z_i}\right) \\
& + \sum_{i=1}^{n} (\delta_i + 1) \ln\left(1 - \theta e^{-\lambda z_i}\right).
\end{aligned} \tag{16}$$

Maximum likelihood estimation (MLE) may be performed by directly maximization of (16). The advantage of this procedure is that it runs immediately using existing statistical packages. We have considered the optim routine of the R (R Development Core Team, 2008), which is a general purpose optimization routine. An important aspect of implementing the estimation procedure concerns convergence or avoid end on multiple maxima. In our numerical examples and simulation studies we have not faced numerical problems, such as evidence of failure of convergence or end on multiple maxima.

Large-sample inference for the parameters are based on the MLEs and their estimated standard errors in an asymptotic fashion. The asymptotic normality is also useful for testing goodness of fit of the LEG distribution and for comparing it with some of its special sub-models, namely, the EG, the long-term exponential (LE) and the simple exponential (E) distributions, via the likelihood ratio statistic (LRS). For comparison of nested models, we can compute the maximum values

of the unrestricted and restricted log-likelihoods to obtain the LRS for the three tests. For instance, to test $H_0 : p = 0$ versus $H_1 : p > 0$ we consider the LRS, $\omega_n = 2(\ell_{LEG} - \ell_{EG})$, where ℓ_{LEG} and ℓ_{EG} are the maximum value of log-likelihoods for the model under the unrestricted hypothesis H_1 and under the restricted hypothesis H_0 under a sample of size n, respectively. Taking into account that the test is performed in the boundary of the parameter space, following Maller and Zhou (1995), the LRS, ω_n, is assumed to be asymptotically distributed as a symmetric mixture of a chi-squared distribution with one degree of freedom and a point-mass at zero. Then, $\lim_{n\to\infty} P(\omega_n \leq c) = 1/2 + 1/2P(\chi^2_1 \leq c)$, where $P(\chi^2_1 \leq c)$ denotes a random variable with a chi-square distribution with one degree of freedom. Large positive values of ω_n give favorable evidence to the full model.

5. Cancer Data Application

In this section, we compare the proposed LEG distribution with its particular case (the LE distribution) fits, as well as with the long-term Weibull (LW) distribution, on two data sets extracted from the literature. The idea is to show the applicability of the new distribution and the direct possibility of choosing between it or its particular cases, as well as its competitiveness in terms of fitting related to an usual survival distribution, such as the LW.

The first data set, the Myelomatosis Data, is extracted from Allison (1995). Myelomatosis is a malignant neoplasm of plasma cells in which the plasma cells proliferate and invade the bone marrow, causing destruction of the bone and resulting in pathologic fracture and bone pain[2]. The data set consists of lifetimes of 25 patients diagnosed with myelomatosis recorded in days from the point of randomization to either death or censoring (which could occur either by loss to follow up or end of the follow up). Censuring is observed for 22% of the lifetimes. The second data set, the Leukemia Data, is extracted from Kersey *et al.* (1987). The data set consists of lifetimes up to recurrence of leukemia, in years, for a group of 46 patients who received autologous marrow. The authors reported that the fraction of cured patients was estimated to be 20%.

Firstly, in order to verify the shape of the hazard rate function, we follow a standard graphical methodology for data analysis, we use the total time on test (TTT) plot, which is described by Chambers *et al.* (1983). It allows to identify the shape of a lifetime data hazard rate function graphically. According to Aarset (1987), in its empirical version the TTT plot is given by $G(r/n) = [(\sum_{i=1}^{r} Y_{i:n}) - (n-r)Y_{r:n}]/(\sum_{i=1}^{r} Y_{i:n})$, where $r = 1, \cdots, n$ and $Y_{i:n}$ represent the order statistics of the sample. It has been shown that the hazard function is

[2]http://medical-dictionary.thefreedictionary.com/myelomatosis

increasing (decreasing) if the TTT plot is concave (convex). Both left panels of Figure 4 show the TTT plot for the Myelomatosis Data (upper left) and Leukemia Data (lower left), implying in decreasing hazard functions.

