Revista Eletrônica Paulista de Matemática
ISSN2316-9664
Volume 14, fev. 2018
Edição Ermac

Thalita Cristina da Costa
UNESP – Universidade Estadual Paulista “Júlio de Mesquita Filho”
thalita_cris91@hotmail.com

Luís Fernando Jorge
UNESP – Universidade Estadual Paulista “Júlio de Mesquita Filho”
jorge.luis.fernando@hotmail.com

Fernando Antônio Moala
UNESP – Universidade Estadual Paulista “Júlio de Mesquita Filho”
f.moala@unesp.br

Sérgio Minoru Oikawa
UNESP – Universidade Estadual Paulista “Júlio de Mesquita Filho”
smoikawa@fct.unesp.br

Análise bayesiana para a distribuição exponencial generalizada geométrica

Resumo

Neste trabalho será realizado um estudo da análise bayesiana para a nova distribuição exponencial generalizada geométrica proposta por Bidram, Behboodian e Towhidi (2013) e obtida através do processo de composição das distribuições exponencial generalizada e geométrica. Serão apresentadas as principais características desta nova distribuição e será verificado se as estimativas dos seus parâmetros, obtidas através do método de máxima verossimilhança e pelo método bayesiano, são semelhantes. Além disso, para verificar se tal distribuição é a melhor ajustada aos dados será comparada a distribuição em estudo com outras distribuições. Este trabalho traz uma visão completa: teoria, estudo de simulação e aplicação sobre tal distribuição, diferente do apresentado no ERMAC (Costa; Jorge; Moala, 2018), onde o mesmo contém apenas o exemplo de aplicação.
Palavras-chave: Distribuição exponencial generalizada geométrica. Análise bayesiana. Máxima verossimilhança. MCMC.

Abstract

In this work, a study of the Bayesian analysis for the new generalized exponential geometric distribution proposed by Bidram, Behboodian and Towhidi (2013) will be carried out and obtained through the process of generalized exponential and geometric distributions. The main characteristics of this new distribution will be presented and it will be verified if the estimates of its parameters, obtained through the maximum likelihood method and by the Bayesian method, are similar. In addition, to verify if such distribution is the best fit to the data will be compared the distribution under study with other distributions. This work brings a complete view: theory, simulation study and application on such distribution, different from the one presented in ERMAC (Costa; Jorge; Moala, 2018), where it contains only the application example.
Keywords: Generalized geometric exponential distribution. Bayesian analysis. Maximum likelihood. MCMC.

Artigo recebido em set. 2018 e aceito em fev. 2019.

1 Introdução

Na literatura existem várias distribuições de probabilidade para modelar os tempos de vida de equipamentos ou problemas individuais na análise de sobrevivência. Entre as famílias de distribuições utilizadas para esse fim, a mais popular é a distribuição de weibull, cuja função de risco apresenta formas constante, crescente e decrescente. No entanto, quando a função de risco é unimodal ou em forma de banheira, a distribuição de weibull não é apropriada. Assim, nos últimos anos, têm sido propostas novas distribuições que se ajustam a várias formas da função de risco e, conseqüentemente, ajustam-se a um número maior de problemas práticos.

Recentemente, algumas novas distribuições aplicadas à análise de confiabilidade e sobrevivência tem sido introduzidas compondo uma distribuição absolutamente contínua com uma distribuição discreta truncadas para n = 0. Por esse mecanismo, obtém-se uma nova distribuição com mais parâmetros, que geralmente tem uma taxa de falhas mais flexível ajustando-se a vários tipos de dados (LEMONTE, 2013).

Bidram, Behboodian e Towhidi (2013) propuseram uma nova distribuição de probabilidade denominada exponencial generalizada geométrica (EGG) obtida através do processo de composição das distribuições exponencial generalizada, proposta por Gupta e Kundu (1999), com uma distribuição geométrica. Esta distribuição tem três parâmetros apresenta função de risco com diferentes formas como decrescente, crescente e tipo banheira. Esta distribuição, devido a flexibilidade em acomodar diferentes formas de riscos, apresenta-se como uma importante distribuição, pois pode ser utilizada nos mais variados problemas de modelagem de dados na análise de sobrevivência.

Bidram, Behboodian e Towhidi (2013) apresentam várias propriedades da distribuição EGG e a estimação de máxima verossimilhança é realizada para estimar os três parâmetros.

Neste artigo, o objetivo principal é a estimação dos parâmetros da EGG através da análise bayesiana. Entre os argumentos a favor da abordagem bayesiana, podemos enfatizar a possibilidade de incorporar informações a priori na análise. Outra consideração a ser destacada é que os métodos baseados na inferência bayesiana quase sempre requerem menos dados amostrais para alcançar melhores resultados. Essa é uma consideração muito importante nas áreas de aplicação em que os dados da amostra podem ser caros e difíceis de obter, como a análise de confiabilidade e sobrevivência. Se o tamanho da amostra é pequeno, os intervalos de confiança construídos usando a abordagem de máxima verossimilhança e a aproximação normal são inapropriados. Ao usar a abordagem bayesiana, deve-se especificar uma distribuição a priori que descreva o conhecimento prévio sobre os parâmetros da distribuição, utilizando a informação de um especialista ou estudos passados. Porém no caso de ausência de qualquer informação, utilizam-se prioris não-informativas para os parâmetros, permitindo que os dados sejam muito mais informativos que a priori.

Neste artigo, considera-se uma análise Bayesiana de estimação, assumindo prioris não-informativas para os parâmetros desconhecidos, e esta é comparada com o método de máxima verossimilhança. Apresentam-se os resultados de um estudo de simulação via Monte Carlo realizado para verificar o desempenho dos dois métodos de estimação e para isto introduzimos os algoritmos MCMC para gerar amostras das distribuições a posteriori.

