REFERÊNCIA

BRUNELLO, Gabriel Hideki Vatanabe; NAKANO, Eduardo Yoshio. Inferência bayesiana no modelo Weibull discreto em dados com presença de censura. **TEMA** (São Carlos), São Carlos, v. 16, n. 2, p. 97-110, maio/ago. 2015. Disponível em: <http://www.scielo.br/scielo.php?script=sci_arttext&pid=S2179-84512015000200097&lng=en&nrm=iso>. Acesso em: 20 mar. 2018. doi: http://dx.doi.org/10.5540/tema.2015.016.02.0097.

Tema

Tendências em Matemática Aplicada e Computacional, **16**, N. 2 (2015), 97-110

www.scielo.br/tema
doi: 10.5540/tema.2015.016.02.0097

Inferência Bayesiana no Modelo Weibull Discreto em Dados com Presença de Censura

G.H.V. BRUNELLO* e E.Y. NAKANO

Recebido em 6 novembro, 2013 / Aceito em 11 março, 2015

RESUMO. Este trabalho apresenta uma inferência bayesiana da distribuição Weibull discreta em dados com presença de censuras. Foi proposto também um teste de significância genuinamente bayesiano (FBST – *Full Bayesian Significance Test*) para testar seu parâmetro de forma. Amostras da distribuição *a posteriori* dos parâmetros foram obtidas por meio de simulações via *Markov Chain Monte Carlo* (MCMC). A metodologia desenvolvida foi ilustrada em simulações e aplicada em um conjunto de dados sobre o tempo de sobrevivência de homens diagnosticados com AIDS. Todas as simulações e obtenções das estimativas foram realizadas com a linguagem R.

Palavras-chave: análise de sobrevivência, inferência bayesiana, testes de hipóteses, MCMC.

1 INTRODUÇÃO

A distribuição Weibull [22] é uma distribuição muito utilizada na modelagem de dados que representam o tempo até a ocorrência de um evento de interesse devido a sua versatilidade e relativa simplicidade [19]. Esse evento de interesse pode ser a morte de um paciente, remissão de uma doença, reação de um medicamento, quebra de um equipamento eletrônico, queima de uma lâmpada, dentre outros eventos. Em geral esses dados são analisados através de técnicas de Análise de Sobrevivência ou Confiabilidade, cuja principal característica é a presença de censuras, que consiste na observação parcial da resposta. Essa informação censurada, apesar de incompleta, é útil e importante para a análise. A distribuição Weibull é utilizada na análise de dados de sobrevivência quando os mesmos são contínuos. No entanto, em muitos casos os dados de sobrevivência não são contínuos. Os dados discretos surgem, por exemplo, quando o tempo de sobrevivência é medido em meses, ciclos ou intervalos. Foram estudadas em [15] as consequências do uso de um modelo contínuo em um conjunto de dados discretos, mostrando que nem sempre é razoável usar um modelo contínuo quando os dados são discretos. Os modelos discretos podem ser facilmente obtidos através de modelos contínuos, agrupando os tempos

*Autor correspondente: Gabriel Hideki Vatanabe Brunello
Departamento de Estatística, Universidade de Brasília – UnB, 70910-900 Brasília, DF, Brasil.
E-mails: ghvbrunello@gmail.com; nakano@unb.br

em intervalos unitários. Extensões desses modelos podem ser vistos em [2], que apresentam um modelo inflacionado de zeros aplicado em dados de confiabilidade.

O foco deste trabalho foi a inferência de dados de sobrevivência discretos dentro de um contexto bayesiano. Neste trabalho o modelo proposto por [14] foi aplicado. Esse modelo Weibull discreto é o correspondente discreto do modelo Weibull contínuo, tendo como caso especial a distribuição Geométrica (que é o correspondente discreto do modelo Exponencial) quando o parâmetro de forma é igual a 1. Procedimentos frequentistas para a estimação dos parâmetros da distribuição Weibull discreta podem ser vistos em [1].

A densidade *a posteriori* dos parâmetros do modelo foi obtida por meio de simulações via MCMC – *Markov Chain Monte Carlo* [3]. Foi proposto também um teste de significância genuinamente bayesiano (FBST – *Full Bayesian Significance Test*) para testar seu parâmetro de forma. O FBST é dito ser um teste genuinamente bayesiano, pois depende exclusivamente da distribuição *a posteriori* dos parâmetros [16] e também pelo mesmo ser caracterizado dentro da abordagem de Teoria da Decisão [11]. Mais especificamente, o interesse foi testar a hipótese do parâmetro de forma da distribuição Weibull discreta ser igual a 1. Casos em que essa hipótese não pode ser rejeitada indicam que um modelo mais simples (modelo Geométrico) pode ser utilizado.

A metodologia desenvolvida neste trabalho foi ilustrada em simulações e aplicada em um conjunto de dados sobre o tempo, em meses, até a morte de homens diagnosticados com AIDS [21]. Todas as simulações e obtenções das estimativas foram realizadas com a linguagem R [18].

