STATISTICS, OPTIMIZATION AND INFORMATION COMPUTING
Stat., Optim. Inf. Comput., Vol. **10**, September 2022, pp 1168–1186.
Published online in International Academic Press (www.IAPress.org)

Discrete Bilal Distribution in the Presence of Right-Censored Data and a Cure Fraction

Bruno Caparroz Lopes de Freitas [1], Jorge Alberto Achcar [2],
Marcos Vinicius de Oliveira Peres [1], Edson Zangiacomi Martinez [2*]

[1]*State University of Maringá, Master Program in Biostatistics, Maringá, Brazil.*
[2]*Ribeirão Preto Medical School, University of São Paulo, Ribeirão Preto, Brazil.*

Abstract The statistical literature presents many continuous probability distributions with only one parameter, which are extensively used in the analysis of lifetime data, such as the exponential, the Lindley, and the Rayleigh distributions. Alternatively, the use of discretized versions of these distributions can provide a better fit for the data in many applications. As the novelty of this study, we present inferences for the discrete Bilal distribution (DB) with one parameter introduced by Altun et al. (2020) in the presence of right-censored data and cure fraction. We assume standard maximum likelihood methods based on asymptotic normality of the maximum likelihood estimators and also a Bayesian approach based on MCMC (Markov Chain Monte Carlo) simulation methods to get inferences for the parameters of the discrete BD distribution. The use of the proposed model was illustrated with three examples considering real medical lifetime data sets. From these applications, we concluded that the proposed model based on the discrete DB distribution has good performance even with the inclusion of a cure fraction in comparison to other existing discrete models, such as the DsFx-I, Lindley, Rayleigh, and Burr-Hatke probability distributions. Moreover, the model can be easily implemented in standard existing software, such as the R package. Under a Bayesian approach, we assumed a gamma prior distribution for the parameter of the DB discrete distribution. We also provided a brief sensitivity analysis assuming the half-normal distribution in place of the gamma distribution for the parameter of the DB distribution. From the obtained results of this study, we can conclude that the proposed methodology can be very useful for researchers dealing with medical discrete lifetime data in the presence of right-censored data and cure fraction.

Keywords Survival analysis, Maximum likelihood estimation, Cure fraction, Bayesian inference, Discrete distributions, Censored data

AMS 2010 subject classifications 60E05, 62N02, 62P10

DOI: 10.19139/soic-2310-5070-1414

1. Introduction

Lifetime data analysis is a statistical methodology extensively used in medical or engineering studies [19]. This class of statistical methods is applied when the response variable is given by the time of the occurrence of an event of interest. Examples in medical research include the time of response to a given treatment, the relapse-free survival time, the time to death, the time to a medical device failure, and the time to regain mobility [36]. Kaplan-Meier plots of the non-parametric estimates of the survival function, non-parametric log-rank tests, and Cox (proportional hazards) regression models are the most widely used survival analysis techniques used in medical studies [31]. As an alternative to the traditional proportional hazards model used in the presence of a vector of covariates, parametric

[*]Correspondence to: Edson Zangiacomi Martinez (Email: edson@fmrp.usp.br). Ribeirão Preto Medical School, University of São Paulo, Av. Bandeirantes 3900, Vila Monte Alegre, Ribeirão Preto, SP, 14049-900, Brazil.

ISSN 2310-5070 (online) ISSN 2311-004X (print)

regression models have also been extensively used in medical studies. Parametric models assume that the time-to-event variable follows a known probability distribution, such as the Weibull, the gamma, or the log-normal distributions. Among the discrete distributions proposed in the statistical literature to model time-to-event data, we can mention the discrete Weibull distribution [37], the discrete Lindley distribution [27], the exponentiated discrete Weibull distribution [8, 38, 21], the discrete generalized Rayleigh distribution [5], the discrete generalized Sibuya distribution [22], and the discrete Sushila distribution [39].

Let X be a random variable denoting a survival time, and let x be an observation of X. The continuous Bilal probability distribution introduced by Abd-Elrahman [1] has a probability density function (pdf) given by

$$f_X(x) = \frac{6}{\theta} e^{-\frac{2x}{\theta}} \left(1 - e^{-\frac{x}{\theta}}\right), \;\; x \geq 0, \;\theta > 0$$

and probability accumulated distribution function given by

$$F_X(x) = 1 - e^{-\frac{2x}{\theta}} \left(3 - 2e^{-\frac{x}{\theta}}\right).$$

The survival function, that is, the probability that an individual survives at least until a time x, is given by $S_X(x) = 1 - F_X(x)$. Abd-Elrahman [1] named the new proposed distribution as Bilal distribution as a tribute to his youngest son. Classical and Bayesian approaches to find an estimator for the parameter θ of a Bilal distribution based on a sample in presence of type-2 censoring data are introduced by Abd-Elrahman and Niazi [3]. Generalizations of the Bilal distribution are found in the works of Abd-Elrahman [2], Akhter et al. [4], Riad et al. [45], and Shi et al. [48].

To obtain a discrete version of the Bilal distribution, let us consider that the random variable T has a probability mass function (pmf) given by

$$\begin{aligned} P(T = t) &= P(t-1 < X \leq t) = P(X \leq t) - P(X \leq t-1) \\ &= F_X(t) - F_X(t-1) = S_X(t-1) - S_X(t), \;\; t \in \mathbb{N}^*, \end{aligned}$$

where X is the underlying continuous random variable, $T = \lceil X \rceil$ (the smallest integer greater than or equal to X), $S_X(x) = P(X > x)$, and $\mathbb{N}^*$ is the set of natural numbers not including the zero value (see Methodology-IV in the article by Chakraborty [9]). Thus, replacing $S_X(t-1)$ by $e^{-\frac{2(t-1)}{\theta}}\left(3 - 2e^{-\frac{t-1}{\theta}}\right)$ and $S_X(t)$ by $e^{-\frac{2t}{\theta}}\left(3 - 2e^{-\frac{t}{\theta}}\right)$ in the expression (1), we have

$$\begin{aligned} P(T = t) &= e^{-\frac{2(t-1)}{\theta}}\left(3 - 2e^{-\frac{(t-1)}{\theta}}\right) - e^{-\frac{2t}{\theta}}\left(3 - 2e^{-\frac{t}{\theta}}\right) \\ &= 2\left(e^{-\frac{3}{\theta}} - 1\right) e^{-\frac{3(t-1)}{\theta}} - 3e^{-\frac{2(t-1)}{\theta}}\left(e^{-\frac{2}{\theta}} - 1\right). \end{aligned} \tag{1}$$

In (1), let us assume the parameter transformation $p = e^{-\frac{1}{\theta}}$, $0 < p < 1$. In this way, following the notation of [6], the pmf of the discrete Bilal distribution is given by

$$f(t) = P(T = t) = 2(p^3 - 1)p^{3(t-1)} - 3(p^2 - 1)p^{2(t-1)}, \; t \in \mathbb{N}^*, \tag{2}$$

where $0 < p < 1$. The corresponding cumulative distribution function is given by

$$F(t) = P(T \leq t) = 1 - (3 - 2p^t)p^{2t}, \; t \in \mathbb{N}^*,$$

and the survival function is thus given by

$$S(t) = 1 - F(t) = P(T > t) = (3 - 2p^t)p^{2t}, \; t \in \mathbb{N}^*. \tag{3}$$

From the expression of the survival function (3), we observe that $S(t) = 1$ when $t = 0$. Altun et al. [6] analogously obtained a discrete version of the Bilal distribution, but these authors considered a pmf given by $P(T = t) = P(t \leq X \leq t+1)$ instead of $P(T = t) = P(t-1 < X \leq t)$, where X is the underlying continuous random variable.

However, an important limitation of the method used by Altun et al. [6] is that the survival function $S(t)$ is smaller than 1 when $t = 0$, which can be undesirable when the model is applied to real data.

To simplify obtaining the estimator for the parameter of the discrete Bilal distribution, we consider the reparameterization $p = e^{-\beta}$, where $\beta > 0$. Thus, the pmf is given by

$$f(t) = P(T = t) = 2(e^{-3\beta} - 1)e^{-3\beta(t-1)} - 3(e^{-2\beta} - 1)e^{-2\beta(t-1)},\ t \in \mathbb{N}^*,$$

where the corresponding cumulative distribution function is given by

$$F(t) = P(T \leq t) = 1 - \left[3 - 2e^{-\beta t}\right] e^{-2\beta t},\ t \in \mathbb{N}^*,$$

and the survival function is

$$S(t) = \left[3 - 2e^{-\beta t}\right] e^{-2\beta t},\ t \in \mathbb{N}^*.$$

The corresponding hazard function is given by

$$h(t) = P(T = t | T \geq t) = \frac{P(T = t)}{P(T \geq t)} = \frac{f(t)}{S(t-1)} = \frac{2(e^{-3\beta} - 1)e^{-\beta(t-1)} - 3(e^{-2\beta} - 1)}{3 - 2e^{-\beta(t-1)}}.$$

The probability generating function (pgf) of the discrete Bilal (DB) distribution is derived as follows:

$$\begin{aligned} E(s^T) &= 2(p^3 - 1)\sum_{t=0}^{\infty} p^{3(t-1)}s^t - 3(p^2 - 1)\sum_{t=0}^{\infty} p^{2(t-1)}s^t \\ &= 2(p^3 - 1)p^{-3}\sum_{t=0}^{\infty}\left(sp^3\right)^t - 3(p^2 - 1)p^{-2}\sum_{t=0}^{\infty}\left(sp^2\right)^t \\ &= \frac{2(p^3 - 1)}{p^3\left(1 - sp^3\right)} - \frac{3(p^2 - 1)}{p^2\left(1 - sp^2\right)}, \end{aligned} \tag{4}$$

where $p = e^{-\beta}$ and s is a real number. Replacing s with e^s in (4), the moment generating function (mgf) of the DB distribution is given by

$$M_T(s) = E(e^{sT}) = \frac{2(p^3 - 1)}{p^3\left(1 - e^s p^3\right)} - \frac{3(p^2 - 1)}{p^2\left(1 - e^s p^2\right)}. \tag{5}$$