O artigo está organizado da seguinte forma. Na Seção 2, apresentamos a distribuição exponencial generalizada geométrica e suas propriedades. Na seção 3, realizamos os cálculos para estimação de máxima verossimilhança dos parâmetros e matriz de informação de Fisher. Na seção 4, introduzimos as motivações, aplicação da inferência bayesiana e proposta de prioris utilizadas na estimação. Na seção 5, apresentamos os resultados de um estudo de simulação e na seção 6, a metodologia é ilustrada por um conjunto de dados reais que representa os tempos de remissão (em meses) de uma amostra aleatória de pacientes com câncer de bexiga.

Na seção 7, é apresentada a conclusão.

2 Distribuição exponencial generalizada geométrica

Seja X representando o tempo de vida de um componente com uma distribuição exponencial generalizada geométrica, denotada por *EGG (α, β, p)*, cuja densidade é dada por

$$f(x;\alpha,\beta,p)=\frac{(1-p)\alpha\beta e^{-\beta x}(1-e^{-\beta x})^{\alpha-1}}{\{1-p(1-e^{-\beta x})^{\alpha}\}^{2}},\ x>0, \quad (1)$$

onde $\alpha > 0$, $\beta > 0$ e $p \in (0,1)$ (BIDRAM; BEHBOODIAN; TOWHIDI, 2013).

A Figura 1 apresenta o gráfico da função densidade de probabilidade da distribuição exponencial generalizada geométrica para diferentes valores de parâmetros.

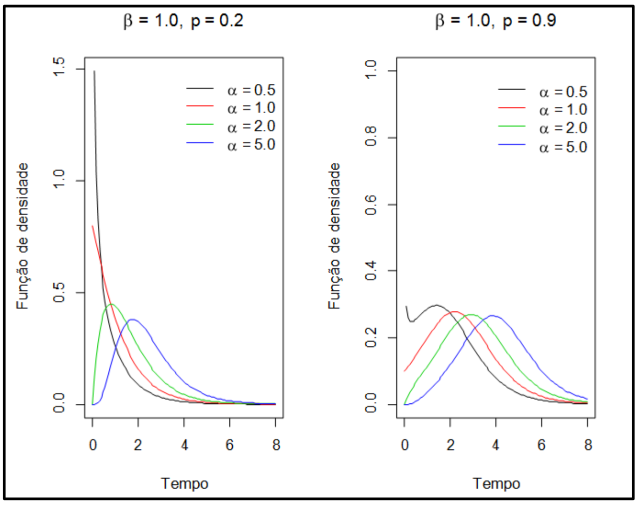

Figura 1 - Gráfico da função densidade da distribuição exponencial generalizada geométrica para diferentes valores de parâmetros.

A função distribuição acumulada de *X* é dada por

$$F(x;\alpha,\beta,p)=\frac{(1-p)(1-e^{-\beta x})^{\alpha}}{1-p(1-e^{-\beta x})^{\alpha}},\ x>0, \quad (2)$$

e, portanto, as funções de sobrevivência e de risco associadas a (1) são dadas, respectivamente, por

$$S(x;\alpha,\beta,p)=\frac{1-(1-e^{-\beta x})^{\alpha}}{1-p(1-e^{-\beta x})^{\alpha}},\ x>0. \quad (3)$$

e

$$h(x;\alpha,\beta,p)=\frac{(1-p)\alpha\beta e^{-\beta x}(1-e^{-\beta x})^{\alpha-1}}{\{1-p(1-e^{-\beta x})^{\alpha}\}\{1-(1-e^{-\beta x})^{\alpha}\}},\ \mathrm{x}>0. \quad (4)$$

As Figuras 2 e 3 apresentam os gráficos da função de sobrevivência e de risco da distribuição exponencial generalizada geométrica, respectivamente, para diferentes valores de parâmetros.

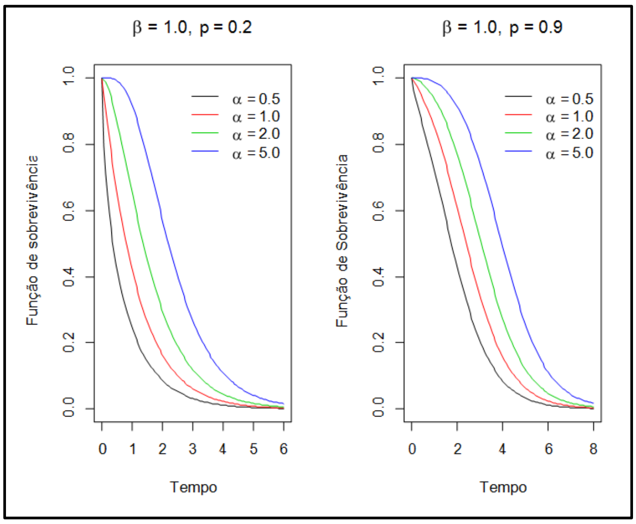

Figura 2-Gráfico da função de sobrevivência da distribuição exponencial generalizada geométrica para diferentes valores de parâmetros.

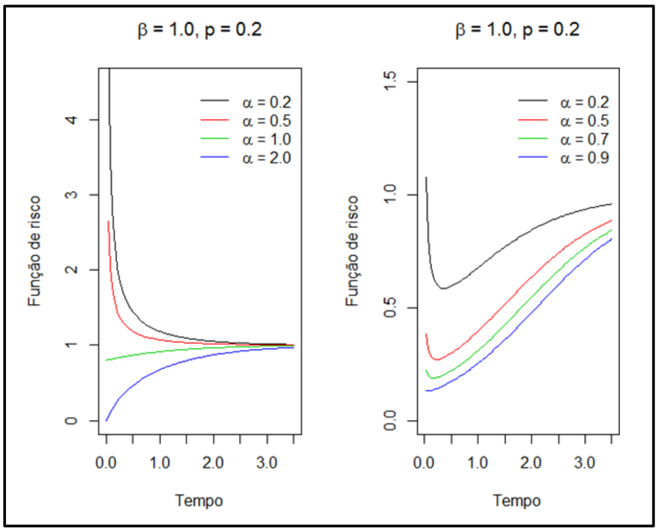

Figura 3 -Funções de risco da distribuição exponencial generalizada geométrica.

