

# INFERÊNCIA CLÁSSICA E BAYESIANA PARA O MODELO DE FRAÇÃO DE CURA GOMPERTZ DEFEITUOSO

Ricardo Ferreira da ROCHA<sup>1</sup>  
Vera Lucia Damasceno TOMAZELLA<sup>1</sup>  
Francisco LOUZADA<sup>2</sup>

- RESUMO: Em dados de sobrevivência é comum a presença de imunes, gerando assim uma fração de cura. Esses casos são usualmente modelados via modelo de mistura padrão, no entanto, aqui utilizamos uma metodologia alternativa, baseada em modelos defeituosos. Modelos defeituosos são aqueles que permitem mudar o domínio em que está definido um certo parâmetro de uma distribuição, gerando uma distribuição imprópria que se comporta naturalmente como um modelo de fração de cura. Na literatura duas distribuições são encontradas com essa propriedade, a inversa Gaussiana e a Gompertz. Neste artigo propomos uma abordagem bayesiana para o modelo Gompertz Defeituoso, juntamente com uma comparação com o estimador de máxima verossimilhança.
- PALAVRAS-CHAVE: Distribuição Gompertz; fração de cura; modelos defeituosos.

## 1 Introdução

A teoria de sobrevivência tem sido bastante explorada por diversos pesquisadores nas mais diversas áreas, sendo uma das principais vertentes a análise de dados clínicos, sendo o objetivo estimar o tempo de vida que um certo indivíduo possui sob determinado cenário. A função de sobrevivência,  $S(t) = P(T > t)$ , é a função utilizada para representar o comportamento aleatório de  $T$ , onde  $T$  é uma variável aleatória não-negativa que representa a probabilidade de um determinado evento de interesse não ocorrer até o tempo  $t$ . Uma propriedade da função  $S(t)$  é que ela tende a 0 à medida que o tempo avança, o que caracteriza um evento de interesse ao qual todos os indivíduos estudados estão sujeitos.

---

<sup>1</sup>Universidade Federal de São Carlos – UFSCar, Departamento de Estatística, CEP: 13565-905, São Carlos, SP, Brasil. E-mail: [ricardo8610@gmail.com](mailto:ricardo8610@gmail.com) / [vera@ufscar.br](mailto:vera@ufscar.br).

<sup>2</sup>Universidade de São Paulo – USP, Instituto de Ciências Matemáticas e de Computação - ICMC, CEP: 13566-590, São Carlos, SP, Brasil. E-mail: [louzada@icmc.usp.br](mailto:louzada@icmc.usp.br).

No entanto, existem situações em que uma parcela da população em estudo torna-se imune ao evento de interesse, por ter obtido a cura da doença estudada ou por outro motivo. Para resolver problemas desse tipo, Boag (1949) propôs o modelo de mistura padrão para fração de cura. A função de sobrevivência é ajustada para

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

com  $p \in (0, 1)$ , de forma que  $S_0(t)$  é uma função de sobrevivência basal conhecida. Dessa forma, tem-se que  $S(t)$  converge para  $p$  na medida em que o tempo aumenta. Em Berkson e Cage (1952) é feita uma análise em pacientes com câncer de estômago e, a partir daí, diversos estudos de fração de cura foram propostos na literatura, concentrando-se nesse modelo de mistura padrão. As escolhas mais comuns para  $S_0(t)$  são as funções de sobrevivência das distribuições Weibull, log-logística e lognormal.

No entanto, neste texto trabalhamos em uma abordagem alternativa para estimar uma proporção de cura. Ao invés de estimar a proporção  $p$  diretamente, como num modelo de mistura, utilizamos uma distribuição que, ao alterar o domínio de seus parâmetros, torna-se um modelo de sobrevivência natural. Obviamente, a distribuição não é mais própria, mas tal alteração faz com que a função de sobrevivência convirja para um valor  $p \in (0, 1)$ . Quando usado como modelo de taxa de cura, a proporção de imunes na população é facilmente obtida. Na literatura, existem duas distribuições utilizadas para esse propósito: a inversa Gaussiana e a Gompertz. Em Haybittle (1959) é feito um ajuste do modelo Gompertz a dados de câncer de mama. Mais adiante, Cantor e Shuster (1992) utilizam o mesmo modelo numa versão modificada para dados de câncer pediátrico. Em Gieser et al (1998) é apresentado o modelo com extensão para covariadas.

