MÉTODO EFICIENTE PARA CALCULAR OS ESTIMADORES DE MÁXIMA VEROSSIMILHANÇA DA DISTRIBUIÇÃO GAMA GENERALIZADA

Pedro Luiz RAMOS[1]
Jorge Alberto ACHCAR[2]
Eduardo RAMOS[3]

▪ RESUMO: Neste artigo, será mostrado que utilizando uma forma simplificada das equações de verossimilhança, apresentada por Hager e Bain (1970), consegue-se reduzir a instabilidade do método de estimação obtendo-se boas estimativas para os parâmetros de interesse. Será proposto um método a fim de se obter bons valores iniciais para serem utilizados nos procedimentos de estimação dos parâmetros, considerando-se dados censurados (censura aleatória). Por fim, serão apresentados alguns exemplos de aplicações, utilizando-se dados da literatura e dados reais.

▪ PALAVRAS-CHAVE: Distribuição gama generalizada; estimador de máxima verossimilhança; censura aleatória; dados de sobrevivência.

1 Introdução

A distribuição Gama Generalizada com três parâmetros, tem se mostrado muito flexível para modelar dados de confiabilidade pois acomoda várias formas da função de risco. Introduzida por Stacy (1962), uma variável aleatória T tem distribuição Gama Generalizada (GG) se sua função densidade de probabilidade é dada por

$$f(t|\boldsymbol{\theta}) = \frac{\alpha}{\Gamma(\phi)}\mu^{\alpha\phi}t^{\alpha\phi-1}\exp\left(-(\mu t)^{\alpha}\right), \quad (1)$$

[1]Universidade de São Paulo – USP, ICMC, Programa de Pós-Graduação em Estatística, CEP: 13566-590, São Carlos, SP, Brasil. E-mail: *pedrolramos@hotmail.com*
[2]Universidade de São Paulo – USP, Faculdade de Medicina de Ribeirão Preto – FMRP, Departamento de Medicina Social, CEP: 14049-900, Ribeirão Preto, SP, Brasil. E-mail: *achcar@fmrp.usp.br*
[3]Universidade Estadual Paulista – UNESP, Faculdade de Ciências e Tecnologia – FCT, CEP: 19060-900, Presidente Prudente, SP, Brasil. E-mail: *edudkm@hotmail.com*

sendo $t > 0$, $\boldsymbol{\theta} = (\phi, \mu, \alpha)$. $\alpha > 0$ e $\phi > 0$ são respectivamente, dois parâmetros de forma e $\mu > 0$ é um parâmetro de escala.

A função distribuição acumulada é dada por

$$F(t|\boldsymbol{\theta}) = \int_0^{(\mu t)^\alpha} \frac{1}{\Gamma(\phi)} w^{\phi-1} e^{-w} dw = \frac{\gamma\left[\phi, (\mu t)^\alpha\right]}{\Gamma(\phi)} \ , \tag{2}$$

sendo $\gamma[y, x] = \int_0^x w^{y-1} e^{-w} dw$ denominada função gama incompleta inferior.

Distribuições de probabilidade usuais podem ser obtidas a partir da distribuição GG como a distribuição Weibull (quando $\phi = 1$), a distribuição Gama ($\alpha = 1$), Log-Normal (caso limite quando $\phi \to \infty$) e a distribuição Normal Generalizada ($\alpha = 2$, veja Nadarajah, 2005). A distribuição Normal generalizada é também uma distribuição que inclui várias distribuições conhecidas como, half-normal ($\phi = 1/2, \mu = 1/\sqrt{2}\sigma$), Rayleigh ($\phi = 1, \mu = 1/\sqrt{2}\sigma$), Maxwell-Boltzmann ($\phi = 3/2$) e chi ($\phi = k/2, k = 1, 2, \ldots$).

Algumas propriedades deste modelo foram obtidas por Stacy e Mihram (1965), assim como a estimação dos parâmetros, baseado no método da máxima verossimilhança. Já Hager e Bain (1970) mostram que as três equações não-lineares obtidas por Stacy e Mihram (1965) são muito instáveis, e ao isolarem dois parâmetros, reduzem o problema para apenas uma equação não-linear.

Huang e Hwang (2006) utilizam o método dos momentos para se obter estimadores dos parâmetros da distribuição GG. Khodabin e Ahmadabadi (2010) comparam este método com o método da máxima verossimilhança e concluem que, em geral, os EMV's possuem melhor desempenho. No entanto, ambos os métodos são instáveis e seus valores dependem do valor inicial, escolhido nos métodos de iteração.

Neste trabalho, será mostrado inicialmente que utilizando uma forma simplificada das equações de verossimilhança, apresentada por Hager e Bain (1970), consegue-se reduzir a instabilidade do método e também a necessidade de se utilizar bons valores iniciais nos três parâmetros da distribuição GG, obtendo-se facilmente boas estimativas para os parâmetros de interesse.

Em análise de confiabilidade, ao analisar dados de falha, certamente haverá unidades na amostra que podem não ter falhado, isto é, o tempo exato de falha das unidades não é conhecido. Esses tipos de observações são chamados de dados censurados (Hamada et al., 2008).

Desta forma, será proposto um método a fim de se obter bons valores iniciais para os procedimentos de estimação dos parâmetros, considerando-se dados censurados (censura aleatória), que possibilitará a obtenção de boas estimativas para os parâmetros de interesse. Esses resultados são de grande interesse prático, pois possibilitarão o uso da distribuição Gama Generalizada em diversas áreas de aplicação como medicina, engenharia e climatologia.

Por fim, serão apresentados alguns exemplos de aplicações, utilizando-se dados da literatura e dados reais.