Um método muito simples para gerar amostras da distribuição *EGG* é baseado na transformação inversa da acumulada. Dada uma variável aleatória U sob distribuição *uniforme (0,1)*, então a variável

$$x = -\frac{1}{\beta}\log\left[1-\left(\frac{U}{1-p+pU}\right)^{\frac{1}{\alpha}}\right], \tag{5}$$

tem a distribuição *EGG* com parâmetros *(α, β, p)* (BIDRAM; BEHBOODIAN; TOWHIDI, 2013).

O r-ésimo momento da distribuição exponencial generalizada geométrica é dado por

$$E(X^r) = \sum_{j=0}^{\infty} \omega_j\, \alpha(j+1) \frac{(-1)^r}{\beta^r B(a,b)} \frac{d^r}{dt^r}\;\; B\big(a(j+1), 1+t-a(j+1)\big)\big|_{t=a(j+1)}. \qquad (6)$$

Em particular, a média e a variância de X são dadas por

$$E(X) = \frac{1}{\beta} \sum_{j=0}^{\infty} \omega_j \{\psi(\alpha j+\alpha+1) - \psi(1)\}, \qquad (7)$$

e

$$Var(X) = \frac{1}{\beta^2} \sum_{j=0}^{\infty} \omega_j \{[\psi(\alpha j+\alpha+1) - \psi(1)]^2 - \psi'(1) - \psi'(\alpha j+\alpha+1)\}$$

$$- \left[\frac{1}{\beta} \sum_{j=0}^{\infty} \omega_j \{\psi(\alpha j+\alpha+1) - \psi(1)\}\right]^2 \qquad (8)$$

respectivamente, onde $\psi(t) = \frac{d}{dt} log\, \Gamma(t)$é a função digamma e $\psi'(t)$ é a sua primeira derivada (BIDRAM; BEHBOODIAN; TOWHIDI, 2013).

A função geradora de momentos e a função característica da distribuição exponencial generalizada geométrica são dadas, respectivamente, por

$$M_X(t) = \sum_{j=0}^{\infty} \omega_j\, \alpha(j+1) B\left(1-\frac{t}{\beta},\;\; \alpha(j+1)\right),\; t < \beta, \qquad (9)$$

e

$$\phi_X(t) = \sum_{j=0}^{\infty} \omega_j\, \alpha(j+1) B\left(1-\frac{it}{\beta},\;\; \alpha(j+1)\right). \qquad (10)$$

3 Estimação de máxima verossimilhança

Seja $x_1, x_2, \ldots, x_n$ uma amostra aleatória de tamanho n da distribuição exponencial generalizada geométrica, com densidade dada na equação (1), então, a função de verossimilhança para os parâmetros α, β e p é dada por

$$L(\alpha,\beta,p|x) = \prod_{i=1}^{n} f(x_i,\alpha,\beta,p) = \frac{\alpha^n \beta^n (1-\mathrm{p})^{\mathrm{n}} \exp\{-\beta \sum_{i=1}^{n} x_i\} \left[\prod_{i=1}^{n} (1-\mathrm{e}^{-\beta x_i})^{\alpha-1}\right]}{\prod_{i=1}^{n} (1-p(1-\mathrm{e}^{-\beta x_i})^{\alpha})^2}. \qquad (11)$$

A função log-verossimilhança baseada na amostra observada é dada por

$$l(\theta) = n\, log\{\alpha\beta(1-p)\} - \beta \sum_{i=1}^{n} x_i + (\alpha-1) \sum_{i=1}^{n} \log(1-e^{-\beta x_i})$$

$$-2\sum_{i=1}^{n}\log\left\{1-p\left(1-e^{-\beta x_i}\right)^{\alpha}\right\} \tag{12}$$

Os estimadores de máxima verossimilhança (EMV) dos parâmetros α, β e p que maximizam $l(\theta)$ devem satisfazer o sistema de equações dadas por:

$$\frac{\partial l(\boldsymbol{\theta})}{\partial\alpha}=\frac{n}{\alpha}+\sum_{i=1}^{n}\log(1-\mathrm{e}^{-\beta x_i})+2p\sum_{i=1}^{n}\frac{(1-\mathrm{e}^{-\beta x_i})^{\alpha}\log(1-\mathrm{e}^{-\beta x_i})}{1-p(1-\mathrm{e}^{-\beta x_i})^{\alpha}} \tag{13}$$

$$\frac{\partial l(\boldsymbol{\theta})}{\partial\beta}=\frac{n}{\beta}-\sum_{i=1}^{n}x_i+(\alpha-1)\sum_{i=1}^{n}\frac{x_i\,\mathrm{e}^{-\beta x_i}}{1-\mathrm{e}^{-\beta x_i}}+2\alpha p\sum_{i=1}^{n}\frac{x_i\,\mathrm{e}^{-\beta x_i}(1-\mathrm{e}^{-\beta x_i})^{\alpha-1}}{1-p(1-\mathrm{e}^{-\beta x_i})^{\alpha}} \tag{14}$$

$$\frac{\partial l(\boldsymbol{\theta})}{\partial p}=\frac{-n}{1-p}+2\sum_{i=1}^{n}\frac{(1-\mathrm{e}^{-\beta x_i})^{\alpha}}{1-p(1-\mathrm{e}^{-\beta x_i})^{\alpha}} \tag{15}$$

As soluções das equações não lineares acima, obtida através de métodos numéricos de Newton-Raphson, fornecem as estimativas de máxima verossimilhança dos parâmetros, fazendo:

$$\frac{\partial l(\boldsymbol{\theta})}{\partial\alpha}=\frac{\partial l(\boldsymbol{\theta})}{\partial\beta}=\frac{\partial l(\boldsymbol{\theta})}{\partial p}=0. \tag{16}$$