O modelo Gaussiano inverso é derivado do tempo de primeira passagem de um processo de Wiener e, portanto, possui essa interpretação extra em termos de processos estocásticos. Nos trabalhos de Balka, Desmond e McNicholas (2009) e Balka, Desmond e McNicholas (2011) são apresentados um estudo dessa distribuição como modelo de fração de cura.

No trabalho de Gieser et al (1998) é apresentada uma abordagem clássica do modelo Gompertz Defeituoso. Neste trabalho o objetivo é propor a análise desse modelo via estimação bayesiana, de maneira análoga a de Balka, Desmond e McNicholas (2011). Na próxima seção deste texto descreveremos os detalhes do modelo Gompertz Defeituoso, bem como a descrição dos métodos de estimação utilizados. Uma seção com duas aplicações em dados de leucemia e melanoma ilustram as metodologias implementadas. Por final, seguem as conclusões, com uma breve discussão sobre os resultados obtidos.

## 2 Metodologia

As distribuições defeituosas caracterizam-se por possuírem funções densidade de probabilidade tal que a integral destas funções sobre todo o domínio da variável aleatória em questão é menor que 1 quando o domínio de algum de seus parâmetros

é diferente daquele usualmente definido. (Balka, Desmond e McNicholas (2009)). Desta maneira, tais modelos apresentam potencial para o tratamento de dados de sobrevivência com fração de cura, pois sua função de sobrevivência agrega naturalmente os elementos não suscetíveis na modelagem. Duas distribuições conhecidas que possuem essa propriedade são a inversa Gaussiana e a Gompertz.

A distribuição inversa Gaussiana possui a propriedade de modelar o tempo de primeira passagem de um processo de Wiener (Borgan, Gjessing e Gjessing (2008)). Desta maneira, é possível construir uma modelagem utilizando propriedade e interpretações provindas da teoria de processos estocásticos. O modelo Gompertz, em princípio, não possui nenhuma interpretação em termos de algum processo adjacente. Em Lee e Whitmore (2004) e Lee e Whitmore (2006) é feita uma revisão abrangente de uma variedade de modelos baseado na primeira passagem do tempo e discutem brevemente o seu potencial como modelos de taxa de cura.

Em Balka, Desmond e McNicholas (2011) é feito um tratamento clássico e bayesiano do modelo Inverso Gaussiano Defeituoso. De maneira similar, apresentaremos aqui um tratamento do modelo Gompertz Defeituoso via inferência clássica, como em Gieser et al (1998), juntamente com a nossa proposta de inferência bayesiana, com prioris não informativas. Os detalhes dos modelos e da implementação são descritos a seguir.

## 2.1 Distribuição Gompertz Defeituosa

O modelo de Gompertz é frequentemente utilizado para a modelagem de dados de sobrevivência em diversas áreas do conhecimento (Gieser (1998)), principalmente onde há a suspeita de um risco aproximadamente exponencial. A função de distribuição de probabilidade é dada por

$$f(t) = be^{at}e^{-\frac{b}{a}(e^{at}-1)},$$

para  $a > 0$ ,  $b > 0$  e  $t > 0$ .

Nessa distribuição,  $a$  é o parâmetro de escala e  $b$  é o parâmetro de forma. Sua função de sobrevivência é dada por

$$S(t) = P(T > t) = e^{-\frac{b}{a}(e^{at}-1)}, \quad (2)$$

e sua função de risco é dada por

$$h(t) = be^{at}. \quad (3)$$

A distribuição Gompertz Defeituosa é a distribuição Gompertz que permite valores negativos para o parâmetro  $a$ . Note que nesse cenário, a função de risco muda sua propriedade de crescimento exponencial para um decrescimento exponencial. Isso sugere seu uso em modelagens onde a chance do evento de interesse ocorra diminua com o decorrer do tempo. Comumente, um cenário que envolve indivíduos não susceptíveis possui essa propriedade. A Figura 1, à direita, ilustra a situação.