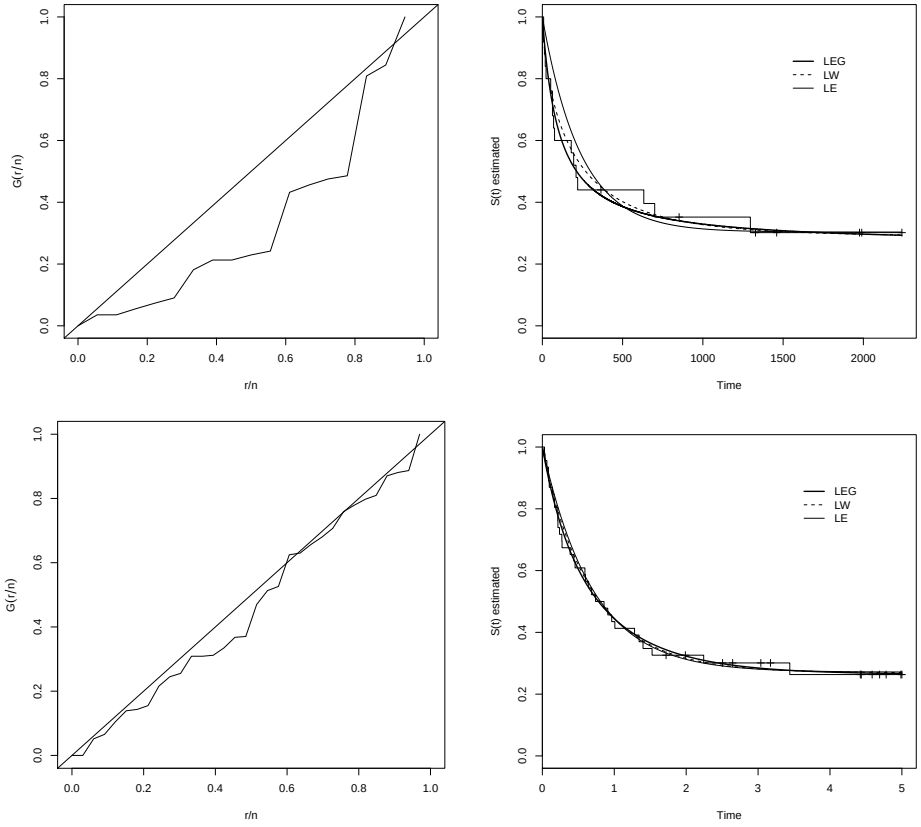

Figure 4: Left panel: TTT plot; right panel: Kaplan Meier curve with estimated survival function via LEG, LW and LE distributions, for the Heart Data (upper panels) and for the Leukaemia Data (lower panels)

We then fit the three distributions for the datasets. Table 1 provides the MLEs (and their corresponding standard errors in parentheses) for the parameters of the fitted distributions. From (8), the estimated medians for the Myelomotosis and Leukemia Data are equal to 150.3636 days and 0.7798 years, respectively.

Table 2 shows $-\ell(\widehat{\boldsymbol{\psi}}_g)$, AIC and BIC criterion values for the three distributions, where $\widehat{\boldsymbol{\psi}}_g$ denotes the MLE vector related to the distribution g, providing evidence in favor of our LEG distribution for both datasets. Besides, we compare the LEG distribution fitting with the LE distribution fittings by considering the test procedure presented in Section 4. The ω_n equals to 6.526, is much greater than $1/2+1/2\ P(\chi^2_1 \leq c) = 2.421$, leading to strong evidence in favor of the LEG

distribution for the Myelomotosis Data. However, there is no evidence that the LEG distribution is better than LE for the Leukemia Data at 5% of significance. Comparing the LED distribution with the usual survival distribution LW, both AIC and BIC criterion values provide evidence to the LEG distribution for both datasets. This results are corroborated by the empirical Kaplan-Meier survival functions superimposed by the fitted survival functions obtained via the LEG, LW and LE distributions shown in Figure 4. Considering the LEG distribution, the long-term parameter is equal to 0.2658 with standard error of 0.1229 for the Myelomotosis Data, and equal to 0.2636 with standard error of 0.0712 for the Leukemia Data.