Para obter os intervalos de confiança, usamos os resultados de normalidade assintótica. Temos que, se $\hat{\boldsymbol{\theta}}=(\hat{\alpha},\hat{\beta},\hat{p})'$ denota o estimador de máxima verossimilhança de $\boldsymbol{\theta}=(\alpha,\beta,p)'$, então:

$$\sqrt{\mathbf{n}}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta})\to \mathrm{N}_3(\mathbf{0},\mathrm{I}^{-1}(\boldsymbol{\theta})) \tag{17}$$

onde $\mathrm{I}^{-1}(\boldsymbol{\theta})$ é a matriz de informação de Fisher. A obtenção da matriz de informação requer o cálculo das esperanças das segundas derivadas parciais que não são obtidas facilmente. No entanto, a matriz de informação de Fisher pode ser aproximada por:

$$\mathrm{I}(\hat{\boldsymbol{\theta}})\approx\begin{pmatrix} -\left.\frac{\partial^2 l(\boldsymbol{\theta})}{\partial\alpha^2}\right|_{(\hat{\alpha},\hat{\beta},\hat{p})} & -\left.\frac{\partial^2 l(\boldsymbol{\theta})}{\partial\alpha\partial\beta}\right|_{(\hat{\alpha},\hat{\beta},\hat{p})} & -\left.\frac{\partial^2 l(\boldsymbol{\theta})}{\partial\alpha\partial p}\right|_{(\hat{\alpha},\hat{\beta},\hat{p})} \\ -\left.\frac{\partial^2 l(\boldsymbol{\theta})}{\partial\beta\partial\alpha}\right|_{(\hat{\alpha},\hat{\beta},\hat{p})} & -\left.\frac{\partial^2 l(\boldsymbol{\theta})}{\partial^2\beta^2}\right|_{(\hat{\alpha},\hat{\beta},\hat{p})} & -\left.\frac{\partial^2 l(\boldsymbol{\theta})}{\partial\beta\partial p}\right|_{(\hat{\alpha},\hat{\beta},\hat{p})} \\ -\left.\frac{\partial^2 l(\boldsymbol{\theta})}{\partial p\partial\alpha}\right|_{(\hat{\alpha},\hat{\beta},\hat{p})} & -\left.\frac{\partial^2 l(\boldsymbol{\theta})}{\partial p\partial\beta}\right|_{(\hat{\alpha},\hat{\beta},\hat{p})} & -\left.\frac{\partial^2 l(\boldsymbol{\theta})}{\partial^2 p^2}\right|_{(\hat{\alpha},\hat{\beta},\hat{p})} \end{pmatrix} \tag{18}$$

Um intervalo de confiança assintótico, com coeficiente de confiança $1 - \gamma$ $(0 < \gamma < 0.5)$ pode ser obtido como

$$\mathrm{IC}(\boldsymbol{\theta}): (\theta_j \pm z_{\left[1-\frac{\gamma}{2}\right]}\sqrt{\hat{\kappa}^{\theta_j,\theta_j}}]) \text{ , j = 1, 2, 3} \quad (19)$$

onde $\hat{\kappa}^{\theta_j,\theta_j}$ indica o j-ésimo elemento da diagonal de $I^{-1}(\boldsymbol{\theta})$, para j= 1, 2, 3 e $z_{\left[1-\frac{\gamma}{2}\right]}$ é o quantil da distribuição normal padrão (BIDRAM; BEHBOODIAN; TOWHIDI, 2013).

4 Análise bayesiana

A análise bayesiana para estimação de parâmetros é uma metodologia muito atrativa em problemas práticos e cresceu em popularidade nos últimos anos. Em uma análise bayesiana, a inferência é baseada na distribuição a posteriori dos parâmetros e, denotada por $\pi(\alpha,\beta,p\,|\,\mathrm{x})$, que por sua vez é usada para inferências e decisões envolvendo os parâmetros.

A distribuição posterior $\pi(\alpha,\beta,p\,|\,\mathrm{x})$ é obtida a partir da combinação das informações fornecidas por uma distribuição a priori $\pi(\alpha,\beta,p)$ e as informações fornecidas pelos dados através da verossimilhança $L(\alpha,\beta,p|x)$. Assim, usando o teorema de Bayes, a distribuição posterior é dada por

$$\pi(\alpha,\beta,p\,|\,\mathrm{x}) \propto \pi(\alpha,\beta,p)L(\alpha,\beta,p|x) \quad (20)$$

A distribuição a priori $\pi(\alpha,\beta,p)$ representa o conhecimento prévio do pesquisador antes de se observar os dados. Essa informação é uma característica distintiva da inferência bayesiana na qual é parte fundamental da sua análise, pois na inferência clássica é utilizada somente a informação proveniente dos dados observados.

Para realizar uma análise bayesiana, é necessário especificar uma distribuição a priori para os parâmetros desconhecidos. Diferentes distribuições a priori podem ser usadas em nosso estudo de acordo com todas as informações atualmente disponíveis.

Se a informação a priori sobre os parâmetros do estudo não está disponível ou não existe para um dispositivo, então a incerteza inicial sobre os parâmetros pode ser quantificada com uma distribuição a priori não-informativa. Isso é o mesmo que incluir na análise apenas as informações provenientes dos dados.

No estudo da distribuição exponencial generalizada geométrica considerada neste artigo, assumimos que os parâmetros sejam independentes a priori e a seguinte densidade conjunta a priori é então proposta

$$\pi(\alpha,\beta,p) \propto \pi(\alpha)\pi(\beta)\pi(p) \ . \quad (21)$$

Para o parâmetro p atribuímos uma distribuição $\mathrm{Beta}(a_1,\ b_1)$ com densidade dada por

$$\pi(p;a_1,b_1)=\frac{1}{\mathrm{B}(a_1,b_1)}p^{a_1-1}(1-p)^{b_1-1}\ ,\ p\in(0,1)\ , \tag{22}$$

e média $a/(a+b)$, variância $ab/(a+b)^2(a+b+1)$.