2 Estimador de máxima verossimilhança

Dentre os métodos estatísticos de inferência clássica, o método de máxima verossimilhança é preferível devido às suas melhores propriedades assintóticas. No método de máxima verossimilhança os estimadores são obtidos a partir da maximização da função de verossimilhança (ver, por exemplo, Casella e Berger, 2002)

Nesta seção, apresentaremos os Estimadores de Máxima Verossimilhança da distribuição gama generalizada (1), considerando dados completos e com censura tipo I.

2.1 Observações completas

Seja $T_1, \ldots, T_n$ uma amostra aleatória tal que $T \sim \mathrm{GG}(\alpha, \mu, \phi)$. Neste caso, a função de verossimilhança de (1) é

$$L(\boldsymbol{\theta}; \boldsymbol{t}) = \frac{\alpha^n}{\Gamma(\phi)^n} \mu^{n\alpha\phi} \left\{ \prod_{i=1}^{n} t_i^{\alpha\phi-1} \right\} \exp\left\{ -\mu^\alpha \sum_{i=1}^{n} t_i^\alpha \right\}, \tag{3}$$

onde $\boldsymbol{\theta} = (\phi, \mu, \alpha)$.

Das expressões $\frac{\partial}{\partial\alpha} \log(L(\boldsymbol{\theta}; \boldsymbol{t}))$, $\frac{\partial}{\partial\mu} \log(L(\boldsymbol{\theta}; \boldsymbol{t}))$ e $\frac{\partial}{\partial\phi} \log(L(\boldsymbol{\theta}; \boldsymbol{t}))$ iguais a 0, obtemos as equações de verossimilhança dadas por

$$n\,\psi(\hat{\phi}) = n\hat{\alpha} \log(\hat{\mu}) + \hat{\alpha} \sum_{i=1}^{n} \log(t_i) \tag{4}$$

$$n\hat{\phi} = \hat{\mu}^{\hat{\alpha}} \sum_{i=1}^{n} {t_i}^{\hat{\alpha}} \tag{5}$$

$$\frac{n}{\hat{\alpha}} + n\hat{\phi} \log(\mu) + \phi \sum_{i=1}^{n} \log(t_i) = \hat{\mu}^{\hat{\alpha}} \sum_{i=1}^{n} {t_i}^{\hat{\alpha}} \log(\hat{\mu} t_i), \tag{6}$$

em que $\psi(k) = \frac{\partial}{\partial k} \log\Gamma(k) = \frac{\Gamma'(k)}{\Gamma(k)}$. As soluções de (4, 5, 6) fornecem os estimadores de máxima verossimilhança (veja, por exemplo, Stacy e Mihram, 1965, Hager e Bain, 1970). Métodos numéricos como o de Newton-Rapshon são necessários para encontrar-se a solução deste sistema não linear.

Os estimadores de máxima verossimilhança de $\boldsymbol{\theta}$ são viciados para amostras pequenas. No caso de grandes amostras, tornam-se não-viciados e assintoticamente eficientes. Tais estimadores têm distribuição conjunta assintoticamente normal dada por

$$(\hat{\boldsymbol{\theta}}) \sim N_k[(\boldsymbol{\theta}), I^{-1}(\boldsymbol{\theta})] \text{ para } n \to \infty, \tag{7}$$

onde $I(\boldsymbol{\theta})$ é a matriz de Informação de Fisher dada por,

$$I(\alpha,\mu,\phi)=\begin{bmatrix} \frac{1+2\psi(\phi)+\phi\psi'(\phi)+\phi\psi(\phi)^2}{\alpha^2} & -\frac{1+\phi\psi(\phi)}{\mu} & -\frac{\psi(\phi)}{\alpha} \\ -\frac{1+\phi\psi(\phi)}{\mu} & \frac{\phi\alpha^2}{\mu^2} & \frac{\alpha}{\mu} \\ -\frac{\psi(\phi)}{\alpha} & \frac{\alpha}{\mu} & \psi'(\phi) \end{bmatrix}, \quad (8)$$

em que $\psi'(k)=\frac{\partial}{\partial k}\psi(k)$ é denominada função trigama.

Quando o tamanho amostral é grande, pode-se construir intervalos de confiança aproximados para os parâmetros individuais θ_i, com coeficiente de confiança $100(1-\gamma)\%$, através das distribuições marginais dadas por

$$(\hat{\theta}_i) \sim N[(\theta_i), I_{ii}^{-1}(\boldsymbol{\theta})] \text{ para } n \to \infty. \quad (9)$$

Com estes resultados é possível construir intervalos de confiança individuais aproximados.

2.2 Censura tipo aleatória

Seja $T_1,\ldots,T_n$ uma amostra aleatória com $T \sim \text{GG}(\alpha,\mu,\phi)$, considerando observações com censura tipo aleatória, a função de verossimilhança é dada por

$$L(\boldsymbol{\theta};\boldsymbol{t},\boldsymbol{\delta})=\frac{\alpha^d\mu^{d\alpha\phi}}{\Gamma(\phi)^n}\left\{\prod_{i=1}^n t_i^{\delta_i\alpha\phi-1}\right\}\exp\left\{-\mu^\alpha\sum_{i=1}^n \delta_i {t_i}^\alpha\right\}\prod_{i=1}^n(\Gamma[\phi,(\mu t_i)^\alpha])^{1-\delta_i} \quad (10)$$

onde δ_i é um indicador de censuras, isto é, $\delta_i=1$ para observações completas e $\delta_i=0$ para observações censuradas.Além disso, $d=\sum_{i=1}^n\delta_i$, isto é, o números de observações não censuradas.