A proporção de imunes na população é calculada como o limite da função de sobrevivência, considerando  $a < 0$ ,

$$\lim_{t \rightarrow \infty} S(t) = \lim_{t \rightarrow \infty} e^{-\frac{b}{a}(e^{at}-1)} = e^{\frac{b}{a}}, \quad (4)$$

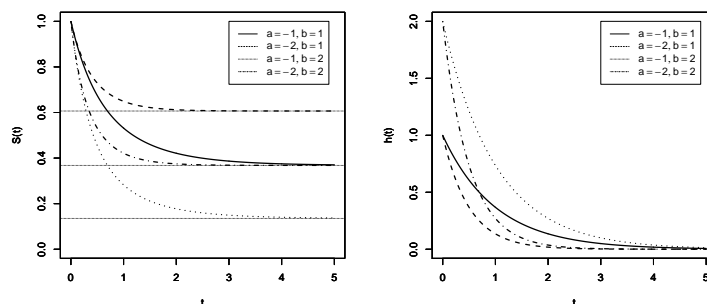


Figura 1 - Função de sobrevivência e de risco para a distribuição Gompertz Defeituosa.

Desta forma, uma vez que os parâmetros  $a$  e  $b$  são estimados, podemos calcular a proporção estimada de elementos não suscetíveis através da equação (4). Na Figura 1, à esquerda, mostramos diversas curvas de sobrevivência para valores negativos de  $a$  e positivos de  $b$ . Note que as curvas atingem suas assíntotas (fração de cura) nos valores  $e^{-1/2}$ ,  $e^{-1}$  e  $e^{-2}$ .

## 2.2 Inferência clássica

Neste trabalho será feita a análise clássica e bayesiana do modelo Gompertz Defeituoso. A estimação clássica foi feita, como usualmente, via máxima verossimilhança. Dada uma amostra aleatória tamanho  $n$ , considere  $\mathbf{t} = (t_1, \dots, t_n)$  e  $\boldsymbol{\delta} = (\delta_1, \dots, \delta_n)$  os vetores de tempos e de censuras, respectivamente. O valor de  $\delta_i = 0$  ocorre quando a  $i$ -ésima observação é censurada,  $\delta_i = 1$  ocorre quando a observação é completa, para  $i = 1, \dots, n$ . A função de verossimilhança (Colosimo e Giolo (2006)) é dada por:

$$L(a, b|\mathbf{t}, \boldsymbol{\delta}) = \prod_{i=1}^n h(t_i|a, b)^{\delta_i} S(t_i|a, b). \quad (5)$$

Combinando as equações (2) e (3) em (5), podemos reescrever a função de verossimilhança por:

$$L(a, b|\mathbf{t}, \boldsymbol{\delta}) = \prod_{i=1}^n \left[ (be^{at_i})^{\delta_i} e^{-\frac{b}{a}(e^{at_i}-1)} \right],$$

que, depois de algumas simplificações algébricas, pode ser escrita como

$$L(a, b|\mathbf{t}, \boldsymbol{\delta}) = b^{\sum \delta_i} e^{a \sum \delta_i t_i} e^{-\frac{b}{a} \sum (e^{at_i} - 1)}. \quad (6)$$

A função de log-verossimilhança é dada por:

$$l(a, b|\mathbf{t}, \boldsymbol{\delta}) = \ln(b) \sum_{i=1}^n \delta_i + a \sum_{i=1}^n \delta_i t_i - \frac{b}{a} \sum_{i=1}^n (e^{at_i} - 1). \quad (7)$$