Table 1: MLEs and their standard error for Myelomatosis and Leukemia Data

Distribution	λ	θ	ϕ	p
Myelomatosis				
LEG	0.0002 (0.0084)	0.9857 (0.0659)	-	0.2658 (0.1229)
LW	0.0049 (0.0022)	-	0.6749 (0.1431)	0.2901 (0.1019)
LE	0.0042 (0.0011)	-	-	0.3035 (0.0966)
Leukemia				
LEG	0.9980 (0.6723)	0.4432 (0.4710)	-	0.2636 (0.0712)
LW	1.4517 (0.3078)	-	0.9452 (0.1377)	0.2688 (0.0690)
LE	1.4331 (0.2795)	-	-	0.2710 (0.0684)

Table 2: The $\ell(\widehat{\boldsymbol{\psi}}_g)$, AIC and BIC values

Model	Myelomatosis			Leukaemia		
	$\ell(\cdot)$	AIC	BIC	$\ell(\cdot)$	AIC	BIC
LEG	-121.0445	248.0890	251.7456	-45.90177	97.80355	103.2895
LW	-121.9174	249.8348	253.4915	-46.14845	98.29691	103.7828
LE	-124.3070	252.6141	255.0518	-46.22798	96.45596	100.1132

6. Concluding Remarks

In this paper we provided the LEG distribution as an extension of the EG distribution proposed by Adamidis and Loukas (1998), which arises on a latent competing risks scenarios, where the lifetime associated with a particular risk is not observable, but only the minimum lifetime value among all risks, and there is presence of long-term survivals. The properties of the proposed distribution are discussed, including its probability density function and explicit algebraic formulas for its pdf, survival, hazard and quantile functions. The Bonferroni function

and the Lorenz curve are provided. MLE is implemented straightforwardly. The practical importance of the LEG distribution was demonstrated in two applications where the LEG distribution provided competitive fitting in comparison with its particular case and with the usual LW lifetime distribution. The performance of the MLE procedure as well as the LR testing considered here, which may be evaluated by Monte Carlo simulation, will be considered elsewhere.

Appendix A

We obtain the order statistic for the EG distribution, given the density, survival and distribution functions in Adamidis and Loukas (1998). Then, we have

$$g_{k:n}(y) = \frac{1}{B(k, n-k+1)} f(y)(F(y))^{k-1}(S(y))^{n-k},$$

$$g_{k:n}(y) = \frac{1}{B(k, n-k+1)} \frac{\lambda(1-\theta)e^{-\lambda y}}{(1-\theta e^{-\lambda y})^2} \left(\frac{1-e^{-\lambda y}}{1-\theta e^{-\lambda y}}\right)^{k-1} \left(\frac{(1-\theta)e^{-\lambda y}}{1-\theta e^{-\lambda y}}\right)^{n-k}$$

$$= \frac{1}{B(k, n-k+1)} \frac{\lambda(1-\theta)e^{-\lambda y}}{(1-\theta e^{-\lambda y})^{n+1}} \left(1-e^{-\lambda y}\right)^{k-1} \left((1-\theta)e^{-\lambda y}\right)^{n-k}.$$

Acknowledgements

The authors thank the reviewers for their comments and criticisms, which led to substantial improvement on the manuscript. The research is supported by the Brazilian organizations CAPES and CNPq.

References

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

Adamidis and Loukas (1998). A lifetime distribution with decreasing failure rate. *Statistics Probability Letters* **39**, 35-42.

Allison, P. D. (1995). *Survival Analysis Using the SAS System*: *A Practical Guide.* SAS Institute Inc., Cary, North Carolina.

Berkson, J. and Gage, R. P. (1952). Survival curve for cancer patients following treatment. *Journal of the American Statistical Association* **47**, 501-515.

Boag, J. W. (1949). Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society, Series B* **11**, 15-44.

Cancho, V. G., Ortega, E. M. M. and Bolfarine, H. (2009). The log-exponentiated-Weibull regression models with cure rate: local influence and residual analysis. *Journal of Data Science* **7**, 433-458.

Chambers, J. M., Cleveland, W. S., Kleiner, B. and Tukey, P. A. (1983). *Graphical Methods for Data Analysis.* Chapman and Hall, New York.

Chen, H. M. and Ibrahim, J. G. (2001). Maximum likelihood methods for cure rate models with missing covariates. *Biometrics* **57**, 43-52.

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data.* Chapman and Hall, London.

Crowder, M., Kimber, A., Smith, R. and Sweeting, T. (1991). *Statistical Analysis of Reliability Data.* Chapman and Hall, London.

Farewell, V. T. (1977). A model for a binary variable with time-censored observations. *Biometrics* **64**, 43-46.

Farewell, V. T. (1982). The use mixture models for the analysis of survival data with long-term survivors. *Biometrics* **38**, 1041-1046.