Uma escolha natural para as distribuições a priori para os parâmetros α e β seria assumir os dois parâmetros com distribuição a priori gama $\alpha:\ \text{Gama}(a_2,\ b_2)$ e $\beta:\ \text{Gama}(a_3,\ b_3)$, com densidade dada por

$$\pi(\nu;a_i,b_i)=\frac{b_i^{a_i}\ \nu^{a_i-1}\,exp(-\nu b_i)}{\Gamma(a_i)}, \nu > 0, \tag{23}$$

com média a_i/b_i, variância a_i/b_i^2. Os hiperparâmetros são valores $a_i>0$ e $b_i>0$, $i=1,2,3,$ e são escolhidos de forma que as prioris proporcionem ausência de informação, prioris "vagas".

Portanto, a distribuição conjunta a posteriori para os parâmetros α, β e p é proporcional ao produto entre a função de verossimilhança (11) e a priori conjunta, ou seja,

$$\pi(\alpha,\beta,p\,|\,x)\propto\frac{\alpha^n\beta^n(1-p)^n\exp\left(-\beta\sum_{i=1}^{n}x_i\right)\left[\prod_{i=1}^{n}(1-e^{-\beta x_i})^{\alpha-1}\right]}{\prod_{i=1}^{n}\{1-p(1-e^{-\beta x_i})^{\alpha}\}^2}\pi(\alpha,\beta,p)\ , \tag{24}$$

Como não somos capazes de encontrar uma expressão analítica para distribuições marginais a posteriori e, consequentemente, extrair características de parâmetros como o estimador de Bayes e intervalos de credibilidade, então uma alternativa para esta situação é a utilização do algoritmo de Metropolis-Hastings para obter uma amostra de valores de p,α e β e da posteriori conjunta (CHIB; GREENBERG, 1995).

O algoritmo de Metropolis-Hastings é um método de Monte Carlo via Cadeias de Markov (MCMC). Tal algoritmo segue os seguintes passos:

i.) Escolha valores iniciais p_0, α_0 e β_0;

ii.) No passo $i+1$, gere um novo valor p_{i+1} condicional ao atual p_i a partir da distribuição Beta$\left(\frac{bp_i}{(1-p_i)},\ b\right)$;

iii.) O candidato p_{i+1} pode ser aceito com a probabilidade dada pela razão

$$\alpha(p_i,p_{i+1}) = min\left\{1,\frac{B\left(\frac{bp_i}{(1-p_i)},\ b\right)\,p(p_{i+1},\ \alpha_i,\ \beta_i|\ \boldsymbol{t})}{B\left(\frac{bp_{i+1}}{(1-p_{i+1})},\ b\right)\,p(p_i,\ \alpha_i,\ \beta_i|\ \boldsymbol{t})}\right\}$$

iv.) Gere um valor para α_{i+1} a partir da distribuição Gama$\left(\frac{\alpha_i}{d},\ d\right)$;

v.) O candidato α_{i+1} pode ser aceito com a probabilidade dada pela razão

$$\alpha(\alpha_i,\alpha_{i+1}) = min\left\{1,\frac{G\left(\frac{\alpha_i}{d},\ d\right)\,p(p_{i+1},\ \alpha_{i+1},\ \beta_i|\ \boldsymbol{t})}{G\left(\frac{\alpha_{i+1}}{d},\ d\right)\,p(p_{i+1},\ \alpha_i,\ \beta_i|\ \boldsymbol{t})}\right\}$$

vi.) Gere um valor também para β_{i+1} a partir da distribuição Gama$\left(\frac{\beta_i}{k},\ k\right)$;

vii.) O candidato β_{i+1} pode ser aceito com a probabilidade dada pela razão

$$\alpha(\beta_i, \beta_{i+1}) = min\left\{1, \frac{G\left(\frac{\beta_i}{k},\ k\right)\ p(p_{i+1},\ \alpha_{i+1},\ \beta_{i+1}|\ \boldsymbol{t})}{G\left(\frac{\beta_{i+1}}{k},\ k\right)\ p(p_{i+1},\ \alpha_{i+1},\ \beta_i|\ \boldsymbol{t})}\right\}$$

As supostas distribuições dos parâmetros b, d e k foram escolhidas para obter uma boa mistura das cadeias e a convergência das amostras MCMC dos parâmetros sendo avaliada usando os critérios de Raftery e Lewis (1992).

5 Estudo de simulação

Nesta seção foi realizada uma simulação de Monte Carlo para comparar o desempenho dos métodos de estimação de máxima verossimilhança e bayesiano proposto nas seções anteriores.

A simulação é baseada em 1000 amostras geradas da distribuição exponencial generalizada geométrica geradas no software R (TEAM, 2013) para diferentes tamanhos amostrais, n=10, 50 e 150, para diferentes valores dos parâmetros.

O foco é o estimador, o desvio padrão para diferentes tamanhos amostrais e valores dos parâmetros.

Nas Tabelas 1 e 2 são apresentadas as estimativas de máxima verossimilhança e bayesiana, respectivamente, para diferentes tamanhos amostrais, n = 10, 50 e 150.

Tabela 1 - Estimadores de máxima verossimilhança.

	(α, β, p)	**Média**	**Desvio padrão**
	(0.5, 0.5, 0.5)	(0.4742, 0.5665, 0.5972)	(0.0071, 0.0173, 0.0038)
n = 10	(0.5, 1.0, 0.2)	(0.3842, 1.2258, 0.2011)	(0.0091, 0.0391, 0.0073)
	(1.0, 2.0, 0.5)	(0.9150, 2.0825, 0.5969)	(0.0089, 0.0073, 0.0089)
	(2.0, 2.0, 0.5)	(1.9494, 2.1063, 0.5962)	(0.0071, 0.0122, 0.0104)
	(5.0, 3.0, 0.7)	(4.9912, 3.0453, 0.5903)	(0.0013, 0.0058, 0.0114)
	(0.5, 0.5, 0.5)	(0.4834, 0.5338, 0.5955)	(0.0041, 0.0120, 0.0032)
n = 50	(0.5, 1.0, 0.2)	(0.3778, 1.1995, 0.2005)	(0.0111, 0.0264, 0.0068)
	(1.0, 2.0, 0.5)	(0.9143, 2.0839, 0.5892)	(0.0045, 0.0037, 0.0082)
	(2.0, 2.0, 0.5)	(1.9489, 2.1069, 0.5867)	(0.0074, 0.1305, 0.0109)
	(5.0, 3.0, 0.7)	(4.9908, 3.0466, 0.6246)	(0.0011, 0.0048, 0.1240)
	(0.5, 0.5, 0.5)	(0.4953, 0.5123, 0.5732)	(0.0035, 0.0101, 0.0024)
n = 150	(0.5, 1.0, 0.2)	(0.3965, 1.1783, 0.2012)	(0.0091, 0.0237, 0.0057)
	(1.0, 2.0, 0.5)	(0.9398, 2.0560, 0.5679)	(0.0032, 0.0024, 0.0069)
	(2.0, 2.0, 0.5)	(1.9605, 2.1011, 0.5798)	(0.0063, 0.1206, 0.0103)

