UNIVERSIDADE ESTADUAL PAULISTA
"JÚLIO DE MESQUITA FILHO"
Câmpus de Presidente Prudente

ALLAN TAILON LOPES DELFINO

ANÁLISE BAYESIANA DA DISTRIBUIÇAO GAMA EXPONENCIADA SOB CENSURA A ESQUERDA

PRESIDENTE PRUDENTE
2023

ALLAN TAILON LOPES DELFINO

ANÁLISE BAYESIANA DA DISTRIBUIÇAO GAMA EXPONENCIADA SOB CENSURA A ESQUERDA

Trabalho de Conclusão de Curso apresentado ao Curso de Graduação em Estatística da FCT/Unesp para aproveitamento na disciplina de TCC. Orientador(a): Prof. Dr. Fernando Antônio Moala.

PRESIDENTE PRUDENTE
2023

D349a

Delfino, Allan Tailon Lopes
Análise bayesiana da distribuição gama exponenciada sob censura à esquerda / Allan Tailon Lopes Delfino. -- Presidente Prudente, 2023
26 p.

Trabalho de conclusão de curso (Bacharelado - Estatística) - Universidade Estadual Paulista (Unesp), Faculdade de Ciências e Tecnologia, Presidente Prudente
Orientador: Fernando Antônio Moala

1. Analise bayesiana. 2. Distribuição gama exponenciada. 3. Análise de sobrevivência. I. Título.

Sistema de geração automática de fichas catalográficas da Unesp. Biblioteca da Faculdade de Ciências e Tecnologia, Presidente Prudente. Dados fornecidos pelo autor(a).

Essa ficha não pode ser modificada.

TERMO DE APROVAÇÃO

Allan Tailon Lopes Delfino

ANÁLISE BAYESIANA DA DISTRIBUIÇÃO GAMA EXPONENCIADA SOB CENSURA À ESQUERDA

Relatório de Final de Trabalho de Conclusão de Curso aprovado como requisito para obtenção de créditos na disciplina Trabalho de Conclusão do curso de graduação em Estatística da Faculdade de Ciências e Tecnologia da Unesp, pela seguinte banca examinadora:

Orientador: _____

Prof. Dr.Fernando Antonio Moala
Departamento de Estatística

_____

Prof. Dr. Sergio Minoru Oikawa
Departamento de Estatística

_____

Prof. Dr. Mário Hissamitsu Tarumoto
Departamento de Estatística

Presidente Prudente, 06 de fevereiro de 2023.

RESUMO

A distribuição gama exponenciada é muito utilizada como modelo para análise de sobrevivência, sendo seu principal atrativo sua função de risco que assume diversas formas, tais como crescente e curva de banheira. Com aplicabilidade em diversas áreas como das ciências biológicas. No presente trabalho estudou-se o uso de censura à esquerda através de técnicas estatísticas, inferências sobre o parâmetro da distribuição, foram feitas com enfoque bayesiano e clássico em dados completos e com censura simulada aos dados. Foi utilizado uma priori gama para utilizar de sua família conjugada, desse modo foi possível chegar ao estimador do parâmetro analiticamente e aplicar a simulação, chegando em resultado semelhantes para os dois tipos de inferências usadas.

Palavras-chaves: Censura a esquerda, Inferência Bayesiana, Verossimilhança, Simulação, priori conjugada.

ABSTRACT

The exponential gamma distribution is widely used as a model for survival analysis, having as its main attraction its risk function, which assumes different forms, such as the rising curve and the bathtub curve. With applicability in several areas such as biological sciences. In the present work, the use of left censoring was studied using statistical techniques, inferences about the distribution parameter were made with Bayesian and classical focus on complete data and with simulated data censoring. An a priori gamma was used to use its conjugate family, so it was possible to arrive analytically at the estimator of the parameter and apply the simulation, reaching similar results for the two types of inferences used.

