TEMA Tend. Mat. Apl. Comput., **13**, No. 3 (2012), 247-255.
doi: 10.5540/tema.2012.013.03.247

Intervalos de Confiança para os Parâmetros do Modelo Geométrico com Inflação de Zeros

C.G. CARRASCO[1], Unidade Universitária de Ciências Exatas e Tecnológicas, Universidade Estadual de Goiás, UEG, 75132-903 Anápolis, GO, Brasil.

M.H. TUTIA[2], Faculdade de Tecnologia de Ourinhos, FATEC - Ourinhos, 19910-206 Ourinhos, SP, Brasil.

E.Y. NAKANO[3], Departamento de Estatística, Universidade de Brasília - UnB, 70910-900 Brasília, DF, Brasil.

Resumo. Propomos neste trabalho a utilização do modelo Geométrico com inflação de zeros, que é uma generalização do modelo Geométrico, na análise de dados de sobrevivência e confiabilidade. O uso deste modelo se faz necessário principalmente quando os dados apresentam um número excessivo de zeros. Estimativas de máxima verossimilhança dos parâmetros do modelo foram obtidas, assim como seus intervalos de confiança baseados na teoria assintótica. Ademais, usamos a técnica de reamostragem *bootstrap* como um procedimento alternativo adequado para construção de intervalos de confiança para os parâmetros do modelo Geométrico com inflação de zeros.

Palavras-chave. Estimação intervalar, probabilidade de cobertura, técnica *bootstrap*.

1. Introdução

A distribuição Geométrica [7] é utilizada para contar o número de fracassos que precedem o primeiro sucesso, e a mesma pode ser vista como uma versão discreta da distribuição Exponencial [11]. Em particular, na análise de confiabilidade, podemos estar interessados no número de impactos termo-elétricos recebidos por um equipamento eletrônico antes do mesmo falhar. Nos casos onde, a chance de falha desse equipamento no primeiro impacto é muito alta, podemos ter uma ocorrência muito grande de números zeros nesse conjunto de dados. Nesta situação, o modelo Geométrico com inflação de zeros (ZIG) será mais apropriado do que o modelo Geométrico padrão para o ajuste desses dados. Os modelos com inflação de zeros como o modelo Binomial Negativo ou Poisson com inflação de zeros já foram amplamente discutidos por [2] e [8].

[1] cleber.carrasco@ueg.br
[2] marcelo.tutia@fatec.sp.gov.br
[3] nakano@unb.br

Recebido em 20 Outubro 2011; Aceito em 19 Novembro 2012.

Neste artigo, propomos trabalhar com a distribuição Geométrica com inflação de zeros na análise de dados de sobrevivência e confiabilidade, a qual é uma generalização da distribuição Geométrica. O objetivo deste trabalho é apresentar a técnica de reamostragem *bootstrap* [3] como um procedimento alternativo na construção de intervalos de confiança para os parâmetros do modelo ZIG, uma vez que os procedimentos usuais podem não ser válidos [1].

Essa técnica de reamostragem foi proposta primeiramente por Efron [5], e visa a obtenção de estimativas intervalares empíricas para os estimadores dos parâmetros de interesse por meio da reamostragem do conjunto de dados original. Existem basicamente dois tipos de *bootstrap*: o paramétrico, no qual os estimadores de máxima verossimilhança (EMV) são obtidos através do modelo ajustado, isto é, geramos dados do modelo ajustado com os valores dos parâmetros fixados nos EMV obtidos da amostra original; e o *bootstrap* não paramétrico, onde os EMV são baseados em B reamostras com reposição obtidas da amostra original.

Os parâmetros do modelo ZIG são estimados através do logaritmo da função de verossimilhança, utilizando-se de um algoritmo do tipo Newton implementado na função *nlm* (*non linear minimization*) do pacote *stats* disponível no *software* R [4]. Do mesmo modo, intervalos de confiança são construídos para os parâmetros do modelo ZIG através da teoria assintótica usual, utilizando-se das propriedades assintóticas dos estimadores de máxima verossimilhança. Também são construídos intervalos de confiança *bootstrap* para os parâmetros do modelo ZIG como uma alternativa adequada aos métodos usuais de estimação intervalar. Para comparar os procedimentos de construção de intervalos de confiança, calculamos a probabilidade de cobertura e as amplitudes médias desses intervalos.

Para ilustrarmos a metodologia adotada neste trabalho, geramos no *software* R um conjunto de dados com muitos zeros e, a esses dados, ajustamos os modelos ZIG e Geométrico através das suas respectivas função de sobrevivência, comparando-as com as estimativas de Kaplan-Meier.