2 METODOLOGIA

Os modelos de variáveis contínuas podem ser usados para gerar modelos discretos agrupando os tempos em intervalos unitários [15]. A variável discreta é obtida por $T = [X]$, em que $[X]$ representa a parte inteira de X (maior inteiro menor ou igual a X). No caso em que $X \sim Weibull(\beta, \lambda)$, com função de distribuição $F(x) = 1 - e^{(\frac{-x}{\lambda})^{\beta}}$, tem-se que a distribuição de probabilidades de T pode ser escrita como:

$$P[T = t] = P\,[t \leq X < t + 1] = q^{t^{\beta}} - q^{(t+1)^{\beta}},\ t = 0, 1, 2, \ldots \tag{2.1}$$

em que $q = e^{\left(\frac{-1}{\lambda}\right)^{\beta}}$.

A formulação apresentada em (2.1) resulta no modelo Weibull discreto proposto por [14] e se reduz a uma distribuição Geométrica com parâmetro $p = 1 - q$ quando $\beta = 1$.

As funções de sobrevivência e de risco do modelo Weibull discreto são dadas respectivamente por:

$$S(t) = P[T > t] = q^{(t+1)^{\beta}},\ t = 0, 1, 2, \ldots \tag{2.2}$$

e

$$h(t) = P[T = t|T \geq t] = \frac{q^{t^{\beta}} - q^{(t+1)^{\beta}}}{q^{t^{\beta}}},\ t = 0, 1, 2, \ldots \tag{2.3}$$

Note que a função de risco (2.3) é crescente se $\beta > 1$, decrescente se $\beta < 1$ e constante se $\beta = 1$.

Assumindo a presença de censuras à direita [10], a contribuição para a verossimilhança do tempo censurado em t é dada pela função de sobrevivência apresentada em (2.2). Assim, a função de verossimilhança apresenta a seguinte forma:

$$\begin{aligned} L(\mathcal{D}|q,\beta) &\propto \prod_{i=1}^{n}\Big[P(T=t)\Big]^{\delta_i}\Big[S(t)\Big]^{1-\delta_i} \\ &= \prod_{i=1}^{n}\left(q^{t_i^{\beta}}-q^{(t_i+1)^{\beta}}\right)^{\delta_i}\ \left[q^{(t_i+1)^{\beta}}\right]^{(1-\delta_i)} \\ &= q^{\sum_{i=1}^{n}\left[\left(1-\delta_i\right)\left(t_i+1\right)^{\beta}+\delta_i t_i^{\beta}\right]}e^{\sum_{i=1}^{n}\delta_i\log\left(1-q^{(t+1)^{\beta}+t_i^{\beta}}\right)}, \end{aligned}$$

em que $0 < q < 1$ e $\beta > 0$ são os parâmetros a serem estimados e $\mathcal{D} = (\mathbf{t}, \delta)$. Aqui, $\mathbf{t} = (t_1, \ldots, t_n)$ é o vetor dos valores observados, com seus respectivos indicadores de censuras dado por $\delta = (\delta_1, \ldots, \delta_n)$, em que $\delta_i = 0$ indica que a observação t_i é censurada, $i = 1, 2, \ldots, n$.

Como $\beta > 0$ e $0 < q < 1$, foram consideradas as seguintes distribuições *a priori* para os parâmetros: $q \sim Beta(a_1, b_1)$ e $\beta \sim Gama(a_2, b_2)$, em que a_1, a_2, b_1 e b_2 são hiper-parâmetros positivos conhecidos. Assim, a distribuição *a posteriori* dos parâmetros é proporcional a:

$$\begin{aligned} \pi(q,\beta|\mathcal{D}) \propto\ & q^{a_1+\sum_{i=1}^{n}\left[\left(1-\delta_i\right)\left(t_i+1\right)^{\beta}+\delta_i t_i^{\beta}\right]-1}(1-q)^{b_1-1}\beta^{a_2-1} \\ & \times e^{-b_2\beta+\sum_{i=1}^{n}\delta_i\log\left(1-q^{(t_i)^{\beta}+t_i^{\beta}}\right)}. \end{aligned} \tag{2.4}$$

As distribuições condicionais *a posteriori* são dadas por:

$$\pi(q|\beta,\mathcal{D}) \propto Beta\left(a_1+\sum_{i=1}^{n}\left[(1-\delta_i)(t_i+1)^{\beta}+\delta_i t_i^{\beta}\right], b_1\right)\times\Psi(q,\beta,\mathcal{D}) \tag{2.5}$$

e

$$\pi(\beta|q,\mathcal{D}) \propto Gama\,(a_2,b_2)\,\Psi(q,\beta,\mathcal{D}), \tag{2.6}$$

com

$$\Psi(q,\beta,\mathcal{D}) = e^{\sum_{i=1}^{n}\delta_i\log\left(1-q^{(t_i+1)^{\beta}+t_i^{\beta}}\right)}. \tag{2.7}$$

A distribuição *a posteriori* (2.4) não pode ser obtida analiticamente, mas amostras da mesma podem ser obtidas numericamente. Um algoritmo para gerar numericamente os valores de q e β é descrito no Apêndice A. Os valores da distribuição *a posteriori* (2.4) também podem ser facilmente gerados através do comando "MCMCmetrop1R" da biblioteca MCMCpack do software R [18]. Esse comando gera os valores através do algoritmo Metropolis adotando-se a técnica do passeio aleatório com a distribuição Normal multivariada como densidade proposta [12].