	(5.0, 3.0, 0.7)	(4.9996, 3.0431, 0.6309)	(0.0009, 0.0037, 0.0098)

Tabela 2 - Sumários a posteriori de interesse para α, β e p.

	(α, β , p)	**Média**	**Desvio padrão**
n = 10	(0.5, 0.5, 0.5)	(0.4308, 0.5318, 0.5559)	(0.0068, 0.0129, 0.0055)
	(0.5, 1.0, 0.2)	(0.3209, 1.2566, 0.2112)	(0.0008, 0.0322, 0.0072)
	(1.0, 2.0, 0.5)	(0.9509, 2.0123, 0.5345)	(0.0105, 0.0071, 0.0005)
	(2.0, 2.0, 0.5)	(1.9609, 2.1108, 0.5576)	(0.0089, 0.0210, 0.0157)
	(5.0, 3.0, 0.7)	(4.8765, 3.0026, 0.6209)	(0.0003, 0.0045, 0.0233)
n = 50	(0.5, 0.5, 0.5)	(0.4798, 0.5209, 0.5213)	(0.0032, 0.0119, 0.0078)
	(0.5, 1.0, 0.2)	(0.4609, 1.1999, 0.2209)	(0.0309, 0.0213, 0.0012)
	(1.0, 2.0, 0.5)	(0.9589, 2.0334, 0.5322)	(0.0001, 0.0032, 0.0009)
	(2.0, 2.0, 0.5)	(1.9532, 2.1109, 0.5760)	(0.0060, 0.1309, 0.0738)
	(5.0, 3.0, 0.7)	(4.7623, 3.0067, 0.6521)	(0.0009, 0.0032, 0.0989)
n = 150	(0.5, 0.5, 0.5)	(0.4890, 0.5008, 0.5602)	(0.0106, 0.0163, 0.0018)
	(0.5, 1.0, 0.2)	(0.4176, 1.0897, 0.2100)	(0.0078, 0.0237, 0.0076)
	(1.0, 2.0, 0.5)	(0.9498, 2.1230, 0.5219)	(0.0059, 0.0024, 0.0050)
	(2.0, 2.0, 0.5)	(1.9712, 2.4093, 0.5247)	(0.0056, 0.1091, 0.0110)
	(5.0, 3.0, 0.7)	(4.8754, 3.0780, 0.6448)	(0.0011, 0.0030, 0.0056)

De acordo com os resultados das Tabelas 1 e 2, ambos estimadores apresentam um comportamento bastante semelhante tanto usando o método de máxima verossimilhança, como o bayesiano.

6 Aplicação a dados da literatura

Nessa seção é realizada uma exploração do modelo exponencial generalizado geométrico em dados reais de análise de sobrevivência. Este conjunto de dados reais representa os tempos de remissão (em meses) de uma amostra aleatória de 128 pacientes com câncer de bexiga (LEE; WANG, 2003). Este conjunto de dados foi recentemente analisado por Lemonte (2013).

Barlow e Campo (1975) propuseram uma técnica gráfica para verificar o comportamento da função de risco, que é chamada de TTT-plot (tempo total em teste). Tal gráfico é construído com o plot consecutivo das quantidades [r/n, G(r/n)], sendo G(r/n) a função dada por

$$G\left(\frac{r}{n}\right) = \frac{\sum_{i=1}^{r} t_i \; + (n-r)t_{(r)}}{\sum_{i=1}^{n} t_i}\,, \qquad (25)$$

onde $r = 1, \dots, n$, $i = 1, \dots, n$ sendo $t_{(i)}$ as estatísticas de ordem da amostra.

Através do TTT-plot podemos escolher o modelo que melhor se adequa aos dados, isto é, de acordo com a forma obtida através do TTT-plot podemos encontrar o modelo que apresenta a forma da função de risco. Para este conjunto de dados, podemos ver que o TTT-plot tem forma côncava e depois convexa, o que indica que a função de risco é em forma de banheira invertida. Assim, como o modelo exponencial generalizado geométrico possui função de risco com essa forma então podemos concluir que os dados se ajustaram bem à distribuição sob estudo.

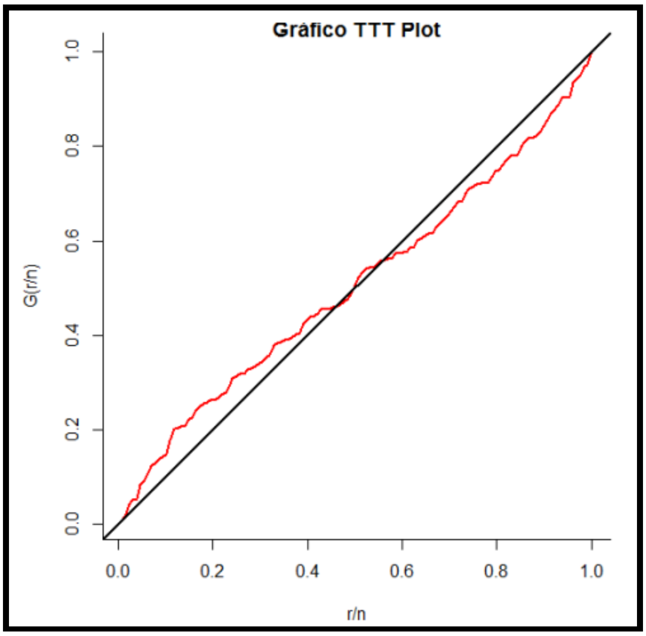

Figura 4 -TTT – plot do conjunto de dados em estudo.