Os parâmetros são estimados via maximização direta de (7) considerando algum método computacional. Desta forma, a inferência da fração de cura é dada ao calcular a equação (4) com os devidos parâmetros estimados, ou melhor

$$\hat{p} = e^{\frac{\hat{b}}{\hat{a}}}. \quad (8)$$

### 2.3 Inferência Bayesiana

Considere uma distribuição a priori  $\pi$  para os parâmetros  $a$  e  $b$ . A distribuição a posteriori, dadas as informações coletadas, é dada por:

$$p(a, b|\mathbf{t}, \boldsymbol{\delta}) \propto \pi(a, b) \cdot L(a, b|\mathbf{t}, \boldsymbol{\delta}). \quad (9)$$

A função de verossimilhança é similar àquela apresentada em (6), enquanto as distribuições a priori para os parâmetros do modelo serão escolhidas de maneira que sejam não informativas, de maneira similar a de Balka, Desmond e McNicholas (2011). Sendo assim, supondo independência entre os parâmetros, tem-se:

$$\pi(a) = \text{constante}, \quad a \in \mathbb{R}$$

$$\pi(b) = \frac{1}{b}, \quad b > 0$$

$$\pi(a, b) = \pi(a) \cdot \pi(b) = \frac{1}{b}, \quad a \in \mathbb{R}, \quad b > 0. \quad (10)$$

Logo, a distribuição a posteriori dos parâmetros, a menos de proporcionalidade, é dada pela aplicação de (6) e (10) em (9), que resulta em:

$$p(a, b|\mathbf{t}, \boldsymbol{\delta}) \propto b^{(\sum \delta_i) - 1} e^{a \sum \delta_i t_i} e^{-\frac{b}{a} \sum (e^{at_i} - 1)}. \quad (11)$$

E desta forma, temos as distribuições condicionais dadas por:

$$p(a|b, \mathbf{t}, \boldsymbol{\delta}) \propto e^{a \sum \delta_i t_i} e^{-\frac{b}{a} \sum (e^{at_i} - 1)}$$

e

$$p(b|a, \mathbf{t}, \boldsymbol{\delta}) \propto b^{(\sum \delta_i) - 1} e^{-\frac{b}{a} \sum (e^{at_i} - 1)}.$$

Para gerar valores a posteriori de  $a$  e  $b$  utilizamos o algoritmo Metropolis-Hastings. Com cada valor gerado para  $a$  e  $b$ , é calculado o valor de  $p$ , de acordo com a equação (4).

### 3 Aplicações

Faremos aqui duas aplicações em conjuntos de dados reais, compararemos as estimativas do modelo bayesiano proposto com as estimativas clássicas de verossimilhança. Em todos os cálculos e gráficos aqui apresentados, foi utilizado o software *R*. Para o cálculo dos desvios padrões de  $p$  foi utilizado o método delta.

A estimação bayesiana se deu através do uso do algoritmo Metropolis-Hastings. Foram geradas 110000 amostras, de forma que 10000 são descartadas como *burn-in*, e as demais são tomadas de 10 em 10, para evitar a auto correlação, totalizando então uma amostra final de tamanho 10000. Nas duas aplicações, as taxas de aceitação da geração das cadeias ficaram em torno de 60%.

#### 3.1 Leucemia

O primeiro banco de dados analisado é referente a um estudo sobre recorrência de leucemia em pacientes que foram submetidos a um certo tipo de transplante. A leucemia é um tipo de câncer que compromete os glóbulos brancos produzidos pela medula óssea e pode se apresentar de diversas formas. O conjunto possui 44 observações, com 20,45% de censuras (9 no total). O tempo máximo de observação foi de aproximadamente 5 anos. Mais detalhes em Kersei et al. (1987).

Na Tabela 1 estão as estimativas encontradas via máxima verossimilhança, para o caso clássico, e a média a posteriori, no caso do procedimento bayesiano. Os intervalos construídos correspondem ao intervalo de confiança de 95%, e no caso bayesiano, ao intervalo de credibilidade, também de 95%. Vemos que as estimativas dos parâmetros estão relativamente próximas, no entanto, os desvios padrões observados na inferência bayesiana estão menores do que na clássica. Fato este que se reflete no tamanho dos intervalos de confiança, sendo o bayesiano menor que o clássico.