Keywords: Left Censorship, Bayesian Inference, Likelihood, Simulation, Conjugate Priori

SUMÁRIO

1 INTRODUÇÃO

A análise de sobrevivência é uma das áreas da estatística que mais vem crescendo nas últimas décadas. A razão deste crescimento e desenvolvimento é o aprimoramento de técnicas estatísticas combinado com computadores cada vez mais velozes e potentes. Além disso, outro indicador do crescimento é o número de citações de artigos na literatura sobre o estimador de Kaplan-Meier (Kaplan e Meier,1958) junto ao modelo de Cox (Cox,1972).

O foco desse projeto é o estudo de dados de tempo de vida, muito utilizados em áreas como engenharia para teste de tempo de vida de componentes ou itens, na medicina e biologia para tempo de vida de indivíduos e até mesmo para astronomia em estudos de ondas gravitacionais e tempo de vida de estrelas.

Os dados em Análise de Sobrevivência referem-se ao tempo até a ocorrência de um evento de interesse. O evento de estudo é denominado falha e os dados coletados cujo evento não foi observado são denominados censura, que são observações parciais do evento. Todos os tipos de censuras serão comentados posteriormente.

Conhecendo-se a distribuição de probabilidade que melhor se ajusta a esses tempos, é possível estimar a probabilidade de sobrevivência do objeto de estudo. Desse modo, com o passar dos anos foi se aprimorando os estudos de análise de sobrevivência necessitando cada vez mais de novos e sofisticados modelos probabilísticos para a modelagem de dados.

Com isso, além do uso de diferentes distribuições, também é possível utilizar outros métodos eficazes para trabalhar com dados censurados, como o estimador de Kaplan-Meier, que é utilizado para estimar a função de sobrevivência de forma não-paramétrica. Também com o uso de inferência bayesiana através de métodos computacionais como Monte Carlo via Cadeias de Markov(MCMC).

Os Métodos de MCMC são técnicas iterativas de simulação de cadeia de markov, no qual as amostras obtidas em vários passos visam a convergência para uma nova distribuição estacionária. Atualmente os algoritmos mais utilizados são Amostrador de Gibbs e Metropolis-Hastings, em que, de acordo com Dogarra e Sulivan

(2000), o algoritmo de Metropolis-Hastings está entre os dez algoritmos que mais influenciaram o desenvolvimento da ciência e engenharia no século XX.

Por fim, o objetivo geral deste trabalho é analisar e observar através de métodos de análise Bayesiana o ajuste da distribuição gama exponenciada para dados com censura a esquerda. A fim de observar e comparar a sua eficácia diante dos métodos comumente utilizados.

A escolha do uso da censura a esquerda se dá pelo fato de que o campo de pesquisa nesse tema não ser tão abordado, o que torna a disseminação do conhecimento menos difundido no meio acadêmico. MITRA e KUNDU (2008) contribuíram para essa área analisando a distribuição exponencial generalizada com presença de censura a esquerda. Contudo, o desenvolvimento desse tipo de censura se torna essencial, visto que, em áreas como astronomia, os dados coletados são comumente contidos de censura a esquerda.

2 ANÁLISE DE SOBREVIVÊNCIA

A análise de sobrevivência, também chamada de análise de sobrevida, é utilizada quando o tempo for o objeto de interesse, seja este interpretado como o tempo até a ocorrência de um evento podendo chegar no risco de ocorrência de um evento por unidade de tempo. Esse tipo de análise é baseado em dados que eventualmente possuem algum tipo de censura.

2.1 Censura

De acordo com Colosimo e Giolo (2006) a censura pode ser classificada em três tipos, sendo eles: censura à direta, censura à esquerda e censura intervalar.