Ghitany, M. and Maller, R. (1992). Asymptotic results for exponential mixture models with long with long-term survivors. *Statistics* **23**, 321-336.

Ghitany, M. E., Maller, R. A. and Zhou, S. (1994). Exponential mixture models with long-term survivors and covariates. *Journal of multivariate Analysis* **49**, 218-241.

Glaser, R. A. (1980). Bathtub and related failure rate characterization. *Journal of the American Statistical Association* **75**, 667-672.

Goetghebeur, E. and Ryan, L. (1995). A modified log rank test for competing risks with missing failure type. *Biometrika* **77**, 207-211.

Greenhouse, J. B. and Wolf, R. A. (1984). A competing risk deviation of a mixture model for the analysis of survival data. *Communications in Statistics - Theory and Methods* **13**, 3133-3154.

Kersey, J. H., Weisdorf, D., Nesbit, M. E., LeBien, T. W., Woods, W. G., McGlave, P. B., Kim, T., Vallera, D. A., Goldman, A. I., Bostrom, B. and Ramsay, N. K. C. (1987). Comparison of autologous and allogeneic bone marrow transplantation for treatment of high-risk refractory acute lymphoblastic leukemia. *New England Journal of Medicine* **317**, 461-467.

Klein, J. and Moeschberger, M. (2003). *Survival Analysis*: *Techniques for Censored and Truncated Data.* Springer, New York.

Kus, C. (2007). A new lifetime distribution. *Computational Statistics and Data Analysis* **51**, 4497-4509.

Lawless, J. F. (2003). *Statistical Models and Methods for Lifetime Data*, 2nd edition. Wiley, New York.

Louzada-Neto, F. (1999). Polyhazard models for lifetime data. *Biometrics* **55**, 1281-1285.

Lu, K. and Tsiatis, A. A. (2001). Multiple imputation methods for estimating regression coefficients in the competing risks model with missing cause of failure. *Biometrics* **54**, 1191-1197.

Lu, K. and Tsiatis, A. A. (2005). Comparison between two partial likelihood approaches for the competing risks model with missing cause of failure. *Lifetime Data Analysis* **11**, 29-40.

MacKenzie, G. (1996). Regression models for survival data. *Journal of the Royal Statistical Society, Series B* **45**, 21-34.

Maller, R. A. and Zhou, S. (1995). Testing for the presence of immune or cured individuals in censored survival data. *Biometrics* **51**, 1197-1205.

Maller, R. and Zhou, X. (1996). *Survival Analysis with Long-Term Survivors.* John Wiley and Sons Chichester, United Kingdom.

Perdoná G. C. and Louzada-Neto, F. (2011). A general hazard model for lifetime data in the presence of cure rate. *Journal of Applied Statistics* **38**, 1395-1405.

Perperoglou, A., Keramopoullos, A. and van Houwelingen, H. C. (2007). Approaches in modelling long-term survival: an application to breast cancer. *Statistics in Medicine* **26**, 2666-2685.

Pons, O. and Lemdani, M. (2003). Estimation and test in long-term survival mixture models. *Computational Statistics and Data Analysis* **41**, 465-479.

Pundir, S., Arora, S. and Jain, K. (2005). Bonferroni curve and the related statistical inference. *Statistics and Probability Letters* **75**, 140-150.

R Development Core Team. (2008). *R*: *A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria.

Reiser, B., Guttman, I., Lin, D. K. J., Guess, F. M. and Usher, J. S. (1995). Bayesian inference for masked system lifetime data. *Applied Statistics* **44**, 79-90.

Received May 10, 2011; accepted November 4, 2011.

Mari Roman
Department of Statistics
Universidade Federal de São Carlos
DEs, UFSCar, CP 676, São Carlos, SP, 13.565-905, Brazil
mari.roman19@hotmail.com

Francisco Louzada
Department of Applied Mathematics and Statistics
Universidade de São Paulo
ICMC, USP, CP 668, São Carlos, SP,13.566-590, Brazil
louzada@icmc.usp.br

Vicente G. Cancho
Department of Applied Mathematics and Statistics
Universidade de São Paulo
ICMC, USP, CP 668, São Carlos, SP,13.566-590, Brazil
garibay@icmc.usp.br

José G. Leite
Department of Statistics
Universidade Federal de São Carlos
DEs, UFSCar, CP 676, São Carlos, SP, 13.565-905, Brazil
leite@ufscar.br