Das expressões $\frac{\partial}{\partial\alpha}\log(L(\boldsymbol{\theta};\boldsymbol{t},\boldsymbol{\delta}))$, $\frac{\partial}{\partial\mu}\log(L(\boldsymbol{\theta};\boldsymbol{t},\boldsymbol{\delta}))$ e $\frac{\partial}{\partial\phi}\log(L(\boldsymbol{\theta};\boldsymbol{t},\boldsymbol{\delta}))$ iguais a zero, obtemos as equações de verossimilhança dadas por

$$\begin{aligned}\sum_{i=1}^n\left\{(1-\delta_i)\left[\frac{(\mu t_i)^{\alpha\phi}e^{-(\mu t_i)^\alpha}\log(\mu t_i)}{\Gamma\left[\phi,\mu t_i^\alpha\right]}\right]\right\} &= \frac{d}{\alpha}+d\phi\log(\mu) \\ &+\phi\sum_{i=1}^n\delta_i\log(t_i)-\mu^\alpha\sum_{i=1}^n\delta_i t_i^\alpha\log(\mu t_i)\end{aligned} \quad (11)$$

$$\frac{d\alpha\phi}{\mu}-\alpha\mu^{\alpha-1}\sum_{i=1}^n\delta_i t_i^\alpha=\sum_{i=1}^n\left\{(1-\delta_i)\left[\frac{\alpha t_i(\mu t_i)^{\alpha\phi-1}e^{-(\mu t_i)^\alpha}}{\Gamma\left[\phi,\mu t_i^\alpha\right]}\right]\right\} \quad (12)$$

$$d\alpha\log(\mu)-n\psi(\phi)+\alpha\sum_{i=1}^n\delta_i\log(t_i)=-\sum_{i=1}^n\left\{(1-\delta_i)\left[\frac{\Psi\left[\phi,\mu t_i^\alpha\right]}{\Gamma\left[\phi,\mu t_i^\alpha\right]}\right]\right\} \quad (13)$$

sendo $\Gamma[y,x]=\int_x^\infty w^{y-1}e^{-w}dw$ denominada de gama incompleta superior e $\Psi(k,x)=\frac{\partial}{\partial k}\Gamma[k,x]=\int_x^\infty w^{k-1}\log(w)e^{-w}dw$.

As soluções deste sistema não-linear fornecem os estimadores de máxima verossimilhança de ϕ, μ e α. Métodos numéricos como o de Newton-Rapshon são necessários para encontrar tais soluções.

Outros tipos de censura podem ser considerados ao se realizar inferências nos parâmetros da distribuição GG. Chen, e Lio (2009) obtêm as equações de verossimilhança para os casos em que as observações possuem censura progressiva tipo II (veja, por exemplo, Balakrishnan e Aggarwala, 2000).

2.3 Equações úteis para encontrar bons valores iniciais

É comum se deparar com dificuldades computacionais ao se realizar inferências nos parâmetros da distribuição Gama Generalizada. As estimativas dos parâmetros obtidas considerando dados completos e censurados, podem depender dos valores iniciais utilizados nos métodos computacionais iterativos.

Algumas manipulações algébricas podem ser feitas nas equações de verossimilhança buscando-se simplificar a solução do sistema. A equação (6) pode ser rescrita através de

$$\frac{n}{\hat{\alpha}} + \phi \sum_{i=1}^{n} \log(t_i) = \hat{\mu}^{\hat{\alpha}} \sum_{i=1}^{n} {t_i}^{\hat{\alpha}} \log(t_i), \tag{14}$$

Utilizando a expressão (5), após algumas manipulações algébricas, temos que

$$\hat{\phi} = \hat{\mu}^{\hat{\alpha}} \frac{\sum_{i=1}^{n} {t_i}^{\hat{\alpha}}}{n}. \tag{15}$$

Substituindo (15) em (14) e após algumas manipulações algébrica temos que

$$\hat{\mu} = \left(\frac{1}{\hat{\alpha}} \frac{n}{\sum_{i=1}^{n} {t_i}^{\hat{\alpha}} \log(t_i) - \frac{\sum_{i=1}^{n} {t_i}^{\hat{\alpha}}}{n} \sum_{i=1}^{n} \log(t_i)} \right)^{\frac{1}{\hat{\alpha}}}. \tag{16}$$

Substituindo (16) em (15) temos que

$$\hat{\phi} = \left(\frac{1}{\hat{\alpha}} \frac{\sum_{i=1}^{n} {t_i}^{\hat{\alpha}}}{\sum_{i=1}^{n} {t_i}^{\hat{\alpha}} \log(t_i) - \frac{\sum_{i=1}^{n} {t_i}^{\hat{\alpha}}}{n} \sum_{i=1}^{n} \log(t_i)} \right). \tag{17}$$

Desta forma utilizando os resultados (16) e (17) na equação (4) temos que

$$\begin{aligned} \mathrm{h}(\hat{\alpha}) = \psi \left(\frac{1}{\hat{\alpha}} \frac{\sum_{i=1}^{n} {t_i}^{\hat{\alpha}}}{\sum_{i=1}^{n} {t_i}^{\hat{\alpha}} \log(t_i) - \frac{\sum_{i=1}^{n} {t_i}^{\hat{\alpha}}}{n} \sum_{i=1}^{n} \log(t_i)} \right) - \frac{\hat{\alpha} \sum_{i=1}^{n} \log(t_i)}{n} \\ - \log \left(\frac{1}{\hat{\alpha}} \frac{n}{\sum_{i=1}^{n} {t_i}^{\hat{\alpha}} \log(t_i) - \frac{\sum_{i=1}^{n} {t_i}^{\hat{\alpha}}}{n} \sum_{i=1}^{n} \log(t_i)} \right) = 0. \end{aligned} \tag{18}$$

Estes resultados, apresentados por Hager e Bain (1970), são de grande interesse prático, pois resolvendo a equação (18), encontramos o EMV de α, com este

resultado encontramos respectivamente os EMV's de ϕ e μ. Este procedimento não elimina totalmente o problema de instabilidade dos EMV, mas o reduz, possibilitando a obtenção de estimativas mais precisas.