Censura à direita, acontece quando o tempo até a ocorrência do evento de interesse está à direita do tempo observado como, por exemplo, em um estudo sobre o tempo entre o diagnostico até a cura de determinada doença por um tempo determinado, alguns indivíduos podem chegar ao fim do estudo sem vivenciar o evento de interesse. Dentre dessa censura, existe mais três tipos de classificação:

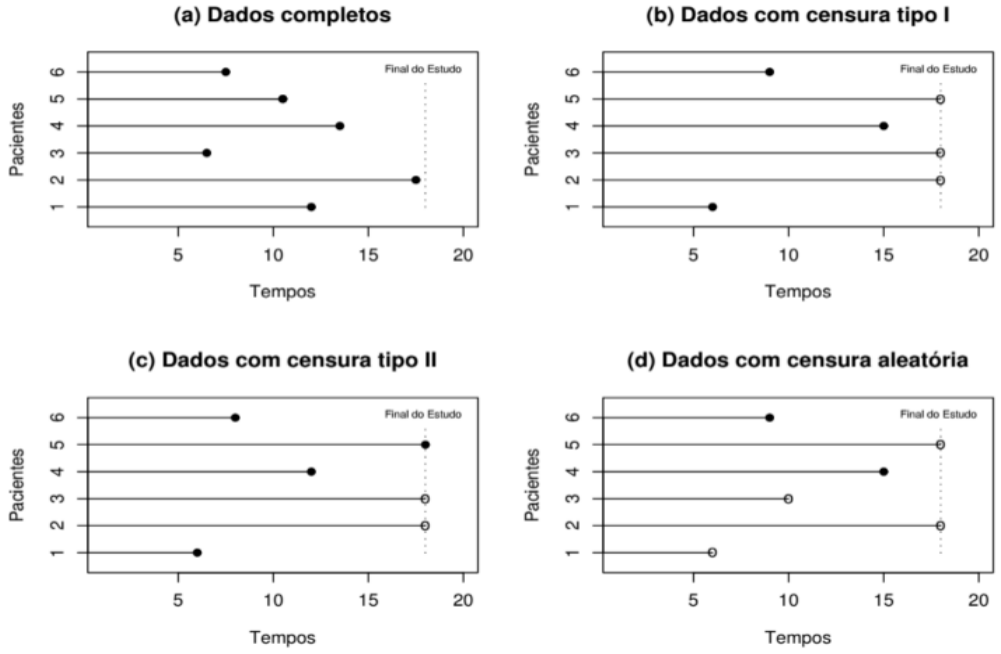
Censura tipo I - é aquela em que o estudo será terminado após um período pré-estabelecido de tempo.

Censura tipo II - é aquela em que o estudo será terminado após ter ocorrido o evento de interesse em um número ´ pré-estabelecido de indivíduos.

Censura aleatória – acontece quando um paciente é retirado no decorrer do estudo sem ter ocorrido a falha. Isto também ocorre, por exemplo, se o paciente morrer por uma razão diferente da estudada.

A figura a seguir mostra os três tipos de censura.