As estimativas dos parâmetros foram obtidas através da média da distribuição *a posteriori* e pelos intervalos HPD (*Highest Posterior Density*). Para uma dada credibilidade, o intervalo HPD

é o intervalo que apresenta a menor amplitude dentre todos os possíveis intervalos de credibilidade. Os valores preditivos da função de sobrevivência do modelo Weibull discreto são obtidos por:

$$\begin{aligned}\widehat{S(t)} &= P[T > t|\mathcal{D}] = \int_0^\infty \int_0^1 P[T > t, q, \beta|\mathcal{D}]dqd\beta \\ &= \int_0^\infty \int_0^1 q^{(t+1)^\beta}\pi(q, \beta|\mathcal{D})dqd\beta \\ &= E_{q,\beta|\mathcal{D}}\left[q^{(t+1)^\beta}\right],\ t = 0, 1, 2, \ldots\end{aligned} \tag{2.8}$$

A esperança em (2.8) é de difícil solução analítica, porém, seu valor pode ser facilmente obtido numericamente a partir dos valores gerados da distribuição *a posteriori* (2.4).

O Teste de Significância Genuinamente Bayesiano (FBST) é baseado no valor-e (evidência da hipótese nula, H), descrito a seguir. Seja $\Theta = (q, \beta)$ o vetor de parâmetros e $\pi(\Theta|\mathcal{D})$ a densidade *a posteriori* dada por (2.4). O interesse é verificar se a distribuição pode ser reduzida para a Geométrica, isto é, $H : \beta = 1$. O conjunto tangente à hipótese nula é:

$$T_H = \{\Theta = (q, \beta) \in (0, 1) \times \Re_+ : \pi(\Theta|\mathcal{D}) > \pi(\Theta^*_H|\mathcal{D})\},$$

em que $\Theta^*_H = (q^*, \beta = 1)$ é o vetor de valores dos parâmetros que maximizam (2.4) sob $H : \beta = 1$. É fácil mostrar que o valor de q^* é dado por:

$$q^* = \frac{a_1 + \sum_{i=1}^n t_i + n - \sum_{i=1}^n \delta_i - 1}{a_1 + b_1 + \sum_{i=1}^n t_i + n - 2}.$$

Assim, a medida de evidência proposta (o valor-e) é definida por:

$$\text{valor-}e(H) = 1 - P(\Theta \in T_H|\mathcal{D}). \tag{2.9}$$

O valor-e é a área (probabilidade) da densidade *a posteriori* no conjunto do espaço paramétrico que consiste nos pontos com densidade *a posteriori* menor do que o ponto máximo da densidade sob H. Note que (2.9) é de difícil solução analítica. No entanto, o valor-e pode ser facilmente obtido através da geração de valores, via métodos MCMC, da distribuição *a posteriori* (2.4). Mais detalhes sobre o FBST podem ser encontrados em [16].

3 SIMULAÇÕES

O modelo Weibull discreto foi ilustrado em uma aplicação numérica, considerando dados simulados no software R [18]. Foram gerados tempos de sobrevivência da distribuição Weibull discreta pelo método da transformação inversa [20] e considerando o mecanismo de censura aleatório. Foram geradas amostras de tamanho $n = 25$, 50, 100 e 200 com níveis de censura de 0%, 10%, 20% e 30%. A censura foi incorporada nas amostras independentemente do tempo de sobrevivência através de uma variável indicadora de censura gerada por uma distribuição Bernoulli,

cujo parâmetro foi fixado de acordo com os percentuais de censura descritos acima. Amostras dos tempos de sobrevivência foram geradas considerando os parâmetros $q = 0{,}9$ e $\beta = 1{,}3$ da distribuição Weibull discreta e estimativas bayesianas para os parâmetros foram consideradas. Para tanto foi adotada uma distribuição *a priori* não informativa $Beta(1, 1)$ para o parâmetro q e distribuições *a priori* Gama para o parâmetro β. Foram considerados diversos valores para os hiper-parâmetros da distribuição Gama para verificar a sensibilidade da escolha da distribuição *a priori* na estimação e no teste de hipótese do parâmetro β. Toda inferência dos parâmetros foi realizada via MCMC (Apêndices A e B). A convergência das cadeias foi verificada a partir da biblioteca CODA [17] do software R. O critério adotado para o diagnóstico de convergência foi o de Gelman e Rubin [5] e a dependência entre os valores gerados foi verificada através dos gráficos de autocorrelação. Os resultados das estimativas dos parâmetros e do valor-*e* do teste $H : \beta = 1$ são apresentados nas Tabelas 1 e 2. Todos os resultados apresentados foram baseados em uma cadeia de tamanho 3.000.000, considerando um *burn-in* de 10.000 e saltos de 3 passos, garantindo a convergência da cadeia (medida de Gelman e Rubin próxima de 1) e baixa correlação entre os valores gerados.