Os estimadores de máxima verossimilhança para a distribuição exponencial generalizada geométrica ajustada aos dados foram obtidos através do pacote maxLik do software R, sob método de Newton-Raphson. Os EMV que melhor se adequam foram $\hat{\alpha}$ = 1.6511, $\hat{\beta}$ = 0.0708 e $\hat{p}$ = -3.5372.

Consideraremos agora uma comparação do ajustamento da distribuição exponencial generalizada geométrica para análise deste conjunto de dados com outras possíveis distribuições que poderiam ser ajustadas, como as distribuições weibull, exponencial generalizada, exponencial power, exponencial geométrica generalizada e weibull modificada. Para isto usaremos os critérios log verossimilhança, AIC (Akaike information criterion) e BIC (Bayes information criterion).

Quanto menor o valor obtido pela métrica, mais adequado é o modelo em questão, esperamos que o valor da métrica, para o modelo a partir do qual gerou-se os dados, seja menor do que dos outros modelos e, a razão obtida deverá ser menor do que 1. Quando isto não ocorre significa que o critério de seleção, métrica, não foi capaz de identificar o modelo que gerou os dados.

A Tabela 3 mostra os resultados dos critérios log verossimilhança, AIC e BIC para diferentes distribuições de probabilidades, considerando os dados em análise (128 pacientes com câncer de bexiga).

Tabela 3 -Resultados da log verossimilhança, AIC e BIC para diferentes modelos.

Modelo	$l(\theta)$	AIC	BIC
exponencial generalizada geométrica	-410.3431	826.6861	849.7983
weibull	-414.0869	832.1738	847.5819
exponencial generalizada	-413.0776	830.1552	845.5633
exponencial power	-426.6474	857.2948	882.4070
exponencial geométrica generalizada	-411.3794	828.7587	851.8709
weibull modificada	-411.0800	828.1601	851.2723

Os resultados, especialmente o AIC e BIC, mostram que a distribuição exponencial generalizada geométrica produz um ajuste melhor do que outras distribuições. Estas conclusões são igualmente confirmadas na Figura 5.

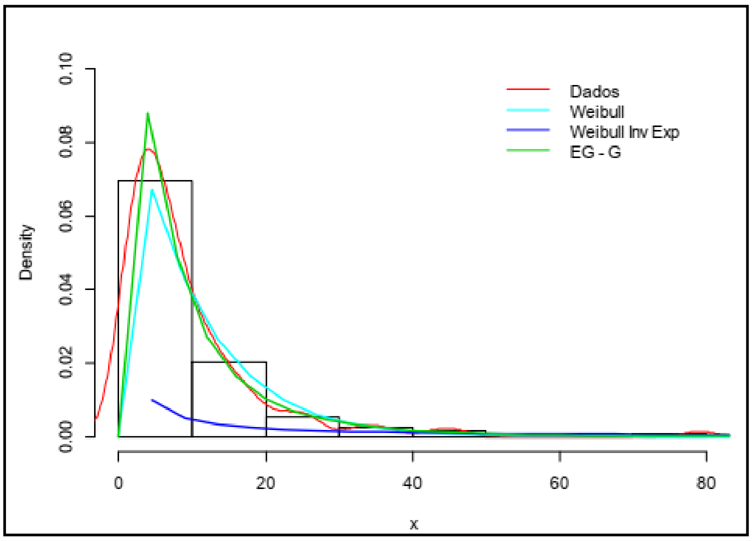

Figura 5 -Densidades para o conjunto de dados.

Para a análise bayesiana dos dados, assumimos prioris não-informativas e independência para os parâmetros sendo

$$p:\ \text{Beta}(0.5, 0.5)\ ,\ \ \alpha:\ \text{Gama}(0.01, 0.01)\ \text{e}\ \ \beta:\ \text{Gama}(0.01, 0.01)$$

Utilizando o software R, realizamos uma simulação MCMC com 10000 iterações e descartamos as primeiras 5000 como burn-in. O diagnóstico de Raftery-Lewis detectou convergência do MCMC e o algoritmo mostrou uma taxa de aceitação em torno de 28% para os três parâmetros.

Os resultados da estimação para os dois métodos de estimação empregados são mostrados nas Tabelas 4-5. A Tabela 5 mostra os estimadores de máxima verossimilhança e bayesianos com seus respectivos desvios padrões (entre parênteses).

Tabela 4 - Estimativas de máxima verossimilhança, desvios padrões e intervalos de 95% de confiança.

Parâmetro	EMV	Bayesiana
α	1.6511 (0.1944)	3.3597 (0.1705)
β	0.0708 (0.0231)	0.2581 (0.0099)
p	-3.5373 (2.9859)	0.8926 (0.0037)

Note que p∈(0,1), no entanto o EMV obtido é p = -3.5373 < 0. Importante destacar que mesmo p < 0 ainda tem-se uma distribuição de probabilidade como observado por Bidram, Behboodian e Towhidi (2013).

A Tabela 5 mostra os intervalos de confiança e credibilidade 95% para os parâmetros da distribuição exponencial generalizada geométrica.

Tabela 5 - Estimativas de máxima verossimilhança, desvios padrões e intervalos de 95% de confiança.

Parâmetro	EMV	Bayesiana
α	(1.6174; 1.6880)	(0.8850; 0.8995)
β	(0.0667; 0.0748)	(3.0213; 3.6891)
p	(-4.0546; -3.0200)	(0.2387; 0.2780)

Nós plotamos as densidades marginais a posteriori obtidas para os três parâmetros na Figura 6.