In order to find the mean and variance of T, we get the first and second derivatives of the mgf (5) with respect to s given by

$$\frac{d}{ds}M_T(s) = \frac{2e^s(p^3 - 1)}{\left(p^3 e^s - 1\right)^2} - \frac{3e^s(p^2 - 1)}{\left(p^2 e^s - 1\right)^2}$$

and

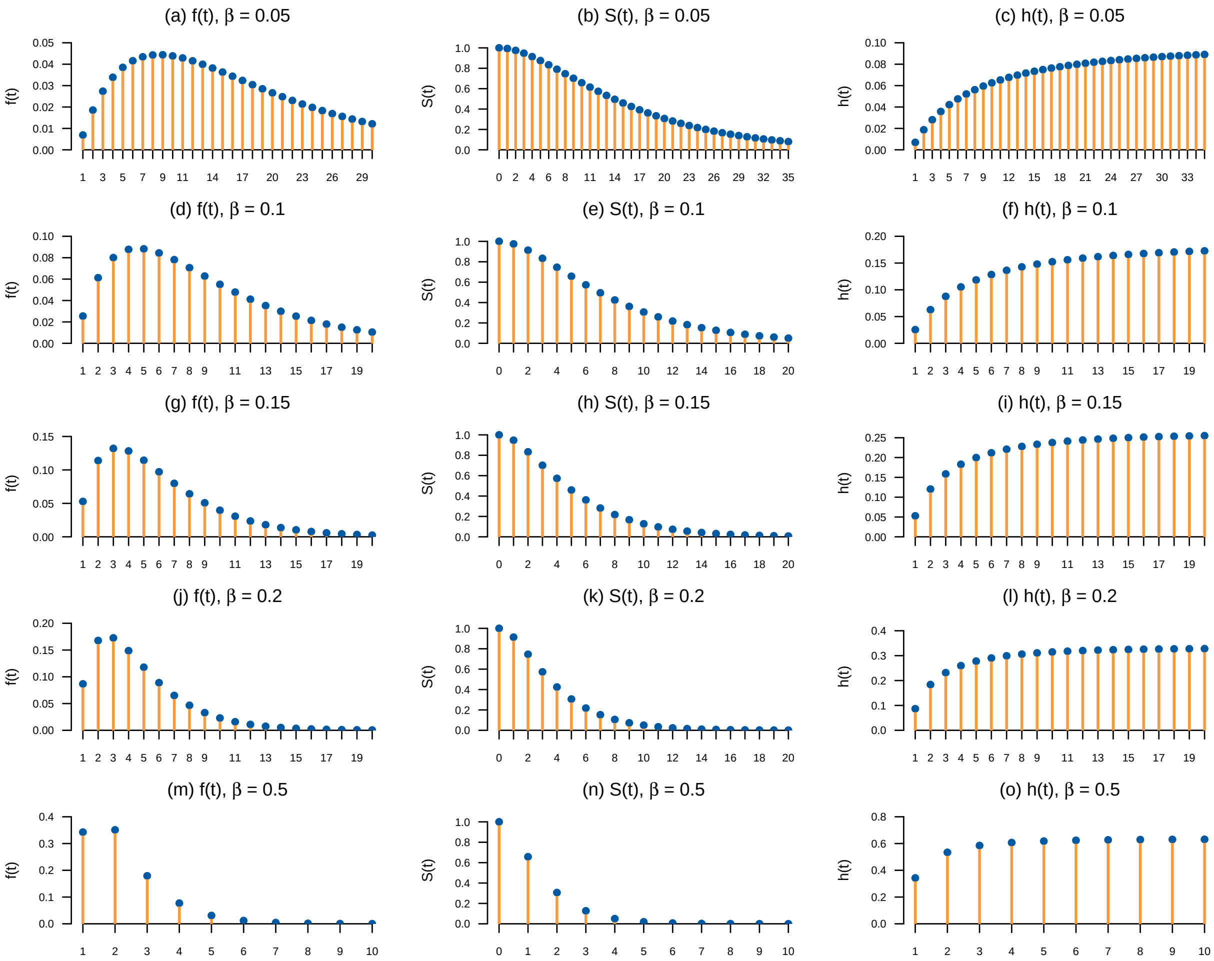
$$\frac{d^2}{ds^2}M_T(s) = -\frac{2e^s\left(e^s p^3 + 1\right)\left(p^3 - 1\right)}{\left(p^3 e^s - 1\right)^3} + \frac{3e^s\left(e^s p^2 + 1\right)\left(p^2 - 1\right)}{\left(p^2 e^s - 1\right)^3},$$

respectively. Next, we evaluate the derivatives at $t = 0$ to find the first and second moments of T. The first moment $E(T)$ corresponds to the mean of the distribution and is given by

$$E(T) = \frac{d}{ds}\left[M_T(s)\right]_{s=0} = \frac{2(p^3 - 1)}{\left(p^3 - 1\right)^2} - \frac{3(p^2 - 1)}{\left(p^2 - 1\right)^2} = \frac{3p^3 - 2p^2 - 1}{\left(p^3 - 1\right)\left(1 - p^2\right)}. \tag{6}$$

The second moment $E(T^2)$ is given by

$$E(T^2) = \frac{d^2}{ds^2}\left[M_T(s)\right]_{s=0} = \frac{3\left(p^2 + 1\right)}{\left(p^2 - 1\right)^2} - \frac{2\left(p^3 + 1\right)}{\left(p^3 - 1\right)^2}. \tag{7}$$



(a) f(t), β = 0.05 (b) S(t), β = 0.05 (c) h(t), β = 0.05

(d) f(t), β = 0.1 (e) S(t), β = 0.1 (f) h(t), β = 0.1

(g) f(t), β = 0.15 (h) S(t), β = 0.15 (i) h(t), β = 0.15

(j) f(t), β = 0.2 (k) S(t), β = 0.2 (l) h(t), β = 0.2

(m) f(t), β = 0.5 (n) S(t), β = 0.5 (o) h(t), β = 0.5

Figure 1. The pmf, survival function and hazard function of the DB distribution DB(β) for different values of β.

The variance of T is obtained from (6) and (7) using the expression $Var(T) = E(T^2) - E^2(T)$. Let us denote a discrete Bilal distribution with parameter β as $DB(\beta)$. Figure 1 shows the graphs of the pmf, survival function, and hazard function of the DB distribution considering different values of the parameter β. From these plots, we observe that the DB distribution has an increasing hazard function.

The novelty of the present article consists in the introduction of a discrete DB distribution to model lifetime data in the presence of right-censored time-to-event data. We also consider the inclusion of a cure fraction in the model. The paper is organized as follows. Section 2 presents the maximum-likelihood (ML) estimator for the parameter of the DB distribution based on complete and censored data. ML estimation in the presence of censored data and a cure fraction is also discussed in this section. In addition, Section 2 also presents a Bayesian approach for the model. Three examples considering real data from the medical literature are used in Section 3 to illustrate the usefulness of this model to a broad range of problems. Finally, in Section 4, some concluding remarks are presented. The computational codes used in this article are provided in an Appendix at the end of the manuscript.

2. Methods

2.1. Maximum likelihood estimation for complete data

After some algebra we can see that the equation (2) is equivalent to

$$f(t) = P(T = t) = p^{2(t-1)}(p-1)\left[2p^{t-1}(p^2+p+1) - 3p - 3\right]. \tag{8}$$

Let T_1,...,T_n be a random sample of failure times from a DB survival distribution. Considering $p = e^{-\beta}$ and the expression (8), the likelihood function for the parameter β is given by

$$L(\beta|\,\mathbf{t}) = \prod_{i=1}^{n} e^{-2\beta(t_i-1)}(e^{-\beta}-1)\left[2e^{-\beta(t_i-1)}(e^{-2\beta}+e^{-\beta}+1) - 3e^{-\beta} - 3\right],$$

and the corresponding log-likelihood function is given by

$$\ell(\beta|\,\mathbf{t}) = -2\beta\sum_{i=1}^{n}(t_i-1) + n\log(e^{-\beta}-1) + \sum_{i=1}^{n}\log\left[2e^{-\beta(t_i-1)}(e^{-2\beta}+e^{-\beta}+1) - 3e^{-\beta} - 3\right].$$

Deriving the log-likelihood function with respect to β, we obtain the following expression:

$$\frac{d\ell}{d\beta} = -2\sum_{i=1}^{n}(t_i-1) - \frac{ne^{-\beta}}{e^{-\beta}-1} - \sum_{i=1}^{n}\frac{2e^{-\beta(t_i-1)}\left[t_i\left(e^{-2\beta}+e^{-\beta}+1\right)+e^{-2\beta}-1\right] - 3e^{-\beta}}{2e^{-\beta(t_i-1)}\left(e^{-2\beta}+e^{-\beta}+1\right) - 3e^{-\beta} - 3}. \tag{9}$$

The maximum-likelihood (ML) estimator $\widehat{\beta}_{ML}$ for β is obtained by equating the right-hand side of (9) to zero and from the obtained equation, we get the estimator for β. Nevertheless, the resulting expression does not have a closed-form solution, and numerical methods are needed to find the ML estimate for β. In this article, we use the maxLik package in the R program to obtain the ML estimate of the parameter β [28]. A confidence interval for β can be constructed from the asymptotic normality of the ML estimate considering large sample sizes, given by

$$\widehat{\beta}_{ML} \sim N\left(\beta, \widehat{Var}(\widehat{\beta}_{ML})\right),$$

where, in the single-parameter case, $\widehat{Var}(\widehat{\beta}_{ML})$ is the estimated variance for $\widehat{\beta}_{ML}$. Therefore, an approximate $100(1-\upsilon)\%$ Wald-type confidence interval (CI) for β is given by

$$\widehat{\beta}_{ML} \mp z_{\upsilon/2}\sqrt{\widehat{Var}(\widehat{\beta}_{ML})},$$

where z_υ denotes the upper υ-th percentile of the standard normal distribution. The asymptotic variance of an ML estimator can be estimated by the negative of the inverse of the second derivative of the log-likelihood function evaluated at $\widehat{\beta}_{ML}$. Thus, the second derivative of the log-likelihood function for β is given by

$$\frac{d^2\ell}{d\beta^2} = -\frac{ne^{-\beta}}{(e^{-\beta}-1)^2} + \sum_{i=1}^{n}\frac{A_i}{\left[2e^{-\beta(t_i-1)}(e^{-2\beta}+e^{-\beta}+1) - 3e^{-\beta} - 3\right]^2}, \tag{10}$$

where

$$\begin{aligned} A_i &= 6e^{-\beta(t_i-1)}\left[t_i^2\left(e^{-3\beta}+2e^{-2\beta}+2e^{-\beta}+1\right) - 2t_i\left(2e^{-\beta}+1\right) + 2e^{-2\beta} + 4e^{-\beta} + 1\right] \\ &\quad -4e^{-2\beta(t_i-1)}\left[e^{-3\beta}+4e^{-2\beta}+e^{-\beta}\right] - 9e^{-\beta}. \end{aligned} \tag{11}$$

The negative second derivative of the log-likelihood function is the observed information denoted by $I_n(\beta)$, that is,

$$I_n(\beta) = -\frac{d^2\ell}{d\beta^2}.$$

The expected value of $I_n(\beta)$, say $i_n(\beta)$, is called the expected Fisher information. The asymptotic variance of $\widehat{\beta}_{ML}$ is given by the inverse of the expected information evaluated at the ML estimate of β, that is,

$$\widehat{Var}(\widehat{\beta}_{ML}) = \left[i_n(\widehat{\beta}_{ML})\right]^{-1}.$$