2. Desenvolvimento

A distribuição Geométrica com inflação de zeros é uma generalização da distribuição Geométrica, onde temos uma combinação da distribuição Geométrica com uma distribuição cuja probabilidade de zero é igual a 1. A distribuição de probabilidades do modelo ZIG pode ser expressa da seguinte forma

$$f(x) = \theta(1-\theta)^x\rho + (1-\rho)I_{(x)}\{0\} = (\theta\rho + (1-\rho))^{I_{(x)}\{0\}}\left(\theta\left(1-\theta\right)^x\rho\right)^{1-I_{(x)}\{0\}} \quad (2.1)$$

onde $x = 0, 1, 2, \ldots$; $0 \leq \theta \leq 1$ é o parâmetro da distribuição Geométrica que representa a probabilidade instantânea do evento de interesse e, $0 \leq \rho \leq 1$ é o parâmetro de mistura da distribuição Geométrica com uma distribuição degenerada em $x = 0$ que modela os excessos de zeros não explicados pelo modelo geométrico. $I_{(x)}\{0\}$ é uma função indicadora, que vale um quando $x = 0$ e zero para $x > 0$.

Em particular se $\rho = 1$, o modelo (2.1) se reduz ao modelo Geométrico.

As funções de sobrevivência e de risco podem ser escritas, respectivamente por

$$S(x) = P\left(X > x\right) = \rho\left(1-\theta\right)^{x-1}$$

e

$$h(x) = (\theta + (1-\rho)(1-\theta))^{I_{(x)}\{0\}} \theta^{\left(1-I_{(x)}\{0\}\right)} = \begin{cases} \theta + (1-\rho)(1-\theta), & se\ x = 0 \\ \theta, & se\ x > 0 \end{cases}$$

Seja X uma variável aleatória com função densidade de probabilidade dada em (2.1), então na ausência de censuras, os parâmetros do modelo ZIG podem ser estimados através da maximização da função de verossimilhança definida como [9]

$$L(X \mid \rho, \theta) = \prod_{i=1}^{n} (\theta\rho + (1-\rho))^{I_{(x_i)}\{0\}} (\theta(1-\theta)^{x_i}\rho)^{1-I_{(x_i)}\{0\}}$$

O logaritmo da função de verossimilhança pode ser apresentado por

$$l(X \mid \rho, \theta) = n_0 \log(\theta\rho + (1-\rho)) + (n-n_0)(\log(\rho) + \log(\theta)) + \log(1-\theta) \sum_{x=1}^{\infty} n_x x \quad (2.2)$$

onde n_x é o número de ocorrências do valor x na amostra, $x = 1, 2, \ldots$ Em particular, n_0 é o número de zeros na amostra. Note que $\sum_{i=1}^{n} x_i = \sum_{x=1}^{\infty} n_x x$.

Os estimadores de máxima verossimilhança dos parâmetros do modelo ZIG podem ser obtidos diretamente através da maximização do logaritmo da função de verossimilhança (2.2) por métodos numéricos.

O vetor gradiente das derivadas parciais de (2.2) é obtido através de

$$\bigtriangledown l(X \mid \rho, \theta) = \begin{pmatrix} \frac{\partial l(X|\rho,\theta)}{\partial \rho} \\ \frac{\partial l(X|\rho,\theta)}{\partial \theta} \end{pmatrix} = \begin{pmatrix} \frac{n-n_0}{\rho} - \frac{n_0(1-\theta)}{\theta\rho+(1-\rho)} \\ \frac{n_0\rho}{\theta\rho+(1-\rho)} + \frac{n-n_0}{\theta} - \frac{\sum_{x=1}^{\infty} n_x x}{1-\theta} \end{pmatrix}$$

e a matriz Jacobiana com as derivadas segundas é dada por

$$\bigtriangledown^2 l(X \mid \rho, \theta) = \begin{pmatrix} \frac{\partial^2 l(X|\rho,\theta)}{\partial \rho^2} & \frac{\partial^2 l(X|\rho,\theta)}{\partial \rho \partial \theta} \\ & \frac{\partial^2 l(X|\rho,\theta)}{\partial \theta^2} \end{pmatrix} =$$

$$= \begin{pmatrix} -\frac{n_0(1-\theta)^2}{[\theta\rho+(1-\rho)]^2} - \frac{n-n_0}{\rho^2} & \frac{n_0}{\theta\rho+(1-\rho)} + \frac{n_0(1-\theta)\rho}{[\theta\rho+(1-\rho)]^2} \\ & -\frac{n_0\rho^2}{[\theta\rho+(1-\rho)]^2} - \frac{n-n_0}{\theta^2} - \frac{\sum_{x=1}^{\infty} n_x x}{(1-\theta)^2} \end{pmatrix}$$