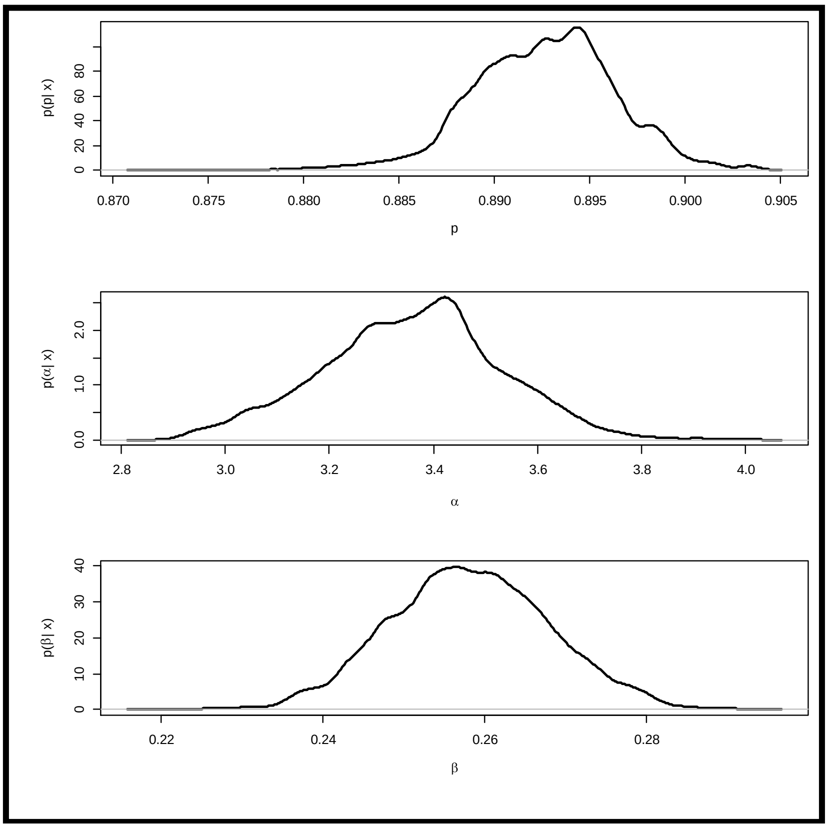

Figura 6 – Plots das densidades a posteriori dos parâmetros para o conjunto de dados.

7 Conclusão

Observamos que a distribuição exponencial generalizada geométrica poderia ser uma boa alternativa para analisar dados de vida útil devido à função de risco apresentar formas bastante variadas como decrescente, crescente e tipo banheira. Além disso, suas funções de sobrevivência e risco possuem forma analítica fechada o que implica em tratabilidade analítica e vantagens computacionais para a análise de inferência. Pudemos comprovar que através do ajuste da distribuição EGG para um conjunto de dados real, esta produz um ajuste melhor do que algumas distribuições de vida conhecidas e já bem estabelecidas. Portanto, a distribuição EGG pode ser considerada adequada na modelagem de dados.

Neste artigo, também consideramos a estimação bayesiana dos parâmetros desconhecidos da distribuição EGG, assumindo algumas distribuições a priori não informativas. Após uma extensa simulação de Monte Carlo, concluímos que a estimação Bayesiana proporciona melhor desempenho do que o método de máxima verossimilhança.

Estes resultados são de grande interesse em aplicações, especialmente nas situações em que não temos opinião de especialistas para construir a distribuição a priori.

8 Referências

BARLOW, R. E.; CAMPO, R. A. Total Time on Test processes and application to failure data analysis. In: Fussell, Jerry B. (Ed.). **Reliability and fault tree analysis**. Pennsylvania, SIAM, 1975.

BIDRAM, H.; BEHBOODIAN, J.; TOWHIDI, M. A new generalized exponential geometric distribution. **Communications in Statistics**: Theory and Methods, v. 42, n. 3, p. 528–542, 2013.

CHIB, S.; GREENBERG, E. Understanding the Metropolis-Hasting algorithm. **The American Statistician**, vol. 49, n. 4, p. 327 – 335, 1995.

COSTA, T. C.; JORGE, L. F.; MOALA; F. A. Análise bayesiana para a distribuição exponencial generalizada geométrica. In: ENCONTRO REGIONAL DE MATEMÁTICA APLICADA E COMPUTACIONAL, 5., 2018, Bauru. **Caderno de trabalhos completos e resumos…** Bauru: Unesp, Faculdade de Ciências, 2018, p. 388-389. Disponível em: <https://www.fc.unesp.br/#!/departamentos/matematica/eventos2341/ermac-2018/caderno-de-trabalhos-e-resumos>. Acesso em: 14 fev. 2019.

GELFAND, A. E.; SMITH, F. M. Sampling-based approaches to calculating marginal densities. **Journal of the American Statistical Association**, n. 85, n. 410, p. 398-409, 1990.

GUPTA, R. D.; KUNDU, D. Generalized exponential distributions. **Australian & New Zealand Journal of Statistics**, v. 41, n. 2, p. 173-188, 1999.

LEE, E. T.; WANG, J. W. **Statistical methods for survival data analysis**. 3rd ed. New York: John Wiley & Sons, 2003.

LEMONTE, A. J. A new exponential-type distribution with constant, decreasing, increasing, upside-down bathtub and bathtub-shaped failure rate function. **Computational Statistics and Data Analysis**, v. 62, p. 149–170, 2013.

RAFTERY, A. E.; LEWIS, S. M. [Practical Markov Chain Monte Carlo]: Comment: One Long Run with Diagnostics: Implementation Strategies for Markov Chain Monte Carlo. **Statistical Science**, v. 7, n. 4, p. 493-497, 1992.

R CORE Team. **R: A language and environment for statistical computing**. R Foundation for Statistical Computing, Vienna, Austria, 2018. Disponível em: <https://www.gbif.org/tool/81287/r-a-language-and-environment-for-statistical-computing>. Acesso em: 30 jun. 2018.