Veja na Tabela 1 que as estimativas dos parâmetros nem sempre são próximas dos verdadeiros valores fixados nas simulações. Isso ocorreu porque na abordagem bayesiana é gerado uma única amostra para cada cenário considerado, podendo a mesma sofrer variações aleatórias. No entanto, apesar de não ser possível controlar esse erro aleatório das amostras (a menos que considere uma amostra muito grande ou uma abordagem frequentista, em que cada cenário é gerado um número muito grande de vezes), é possivel notar que a evidência da hipótese $H : \beta = 1$ (valor-*e*) cai a medida que a estimativa de β se distancia de 1 e essa queda é mais acentuada quanto maior o tamanho da amostra e menor o percentual de censura. Note, por exemplo, que para $n = 50$ e 10% de censura tem-se a estimativa de β igual a 1,272, resultando em valor-*e* $= 0{,}250$. Ou seja, mesmo uma diferença de 0,272 da hipótese não foi suficiente para rejeitar o modelo Geométrico. Entretanto, essa mesma diferença já rejeitaria o modelo Geométrico no caso de uma amostra com $n = 100$ e 10% de censura.

A Tabela 2 apresenta um estudo de sensibilidade da escolha dos hiper-parâmetros da distribuição *a priori* de β. Pode-se notar que, para $n = 100$ e 10% de censura, a inferência é robusta quanto a escolha da distribuição *a priori*. Mesmo adotando uma distribuição *a priori* informativa e divergente da amostra observada, $a_2 = 1$ e $b_2 = 3$ (que resulta em um valor médio de 0,333 e variância 0,111), a estimativa de β não variou muito quando comparada com uma distribuição *a priori* não-informativa.

Resultados similares também foram observados para outras combinações de parâmetros. A Tabela 3 e a Figura 1 apresentam os resultados para três combinações (cenários) de valores dos parâmetros para $n = 100$ e 10% de censura.

No Cenário 1 mostrado na Tabela 3, apesar da estimativa do parâmetro β ser igual a $0,893$, a variabilidade foi grande e portanto esse valor não foi significativamente diferente de 1 (valor-*e* $= 0{,}215$). Neste caso, mesmo o modelo Geométrico já apresenta um bom ajuste da função de sobrevivência (veja Fig. 1(a)). No Cenário 2 a estimativa de β apresentou uma baixa evidência

Tabela 1: Inferência Bayesiana dos parâmetros do modelo Weibull discreto para dados simulados com diferentes tamanhos amostrais e percentuais de censura.

n	Censura	Parâmetro	Estimativa (média *a posteriori*)	Intervalo HPD 95%		valor-e $\boldsymbol{H : \beta = 1}$
25	0%	q	0,789	0,647	0,912	
		β	0,920	0,621	1,239	0,844
	10%	q	0,894	0,796	0,974	
		β	1,164	0,775	1,555	0,761
	20%	q	0,970	0,920	0,999	
		β	1,795	1,101	2,477	0,032
	30%	q	0,907	0,812	0,984	
		β	1,065	0,647	1,489	0,982
50	0%	q	0,963	0,930	0,990	
		β	1,857	1,459	2,274	0,000
	10%	q	0,906	0,840	0,964	
		β	1,272	0,950	1,593	0,250
	20%	q	0,922	0,867	0,969	
		β	1,308	1,023	1,601	0,110
	30%	q	0,920	0,859	0,971	
		β	1,196	0,870	1,524	0,533
100	0%	q	0,891	0,843	0,935	
		β	1,317	1,116	1,530	0,008
	10%	q	0,896	0,848	0,941	
		β	1,271	1,052	1,491	0,029
	20%	q	0,911	0,867	0,952	
		β	1,322	1,088	1,559	0,021
	30%	q	0,932	0,895	0,964	
		β	1,397	1,160	1,644	0,003
200	0%	q	0,912	0,882	0,939	
		β	1,327	1,183	1,479	0,000
	10%	q	0,945	0,924	0,965	
		β	1,528	1,350	1,704	0,000
	20%	q	0,913	0,883	0,943	
		β	1,222	1,065	1,381	0,018
	30%	q	0,952	0,931	0,971	
		β	1,580	1,382	1,787	0,000

Dados simulados de (2.1) com $q = 0{,}9$ e $\beta = 1{,}3$.

As estimativas foram baseadas em uma cadeia de tamanho 3.000.000, considerando um *burn-in* de 10.000 e saltos de 3 passos.

Tabela 2: Influência da escolha da distribuição *a priori* de β na inferência dos parâmetros do modelo Weibull discreto.

a_2	b_2	Parâmetros	Estimativa (média *a posteriori*)	Intervalo HPD 95%		valor-e $\boldsymbol{H:\beta=1}$
		q	0,896	0,848	0,941	
0,001	0,001	β	1,271	1,052	1,491	0,029
		q	0,895	0,845	0,939	
0,01	0,01	β	1,270	1,058	1,497	0,030
		q	0,895	0,846	0,939	
0,1	0,1	β	1,269	1,055	1,493	0,030
		q	0,895	0,846	0,939	
1	1	β	1,268	1,047	1,484	0,029
		q	0,893	0,843	0,937	
1	2	β	1,256	1,048	1,476	0,037
		q	0,890	0,841	0,935	
1	3	β	1,243	1,037	1,458	0,047
		q	0,885	0,833	0,932	
1	5	β	1,219	1,013	1,437	0,087
		q	0,897	0,848	0,940	
2	1	β	1,277	1,058	1,493	0,024
		q	0,899	0,851	0,941	
3	1	β	1,287	1,073	1,508	0,018
		q	0,903	0,858	0,946	
5	1	β	1,308	1,095	1,535	0,010