Note que $n_i \sim Binomial(n, P(X=i))$, assim, $E(n_i) = nP(X=i)$. Dessa forma, temos

$$E(n_0) = nP(X=0) = n(\theta\rho + 1 - \rho)$$

e

$$E\left(\sum_{x=1}^{\infty} n_x x\right) = n \sum_{x=1}^{\infty} xP(X=x) = \frac{n\rho(1-\theta)}{\theta}$$

Logo, é fácil verificarmos que

$$E\left(\bigtriangledown l\left(X \mid \rho,\theta\right)\right) = \begin{pmatrix} E\left(\frac{\partial l(X|\rho,\theta)}{\partial \rho}\right) \\ E\left(\frac{\partial l(X|\rho,\theta)}{\partial \theta}\right) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$$

A matriz de informação de Fisher $I_n(\rho,\theta)$, é dada por

$$I_n\left(\rho,\theta\right) = E\left(-\bigtriangledown^2 l\left(X \mid \rho,\theta\right)\right) = \begin{pmatrix} E\left(-\frac{\partial^2 l(X|\rho,\theta)}{\partial \rho^2}\right) & E\left(-\frac{\partial^2 l(X|\rho,\theta)}{\partial \rho \partial \theta}\right) \\ & E\left(-\frac{\partial^2 l(X|\rho,\theta)}{\partial \theta^2}\right) \end{pmatrix} =$$

$$= \begin{pmatrix} \frac{n(1-\theta)}{\rho(\theta\rho+1-\rho)} & -\frac{n}{\theta\rho+1-\rho} \\ & \frac{n\rho\left(\theta+(1-\rho)(1-\theta)^2\right)}{\theta^2(1-\theta)(\theta\rho+1-\rho)} \end{pmatrix}$$

Invertendo a matriz de informação de Fisher obtemos

$$I_n^{-1}\left(\rho,\theta\right) = \begin{pmatrix} \frac{\rho(1-\rho)}{n} + \frac{\theta\rho}{n(1-\theta)^2} & \frac{\theta^2}{n(1-\theta)} \\ & \frac{\theta^2}{n\rho} \end{pmatrix}$$

Portanto, temos que

$$\begin{pmatrix} \hat{\rho} \\ \hat{\theta} \end{pmatrix} \approx Normal_2\left(\begin{pmatrix} \rho \\ \theta \end{pmatrix}, I_n^{-1}\left(\rho,\theta\right)\right) \tag{2.3}$$

Assim, os intervalos de $100\left(1-\alpha\right)\%$ de confiança para os parâmetros do modelo ZIG (2.1) são dados respectivamente por

$$\hat{\rho} \pm z_{\left(1-\frac{\alpha}{2}\right)}\sqrt{\frac{\rho\left(1-\rho\right)}{n} + \frac{\theta\rho}{n\left(1-\theta\right)^2}} \tag{2.4}$$

e

$$\hat{\theta} \pm z_{\left(1-\frac{\alpha}{2}\right)}\sqrt{\frac{\theta^2}{n\rho}} \tag{2.5}$$

onde $z_{\left(1-\frac{\alpha}{2}\right)}$ é o quantil $\left(1-\frac{\alpha}{2}\right)$ de uma distribuição Normal Padrão. Note que θ e ρ em geral são parâmetros desconhecidos, dessa forma, podemos substituí-los pelos seus estimadores $\hat{\theta}$ e $\hat{\rho}$, respectivamente em (2.4) e (2.5).

A utilização de (2.3) é direcionada pelo tamanho da amostra, que deve ser suficientemente grande. No entanto, em análise de sobrevivência e confiabilidade é comum termos amostras pequenas ou moderadas, onde a aproximação (2.3) pode não ser válida [1]. Nestes casos uma possibilidade é a utilização da técnica de reamostragem *bootstrap* paramétrica e/ou não paramétrica na construção de intervalos de confiança, através da reamostragem do conjunto de dados original [3].