No entanto, quando utilizamos observações censuradas, as equações de verossimilhança (11, 12 e 13) não podem ser simplificadas similarmente, como foi feito nas equações de verossimilhança (4, 5 e 6). Desta forma, as equações obtidas considerando-se observações censuradas são muito instáveis e as estimativas obtidas de ϕ, μ e α podem depender dos valores iniciais, utilizados no método iterativo.

Bons valores iniciais para se utilizar nas equações (11, 12 e 13) podem ser obtidos, removendo-se as observações censuradas e calculando-se $\tilde{\phi}, \tilde{\mu}$ e $\tilde{\alpha}$ através das equações (16, 17 e 18). Embora haja perda de informação ao se remover as observações censuradas, como se trata de uma análise preliminar, os valores $\tilde{\phi}, \tilde{\mu}$ e $\tilde{\alpha}$ podem ser utilizados apenas para se inicializar o método iterativo.

3 Análise numérica

Neste exemplo, serão realizadas simulações via método de Monte de Carlo. O objetivo destas simulações será estudar a eficiência dos estimadores obtidos para os diferentes métodos apresentados. Também será investigada a influência do tamanho da amostra na acurácia dos estimadores. O seguinte procedimento foi adotado no estudo:

1. Determine os valores de $\boldsymbol{\theta} = (\phi, \mu, \alpha)$.
2. Especifique o tamanho da amostra n.
3. Gere valores de uma distribuição GG(ϕ, μ, α) com tamanho n.
4. Utilizando os dados obtidos no passo 3, calcule os EMV's $\hat{\phi}, \hat{\mu}$ e $\hat{\alpha}$.
5. Repita o procedimento 3 e 4 N vezes.

Os valores escolhidos para se realizar tal procedimento serão, $\boldsymbol{\theta} = ((0.5, 0.5, 3), (2, 1, 0.5), (4, 2, 2), (0.4, 1.5, 5))$ e $n = (50, 200, 500)$. A semente utilizada para gerar os valores aleatórios no software R é 2013.

Ao utilizar-se o método de Newton-Raphson, é necessário especificar os valores inicias para inicialização do procedimento iterativo. Espera-se que o método, após algumas iterações, leve aos verdadeiros valores de $\boldsymbol{\theta}$.

A comparação entre os métodos será feita através do cálculo das médias e dos erros-padrão das N estimativas, obtidas através dos EMV's. Espera-se que o melhor método de estimação, possua as médias das N estimativas, mais perto dos verdadeiros valores de $\boldsymbol{\theta}$ com menores erros-padrão.

Serão calculadas também as probabilidades de cobertura dos parâmetros, através dos intervalos de confiança assintóticos (inferência clássica). Para um grande número de experimentos, utilizando-se um nível de confiança de 95%, as frequências dos intervalos que cobriram os verdadeiros valores de $\boldsymbol{\theta}$ devem ser de aproximadamente 95%.

Na ausência de informação sobre quais valores iniciais devem ser utilizados no método iterativo, serão gerados valores aleatórios tais que $\tilde{\phi} \sim U(0,5), \tilde{\mu} \sim U(0,5)$ e $\tilde{\alpha} \sim U(0,5)$. Portanto os valores utilizados para a inicialização do método de Newton-Raphson serão $\boldsymbol{\theta}^{(0)} = (\tilde{\phi}, \tilde{\mu}, \tilde{\alpha})$. Este procedimento será denominado de Método 1. O pacote do software *R* utilizado para resolver o método iterativo será o maxNR.

3.1 Observações completas

Como foi demonstrado, as equações de verossimilhança (4, 5 e 6) podem ser simplificadas através de (16, 17 e 18). Será utilizado o pacote multiroot, disponível no R, para resolver a equação (18). A solução desta equação acarreta na obtenção do EMV de $\hat{\alpha}$ e com este resultado obtêm-se facilmente, através de (16 e 17), os EMV's de $\hat{\mu}$ e $\hat{\phi}$. Este procedimento será denominado de Método 2.

Na Tabela 1 temos as médias e os erros-padrão das estimativas de $N = 100000$ amostras obtidas utilizando os EMV's, calculados através dos Métodos 1 e 2 para diferentes valores de $\boldsymbol{\theta}$ e n. Na Tabela 2, podemos ver as probabilidades de cobertura com o nível de confiança de 95%.

Tabela 1 - Estimativas das médias e erros-padrão dos EMV's encontrados para N amostras de tamanho $n = (50, 100, 200)$, através dos Métodos 1 e 2 para diferentes valores de $\boldsymbol{\theta}$ e n

		Método 1			Método 2	
$\boldsymbol{\theta}$	$n = 50$	$n = 200$	$n = 500$	$n = 50$	$n = 200$	$n = 500$
$\phi = 0.5$	1.94(1.41)	1.79(1.25)	1.72(1.24)	0.72(0.73)	0.52(0.18)	0.51(0.10)
$\mu = 0.5$	1.82(1.99)	1.65(1.84)	1.58(1.73)	0.68(0.88)	0.51(0.07)	0.50(0.04)
$\alpha = 3$	1.83(1.26)	1.74(0.92)	1.78(0.87)	3.32(1.62)	3.16(0.78)	3.06(0.44)
$\phi = 2$	2.56(1.10)	2.52(0.98)	2.46(0.93)	1.68(0.97)	2.02(0.80)	2.09(0.63)
$\mu = 1$	4.52(4.64)	4.13(4.19)	3.80(4.01)	1.75(3.30)	2.11(3.28)	1.86(2.56)
$\alpha = 0.5$	0.56(0.43)	0.52(0.32)	0.51(0.29)	0.70(0.58)	0.54(0.13)	0.51(0.09)
$\phi = 4$	4.72(2.73)	4.70(2.45)	4.74(2.42)	4.66(3.69)	4.77(3.15)	4.59(2.32)
$\mu = 2$	2.63(1.91)	2.62(1.83)	2.66(1.90)	2.96(3.20)	2.89(2.79)	2.55(1.88)
$\alpha = 2$	2.20(0.90)	2.05(0.56)	2.01(0.51)	2.35(0.87)	2.14(0.67)	2.03(0.46)
$\phi = 0.4$	1.12(0.99)	0.98(0.86)	1.02(0.85)	0.69(0.64)	0.44(0.13)	0.41(0.08)
$\mu = 1.5$	2.35(1.69)	2.20(1.58)	2.22(1.51)	1.81(0.93)	1.54(0.11)	1.51(0.06)
$\alpha = 5$	3.52(1.92)	3.47(1.48)	3.38(1.39)	4.35(1.63)	4.93(0.94)	5.05(0.67)