Tabela 1 - Resumo das estimativas clássicas e bayesianas propostas

| Estimação | Parâmetro | Estimação | Desvio Padrão | IC Inf. | IC Sup. |
|-----------|-----------|-----------|---------------|---------|---------|
| Clássica  | $a$       | -1,5101   | 0,3695        | -2,2344 | -0,7857 |
|           | $b$       | 2,3762    | 0,5169        | 1,363   | 3,3894  |
|           | $p$       | 0,2073    | 0,1240        | 0,029   | 0,4570  |
| Bayesiana | $a$       | -1,5551   | 0,2505        | -2,0673 | -1,0844 |
|           | $b$       | 2,4173    | 0,3520        | 1,7546  | 3,1397  |
|           | $p$       | 0,2112    | 0,0453        | 0,1339  | 0,3080  |

Na Figura 2 apresentam-se as distribuições empíricas dos valores gerados na estimação bayesiana. Na Figura 3 está o gráfico das funções de sobrevida com os devidos valores estimados. Note que as duas curvas geradas são bastante parecidas, e apresentam um ajuste relativamente bom em relação à curva de Kaplan-Meier. A fração de cura estimada pelo modelo clássico foi 20,73%, enquanto que o bayesiano foi ligeiramente superior, de 21,12%.

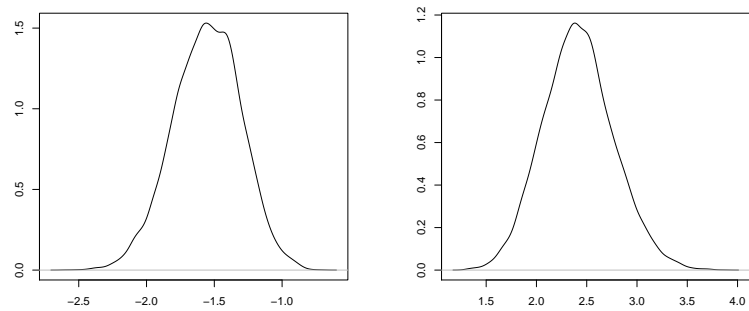


Figura 2 - Densidades empíricas dos parâmetros nos dados de leucemia.

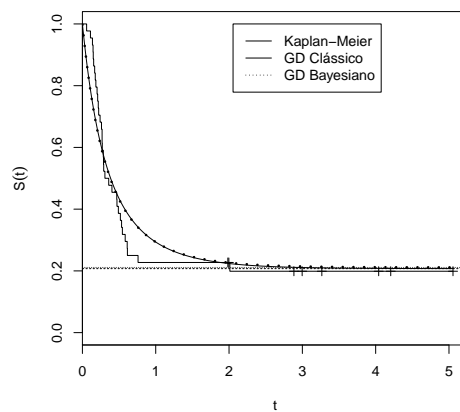


Figura 3 - Sobrevida das estimativas clássicas e bayesianas, juntamente com o estimador de Kaplan-Meyer.

### 3.2 Melanoma

Esse conjunto de dados é referente a um estudo clínico em que os pacientes foram observados quanto a recorrência após a remoção de um melanoma maligno, coletados no período de 1991-1998. Melanoma é um tipo de câncer que se desenvolve nos melanócitos, células responsáveis pela pigmentação da pele. Trata-se de uma lesão maligna potencialmente grave, que pode surgir na pele, nas membranas mucosas, nos olhos e no sistema nervoso central, com grande risco de produzir metástases e com altas taxas de mortalidade nos estágios mais avançados. O

conjunto possui 417 observações, com 55,63% de censuras (232 no total). O tempo máximo de observação foi de 7 anos e a morte do paciente é considerado o evento de interesse. Uma descrição mais detalhada desses dados pode ser encontrada em Ibrahim, Chen e Sinha (2001).