Consideremos θ como o parâmetro de interesse do modelo ZIG. Para cada reamostra da amostra original, calculamos o EMV para θ e temos no final de B reamostragens, $\hat{\theta}_1 < \ldots < \hat{\theta}_B$ valores dos EMV ordenados. Utilizamos então

$$\hat{\theta}_{(B)\left(\frac{\alpha}{2}\right)} \quad e \quad \hat{\theta}_{(B)\left(1-\frac{\alpha}{2}\right)} \tag{2.6}$$

como os limites inferiores e superiores do intervalo $100(1-\alpha)\%$ de confiança para θ, onde α é o nível de significância. Neste trabalho utilizamos $B = 1000$. Intervalos de confiança percentis *bootstrap* $100(1-\alpha)\%$ para o parâmetro ρ do modelo ZIG podem ser obtidos de maneira análoga.

Quando a reamostra for obtida de um modelo probabilístico, utilizando como parâmetros deste modelo as estimativas dos mesmos calculadas através da amostra original, temos o *bootstrap* paramétrico. Agora, se a reamostra for feita com reposição diretamente da amostra original, temos o *bootstrap* não paramétrico. No *bootstrap* paramétrico é feita suposição sobre a distribuição dos dados que gerou a amostra original, isto é, necessita supor ou conhecer a distribuição que gerou a amostra original. No caso do *bootstrap* não paramétrico não precisa supor ou conhecer essa distribuição. Em ambos os casos, é necessário que a amostra seja representativa da população. Maiores detalhes sobre essa técnica podem ser vistos em [3].

Para compararmos os procedimentos de construção de intervalos de confiança para os parâmetros de uma distribuição, é usual o cálculo das probabilidades de cobertura e das amplitudes médias desses intervalos [10]. A probabilidade de cobertura é determinada repetindo o procedimento de construção do intervalo de confiança D vezes, nas quais verificamos em cada uma se o verdadeiro valor do parâmetro pertence ou não ao intervalo de confiança obtido. Assim, a probabilidade de cobertura para um intervalo de confiança pode ser calculada por

$$1 - \frac{\sum_{j=1}^{D} \psi(vp \notin IC_j)}{D} \tag{2.7}$$

onde $\psi(.)$ é uma função indicadora que vale um se $vp \notin IC_j$ e zero caso contrário, vp é o verdadeiro valor do parâmetro e IC_j é o j-ésimo intervalo de confiança construído. Neste trabalho utilizamos $D = 1000$.

A amplitude de um intervalo de confiança é outro critério para comparação de intervalos de confiança. Com a mesma probabilidade de cobertura, procedimentos de intervalos de confiança que possuem menores amplitudes são considerados melhores [6]. Procedimentos de intervalos de confiança conservativos tendem a terem maiores amplitudes do que os procedimentos não conservativos.

3. Resultados e Discussões

A metodologia adotada neste trabalho é aplicada a um conjunto de dados de tamanho 50 gerado no *software* R através do modelo ZIG com os parâmetros fixados em $\theta = 0,15$ e $\rho = 0,4$. O tamanho da amostra foi definido de forma a ser suficiente para representar bem o excesso de zeros e pequeno o bastante para evitar a normalidade assintótica das estimativas. Neste exemplo estamos considerando que esses dados são referentes a tolerância de um equipamento eletrônico ao número de impactos termo-elétricos onde, pelas características de fabricação, sabemos que o primeiro impacto é o mais fatal para o equipamento, isto é, a chance de falha no primeiro impacto é maior do que nos demais. Os dados simulados foram: 0, 0, 0,

20, 0, 2, 0, 0, 2, 0, 0, 0, 1, 0, 0, 2, 0, 0, 0, 0, 0, 17, 4, 0, 0, 0, 0, 4, 0, 11, 0, 0, 0, 0, 0, 2, 0, 15, 0, 0, 14, 5, 0, 0, 0, 11, 0, 5, 19 e 9.

A Figura 1 apresenta as curvas de sobrevivência dos modelos ZIG e Geométrico juntamente com a curva de Kaplan-Meier, observa-se que o modelo Geométrico com inflação de zeros se ajusta melhor a esse conjunto de dados do que o modelo Geométrico simples, o que era esperado devido ao excesso de zeros contido nesses dados.