Dados simulados de (2.1) com $q = 0{,}9$; $\beta = 1{,}3$; $n = 100$ e 10% de censura.
As estimativas foram baseadas em uma cadeia de tamanho 3.000.000, considerando um *burn-in* de 10.000 e saltos de 3 passos.

de ser igual a 1, (valor-$e = 0{,}029$), indicando que o modelo Geométrico pode não ser adequado para o ajuste dos dados (Fig. 1(b)). O Cenário 3 apresentou uma estimativa de β que rejeita fortemente o modelo Geométrico (valor-e <0,001), mostrando que o modelo Geométrico não é um bom modelo para o ajuste desse conjunto de dados, sendo fundamental a utilização do modelo Weibull discreto (Fig. 1(c)).

4 ILUSTRAÇÃO NUMÉRICA

O ajuste e teste de significância do parâmetro de forma do modelo Weibull discreto foram ilustrados através de um conjunto de dados sobre o tempo até a morte de homens diagnosticados com AIDS (Síndrome de Imunodeficiência Adquirida). Os dados foram baseados em uma amostra de 174 homens que viveram em uma região altamente afetada da cidade de São Francisco no estado

Tabela 3: Inferência Bayesiana dos parâmetros do modelo Weibull discreto para amostras de tamanho $n = 100$ e 10% de censura.

Parâmetro	Valor fixado para gerar os dados	Estimativas (média *a posteriori*)	Intervalo HPD 95%	valor-e $\boldsymbol{H : \beta = 1}$
Cenário 1				
q	0,8	0,818	(0,752;0,882)	–
$\boldsymbol{\beta}$	0,9	0,893	(0,801;1,016)	0,215
Cenário 2				
q	0,9	0,896	(0,848;0,941)	–
$\boldsymbol{\beta}$	1,3	1,271	(1,052;1,491)	0,029
Cenário 3				
q	0,95	0,957	(0,930;0,980)	–
$\boldsymbol{\beta}$	2	2,039	(1,691;2,390)	<0,001

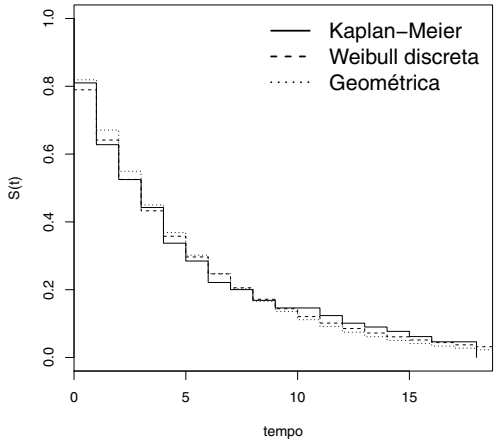
As estimativas foram baseadas em uma cadeia de tamanho 3.000.000, considerando um *burn-in* de 10.000 e saltos de 3 passos.

da Califórnia [21]. Neste exemplo, foi considerado um limite do tempo de estudo de 5 anos. A variável T representa o número de meses desde o diagnóstico da AIDS até a morte do indivíduo. Neste caso, $t = 0$ indica que o indivíduo morreu antes de completar 1 mês de diagnóstico.

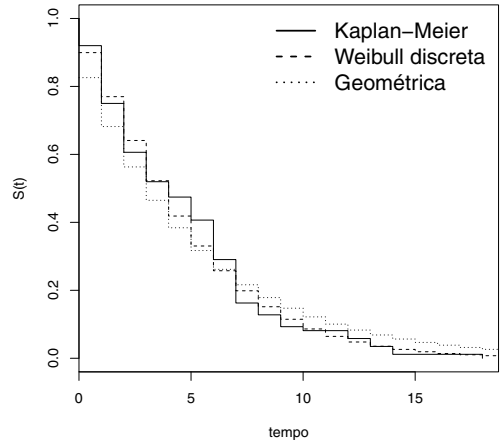
A Figura 2 apresenta as estimativas da função de sobrevivência para os dados sobre o tempo de sobrevivência de homens diagnosticados com AIDS. Considerando o erro máximo cometido na estimação [15], $\varepsilon = \max |\hat{S}(t) - \hat{S}_{KM}(t)|$, tem-se para o modelo Weibull discreto $\varepsilon = 0{,}083$ e para o modelo Geométrico $\varepsilon = 0{,}109$. Esse resultado mostra que as estimativas do modelo Geométrico são semelhantes às estimativas do modelo Weibull discreto. Realizando o FBST de $H : \beta = 1$, tem-se que valor-$e = 0{,}131$, não rejeitando a hipótese do modelo Geométrico ser adequado para a modelagem desses dados (Tabela 4).