2.2. Maximum-likelihood estimation in the presence of censored data

Considering a random sample (t_i, d_i) of size n, $i = 1, \cdots, n$, the contribution of the ith individual to the likelihood function is given by

$$L_i = [f(t_i)]^{d_i} [S(t_i)]^{1-d_i},$$

where d_i is a censoring indicator variable, that is, $d_i = 1$ for an observed survival time and $d_i = 0$ for a right-censored survival time. Assuming the data with a DB distribution, the likelihood function for the parameter β is given by

$$\begin{aligned} L(\beta|\, \mathbf{t}, \mathbf{d}) &= \prod_{i=1}^{n} e^{-2\beta(t_i-1)d_i} (e^{-\beta}-1)^{d_i} \left[2e^{-\beta(t_i-1)}(e^{-2\beta}+e^{-\beta}+1) - 3e^{-\beta} - 3\right]^{d_i} \\ &\quad \times \left[3 - 2e^{-\beta t_i}\right]^{(1-d_i)} e^{-2\beta t_i(1-d_i)}, \end{aligned}$$

and the corresponding log-likelihood function is

$$\begin{aligned} \ell(\beta|\, \mathbf{t}, \mathbf{d}) &= 2\beta \sum_{i=1}^{n}(t_i-1)d_i + \log(e^{-\beta}-1)\sum_{i=1}^{n} d_i \\ &\quad + \sum_{i=1}^{n} d_i \log\left[2e^{-\beta(t_i-1)}(e^{-2\beta}+e^{-\beta}+1) - 3e^{-\beta} - 3\right] \\ &\quad + \sum_{i=1}^{n}(1-d_i)\log\left[3-2e^{-\beta t_i}\right] - 2\beta\sum_{i=1}^{n} t_i\,(1-d_i). \end{aligned} \tag{12}$$

Deriving the log-likelihood function (12) with respect to β, we have

$$\begin{aligned} \frac{d\ell}{d\beta} &= 2\sum_{i=1}^{n}(t_i-1)d_i - \frac{e^{-\beta}}{e^{-\beta}-1}\sum_{i=1}^{n} d_i \\ &\quad - \sum_{i=1}^{n} d_i \frac{2e^{-\beta(t_i-1)}\left(e^{-2\beta}+t_i+t_ie^{-\beta}+t_ie^{-2\beta}-1\right) - 3e^{-\beta}}{2e^{-\beta(t_i-1)}(e^{-2\beta}+e^{-\beta}+1) - 3e^{-\beta} - 3} \\ &\quad - 2\sum_{i=1}^{n}(1-d_i)\frac{t_ie^{-\beta t_i}}{2e^{-\beta t_i}-3} - 2\sum_{i=1}^{n} t_i\,(1-d_i). \end{aligned}$$

Setting this expression equal to zero, we get the corresponding score equation whose numerical solution leads to the ML estimator. The second derivative of the log-likelihood function with respect to β is given by

$$\begin{aligned} \frac{d^2\ell}{d\beta^2} &= -\frac{e^{-\beta}}{(e^{-\beta}-1)^2}\sum_{i=1}^{n} d_i - \sum_{i=1}^{n} d_i \frac{A_i}{\left[2e^{-\beta(t_i-1)}(e^{-2\beta}+e^{-\beta}+1) - 3e^{-\beta} - 3\right]^2} \\ &\quad - 6\sum_{i=1}^{n}(1-d_i)\frac{t_i^2 e^{-\beta t_i}}{(2e^{-\beta t_i}-3)^2}, \end{aligned}$$

where A_i is given by (11). Approximated confidence intervals for β can also be obtained from the asymptotic normality of the ML estimate for β, in a similar way as described in the previous subsection.

Given that the time-to-event variable is discrete, randomized quantile residuals can be used to test model adequacy [14]. These residuals are given by $r_i = \Phi^{-1}(u_i)$, $i = 1, ..., n$, where $\Phi(\cdot)$ is the standard normal distribution function, and u_i is a uniform random variable on the interval $\left[F(t_i - 1, \widehat{\beta}_{ML}), F(t_i, \widehat{\beta}_{ML})\right]$ if $d_i = 1$ and $\left[F(t_i, \widehat{\beta}_{ML}), 1\right]$ if $d_i = 0$. Let us consider that $F(t_i, \widehat{\beta}_{ML})$ is the cumulative distribution function of the DB distribution assuming the ML estimate for β. The randomized quantile residuals are expected to follow the standard normal distribution if the model is correct. Thus, a normal Q-Q plot can be used to visually verify the normality assumption of residuals.

2.3. Maximum-likelihood estimation including censored data and a cure fraction

A fundamental characteristic of the traditional survival analysis methodology is that the survival function $S(t)$ converges to zero when the time variable tends to infinity. In applications of survival analysis to medical research data, this implies assuming that all individuals under study are susceptible to the event of interest. However, there are situations where this assumption is not satisfied [40]. For example, in randomized trials evaluating the efficacy of a treatment for a disease of interest, it is possible that some patients may be cured of the disease due to the treatment under study. If the event of interest is the death due to this disease, then these patients are no longer subject to this event. The presence of cured individuals in a data set is usually suggested by a stable plateau at the right tail of the Kaplan–Meier non-parametric estimator of the survival function, with heavy censoring in this portion of the plot [11]. Different parametric and non-parametric approaches that consider the presence of immune individuals have been proposed in the literature [7, 33, 43]. These approaches include a mixture model, which explicitly includes a parameter accounting for a fraction of immune individuals [30, 35]. This model assumes that the probability of observing a survival time greater than or equal to some fixed value t is given by the survival function

$$S(t) = \eta + (1 - \eta) S_0(t),$$

where η is the proportion of immune, cured or not susceptible individuals, and $S_0(t)$ is the baseline survival function for the susceptible individuals [18]. Considering a random sample (t_i, d_i) of size n, $i = 1, \cdots, n$, the contribution of the ith individual to the likelihood function is given by

$$L_i = \left[f\left(t_i\right)\right]^{d_i} \left[S\left(t_i\right)\right]^{1-d_i} = \left[(1-\eta) f_0\left(t_i\right)\right]^{d_i} \left[\eta + (1-\eta) S_0(t)\right]^{1-d_i},$$

where d_i is a binary censoring indicator variable and $f_0(t)$ is the corresponding baseline probability function. Assuming the mixture model based on the DB distribution, the likelihood function for β and η is given by

$$\begin{aligned} L(\beta, \eta | \, \mathbf{t}, \mathbf{d}) \quad &= \quad \prod_{i=1}^{n} (1-\eta)^{d_i} e^{-2\beta(t_i - 1)d_i} (e^{-\beta} - 1)^{d_i} \left[2e^{-\beta(t_i-1)}(e^{-2\beta} + e^{-\beta} + 1) - 3e^{-\beta} - 3\right]^{d_i} \\ &\quad \times \left[\eta + (1-\eta)\left(3 - 2e^{-\beta t_i}\right) e^{-2\beta t_i}\right]^{1-d_i} \end{aligned}$$

The log-likelihood function in this case is

$$\begin{aligned} \ell(\beta, \eta | \, \mathbf{t}, \mathbf{d}) \quad &= \quad \sum_{i=1}^{n} d_i \log(1-\eta) - 2\beta \sum_{i=1}^{n} (t_i - 1)\, d_i + \sum_{i=1}^{n} d_i \log(e^{-\beta} - 1) \\ &\quad + \sum_{i=1}^{n} d_i \log\left[2e^{-\beta(t_i-1)}(e^{-2\beta} + e^{-\beta} + 1) - 3e^{-\beta} - 3\right] \\ &\quad + \sum_{i=1}^{n} (1 - d_i) \log\left[\eta + (1-\eta)\left(3 - 2e^{-\beta t_i}\right) e^{-2\beta t_i}\right]. \end{aligned} \tag{13}$$

The first derivative of the log-likelihood function (13) with respect to β is given by

$$\begin{aligned}\frac{\partial \ell}{\partial \beta} &= -2\sum_{i=1}^{n}(t_i-1)\,d_i - \sum_{i=1}^{n} d_i \frac{e^{-\beta}}{e^{-\beta}-1} \\ &\quad - \sum_{i=1}^{n} d_i \frac{2e^{-\beta(t_i-1)}\left[e^{-2\beta}+t_i+t_ie^{-\beta}+t_ie^{-2\beta}-1\right]-3e^{-\beta}}{2e^{-\beta(t_i-1)}(e^{-2\beta}+e^{-\beta}+1)-3e^{-\beta}-3} \\ &\quad +6\,(\eta-1)\sum_{i=1}^{n}(1-d_i)\frac{t_ie^{-2\beta t_i}\left(e^{-\beta t_i}-1\right)}{\eta+(1-\eta)\,(3-2e^{-\beta t_i})\,e^{-2\beta t_i}}, \end{aligned} \tag{14}$$

and the first derivative of the log-likelihood function with respect to η is given by

$$\frac{\partial \ell}{\partial \eta} = -\frac{1}{1-\eta}\sum_{i=1}^{n} d_i + \sum_{i=1}^{n}(1-d_i)\frac{\left(2e^{-\beta t_i}+1\right)\left(e^{-\beta t_i}-1\right)^2}{\eta+(1-\eta)\,(3-2e^{-\beta t_i})\,e^{-2\beta t_i}}. \tag{15}$$

From the equations obtained from the expressions (14) and (15) equal to zero, we obtain the maximum likelihood estimators of the parameters β and η. Although we cannot obtain explicit expressions for the ML estimators for these parameters, they can be estimated numerically using iterative algorithms such as the Newton-Raphson method and its variants.