Pode-se observar pelos resultados obtidos e demonstrados na Tabela 1, que as estimativas de máxima verossimilhança são muito influenciadas pelos valores iniciais. Por exemplo, utilizando-se o Método 1, para $\mu = 1$ e $n = 50$, a média das 100000 estimativas é de $\mu = 4.52$, com erro-padrão $\sigma_\mu = 4.64$.

Tabela 2 - Probabilidades de cobertura com o nível de confiança de 95% dos EMV's encontrados para N amostras de tamanho $n = (50, 100, 200)$, com diferentes valores de $\boldsymbol{\theta}$ utilizando-se os Métodos 1 e 2

	Método 1			Método 2		
$\boldsymbol{\theta}$	$n = 50$	$n = 200$	$n = 500$	$n = 50$	$n = 200$	$n = 500$
$\phi = 0.5$	97.94%	98.53%	39.87%	92.31%	91.56%	93.78%
$\mu = 0.5$	98.28%	98.59%	97.85%	92.78%	91.83%	94.00%
$\alpha = 3$	44.64%	29.57%	28.59%	92.80%	94.92%	94.94%
$\phi = 2$	90.23%	89.20%	87.79%	78.67%	88.30%	92.18%
$\mu = 1$	86.27%	86.20%	85.42%	62.87%	77.62%	84.40%
$\alpha = 0.5$	100.00%	93.21%	86.63%	100.00%	98.37%	97.28%
$\phi = 4$	93.11%	92.71%	89.03%	88.66%	87.08%	90.53%
$\mu = 2$	91.12%	90.56%	86.51%	83.63%	83.88%	88.14%
$\alpha = 2$	100.00%	99.31%	95.20%	100.00%	98.82%	95.99%
$\phi = 0.4$	98.74%	98.70%	50.18%	98.96%	98.70%	96.34%
$\mu = 1.5$	99.15%	98.86%	74.34%	99.05%	98.36%	96.40%
$\alpha = 5$	61.85%	46.01%	39.33%	88.39%	94.69%	96.10%

Utilizando-se o Método 2, observa-se que o método proposto retorna em todos os casos ótimas estimativas de ϕ, μ e α. As probabilidades de cobertura dos parâmetros mostram que, embora alguns casos estejam abaixo do ideal, em sua maioria são superiores a 90%.

Com estes resultados pode-se constatar que, ao utilizarmos o Método 2, conseguimos obter facilmente ótimas inferências para os parâmetros da distribuição gama generalizada.

3.2 Observações censuradas

Assim, como foi discutido na seção 2.3, bons valores iniciais para se utilizar nas equações (11, 12 e 13) podem ser obtidos removendo-se as observações censuradas e calculando-se $\tilde{\phi}, \tilde{\mu}$ e $\tilde{\alpha}$ através das equações (16, 17 e 18). Embora haja perda de informação ao se remover as observações censuradas, como se trata de uma análise preliminar, os valores $\tilde{\phi}, \tilde{\mu}$ e $\tilde{\alpha}$ obtidos de (11, 12 e 13) podem ser utilizados apenas para se inicializar o método iterativo. Tal procedimento será denominado de Método 3.

Nas Tabelas 3 e 5 temos as médias e os erros-padrão das estimativas de $N = 10000$ amostras obtidas utilizando os EMV's, calculados através dos Métodos 1 e 3 para diferentes valores de $\boldsymbol{\theta}$ e n considerando 20% e 40% de censura respectivamente. Nas Tabela 4 e 6, podemos ver as probabilidades de cobertura com o nível de confiança de 95% considerando 20% e 40% de censura respectivamente.

Tabela 3 - Estimativas das médias e erros-padrão dos EMV's encontrados para N amostras de tamanho $n = (50, 100, 200)$, através dos Métodos 1 e 3 para diferentes valores de $\boldsymbol{\theta}$, n e 20% de censura

	Método 1			Método 3		
$\boldsymbol{\theta}$	$n = 50$	$n = 100$	$n = 200$	$n = 50$	$n = 100$	$n = 200$
$\phi = 0.5$	1.14(1.42)	1.00(1.21)	0.97(1.17)	0.74(0.79)	0.52(0.22)	0.51(0.12)
$\mu = 0.5$	1.23(2.65)	1.09(2.39)	1.05(2.29)	0.67(1.13)	0.50(0.10)	0.49(0.04)
$\alpha = 3$	3.70(3.39)	3.11(1.95)	3.02(1.82)	3.61(2.42)	3.39(0.98)	3.24(0.54)
$\phi = 2$	1.77(1.18)	1.84(0.99)	1.86(0.90)	1.47(0.96)	1.79(0.80)	1.88(0.62)
$\mu = 1$	2.15(4.50)	1.86(3.83)	1.68(3.39)	1.45(4.94)	1.52(4.05)	1.24(2.92)
$\alpha = 0.5$	0.95(1.55)	0.69(0.78)	0.65(0.61)	0.84(1.02)	0.60(0.16)	0.55(0.10)
$\phi = 4$	3.06(2.81)	3.12(2.57)	3.14(2.41)	3.83(3.46)	4.09(3.10)	4.11(2.29)
$\mu = 2$	1.87(2.11)	1.89(2.15)	1.86(1.97)	2.45(3.34)	2.52(3.17)	2.27(2.01)
$\alpha = 2$	3.52(2.65)	3.03(1.94)	2.91(1.77)	2.62(1.00)	2.37(0.72)	2.19(0.51)
$\phi = 0.4$	0.68(0.92)	0.61(0.79)	0.59(0.74)	0.74(0.72)	0.46(0.16)	0.42(0.09)
$\mu = 1.5$	1.86(1.81)	1.78(1.75)	1.73(1.66)	1.83(1.26)	1.52(0.12)	1.49(0.06)
$\alpha = 5$	6.21(4.40)	5.34(2.57)	5.20(2.25)	4.52(2.47)	5.08(1.08)	5.22(0.77)