5 CONCLUSÕES

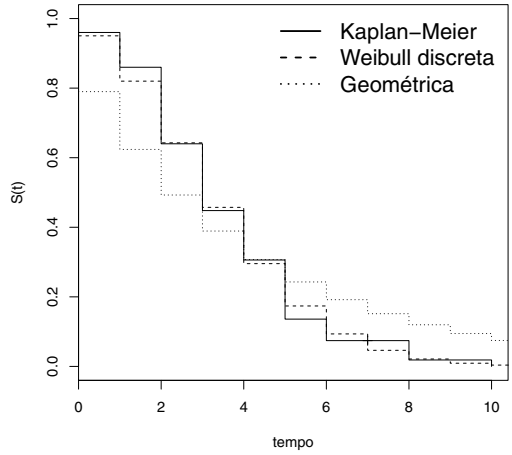
A distribuição Weibull discreta é um modelo bastante flexível para modelar tempos de sobrevivência discretos quando os mesmos não apresentam um risco constante. Apesar da distribuição *a posteriori* conjunta dos parâmetros não ser conhecida, amostras da mesma podem ser facilmente obtidas através dos métodos MCMC. O FBST mostrou-se uma forma simples para testar o parâmetro de forma do modelo Weibull discreto, sendo eficaz para decidir quando um modelo mais simples pode ser utilizado para ajustar os dados. Com base nos resultados apresentados, pôde-se concluir que a metodologia se mostrou robusta quanto a escolha da distribuição *a priori* a ser adotada, sendo eficaz no ajuste e seleção do modelo para representar os dados sobre o tempo até a morte de homens diagnosticados com AIDS.



(a) $\hat{q} = 0,818; \hat{\beta} = 0,893$



(b) $\hat{q} = 0,896; \hat{\beta} = 1,271$



(c) $\hat{q} = 0,957; \hat{\beta} = 2,039$

Figura 1: Gráficos das funções de sobrevivência estimadas por Kaplan-Meier [9], pela distribuição Weibull discreta e pela distribuição Geométrica.

A facilidade computacional da implementação dos métodos MCMC através do pacote MCMCPack do R [12] pode encorajar a utilização de uma abordagem bayesiana na modelagem. Ainda, poder contar com uma medida de evidência para testes de significância dos parâmetros, que mantém as mais desejáveis propriedades do uso prático dos valores-p (sem ser sua simples versão bayesiana) e que seja conceitualmente simples, teoricamente coerente e facilmente implementável pode encurtar ainda mais a distância dos métodos bayesianos na modelagem de dados discretos de sobrevivência e também nas mais diversas áreas.

Tabela 4: Inferência Bayesiana dos parâmetros dos modelos Weibull discreto e Geométrico para os dados sobre o tempo de sobrevivência de homens diagnosticados com AIDS.

Parâmetro	Estimativas (média *a posteriori*)	Intervalo HPD 95%	valor-e $\boldsymbol{H : \beta = 1}$
Weibull discreto			
$\mathbf{q}$	0,970	(0,954;0,984)	–
$\boldsymbol{\beta}$	1,125	(0,990;1,282)	0,131
Geométrico			
$\mathbf{q}$	0,955	(0,948;0,962)	–

Fonte: [21], pág. 248.

Nota: As estimativas foram baseadas em uma cadeia de tamanho 3.000.000, considerando um *burn-in* de 10.000 e saltos de 3 passos.

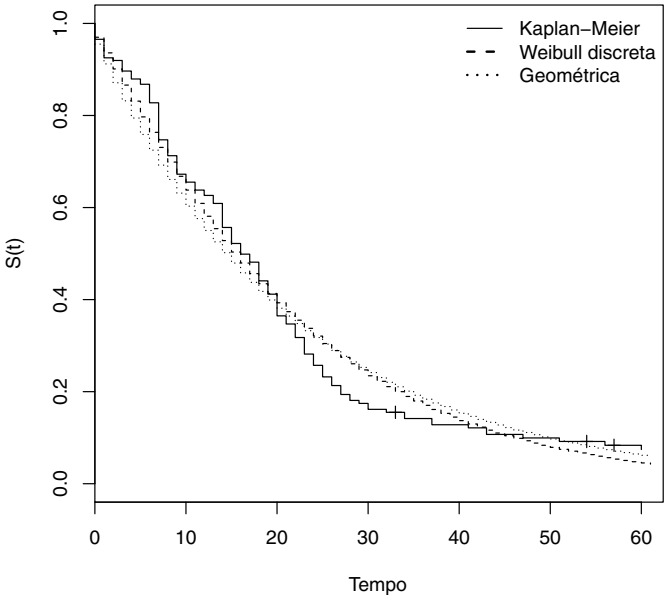


Figura 2: Estimativas da função de sobrevivência para os dados sobre o tempo de sobrevivência de homens diagnosticados com AIDS.

O modelo Weibull discreto se mostrou uma boa opção nos casos em que a função de risco é monótona. Outros modelos para ajustar dados discretos podem ser considerados em futuros trabalhos como, por exemplo, a distribuição Binomial Negativa, que apresenta função de risco não monótona, modelos com excessos de zeros, como apresentados em [2] e modelos discretos com fração de cura.

AGRADECIMENTOS

Os autores agradecem ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq) e ao Decanato de Pesquisa e Pós-Graduação da Universidade de Brasília (DPP-UnB) pelo auxílio concedido para a realização desse trabalho.