The second partial derivatives of the log-likelihood function are given by:

$$\begin{aligned}\frac{\partial^2 \ell}{\partial \beta^2} &= -\sum_{i=1}^{n} d_i \frac{e^{-\beta}}{\left(e^{-\beta}-1\right)^2} - \sum_{i=1}^{n} d_i \frac{A_i}{\left[2e^{-\beta(t_i-1)}(e^{-2\beta}+e^{-\beta}+1)-3e^{-\beta}-3\right]^2} \\ &\quad +6\,(1-\eta)\sum_{i=1}^{n}(1-d_i)t_i^2\frac{\eta e^{-2\beta t_i}\left(3e^{-\beta t_i}-2\right)+e^{-5\beta t_i}\,(1-\eta)}{B_i}, \end{aligned} \tag{16}$$

$$\frac{\partial^2 \ell}{\partial \eta^2} = -\frac{1}{(\eta-1)^2}\sum_{i=1}^{n} d_i - \sum_{i=1}^{n}(1-d_i)\frac{\left(2e^{-\beta t_i}+1\right)^2\left(e^{-\beta t_i}-1\right)^4}{B_i}, \tag{17}$$

and

$$\frac{\partial^2 \ell}{\partial \beta \partial \eta} = -6\sum_{i=1}^{n}(1-d_i)\frac{t_ie^{-2\beta t_i}\left(e^{-\beta t_i}-1\right)}{B_i}, \tag{18}$$

where A_i is given by (11) and

$$B_i = \eta^2 + 2\eta\,(\eta-1)\left(2e^{-3\beta t_i}-3e^{-2\beta t_i}\right)+(\eta-1)^2\left(9e^{-4\beta t_i}+4e^{-6\beta t_i}-12e^{-4\beta t_i}e^{-2\beta t_i}\right).$$

The asymptotic multivariate normal distribution of the ML estimators $\widehat{\beta}_{ML}$ and $\widehat{\eta}_{ML}$ is given by

$$\begin{bmatrix} \widehat{\beta}_{ML} \\ \widehat{\eta}_{ML} \end{bmatrix} \sim N\left(\begin{bmatrix} \beta \\ \eta \end{bmatrix}, \begin{bmatrix} V_{11} & V_{12} \\ V_{21} & V_{22} \end{bmatrix}\right),$$

where V_{11} is the variance of $\widehat{\beta}_{ML}$, V_{22} is the variance of $\widehat{\eta}_{ML}$, and $V_{12}=V_{12}$ is the covariance between $\widehat{\beta}_{ML}$ and $\widehat{\eta}_{ML}$. Approximate $100(1-\upsilon)\%$ Wald-type confidence intervals for β and η are, respectively, given by

$$\widehat{\beta}_{ML} \mp z_{\upsilon/2}\sqrt{\widehat{Var}(\widehat{\beta}_{ML})} \quad \text{and} \quad \widehat{\eta}_{ML} \mp z_{\upsilon/2}\sqrt{\widehat{Var}(\widehat{\eta}_{ML})},$$

where z_υ denotes the upper υ-th percentile of the standard normal distribution. The asymptotic variances of the ML estimators are given by the elements of the inverse of the Fisher's information matrix. The expected Fisher's information matrix is given by

$$i_n(\beta,\eta) = \begin{bmatrix} -E\left(\frac{\partial^2 \ell}{\partial \beta^2}\right) & -E\left(\frac{\partial^2 \ell}{\partial \beta \partial \eta}\right) \\ -E\left(\frac{\partial^2 \ell}{\partial \beta \partial \eta}\right) & -E\left(\frac{\partial^2 \ell}{\partial \eta^2}\right) \end{bmatrix},$$

where the second derivatives with respect to β and η are given by the expressions (16), (17) and (18). The R code for implementing this procedure is presented in an Appendix at the end of the manuscript.

2.4. *A Bayesian approach*

The Bayesian approach is an alternative to the ML estimation method to get the estimators of the parameters of the DB discrete model in presence of right-censored data and cure fraction. Under a Bayesian, approach it is needed to specify a prior distribution for each unknown parameter [25]. From the Bayes' theorem, the posterior distribution of a particular parameter under the model specification is proportional to its prior distribution multiplied by the likelihood of the data. Considering the discrete Bilal distribution, we can assume a gamma prior distribution for the β parameter since it is defined for positive values. Thus, we assume $\beta \sim Gamma(a_\beta, b_\beta)$, where a_β and b_β are known hyperparameters and $Gamma(a, b)$ denotes a gamma distribution with mean a/b and variance a/b^2. In the case of the model in the presence of a cure fraction, since this parameter is defined in the interval $(0, 1)$, we assume a beta prior distribution, that is, $\eta \sim Beta(a_\eta, b_\eta)$, where a_η and b_η are known hyperparameters and $Beta(a, b)$ denotes a beta distribution with mean $a/(a + b)$ and variance $ab/[(a + b)^2(a + b + 1)]$. Further, we assumed prior independence between the parameters β and η.

In this article, posterior summaries of interest were obtained using standard Markov-chain Monte Carlo (MCMC) procedures as the Gibbs sampling and Metropolis-Hastings algorithms [10, 24]. The conditional posterior distributions for each parameter of the model needed for the Gibbs-Metropolis-Hastings algorithms are not presented in this study, since we are using existing Bayesian programs available in R software which only requires to define the likelihood function and the prior distributions for the parameters of the model [10]. From the simulation algorithm, we generated 1,005,000 samples of the joint posterior distribution of interest with a burn-in phase of 5,000 simulated samples that were discarded to eliminate the effect of the initial values in the iterative procedure and considering a thinning interval of size 200 to have approximately independent samples. The Bayes estimates of the parameters were obtained as the mean of the simulated samples drawn from the joint posterior distribution for all parameters of the model, that is, we are assuming a quadratic loss function to get the Bayesian estimators of interest. The convergence of the simulated sequences was monitored by using traceplots and the Geweke diagnostic criteria [12, 26]. The Geweke convergence diagnostic criterion is based on a z score that compares the difference in the two means of non-overlapping sections of a simulated Markov chain, divided by the asymptotic standard error of the difference. This z score asymptotically follows a standardized normal distribution, so we obtain convergence for a chain if its correspondent absolute z score is less than 1.96. Posterior summaries of interest were obtained using the MCMCpack package of the R software [34]. See the Appendix at the end of the manuscript for details about the R code used in this article.

3. Examples

This section illustrates the proposed methodology considering three lifetime data applications introduced in the literature. We also compare the fits of the discrete Bilal distribution to the datasets with some competitive models such as DsFx-I [16], discrete Lindley [27], discrete Rayleigh [47], and discrete Burr-Hatke [17] distributions. These distributions also have only one parameter to be estimated where they were adapted to have $S(t) = 1$ when $t = 0$. The fitted models are compared using Akaike and Bayesian information criteria (AIC and BIC). In the discrimination of the proposed models, we also used the corrected AIC criterion (AICC), suggested for situations with small sample sizes, as a better alternative to the AIC criterion [29].

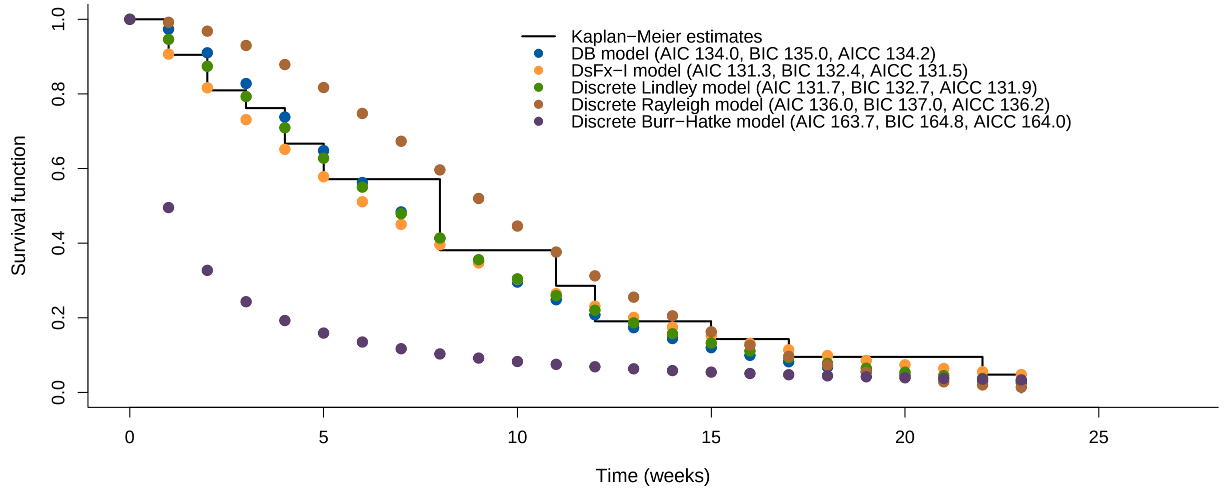


Figure 2. Survival function for the acute leukemia patients' data estimated by the Kaplan-Meier method and by using the models based on the DB, DsFx-I, discrete Lindley, discrete Rayleigh, and discrete Burr-Hatke distributions.

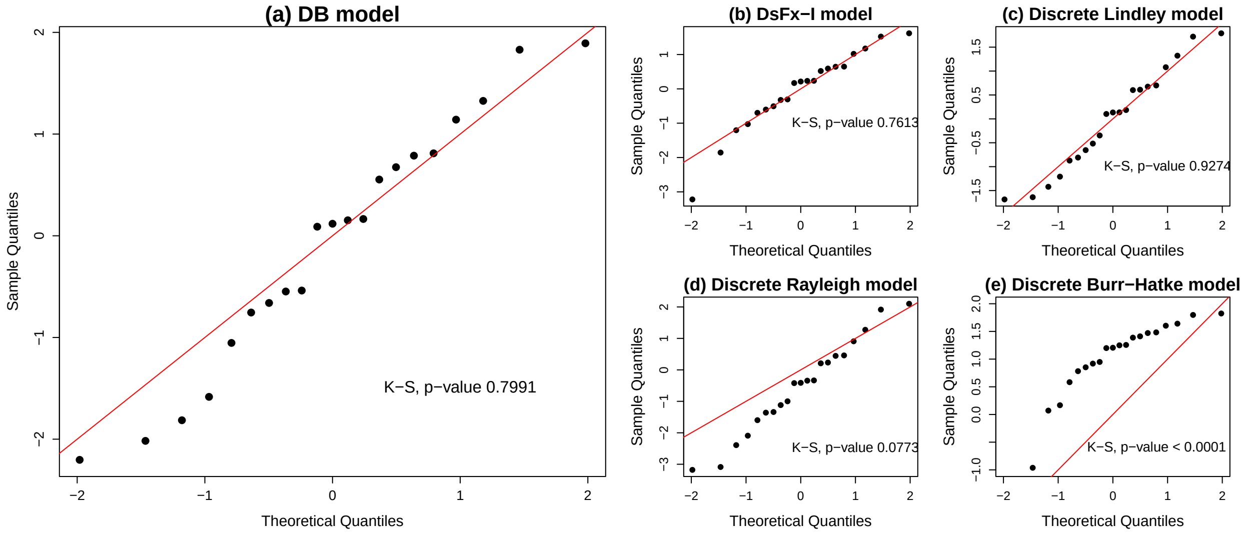


Figure 3. Randomized quantile residuals for the models based on the (a) DB, (b) DsFx-I, (c) discrete Lindley, (d) discrete Rayleigh, and (e) discrete Burr-Hatke distributions, fitted to the data from patients with acute leukemia (placebo group).