Tabela 4 - Probabilidades de cobertura com o nível de confiança de 95% dos EMV's encontrados para N amostras de tamanho $n = (50, 100, 200)$, com diferentes valores de $\boldsymbol{\theta}$, 20% censura e utilizando-se os Métodos 1 e 3

	Método 1			Método 3		
$\boldsymbol{\theta}$	$n = 50$	$n = 100$	$n = 200$	$n = 50$	$n = 100$	$n = 200$
$\phi = 0.5$	69.57%	61.80%	56.37%	93.72%	91.24%	93.10%
$\mu = 0.5$	78.82%	64.27%	56.13%	91.61%	86.02%	86.12%
$\alpha = 3$	70.42%	63.60%	56.95%	95.61%	96.52%	95.56%
$\phi = 2$	60.81%	57.34%	52.15%	70.95%	76.94%	81.13%
$\mu = 1$	51.76%	50.28%	45.52%	49.19%	59.40%	63.81%
$\alpha = 0.5$	80.37%	65.59%	55.87%	94.42%	91.25%	89.95%
$\phi = 4$	56.76%	47.64%	39.96%	81.37%	70.41%	73.82%
$\mu = 2$	52.61%	43.90%	36.56%	76.05%	67.20%	71.04%
$\alpha = 2$	81.10%	61.39%	48.30%	90.22%	85.40%	83.27%
$\phi = 0.4$	77.16%	73.31%	68.69%	97.98%	99.13%	97.74%
$\mu = 1.5$	82.73%	75.01%	68.36%	98.12%	97.77%	93.55%
$\alpha = 5$	78.86%	75.32%	69.62%	92.56%	95.92%	97.43%

Tabela 5 - Estimativas das médias e erros-padrão dos EMV's encontrados para N amostras de tamanho $n = (50, 100, 200)$, através dos Métodos 1 e 3 para diferentes valores de $\boldsymbol{\theta}$, n e 40% de censura

	Método 1			Método 3		
$\boldsymbol{\theta}$	$n = 50$	$n = 100$	$n = 200$	$n = 50$	$n = 100$	$n = 200$
$\phi = 0.5$	1.16(1.46)	1.06(1.31)	1.02(1.26)	1.19(1.18)	0.67(0.26)	0.57(0.12)
$\mu = 0.5$	1.19(2.52)	1.10(2.43)	1.07(2.46)	1.00(2.41)	0.52(0.11)	0.48(0.04)
$\alpha = 3$	4.18(4.05)	3.30(2.14)	3.17(1.83)	2.83(3.26)	2.91(0.82)	3.09(0.43)
$\phi = 2$	1.83(1.29)	1.87(1.06)	1.91(0.96)	1.31(0.94)	1.59(0.78)	1.75(0.67)
$\mu = 1$	2.09(4.52)	1.71(3.76)	1.56(3.44)	0.89(4.41)	0.88(2.81)	0.90(2.58)
$\alpha = 0.5$	1.16(2.15)	0.73(0.71)	0.66(0.55)	1.15(2.26)	0.68(0.23)	0.60(0.13)
$\phi = 4$	3.03(2.79)	3.12(2.60)	3.16(2.50)	3.88(3.67)	3.99(3.15)	3.98(2.40)
$\mu = 2$	1.80(1.95)	1.84(2.02)	1.85(2.08)	2.48(3.68)	2.43(3.24)	2.19(2.25)
$\alpha = 2$	3.82(3.10)	3.13(2.05)	2.96(1.80)	2.74(1.16)	2.46(0.78)	2.29(0.57)
$\phi = 0.4$	0.76(1.04)	0.63(0.85)	0.60(0.77)	1.06(1.06)	0.58(0.21)	0.50(0.10)
$\mu = 1.5$	1.89(1.78)	1.76(1.81)	1.71(1.60)	2.14(2.08)	1.57(0.17)	1.51(0.07)
$\alpha = 5$	6.69(5.07)	5.71(2.86)	5.45(2.30)	4.04(3.45)	4.45(1.02)	4.79(0.71)

Tabela 6 - Probabilidades de cobertura com o nível de confiança de 95% dos EMV's encontrados para N amostras de tamanho $n = (50, 100, 200)$, com diferentes valores de $\boldsymbol{\theta}$, 40% censura e utilizando-se os Métodos 1 e 3