Similarmente a aplicação anterior, a diferença entre as estimativas é pequena, conforme nos ilustra a Tabela 2. A região de credibilidade obtida para os parâmetros é novamente menor que o intervalo de confiança clássico.

Tabela 2 - Resumo das estimativas clássicas e bayesianas propostas

| Estimação | Parâmetro | Estimação | Desvio Padrão | IC Inf. | IC Sup. |
|-----------|-----------|-----------|---------------|---------|---------|
| Clássica  | $a$       | -0,1312   | 0,0540        | -0,2371 | -0,0253 |
|           | $b$       | 0,1791    | 0,0217        | 0,1367  | 0,2216  |
|           | $p$       | 0,2552    | 0,0491        | 0,1594  | 0,3520  |
| Bayesiana | $a$       | -0,1379   | 0,0362        | -0,2081 | -0,0668 |
|           | $b$       | 0,1814    | 0,0147        | 0,1527  | 0,2106  |
|           | $p$       | 0,2682    | 0,0737        | 0,0934  | 0,3830  |

Na Figura 4 estão as distribuições empíricas dos valores gerados na estimação bayesiana. Na Figura 5 está o gráfico das funções de sobrevida com os devidos valores estimados. A fração de cura estimada pelo modelo clássico foi 25,52%, enquanto que no bayesiano foi de 26,82%.

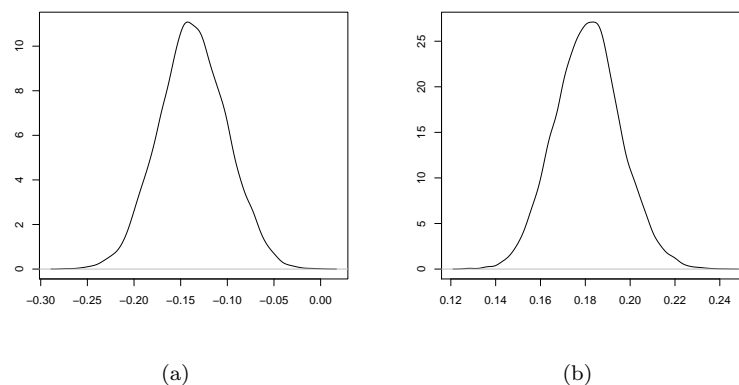


Figura 4 - Sobrevida das estimativas clássicas e bayesianas, juntamente com o estimador de Kaplan-Meier.



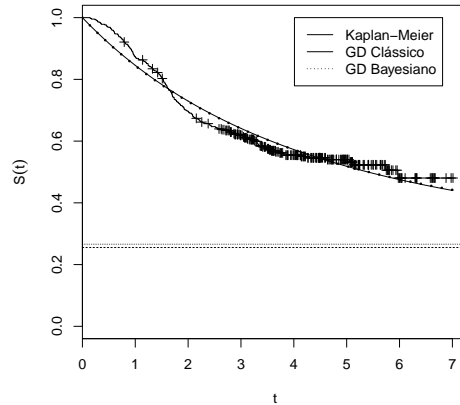


Figura 5 - Densidades empíricas dos parâmetros nos dados de melanoma.

## Conclusões

Neste trabalho, propusemos a utilização do modelo Gompertz Defeituoso na sua versão bayesiana. Comparamos tal abordagem com a estimação clássica do modelo e encontramos estimativas muito próximas para ambas as abordagens propostas, o que é esperado devido ao uso de prioris não informativas. Pode-se verificar que o modelo bayesiano possui intervalos de credibilidade menores para os parâmetros de interesse, inclusive a fração de cura. Além disso, é uma abordagem simples e eficaz do problema, sendo capaz de estimar uma fração de curados facilmente.

Notamos também que, deixando flexível os valores do domínio do parâmetro  $a$ , as estimativas vão naturalmente para os valores negativos, indicando assim a presença de imunes na população. Nos dois casos estudados, o modelo bayesiano apresentou estimativas maiores para a fração de cura.