3.1. *Patients with acute leukemia*

In this subsection, a numerical example with complete data is presented to illustrate the applicability of the discrete Bilal distribution. From a total of 42 patients with acute leukemia in a clinical trial investigating the effect of 6-mercaptopurine to prolong the remission times, divided in two groups (21 patients receiving the 6-mercaptopurine drug and 21 patients receiving placebo) [20], we consider in this illustration only the remission times (uncensored data) for the $n = 21$ patients treated with placebo, given by 1, 1, 2, 2, 3, 4, 4, 5, 5, 8, 8, 8, 8, 11, 11, 12, 12, 15, 17, 22, and 23 weeks. Using the maxLik package of the R software, we obtained an ML estimate $\widehat{\beta}_{ML} = 0.10214$ for the parameter β of the discrete Bilal distribution, with a standard error equal to 0.01611. An approximate 95% Wald-type CI for β is given by $(0.0706, 0.1337)$. Figure 2 compares the survival function estimated by the Kaplan-Meier method and fitted by the parametric models based on the discrete Bilal, DsFx-I, discrete Lindley, discrete Rayleigh, and discrete Burr-Hatke distributions. Figure 2 also shows the corresponding AIC, BIC, and AICC values. Figure 3 shows the resulting residual analysis and the corresponding p-values for the Kolmogorov-Smirnov (K-S) test for normality. The model based on the DB distribution shows good fit for the data, as well as the model based on the Lindley distribution. Figures 2 and 3 suggest that the discrete Rayleigh and discrete Burr-Hatke distributions do not fit the data well.

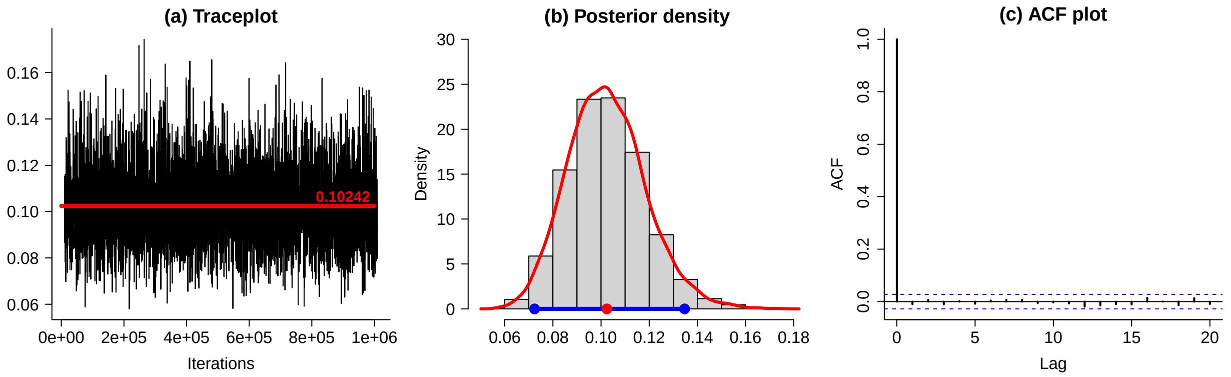


Figure 4. Posterior samples for the model parameters of the DB distribution applied to data from patients with acute leukemia (placebo group). (a) Traceplot of posterior samples, (b) histogram and posterior density for β with the correspondent 95% HDI (blue line), and (c) auto-correlation function (ACF) plot for the posterior samples of the model parameter.

For the Bayesian data analysis, it was assumed an approximately non-informative gamma prior distribution for the parameter β of the DB distribution, that is, $\beta \sim Gamma(0.001, 0.001)$. Posterior samples for β are described in Figure 4. The traceplot presented in panel (a) shows the evolution of the samples generated by the MCMC method over the iterations, indicating good convergence of the simulation algorithm. The plot of the autocorrelation function (ACF) shows that the posterior samples are uncorrelated (panel (c)). The corresponding Geweke z-score is given by 0.209, also suggesting satisfactory convergence of the samples to a stable distribution. The posterior mean for β is estimated by $\widehat{\beta}_{Bayes} = 0.10242$, and the corresponding 95% HDI (highest density interval) is $(0.0723, 0.1347)$. The 95% HDI is plotted on the histogram shown in panel (b) of Figure 4. We observe that the ML and the Bayesian estimates are are very close assuming a non-informative prior distribution for the parameter β. It is important to point out that more accurate inference results could be obtained under a Bayesian approach using informative priors elicited from medical experts.

Alternatively, we also assumed a half-normal prior distribution for β with hyperparameter a_β and density given by

$$f(\beta) = \frac{2a_\beta}{\pi} \exp\left(-\frac{\beta^2 a_\beta^2}{\pi}\right),$$

where $\beta > 0$ and $a_\beta > 0$. The half-normal distribution has mean $1/a_\beta$ and variance $(\pi - 2)/(2a_\beta^2)$. Considering $a_\beta = 0.01$, the posterior mean for β is estimated by $\widehat{\beta}_{Bayes} = 0.1050$, and the corresponding 95% HDI is given by $(0.0726, 0.1359)$. In addition, assuming $a_\beta = 25$, we have $\widehat{\beta}_{Bayes} = 0.09562$ and the corresponding 95% HDI is given by $(0.0675, 0.1225)$. In this brief sensitivity analysis considering different prior distributions for the parameter β, we observe that the posterior mean for β is not very sensitive for the choice of priors, since the obtained inferences are very close.

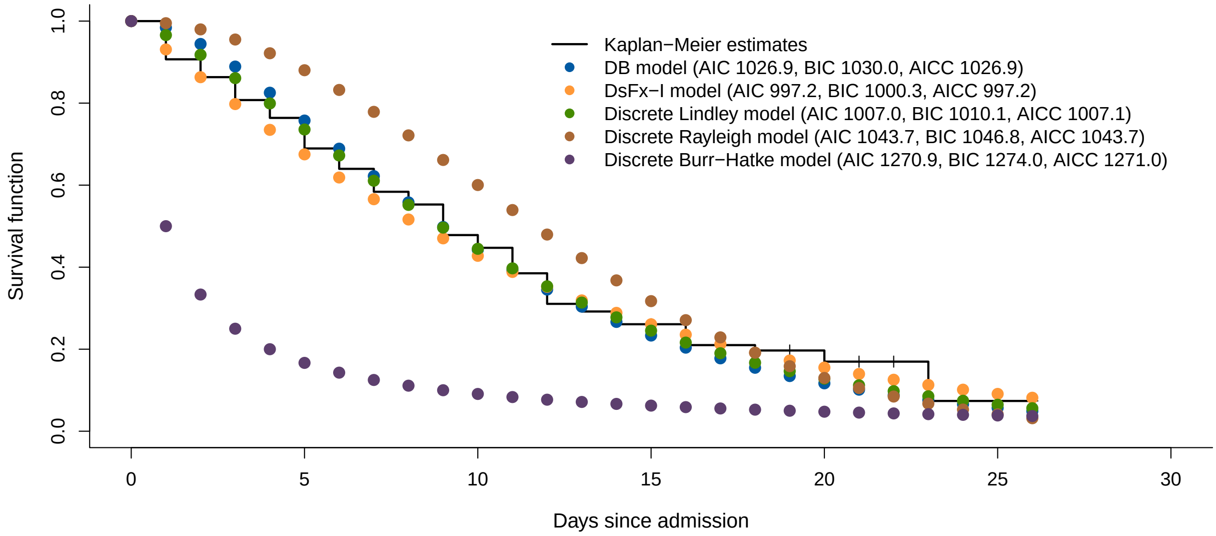


Figure 5. Survival function for the COVID-19 patients' data estimated by the Kaplan-Meier method and by using the models based on the DB, DsFx-I, discrete Lindley, discrete Rayleigh, and discrete Burr-Hatke distributions.

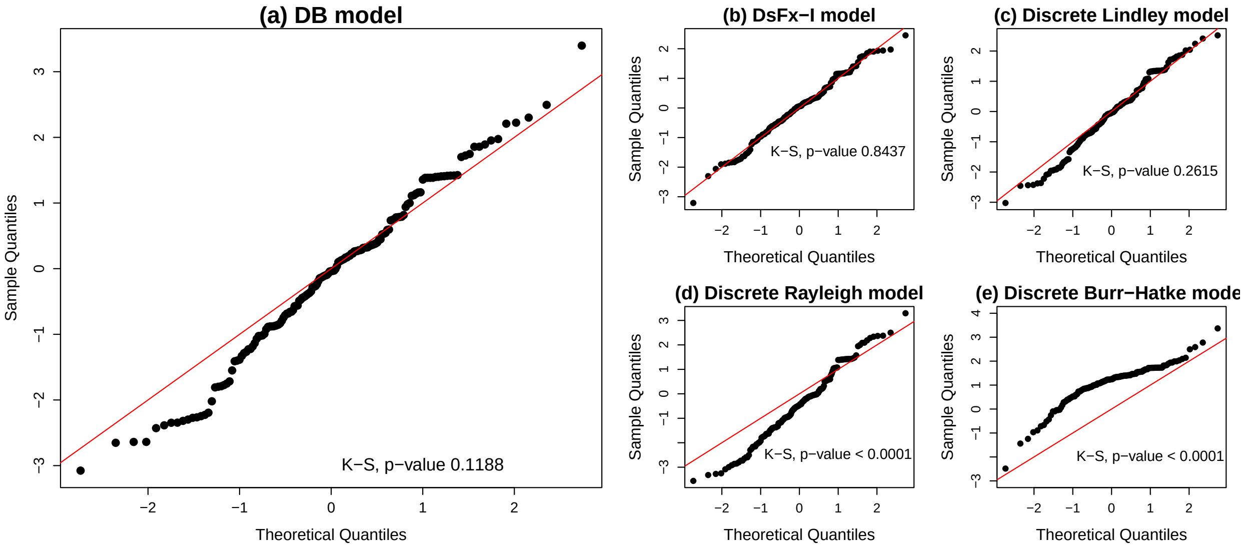


Figure 6. Randomized quantile residuals for the models based on the (a) DB, (b) DsFx-I, (c) discrete Lindley, (d) discrete Rayleigh, and (e) discrete Burr-Hatke distributions, fitted to data from hospitalized patients with COVID-19.