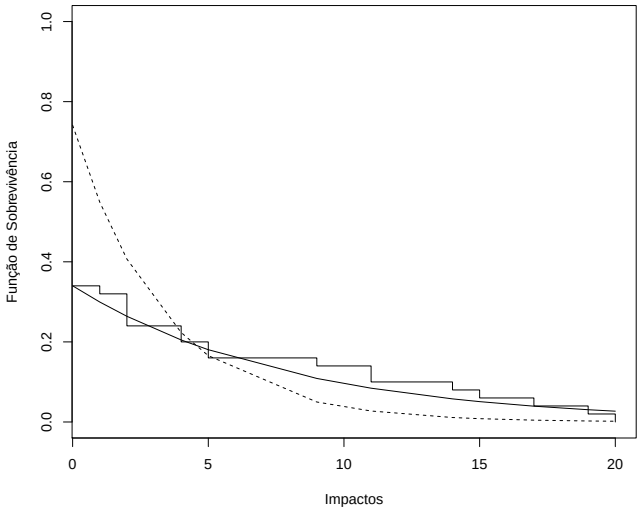

Figura 1: Curvas de sobrevivência dos modelos ZIG (contínua), Geométrico (pontilhada) e de Kaplan-Meier (escada).

As estimativas de máxima verossimilhança para os parâmetros θ e ρ do modelo ZIG, calculadas através da maximização de (2.2) utilizando a função *nlm* do *software* R, são respectivamente $\hat{\theta} = 0,1189$ e $\hat{\rho} = 0,3859$. Os intervalos de confiança para os parâmetros do modelo ZIG construídos utilizando a teoria assintótica usual (2.4) e (2.5), e a técnica *bootstrap* (2.6) estão condensados na Tabela 1, que também apresenta as variâncias e os vícios dos estimadores obtidos pelos métodos descritos neste trabalho. Observando a Tabela 1 percebemos que os intervalos de confiança estão próximos, sendo que o intervalo via técnica *bootstrap* não paramétrica apresenta uma menor amplitude e variância em relação ao *bootstrap* paramétrico. Este fato também ocorre entre os vícios dos estimadores, onde novamente os intervalos não paramétricos apresentam-se menores.

A Figura 2 apresenta os histogramas e os *qq-plots* das distribuições empíricas dos EMV obtidos via *bootstrap* paramétrico e não paramétrico, onde há um indicativo de não normalidade dos EMV, em particular para os estimadores de θ, sugerindo, neste caso, que a teoria usual de verossimilhança (2.3) pode não propiciar resultados suficientemente adequados.

Tabela 1: Intervalos de confiança para os parâmetros do modelo ZIG.

Intervalo de Confiança	Parâmetro	IC(95%)	Variância	Vício
Assintótico		$[0,0658\,;\,0,1719]$	$0,0007$	–
Bootstrap Paramétrico	θ	$[0,0785\,;\,0,1964]$	$0,0010$	$0,0592$
Bootstrap Não Param.		$[0,0876\,;\,0,1842]$	$0,0007$	$0,0468$
Assintótico		$[0,2351\,;\,0,5367]$	$0,0059$	–
Bootstrap Paramétrico	ρ	$[0,2460\,;\,0,5408]$	$0,0060$	$0,0049$
Bootstrap Não Param.		$[0,2578\,;\,0,5410]$	$0,0055$	$0,0060$

Os resultados da Tabela 2 mostram que as probabilidades de cobertura estimadas dos procedimentos de intervalo de confiança calculadas através de (2.7) estão próximas da probabilidade de cobertura nominal fixada em 0,95, exceto para o parâmetro θ do intervalo *bootstrap* não paramétrico (0,895) e, que os procedimentos de construção dos intervalos de confiança são não conservativos, pois as probabilidades de cobertura estimadas estão abaixo da probabilidade de cobertura nominal (0,95). Com relação às amplitudes médias, a Tabela 2 apresenta amplitudes médias próximas entre os procedimentos de intervalo de confiança.

Tabela 2: Probabilidade de cobertura e amplitude média para os parâmetros do modelo ZIG.

Intervalo de Confiança	Probabilidade de Cobertura		Amplitude Média	
Parâmetros	θ	ρ	θ	ρ
Assintótico	$0,944$	$0,945$	$0,139$	$0,316$
Bootstrap Paramétrico	$0,937$	$0,945$	$0,156$	$0,318$
Bootstrap Não Paramétrico	$0,895$	$0,942$	$0,148$	$0,317$

4. Conclusões

Podemos utilizar o modelo Geométrico com inflação de zeros em dados de sobrevivência e confiabilidade, no entanto, é preciso ter cuidado na construção de intervalos de confiança para os parâmetros desse modelo, uma vez que os procedimentos usuais podem não ser válidos, em particular para amostras pequenas. Neste contexto, a técnica de reamostragem *bootstrap* paramétrica e/ou não paramétrica utilizada apresenta-se como um procedimento alternativo de estimação intervalar para os parâmetros deste modelo, possibilitando a obtenção de intervalos de confiança adequados. Para o conjunto de dados simulados, destacamos o intervalo *bootstrap* paramétrico, que apresentou uma probabilidade de cobertura próxima da nominal e maior do que o *bootstrap* não paramétrico, além de uma amplitude média próxima dos demais intervalos de confianca.