	Método 1			Método 3		
$\boldsymbol{\theta}$	$n = 50$	$n = 100$	$n = 200$	$n = 50$	$n = 100$	$n = 200$
$\phi = 0.5$	71.65%	61.66%	57.60%	95.57%	99.55%	99.33%
$\mu = 0.5$	76.36%	61.28%	47.98%	95.59%	96.86%	92.15%
$\alpha = 3$	71.73%	64.46%	58.58%	90.79%	96.66%	98.69%
$\phi = 2$	61.45%	57.62%	53.56%	68.73%	73.03%	75.36%
$\mu = 1$	49.78%	46.02%	42.32%	38.86%	42.83%	43.83%
$\alpha = 0.5$	82.95%	67.58%	58.99%	96.16%	93.97%	89.21%
$\phi = 4$	56.86%	49.11%	42.19%	82.16%	72.09%	75.17%
$\mu = 2$	52.21%	43.86%	36.45%	76.04%	66.65%	69.78%
$\alpha = 2$	83.51%	64.35%	51.74%	90.15%	87.85%	85.61%
$\phi = 0.4$	77.50%	74.11%	70.31%	95.98%	99.69%	99.27%
$\mu = 1.5$	81.19%	71.62%	63.23%	96.84%	98.82%	96.85%
$\alpha = 5$	79.08%	75.77%	71.13%	89.50%	94.11%	96.94%

Através dos resultados demonstrados pode-se concluir que o Método 3 possibilita a obtenção de bons valores iniciais, e que, ao serem utilizados no método iterativo (Newton-Raphson), retornam melhores estimativas para os parâmetros ϕ, μ e α.

É importante salientar que, utilizando-se o Método 3, as probabilidades de cobertura dos parâmetros tende a aumentar quando o tamanho amostral cresce, aproximando-se dos valores esperados (95% de confiança).

4 Aplicações

Nesta seção, serão apresentados dois exemplos de aplicações utilizando-se dados da literatura e dados reais. Iremos considerar a distribuição Gama Generalizada para analisar tais conjuntos de dados e iremos comparar os resultados obtidos com outros modelos, como as distribuições Weibull e Gama através do teste AIC (Akaike, 1974).

4.1 Dados de precipitação pluvial total mensal

Os dados de precipitação pluviométrica utilizados nesta seção, foram obtidos da Estação Meteorológica da Faculdade de Ciências e Tecnologia da Universidade Estadual Paulista, campus de Presidente Prudente (SP), disponíveis na Tabela 7 e na página virtual do SIGRH-SP - Sistema de Informações para o Gerenciamento de Recursos Hídricos do Estado de São Paulo, da estação medidora D8-003, no município de Presidente Prudente - SP. Os valores obtidos são referentes ao mês de maio e compreendem um período de 56 anos (1947 a 2003). Desta forma, o conjunto de dados de precipitação pluvial total mensal (em mm), representando uma amostra, é composto por meio do valor total (em mm), registrado no mês de maio para cada ano.

Tabela 7 - Conjunto de dados de precipitação pluvial total mensal (em mm), referente ao mês de maio, considerando um período de 56 anos (1947 a 2003)

15.5	3.0	23.0	21.8	150.6	31.0	46.7	22.8	39.4	0.1
48.0	238.6	51.4	187.0	19.9	85.5	86.8	51.6	24.3	86.0
22.4	46.8	157.8	101.8	0.1	35.6	54.7	64.2	104.1	83.6
72.0	77.0	13.7	89.6	43.2	110.6	97.2	65.4	0.1	83.3
158.0	29.7	59.3	129.7	160.7	126.6	47.2	75.5	13.0	171.3
70.8	75.5	61.2	41.5	81.4	55.4	37.8	114.7	144.4	34.8
172.1									

Akaike (1974) propõe um critério de comparação de modelos baseado na medida de Informação de Kullback-Leibler. Seja k o número de parâmetros a serem

estimados, n o número de observações de $\boldsymbol{x}$ e $\hat{\boldsymbol{\theta}}$ uma estimativa de $\boldsymbol{\theta}$, o critério de informação de Akaike (AIC) é obtido por

$$AIC = -2\log(L(\hat{\boldsymbol{\theta}}; \boldsymbol{x})) + 2k \ . \tag{19}$$

Dado um conjunto de modelos candidatos para $\boldsymbol{x}$, ajustados os dados, o preferido será o que fornecer o menor AIC. Além de selecionar um ótimo ajuste, o critério penaliza a adição de parâmetros, desencorajando overfitting, ou seja, a seleção de um modelo extremamente complexo e com muitos parâmetros que tenham um pobre desempenho preditivo.

Para se obter as estimativas de ϕ, μ e α, utilizamos as equações (16, 17 e 18). Desta forma, para o conjunto de dados analisado, encontramos que $\hat{\phi} = 0.2761, \hat{\mu} = 0.00628$ e $\hat{\alpha} = 2.9051$.

A Tabela 8 apresenta os resultados do critério AIC de diferentes distribuições de probabilidade, considerando-se os dados de precipitação pluvial total mensal.

Tabela 8 - Resultados do critério AIC de diferentes distribuições de probabilidade, considerando-se os dados de precipitação pluvial total mensal

Critérios	G. Generalizada	Weibull	Gama
AIC	**640,87**	646,34	648,17

Baseado nos resultados obtidos por meio do critério AIC, podemos concluir, através da Tabela 8, que a distribuição Gama Generalizada foi a que melhor se ajustou aos dados de precipitação pluvial total mensal de Presidente Prudente.

4.2 Dados de sobrevida de pacientes com câncer na cabeça e pescoço

Em estudos médicos, é comum se deparar com dados que possuem censura aleatória. Nesta seção iremos analisar os dados disponíveis em Efron (1988). Estes dados consistem nos tempos de vida (em dias) de 51 pacientes com câncer na cabeça e pescoço.

Na Tabela 9, temos os tempos de vida (em dias) de 51 pacientes com câncer na cabeça e pescoço (+ indica a presença de censura).