3.2. Hospitalized patients with COVID-19

The study by Paranjpe et al. [41] assessed the association between administration of in-hospital anticoagulation and survival times in a large cohort of hospitalized patients with COVID-19 in the Mount Sinai Health System, New York City. In this example, we consider a subsample of $n = 161$ patients who required mechanical ventilation and were not treated with in-hospital systemic anticoagulation. The variable of interest is the time from admission to the hospital until time to death or censoring, in days. We have 15 (9.3%) censored observations. As the data were available in figures and not in numerical form, we used the open-source software WebPlotDigitizer, a web-based tool to extract numerical data from images [13, 46]. In this example, the ML estimate for the parameter β of the discrete Bilal distribution is given by $\widehat{\beta}_{ML} = 0.07731$ (standard error given by 0.00454 and a 95% Wald-type CI given by 0.0684 to 0.0862). Figure 5 compares the survival function estimated by the Kaplan-Meier method and fitted by parametric models based on the DB and other lifetime distributions. Figure 6 shows the randomized quantile residuals from the fitted models. We note that the DB distribution fitted the data as well as the DsFx-I and the discrete Lindley distributions, but the models based on the discrete Rayleigh and discrete Burr-Hatke distributions did not fit the data well.

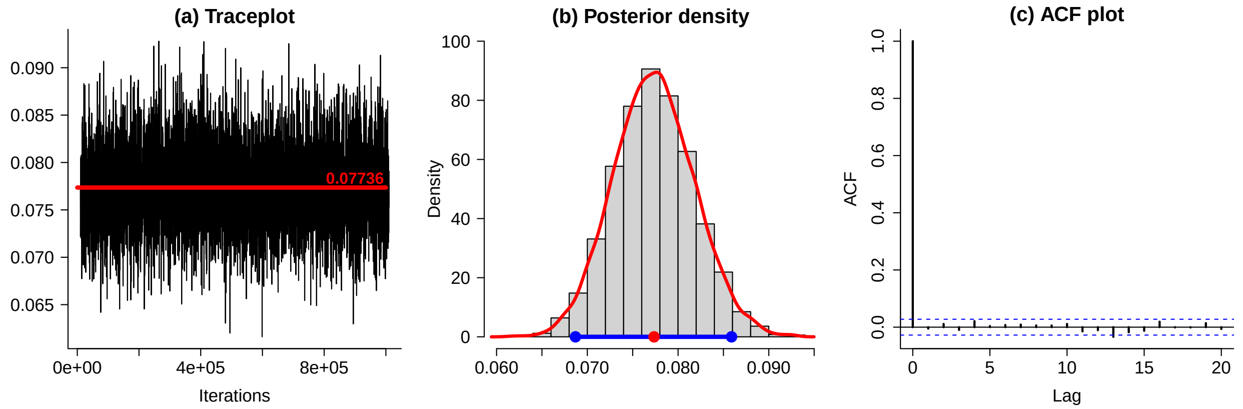


Figure 7. Posterior samples for the parameter of the DB distribution applied to data from hospitalized patients with COVID-19. (a) Traceplot of posterior samples, (b) histogram and posterior density with the correspondent 95% HDI (blue line), and (c) auto-correlation function (ACF) plot for the posterior samples of the model parameter.

Under a Bayesian analysis, as in the previous example, we assumed a gamma prior distribution for the parameter β given by $\beta \sim Gamma(0.001, 0.001)$. Figure 7 describes the posterior samples for β. The traceplot in panel (a) shows that the MCMC algorithm converged, and the corresponding Geweke z-score is given by 0.033, also suggesting satisfactory convergence. Panel (b) describes the shape of the posterior distribution and shows the 95% HDI. The posterior mean for β is estimated by $\widehat{\beta}_{Bayes} = 0.07736$, and the corresponding 95% HDI is $(0.0686, 0.0859)$. This MCMC estimate is closer to the corresponding ML estimate. The ACF plot shows that autocorrelations are not significantly different from zero (panel (c)). We have similar results assuming a prior half-normal distribution for β with hyperparameter equal to 0.01, that is, $\widehat{\beta}_{Bayes} = 0.07762$ with a 95% HDI given by $(0.0687, 0.0861)$.

3.3. *Recurrence rates of pelvic tumors with marginal or intracapsular margins*

In this example, we consider a model for survival data in the presence of a cure fraction based on the DB distribution. Let us consider the data from a study undertaken at the Musculoskeletal Oncology Center of the First Affiliated Hospital of Sun Yat-Sen University, China, between the years 2003 and 2013 [49]. The objective of this study was to evaluate the effectiveness of reconstruction with a modular hemipelvic endoprosthesis after pelvic tumor resection. The recurrence times of pelvic tumors with marginal or intracapsular margins are given by 3, 7, 11+, 18, 22+, 25, 28, 32+, 34+, 35, 35+, 36+, 40+, 40+, 41, 54+, 66+, 76+, 84+, 88+, and 92+ months, where + denotes a censored observation. Assuming the model described in subsection 2.3 to these data, we obtained the ML estimates $\widehat{\beta}_{ML} = 0.03029$ (standard error given by 0.01091 and $95\%CI$ given by 0.0089 to 0.0517) and $\widehat{\eta}_{ML} = 0.58444$ (standard error given by 0.13671 and $95\%CI$ given by 0.3165 to 0.8524). Figure 8 compares the Kaplan-Meier estimates and the ML estimates of the survival function corresponding to the model based on DB distribution to some other assumed discrete distributions. We observe that the obtained inference results provided

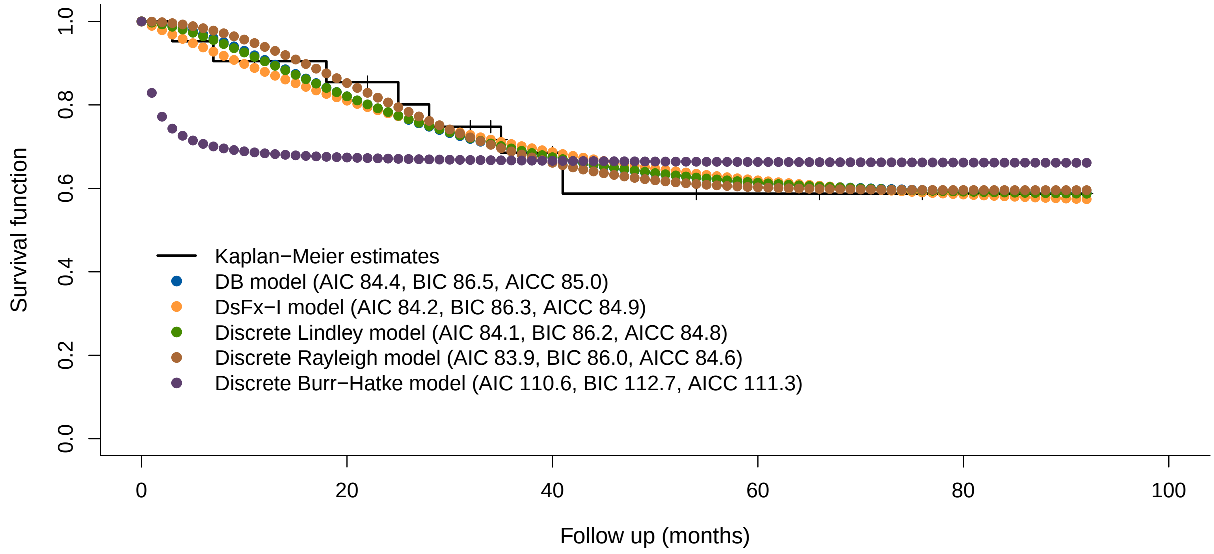


Figure 8. Survival function for the pelvic tumors patients' data estimated by the Kaplan-Meier method and by using the models based on the DB, DsFx-I, discrete Lindley, discrete Rayleigh, and discrete Burr-Hatke distributions.

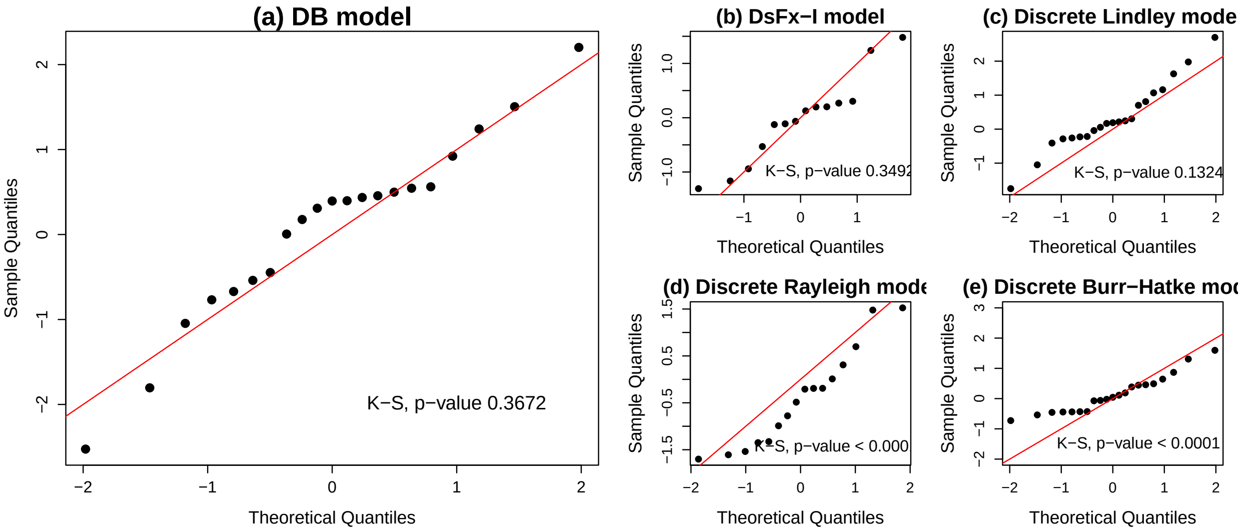


Figure 9. Randomized quantile residuals for the models based on the (a) DB, (b) DsFx-I, (c) discrete Lindley, (d) discrete Rayleigh, and (e) discrete Burr-Hatke distributions, fitted to data from patients with pelvic tumors.

by the models based on DB and discrete Lindley distributions are almost identical and are overlapping on the plots of Figure 8. Figure 9 describes the randomized quantile residuals from the fitted models. The model based on the DB distribution shows a good fit for the data as well as the models based on the DsFx-I and the Lindley distribution. The models based on the discrete Rayleigh and discrete Burr-Hatke distributions did not fit the data well.