ABSTRACT. This work presents a bayesian inference on discrete Weibull distribution with censored data. A Full Bayesian Significance Test (FBST) was proposed to test the shape parameter of model. Samples from the posterior distributions of parameters were numerically obtained by Markov Chain Monte Carlo (MCMC) simulations. This methodology was illustrated using simulated data and by application on a real database of survival times of men diagnosed with AIDS. All simulations and estimates were performed in R language.

Keywords: survival analysis, bayesian inference, hypothesis tests, MCMC.

REFERÊNCIAS

[1] M.S. Ali Khan, A. Khalique & A.M. Abouammoh. On estimating parameters in a discrete Weibull distribution. *IEEE Transactions on Reliability*, **38** (1989), 348–350.

[2] C.G. Carrasco, M.H. Tutia & E.Y. Nakano. Intervalos de confiança para os parâmetros do modelo geométrico com inflação de zeros. *TEMA – Tend. Mat. Apl. Comput.*, **13**(3) (2012), 247–255.

[3] D. Gamerman & H.F. Lopes. "Markov chain monte carlo: Stochastic simulation for bayesian inference", Chapman & Hall, London, v. 1, (2006).

[4] A.E. Gelfand & A.F.M. Smith. Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85** (1990), 398–409.

[5] A. Gelman & D.R. Rubin. Inference from iterative simulation using multiple sequences. *Statistical Science*, **7** (1992), 457–511.

[6] S. Geman & D. Geman. Stochastic relaxation, Gibbs distribution and the Bayesian restoration of images. *IEEE transactions on Pattern Analysis and Machine Intelligence*, **6** (1984), 721–741.

[7] J.E. Gentle. "Random Number Generation and Monte Carlo Methods", Springer-Verlag, New York, 2 ed, (2004).

[8] W.K. Hastings. Monte carlo sampling methods using markov chains and their applications. *Biometrika*, **57** (1970), 97–109.

[9] E.L. Kaplan & P. Meier. Nonparametric estimation from incomplete observations. *J. Am. Stat. Assoc.*, **53** (1958), 457–481.

[10] F.J. Lawless. "Statistical Models and Methods for Lifetime Data", Wiley, New York (1982).

[11] M.R. Madruga, L.G. Esteves & S. Wechsler. On the bayesianity of Pereira-Stern tests. *Test*, **10** (2001), 291–299.

[12] A.D. Martin, K.M. Quinn & J.H. Park. MCMCpack: Markov Chain Monte Carlo in R. *Journal of Statistical Software*, **42**(9) (2011), 1–21.

[13] N. Metropolis, A.W. Rosenbluth, M.N. Rosenbluth, A.H. Teller & E. Teller. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, **21** (1953), 1087–1092.

[14] T. Nakagawa & S. Osaki. The discrete weibull distribution. *IEEE Transactions on Reliability*, **R-24**(5) (1975), 300–301.

[15] 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**(1) (2006), 91–100.

[16] C.A.B. Pereira & J. Stern. Evidence and credibility: full bayesian significance test of precise hypothesis. *Entropy*, **1** (1999), 99–110.

[17] M. Plummer, N. Best, K. Cowles & K. Vines. CODA: Convergence diagnosis and output analysis for MCMC. *R News*, **6** (2006), 7–11.

[18] R Core Team. R: A language and environment for statistical computing, R Foundation for Statistical Computing, Vienna, Austria, (2013).

[19] H. Rinne. "The Weibull distribution: a handbook", Taylor & Francis, (2008).

[20] S.M. Ross. "Simulation", Academic Press, 5 ed, (2012).

[21] E. Selvin. "Survival analysis for epidemiologic and medical research: A practical guide", Cambridge University Press, New York (2008).

[22] W. Weibull. A statistical distribution function of wide applicability. *J. Appl. Mech. Trans.*, **18**(3) (1951), 293–297.

APÊNDICE A – ALGORITMO MCMC PARA GERAÇÃO DE AMOSTRAS *A POSTERIORI* CONJUNTA

É apresentado a seguir um algoritmo MCMC para geração de amostras da distribuição *a posteriori* conjunta (2.4) através das condicionais (2.5) e (2.6). O algoritmo é composto por passos de Gibbs [6, 4] com Metropolis-Hastings [8, 13]. Adotou-se aqui a técnica de cadeias simétricas e considerando a distribuição Beta como densidade proposta para o parâmetro q e a distribuição Gama para o parâmetro β.