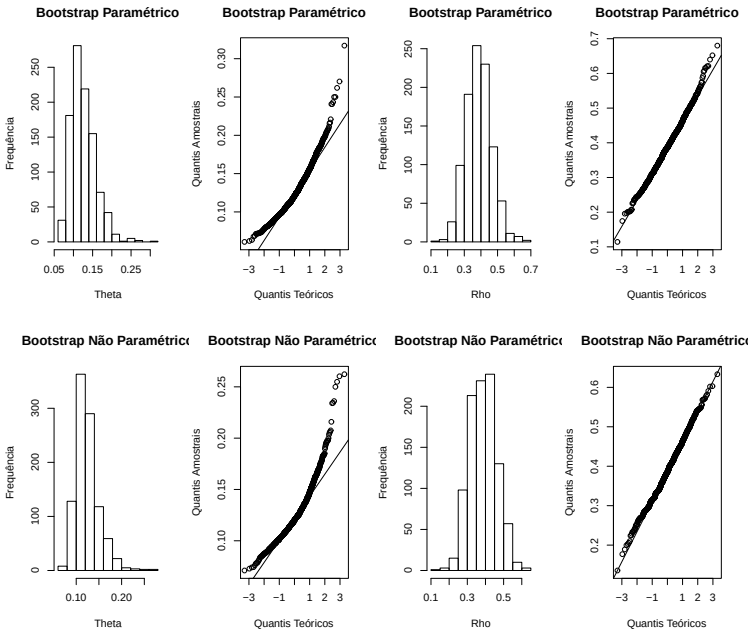

Figura 2: *QQ-plots* e histogramas das distribuições empíricas dos estimadores dos parâmetros do modelo ZIG via *bootstrap* paramétrico e não paramétrico.

Abstract. We propose in this paper the use of zero-inflated Geometric model, which is a generalization of the Geometric model, in the analysis of reliability and survival data. The use of this model is necessary especially when the data presents an excessive number of zeros. Maximum likelihood estimates of parameters were obtained, even as their confidence intervals based on asymptotic theory. In addition, we use the bootstrap resampling technique as an appropriate alternative procedure to construct confidence intervals of parameters of zero-inflated Geometric model.

Keywords. Intervalar estimation, coverage probability, bootstrapping technique.

Referências

[1] C.G. Carrasco, F. Louzada-Neto, Estimação intervalar para os parâmetros do modelo poly-log-logístico, *Rev. Mat. Estat.*, 21, No. 1 (2003), 85-95.

[2] A.C. Cohen, Estimation of mixtures of discrete distributions. em "Proceedings of the International Symposium on Discrete Distributions", pp. 373-378, Montreal, Quebec. 1963.

[3] A.C. Davison, D.V. Hinkley, "Bootstrap Methods and their Application", Cambrige: Cambrige University Press, 1997.

[4] R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

[5] B. Efron, Bootstrap methods: another look at the jacknife, *Annals of Statistics*, **7** (1979), 1–26.

[6] S.L. Jeng, W.Q. Meeker, Comparisons of approximate confidence interval procedures for type I censored data, *Technometrics*, 42 (1999), 135-148.

[7] N.L. Jonhson, S. Kotz, A.W. Kemp, "Univariate Discrete Distributions", second edition, John Wiley and Sons, New York, 1992.

[8] N.L. Jonhson, S. Kotz, A.W. Kemp, "Discrete Distributions: Distributions in Statistics". John Wiley and Sons, New York, 1969.

[9] J.F. Lawless, "Statistical Models and Methods for Lifetime Data", John Wiley and Sons, New York, 1982.

[10] F. Louzada-Neto, G.C. Perdoná, C.G. Carrasco, The Bi-log-logistic Model - A comparison study of some approximate confidence interval procedures, *JSTA - Journal of Statistical Theory and Applications*, 8, No.4 (2009), 478-492.

[11] E.Y. Nakano, C.G. Carrasco, Uma avaliação do uso de um modelo contínuo na análise de dados discretos de sobrevivência, *TEMA - Tend. Mat. Apl. Comput.*, 7, No 1 (2006), 91-100.