## 4 Agradecimentos

Os autores agradecem ao CNPq e FAPESP pelo auxílio financeiro parcial.

R. F. COSTA; V. L. D. TOMAZELLA; F. LOUZADA. Bayesian and classic inference for the Defective Gompertz Cure Rate Model. *Rev. Bras. Biom.*, São Paulo, v.32, n.1, p.104-114, 2014.

- **ABSTRACT:** *In the survival analysis theory, there is the case in which a certain fraction of population are immune. These cases are usually treated using a standard mixture model, however, here we use an alternative methodology based on defective models. Defective models are those which allow to change the domain that is defined to certain distribution parameter, generating an improper distribution that naturally behaves as a cure rate model. In literature, there is two distributions with this property, the inverse Gaussian and Gompertz. We propose a Bayesian approach for Defective Gompertz model, along with a comparison to the maximum likelihood estimator. The results indicate that the bayesian model approximates to the classical, but with some small additional advantages in terms of confidence associated.*
- **KEYWORDS:** *Gompertz distribution; cure rates; defective models.*

## Referências

- BALKA, J.; DESMOND, A. F.; MCNICHOLAS, P. D. Bayesian and likelihood inference for cure rates based on defective inverse gaussian regression models. *Journal of Applied Statistics*, v.38, n.1 , p.127-144, 2011.
- BALKA, J.; DESMOND, A. F.; MCNICHOLAS, P. D. Review and implementation of cure models based on first hitting times for wiener processes. *Lifetime Data Analysis*, v.15, n.2, p.147-176, 2009.
- BERKSON, J.; GAGE, R. P. Survival curve for cancer patients following treatment. *Journal of the American Statistical Association*, v.47, n.259, p.501-515, 1952.
- BOAG, J. W. Maximum likelihood estimates of the proportion of patients cured by cancer therapy. *Journal of the Royal Statistical Society, Series B (Methodological)*, v.11, n.1, p.15-53, 1949.
- BORGAN, O.; GJESSING H. K.; GJESSING, S. *Survival and event history analysis: a process point of view*, Springer, 2008, 540p.
- CANTOR, A. B.; SHUSTER J. J. Parametric versus non-parametric methods for estimating cure rates based on censored survival data. *Statistics in Medicine*, v.11, n.7, p.931-937, 1992.
- COLOSIMO, E. A.; GIOLO, S.R. *Análise de sobrevivência aplicada*, ABE-Projeto Fisher, Edgard Blücher, 2006, 367p.
- GIESER, P. W.; CHANG, M. N.; RAO, P. V.; SHUSTER, J. J.; PULLEN, J. Modelling cure rates using the gompertz model with covariate information. *Statistics in medicine*, v.17, n.8, p.831–839, 1998.
- HAYBITTLE, J. L. The estimation of the proportion of patients cured after treatment for cancer of the breast. *British Journal of Radiology*, v.32, n.382, p.725-733, 1959.
- IBRAHIM, J. G.; CHEN, M. H.; SINHA, D. *Bayesian Survival Analysis*. New York: Springer, 2001, 479p.

KERSEY, J. H.; WEISDORF, D.; NESBIT, M. E.; LEBIEN, T. W.; WOODS, W. G. et al. Comparison of autologous and allogeneic bone marrow transplantation for treatment of high-risk refractory acute lymphoblastic leukemia. *New England Journal of Medicine*, v.8, n.317, p.461-467, 1987.

LEE, M. L. T.; WHITMORE, G. A. First hitting time models for lifetime data, *Handbook of statistics*, v.23, p.537-543, 2004.

LEE, M. L. T.; WHITMORE G. A. Threshold regression for survival analysis: modeling event times by a stochastic process reaching a boundary. *Stat Sci*, v.21, n.4, p.501-513, 2006.

Recebido em 06.11.2013.

Aprovado após revisão em 26.03.2014.