(1) iniciar arbitrariamente em um ponto inicial qualquer $(q^{[0]}, \beta^{[0]})$ e também o contador $k = 1$;

(2) obter um novo valor de $(q^{[k]}, \beta^{[k]})$ a partir de $(q^{[k-1]}, \beta^{[k-1]})$ através de sucessivas gerações de valores:

(a) gerar $q' \sim Beta\left(a_1 + \sum_{i=1}^{n}\left[(1-\delta_i)(t_i+1)^{\beta^{[k-1]}} + \delta_i t_i^{\beta^{[k-1]}}\right], b_1\right)$

(a$_1$) gerar $u \sim Uniforme(0, 1)$;

(a$_2$) se

$$u \leq \frac{\Psi(q', \beta^{[k-1]}, \mathcal{D})}{\Psi(q^{[k-1]}, \beta^{[k-1]}, \mathcal{D})},$$

fazer $q^{[k]} = q'$, caso contrário, fazer $q^{[k]} = q^{[k-1]}$, com $\Psi(q, \beta, \mathcal{D})$ dado por (2.7);

(b) gerar $\beta' \sim Gama\left(\frac{(\beta^{[k-1]})^2}{\sigma_\beta^2}, \frac{\beta^{[k-1]}}{\sigma_\beta^2}\right)$

($\mathbf{b}_1$) gerar $v \sim Uniforme(0, 1)$;

($\mathbf{b}_2$) se

$$v \leq \frac{\Psi(q^{[k]}, \beta', \mathcal{D})}{\Psi(q^{[k]}, \beta^{[k-1]}, \mathcal{D})} \left(\frac{\beta'}{\beta^{[k-1]}}\right)^{a_2 - \frac{(\beta^{[k-1]})^2}{\sigma_\beta^2}} \left(e^{\left(\frac{\beta^{[k-1]}}{\sigma_\beta^2} - b_2\right)(\beta' - \beta^{[k-1]})}\right),$$

fazer $\beta^{[k]} = \beta'$, caso contrário, fazer $\beta^{[k]} = \beta^{[k-1]}$, com $\Psi(q, \beta, \mathcal{D})$ dado por (2.7);

(3) atualizar o contador para $k = k + 1$ e retornar ao passo **(2)** até obter a cadeia desejada.

Nota: As escolhas dos parâmetros da densidade proposta de β foram definidas de forma que a média da distribuição seja o valor do parâmetro no passo anterior $\left(\beta^{[k-1]}\right)$ e a variância seja σ_β^2. Para que o algoritmo apresente um certo grau de eficiência, é necessário que se faça uma boa escolha de σ_β^2. Um valor grande de σ_β^2 pode implicar uma baixa taxa de aceitação do algoritmo e um valor muito baixo aumenta a autocorrelação dos valores da cadeia, diminuindo a sua evolução. Na prática a escolha de σ_β^2 não é imediata e é preciso fazer um estudo preliminar para a sua escolha de forma tornar o método eficiente.

APÊNDICE B – SCRIPT PARA OBTENÇÃO DE ESTIMATIVAS DOS PARÂMETROS

```
##Entrada dos dados
require(MCMCpack)
x<-c(0,0,0,0,0,0,1,1,1,1,1,1,1,2,3,3,3,3,4,4,4,5,5,6,6,6,6,6,6,6,
7,7,7,7,7,7,7,7,7,7,7,7,7,8,8,8,8,8,8,9,9,9,9,9,9,10,10,10,
11,11,11,11,12,12,13,13,13,14,14,14,14,14,14,14,14,15,15,15,
15,15,15,16,16,16,16,17,17,17,18,18,18,18,18,18,18,19,19,19,19,
19,19,20,20,20,20,20,20,20,20,21,21,21,22,22,22,22,22,22,23,23,
23,23,23,23,24,24,24,24,24,25,25,25,25,26,26,26,26,27,27,27,
28,28,29,30,30,32,33,34,34,37,37,41,41,43,43,43,47,51,54,56,57,
60,60,60,60,60,60,60,60,60)
d<-c(1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,
1,1,1,1,1,1,0,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,1,0,1,1,0,1,1,1,1,1,1,
1,0,1,1,1,1,1,1,1,1,1,0,1,1,1,1,1,1,1,1,1,1,1,0,1,1,1,1,1,1,0,1,1,0,1,1,
0,1,0,1,0,0,0,0,0,0,0,0)
prioris<-c(1, 1, 10^-3, 10^-3)
#Log posteriori
posteriori<-function(p,x,a,b,a2,b2,d){
if ( (p[1]>0) && (p[1]<1) && (p[2]>0) )
return ( sum( ( d * ( x^p[2] *log(p[1]) +
```

```
log(1-p[1]^((x+1)^p[2] - x^p[2])) ) + ( (x+1)^(p[2]) *
(1-d) *log(p[1]) ))) + ( (a-1) * log(p[1]) ) + ( (b-1) *
log( (1-p[1]) ) ) + ( (a2-1) * log(p[2]) - (b2*p[2]) ) )
else return (-Inf)
}
amost <- MCMCmetrop1R(posteriori, theta.init=c(.5,1),mcmc=3000000,
burnin=10000,x=x,a=prioris[1],b=prioris[2],a2=(prioris[3]),
b2=(prioris[4]),d=d,thin=3)
FBST<-function(amost,x,d,priorif){
a1<-1;b1<-1;cont<-0
maxqho<-(length(x)+sum(x)-sum(d)+b1-1)/(length(x)+sum(x)+a1+b1-2)
k<-posteriori(c(maxqho,1),x,a1,b1,priorif[1],priorif[2],d)
comparar<- function(w,l){if(posteriori(c(w,l),x,a1,b1,priorif[1],
priorif[2],d)<k) {cont<<-cont+1}}
mapply(comparar,amost[,1],amost[,2])
print(cont/nrow(amost))
}
FBST(amost,x,d,c(prioris[3],prioris[4]))
```