Para se obter os EMV's de ϕ, μ e α, é necessário se encontrar os bons valores iniciais $\tilde{\boldsymbol{\theta}} = (\tilde{\phi}, \tilde{\mu}, \tilde{\alpha})$, sendo estes obtidos utilizando-se (16, 17 e 18). Desta forma, encontramos que $\tilde{\phi} = 2.6193, \tilde{\mu} = 0.0160$ e $\tilde{\alpha} = 0.6062$. Utilizando-se estes resultados em (11, 12 e 13), as estimativas obtidas para o conjunto de dados são $\hat{\phi} = 4.1474, \hat{\mu} = 0.0787$ e $\hat{\alpha} = 0.4339$.

A Tabela 10 apresenta os resultados do critério AIC de diferentes distribuições de probabilidade, considerando-se os dados de tempos de vida de 51 pacientes com câncer na cabeça e pescoço.

Tabela 9 - Os tempos de vida (em dias) de 51 pacientes com câncer na cabeça e pescoço (+ indica a presença de censura)

7	34	42	63	64	74+	83	84	91
108	112	129	133	133	139	140	140	149
154	157	160	160	165	173	176	185+	218
225	241	248	273	277	279+	297	319+	405
417	420	440	523	523+	583	594	1101	1116+
1146	1226+	1349+	1412+	1417				

Tabela 10 - Resultados do critério AIC de diferentes distribuições de probabilidade, considerando-se os dados de tempos de vida (em dias) de 51 pacientes com câncer na cabeça e pescoço

Critérios	G. Generalizada	Weibull	Gama
AIC	**582,14**	585,40	585,01

Baseado nos resultados obtidos do critério AIC, podemos concluir, através da Tabela 10, que a distribuição Gama Generalizada foi a que melhor se ajustou aos tempos de vida de pacientes com câncer na cabeça e pescoço, mostrando-se que este é um excelente modelo para descrever dados médicos.

Conclusões

Neste trabalho, foi proposto um método, a fim de se obter bons valores iniciais para serem utilizados nos procedimentos de estimação dos parâmetros, considerando-se dados censurados (censura aleatória), que possibilitou a obtenção de boas estimativas para os parâmetros de interesse.

Demonstramos também que ao se utilizar a inferência clássica e considerando-se observações completas, por meio das equações obtidas por Harger e Bain (1970) é possível reduzir a instabilidade dos estimadores no método iterativo, possibilitando a obtenção de boas estimativas para os parâmetros da distribuição Gama Generalizada.

Baseado em estudos de simulação utilizando-se o Método de Monte Carlo observou-se que os procedimentos abordados conseguiram estimar com boa precisão os verdadeiros parâmetros, tanto para observações completas quanto censuradas, principalmente quando o tamanho de n aumenta.

Estes resultados mostram-se de grande interesse prático, pois possibilitarão, a fácil utilização da distribuição Gama Generalizada em diversas áreas de aplicação.

RAMOS, P. L.; ACHCAR, J. A.; RAMOS, E. Useful method to calculate the maximum likelihood estimation of generalized Gamma distribution. *Rev. Bras. Biom.,* São Paulo, v.32, n.2, p.267-281, 2014.

▪ *ABSTRACT: In this paper, we introduce an alternative to simplify the likelihood equations for the parameters of the generalized gamma distribution considering complete data sets , in a way to minimize the problem of instability of the iterative procedure used to numerically obtain the maximum likelihood estimates. It is also proposed some ways to get good initial values to be used in the iterative procedure considering censored or uncensored data. Some examples to illustrate the proposed methodology are considered using literature data and real data sets.*

▪ *KEYWORDS: Generalized gamma distribution; maximum likelihood estimates; censored data; lifetime data.*

Referências

AKAIKE, H. A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, v.19, n.6, p.716-723, 1974.

BALAKRISHNAN, N.; AGGARWALA, R. *Progressive Censoring: Theory, Methods and Applications*, Birkhauser, Boston, 2000.

CASELLA, G.; BERGER, R. *Statistical Inference* .2 ed. Belmont, CA: Duxbury, 2002.

CHEN, D. G.; LIO, Y. A Note on the Maximum Likelihood Estimation for the Generalized Gamma Distribution Parameters under Progressive Type-II Censoring. *International Journal of Intelligent Technology and Applied Statistics*, v.2, n.2, p.57-64, 2009.

EFRON, B. Logistic regression, survival analysis, and the Kaplan-Meier curve, *Journal of the American Statistical Association*, v.83, p.415–425, 1988.

HAGER, H. W.; BAIN L. J. Inferential procedures for the generalized gamma distribution. *Journal of the American Statistical Association*, v.65, p.1601-1609, 1970.

HAMADA, M.; WILSON, A. G.; REESE, C. S.; MARTZ, H. F. *Bayesian Reliability*, Springer, New York, 2008.

HUANG, P; HWANG, T. On new moment estimation of parameters of the Generalized Gamma distribution using it's characterization. *Taiwanese journal of Mathematics*, v.10, n.4, p.1083-1093, 2006.

KHODABIN M; AHMADABADI A. Some properties of generalized gamma distribution. *J. Mathematical Sciences*. v.4, p.9-28, 2010.

KUMAR, V.; SHUKLA, G. Maximum Likelihood Estimation in Generalized Gamma Type Model. *Journal of Reliability and Statistical Studies*, v.3, n.1, p.43-51, 2010.

LAWLESS, J. F. *Statistical models and methods for lifetime data*. New York: John Wiley and Sons, 580p, 1982.

NADARAJAH, S. A generalized normal distribution, *Journal of Applied Statistics*, v.32, n.7, p.685-694, 2005.

STACY, E. W. A generalization of the gamma distribution. *Annals of Mathematical Statistics*, v.28, p.1187-1192, 1962.

STACY, E. W.; MIHRAM, G. A. Parameter estimation for a generalized gamma distribution. *Technometrics*, v.7, p.349-358, 1965.

Recebido em 03.02.2014.

Aprovado após revisão em 17.07.2014.