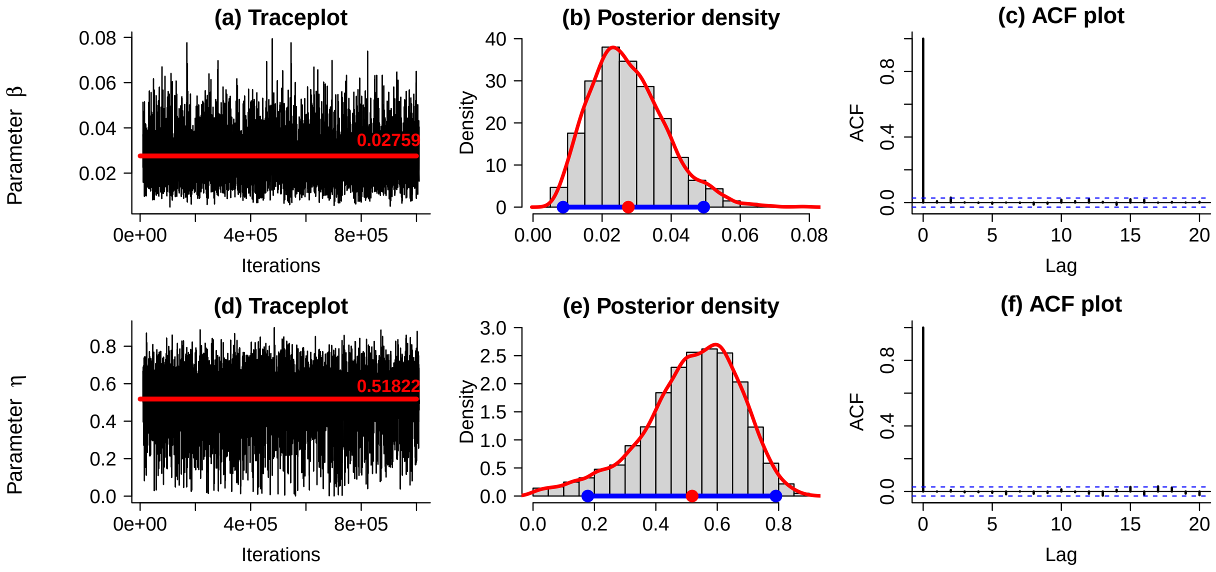


Figure 10. Posterior samples for the model parameters of the DB distribution applied to data from patients with pelvic tumors. (a) Traceplots of posterior samples, (b) histograms and posterior densities with the correspondent 95% HDI (blue lines), and (c) auto-correlation function (ACF) plots for the posterior samples of the model parameters.

Under a Bayesian approach introduced in Section 2.4, we assumed the prior distributions $\beta \sim Gamma(0.001, 0.001)$ and $\eta \sim Beta(1, 1)$, that are approximately non-informative prior distributions for the model parameters. The posterior means for β and η are estimated by $\widehat{\beta}_{Bayes} = 0.02759$ (95% HDI given by 0.0087 to 0.0494) and $\widehat{\eta}_{Bayes} = 0.51822$ (95% HDI 0.1783 to 0.7911), respectively. Figure 10 describes the posterior samples for the parameters β and η. Panels (a) and (d) show that the MCMC chains reached satisfactory convergence for both parameters and the Geweke Z scores for β and η chains are given by 0.3885 and 0.2049, respectively. Panels (b) and (e) describe the posterior densities and the 95% HDI. Panels (c) and (f) shows that autocorrelations within each chain were reasonably low. If we alternatively consider a half-normal prior distribution for β with hyperparameter 0.01, we have $\widehat{\beta}_{Bayes} = 0.03182$ (95% HDI given by 0.0121 to 0.0550) and $\widehat{\eta}_{Bayes} = 0.55061$ (95% HDI given by 0.2712 to 0.8143).

4. Concluding Remarks

The statistical literature does not present many papers on the Discrete Bilal (DB) distribution recently introduced by Altun et al. [6], although this distribution has only one parameter, which makes it easier to obtain inferences of interest, which could be a good alternative in the analysis of lifetime data. The DB distribution was originally proposed for continuous data analysis. The novelty of this study was to extend the DB distribution for discrete data in the presence of censored observations and a cure fraction. Discrete survival data may appear in many applications where the original data are indeed discrete, making continuous survival models inappropriate. Another possibility used in many applications would be to discretize continuous survival data, possibly rounding off responses (times to failure) given on a real scale to discrete data, for example, data rounded to hours, days, weeks, or months in a discrete form. The present study investigated the use of different statistical inference techniques under the classical and Bayesian approaches, assuming the discrete DB distribution, considering data in the presence of right censoring and also a fraction of cure, which are common in medical studies. The classical inference techniques given by the maximum likelihood method assume normal asymptotic approximations for the estimators, which can be a disadvantage given that many medical data sets have small sample sizes, especially in the presence of a large proportion of censored data and cure fraction, common facts in clinical trials. The use of Bayesian methods assuming MCMC techniques to generate samples of the joint posterior distribution of interest can be a good alternative to finding point and interval estimators with good accuracy, as these techniques do not depend on large sample sizes to obtain precise inferences. The use of existing and freely accessible computer programs facilitates the generation of samples of the joint posterior distribution of interest to obtain Monte Carlo estimators for the parameters of interest. Another advantage of the Bayesian methodology could be the use of informative prior distributions with information from medical experts, which can lead to inferences with better and more accuracy. In the three examples considered in this study, we used non-informative prior distributions for the parameters of the proposed model, with results similar to those obtained using maximum likelihood methods.

In future work, we will assume prior distributions for the parameters β and η with a dependence structure, possibly using copula functions or a hierarchical Bayesian analysis. Another possibility for a future study is to consider the presence of a covariate vector $\mathbf{x}$ while assuming a regression structure for the parameter β in the parametrical form $exp(\alpha\mathbf{x})$ and a logistic structure for the cure fraction η. As an alternative to the asymptotic approximation in small samples, we could also use in future work a bootstrap approach to find 95% confidence intervals for the parameters of the discrete DB distribution. When the data is subject to right censoring, Efron [15] proposed using bootstrap techniques to find confidence intervals for the parameters of an unknown distribution. The three applications with real data sets showed that the model based on the DB distribution performs at least as well as some other traditional discrete models such as the DsFx-I, Lindley, Rayleigh, and Burr-Hatke discrete distributions. In general, the model based on a DB distribution was shown to be a good alternative for analyzing discrete survival data in the presence or absence of censored data, even including the presence of immune individuals. Furthermore, the model can be easily implemented in computer programs like R software, as shown in the Appendix. In the Bayesian analysis, the gamma prior distribution was chosen to model the uncertainty about the parameter β, given that β assumes only positive values. We also considered a brief sensitivity analysis assuming a half-normal prior distribution as an alternative to the gamma distribution. From the applications of the model based on the DB distribution in real data, we observed that the posterior mean for β is not very sensitive to the choice of prior distributions for the parameter β. From the results obtained in this study, we can conclude that the proposed methodology can be very useful for researchers working with discrete medical life data in the presence of a cure fraction, a common situation in both medical and engineering applications [23, 32, 42, 44].

Acknowledgement

This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001. The research of EZM and JAA is supported by the Brazilian organization CNPq. A preliminary version of this article was published in arXiv.

REFERENCES

1. A. M. Abd-Elrahman, *Utilizing ordered statistics in lifetime distributions production: A new lifetime distribution and applications*, Journal of Probability and Statistical Science, vol. 11, no. 2, pp. 153–164, 2013.
2. A. M. Abd-Elrahman, *Reliability estimation under type-II censored data from the generalized Bilal distribution*, Journal of the Egyptian Mathematical Society, vol. 27, no. 1, pp. 1–15, 2019.
3. A. M. Abd-Elrahman, and S. F. Niazi, *Approximate Bayes estimators applied to the Bilal model*, Journal of the Egyptian Mathematical Society, vol. 25, no. 1, pp. 65–70, 2017.
4. Z. Akhter, E. M. Almetwally, and C. Chesneau, *On the generalized Bilal distribution: some properties and estimation under ranked set sampling*, Axioms, vol. 11, no. 4, pp. 173, 2022.
5. M. H. Alamatsaz, S. Dey, T. Dey, and S. S. Harandi, *Discrete generalized Rayleigh distribution*, Pakistan Journal of Statistics, vol. 32, no. 1, pp. 1–20, 2016.
6. E. Altun, M. El-Morshedy, and M. S. Eliwa, *A study on discrete Bilal distribution with properties and applications on integer-valued autoregressive process*, Revstat Statistical Journal, vol. 18, pp. 70–99, 2020.
7. M. Amico, and I. Van Keilegom, *Cure models in survival analysis*, Annual Review of Statistics and Its Application, vol. 5, pp. 311–342, 2018.

8. M. R. P. Cardial, J. B. Fachini-Gomes, and E. Y. Nakano, *Exponentiated discrete Weibull distribution for censored data*, Brazilian Journal of Biometrics, vol. 38, no. 1, pp. 35–56, 2020.
9. S. Chakraborty, *Generating discrete analogues of continuous probability distributions - A survey of methods and constructions*, Journal of Statistical Distributions and Applications, vol. 2, no. 1, pp. 1–30, 2015.
10. S. Chib, and E. Chakraborty, *Understanding the Metropolis-Hastings algorithm*, The American Statistician, vol. 49, no. 4, pp. 327–335, 1995.
11. F. Corbière, D. Commenges, J. M. Taylor, and P. Joly, *A penalized likelihood approach for mixture cure models*, Statistics in Medicine, vol. 28, no. 3, pp. 510–524, 2009.
12. M. K. Cowles, and B. P. Carlin, *Markov Chain Monte Carlo convergence diagnostics: A comparative review*, Journal of American Statistical Association, vol. 91, no. 434, pp. 883–904, 1994.
13. D. Drevon, S. R. Fursa, and A. L. Malcolm, *Intercoder reliability and validity of WebPlotDigitizer in extracting graphed data*, Behavior Modification, vol. 41, no. 3, pp. 323–339, 2017.
14. P. K. Dunn, and G. K. Smyth, *Randomized quantile residuals*, Journal of Computational and Graphical Statistics, vol. 5, no. 3, pp. 236–244, 1996.
15. B. Efron, *Censored data and the bootstrap*, Journal of the American Statistical Association, vol. 76, no. 374, pp. 312–319, 1981.
16. M. S. Eliwa, and M. El-Morshedy, *A one-parameter discrete distribution for over-dispersed data: Statistical and reliability properties with applications*, Journal of Applied Statistics, pp. 1–21, 2021.
17. M. El-Morshedy, M. S. Eliwa, and E. Altun, *Discrete Burr-Hatke distribution with properties, estimation methods and regression model*, IEEE Access, vol. 8, pp. 74359–74370, 2020.
18. V. T. Farewell, *The use of mixture models for the analysis of survival data with long-term survivors*, Biometrics, vol. 38, no. 4, pp. 1041–1046, 1982.
19. R. Flynn, *Survival analysis*, Journal of Clinical Nursing, vol. 21, pp. 2789–2797, 2012.
20. E. J. Freireich, E. Gehan, E. Frei-3rd, L. R. Schroeder, I. J. Wolman, R. Anbari, E. O. Burgert, S. D. Mills, D. Pinkel, O. S. Selawry, J. H. Moon, B. R. Gendel, C. L. Spurr, R. Storrs, F. Haurani, B. Hoogstraten, and S. Lee, *The effect of 6-mercaptopurine on the duration of steroid induced remissions in acute leukemia - A model for evaluation of other potentially useful therapy* Blood, vol. 21, no. 6, pp. 699–716, 1963.
21. B. C. L. Freitas, M. V. Oliveira-Peres, J. A. Achcar, and E. Z. Martinez, *Classical and Bayesian inference approaches for the exponentiated discrete Weibull model with censored data and a cure fraction*, Pakistan Journal of Statistics and Operation Research, vol. 17, no. 2, pp. 467–481, 2021.
22. B. C. L. Freitas, M. V. Oliveira-Peres, J. A. Achcar, and E. Z. Martinez, *Bayesian and maximum likelihood inference approaches for the discrete generalized Sibuya distribution with censored datan*, Electronic Journal of Applied Statistical Analysis, vol. 15, no. 1, pp. 50–74, 2022.
23. D. I. Gallardo, M. Castro, and H. W. Gómez, *An alternative promotion time cure model with overdispersed number of competing causes: an application to melanoma data*, Mathematics, vol. 9, no. 15, pp. 1815, 2021.
24. A. E. Gelfand, and A. F. Smith, *Sampling-based approaches to calculating marginal densities*, Journal of the American Statistical Association, vol. 85, no. 410, pp. 398–409, 1990.
25. A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian data analysis*, 3rd ed., Chapman Hall, 2013.
26. J. Geweke, *Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments*, Bayesian Statistics, vol. 4, pp. 641–649, 1992.
27. E. Gómez-Déniz, and E. Calderín-Ojeda, *The discrete Lindley distribution: properties and applications*, Journal of Statistical Computation and Simulation, vol. 81, no. 11, pp. 1405–1416, 2011.
28. A. Henningsen, and O. Toomet, *maxLik: A package for maximum likelihood estimation in R*, Computational Statistics, vol. 26, no. 3, pp. 443–458, 2011.
29. C. M. Hurvich, and C. L. Tsai, *Regression and time series model selection in small samples*, Biometrika, vol. 76, no. 2, pp. 297–307, 1989.
30. P. C. Lambert, *Modeling of the cure fraction in survival studies*, The Stata Journal, vol. 7, no. 3, pp. 351–375, 2007.
31. J. Le Rademacher, and X. Wang, *Time-to-event data: An overview and analysis considerations*, Journal of Thoracic Oncology, vol. 16, no. 7, pp. 1067–1074, 2021.
32. J. Leão, M. Bourguignon, D. I. Gallardo, R. Rocha, and V. Tomazella, *A new cure rate model with flexible competing causes with applications to melanoma and transplantation data*, Statistics in Medicine, vol. 39, no. 24, pp. 3272–3284, 2020.
33. R. A. Maller, and X. Zhou, *Survival analysis with long-term survivors*, John Wiley & Sons, New York, 1996.
34. A. D. Martin, and K. M. Quinn, *Applied Bayesian inference in R using MCMCpack*, R News, vol. 6, pp. 2–7, 2006.
35. E. Z. Martinez, J. A. Achcar, A. A. Jácome, and J. S. Santos, *Mixture and non-mixture cure fraction models based on the generalized modified Weibull distribution with an application to gastric cancer data*, Computer Methods and Programs in Biomedicine, vol. 112, no. 3, pp. 343–355, 2013.
36. J. Myers, *Survival analysis techniques in clinical research*, The Journal of the Kentucky Medical Association, vol. 105, no. 11, pp. 545–550, 2007.
37. T. Nakagawa, and S. Osaki, *The discrete Weibull distribution*, IEEE Transactions on Reliability, vol. 24, no. 5, pp. 300-301, 1975.
38. V. Nekoukhou, and H. Bidram, *The exponentiated discrete Weibull distribution*, SORT Statistics and Operations Research Transactions, vol. 39, pp. 127–146, 2015.
39. R. P. Oliveira, M. V. Oliveira-Peres, E. Z. Martinez, and J. A. Achcar, *Use of a discrete Sushila distribution in the analysis of right-censored lifetime data*, Model Assisted Statistics and Applications, vol. 14, no. 3, pp. 255–2681, 2019.
40. M. Othus, B. Barlogie, M. L. LeBlanc, and J. J. Crowley, *Cure models as a useful statistical tool for analyzing survival*, Clinical Cancer Research, vol. 18, no. 14, pp. 3731–3736, 2012.
41. I. Paranjpe, V. Fuster, A. Lala, A. J. Russak, B. S. Glicksberg, M. A. Levin, A. W. Charney, J. Narula, Z. A. Fayad, E. Bagiella, S. Zhao, and G. N. Nadkarni, *Association of treatment dose anticoagulation with in-hospital survival among hospitalized patients with*

COVID-19, Journal of the American College of Cardiology, vol. 76, no. 1, pp. 122–124, 2020.

42. M. Pedrosa-Laza, A. López-Cheda, R. Cao, *Cure models to estimate time until hospitalization due to COVID-19*, Applied Intelligence, vol. 52, no. 1, pp. 794-807, 2022.
43. Y. Peng, and B. Yu, *Cure models: Methods, applications, and implementation*, CRC Press, 2021.
44. S. Rafati, M. R. Baneshi, and A. Bahrampour, *Factors affecting long-survival of patients with breast cancer by non-mixture and mixture cure models using the Weibull, log-logistic and Dagum distributions: a Bayesian approach*, Asian Pacific Journal of Cancer Prevention: APJCP, vol. 21, no. 2, pp. 485, 2020.
45. F. H. Riad, B. Alruwaili, A. M. Gemeay, and E. Hussam, *Statistical modeling for COVID-19 virus spread in Kingdom of Saudi Arabia and Netherlands*, Alexandria Engineering Journal, vol. 61, no. 12, pp. 9849-9866, 2022.
46. A. Rohatgi, *WebPlotDigitizer version 4.4*, Available from: https://automeris.io/WebPlotDigitizer/.
47. D. Roy, *Discrete Rayleigh distribution*, IEEE Transactions on Reliability, vol. 53, no. 2, pp. 255–260, 2004.
48. X. Shi, Y. Shi, and K. Zhou, *Estimation for entropy and parameters of generalized Bilal distribution under adaptive type II progressive hybrid censoring scheme*, Entropy, vol. 23, no. 2, pp. 206, 2021.
49. B. Wang, X. Xie, J. Yin, C. Zou, J. Wang, G. Huang, Y. Wang, and J. Shen, *Reconstruction with modular hemipelvic endoprosthesis after pelvic tumor resection: a report of 50 consecutive cases*, PLoS One, vol. 10, no. 5, pp. e0127263, 2015.

Appendice: R Codes

Under the frequentist approach, the following R code is used to implement the model for survival data with a cure fraction based on the DB distribution, as presented in subsection 2.3. We used the function `maxLik` of the maxLik package [28] for the maximization of the likelihood function.

```
# Reading data (Wang et al., 2015)
t <- c(3,7,11,18,22,25,28,32,34,35,35,36,40,40,41,54,66,76,84,88,92)
d <- c(1,1,0,1,0,1,1,0,0,1,0,0,0,0,1,0,0,0,0,0,0)
n <- length(t)  # the sample size
K <- 2          # number of parameters

# Loading the maxLik package
library(maxLik)
# The likelihood function
log.f <- function(parms) {
beta  <- parms[1]
eta   <- parms[2]
if (parms[1]<0) return(-Inf)
if (parms[2]<0) return(-Inf)
if (parms[2]>1) return(-Inf)
p    <- exp(-beta)
St0  <- (3-2*p^t)*p^(2*t)
ft0  <- p^(2*(t-1))*(p-1)*(2*p^(t-1)*(p^2+p+1)-3*p-3)
St   <- eta + (1-eta)*St0
ft   <- (1-eta)*ft0
like <- ft^d * St^(1-d)
L    <- sum(log(like))
if (is.na(L)==TRUE) {return(-Inf)} else {return(L)} }
# Obtaining the ML estimates
mle  <- c()
mle  <- maxLik(logLik=log.f,start=c(0.08,0.6))
summary(mle)
betaDB <-mle$estimate[1]
etaDB  <-mle$estimate[2]
s <- vcov(mle)
# The 95% confidence intervals
llimDB  <- round(betaDB - qnorm(0.975) * sqrt(s[1,1]),4)
```

```
ulimDB  <- round(betaDB + qnorm(0.975) * sqrt(s[1,1]),4)
llimDBe <- round(etaDB  - qnorm(0.975) * sqrt(s[2,2]),4)
ulimDBe <- round(etaDB  + qnorm(0.975) * sqrt(s[2,2]),4)
cat("n = ",n,"\n")
cat("Beta  = ",betaDB, "95%CI: (",llimDB,",",ulimDB, ") \n")
cat("Eta   = ",etaDB,  "95%CI: (",llimDBe,",",ulimDBe, ") \n")
# Calculating AIC, BIC and AICC
aic  <- AIC(mle)
bic  <- AIC(mle,k = log(n))
aicc <- aic + (2*K^2+2*K)/(n-K-1)
cat("AIC = ",aic,", BIC = ",bic,", AICC = ",aicc,"\n")
```

This is the R code for the Bayesian model for survival data with a cure fraction based on the DB distribution, as presented in subsection 2.4:

```
# Loading the MCMCpack package
library(MCMCpack)
# The log posterior function
log.post <- function(t,d,parms) {
beta <- parms[1]
eta  <- parms[2]
if (parms[1]<0) return(-Inf)
if (parms[2]<0) return(-Inf)
if (parms[2]>1) return(-Inf)
p    <- exp(-beta)
St0  <- (3-2*p^t)*p^(2*t)
ft0  <- p^(2*(t-1))*(p-1)*(2*p^(t-1)*(p^2+p+1)-3*p-3)
St   <- eta + (1-eta)*St0
ft   <- (1-eta)*ft0
like <- ft^d * St^(1-d)
log.like  <- sum(log(like))
prior     <- dgamma(beta,0.001,0.001)*dbeta(eta,1,1)
log.prior <- log(prior)
L <- log.like + log.prior
if (is.na(L)==TRUE) {return(-Inf)} else {return(L)} }

# Obtaining the MCMC estimates
posterior <- MCMCmetrop1R(log.post,theta.init=c(beta=0.05,eta=0.6),
burnin=10000, mcmc=1000000, thin=200, logfun=T, t=t, d=d, verbose=100000,
tune = 1)
varnames(posterior) <- c("beta","eta")
summary(posterior)
# Obtaining the HPD intervals
HPDinterval(posterior, prob = 0.95)
# Geweke z scores
geweke.diag(posterior)
```