Figura 1 - tipos de censura à direita

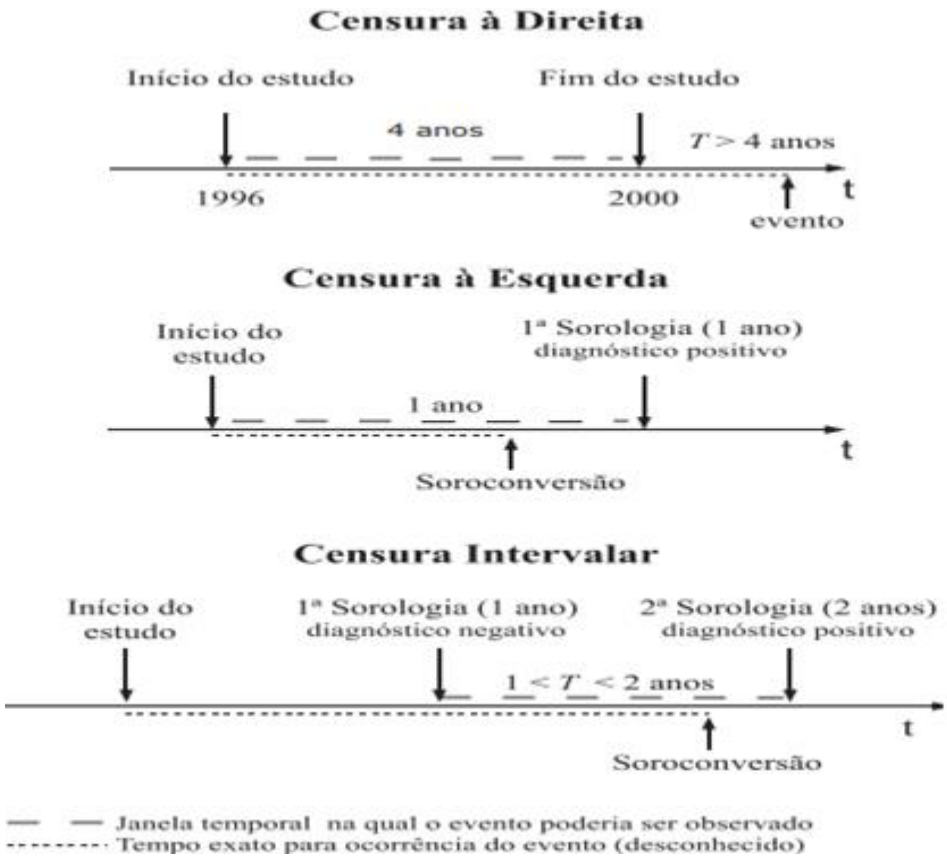


Fonte: Colosimo e Giolo (2006)

Censura à esquerda, ocorre quando o tempo registrado é maior que o tempo de falha. Isto é, o evento de interesse já aconteceu quando o indivíduo foi observado como, por exemplo, ao se realizar um estudo sobre o tempo de alfabetização de crianças em uma comunidade, é possível que na amostra seja selecionada crianças que já aprenderam a ler, não sendo possível determinar quando isso ocorreu;

Censura intervalar ocorre quando a falha ocorre em um intervalo de tempo, mas não é possível especificar o momento exato de sua ocorrência. Por exemplo, em alguns casos, o evento de interesse pode ser o tempo entre os primeiros sintomas e o diagnóstico de determinada doença. Entretanto, o diagnóstico é dado de acordo com o resultado de exame feito mensalmente, assim não se conhece o tempo exato da ocorrência da doença, apenas o intervalo entre os resultados de exame que indicaram a existência da doença.

Figura 2 - diferentes tipos de censura



Fonte: Adaptado de www.sobrevida.fiocruz.br

2.2 Tempo de falha

É possível definir o tempo de falha baseado em três elementos:

Tempo inicial: data de um diagnóstico ou início de um tratamento.

Escala de medida: é geralmente o tempo real, medido em horas, dias ou minutos, podendo ser também medidos em quilometragem, número de ciclo ou qualquer outra medida de carga.

Falha ou evento de interesse: comumente já é bem definido o evento de interesse como, por exemplo, a morte do paciente, a quebra de um equipamento ou a cura de uma doença.

Para descrever o tempo de falha é utilizado a função de sobrevivência que é definida como a probabilidade de o objeto de estudo não falhar até um certo tempo t. Essa função é dada pela equação:

$$S(t) = P\,(T \geq t) \qquad (2.1)$$

A partir da função de sobrevivência é possível chegar na função taxa de falha (função de risco) que descreve a probabilidade de ocorrer uma falha em um determinado intervalo. Assim a taxa de falha no intervalo $[t_1, t_2)$ é expressa por:

$$\frac{S(t_1)-S(t_2)}{(t_2-t_1)S(t_1)} \tag{2.2}$$

A função de risco $\lambda(t)$ é bastante útil para descrever a distribuição do tempo de vida de pacientes. De forma geral, trocando o intervalo como $[t, \Delta t)$, a expressa assuma a seguinte forma:

$$\lambda(t) = \frac{S(t)-S(t+\Delta t)}{\Delta t S(t)} \tag{2.3}$$

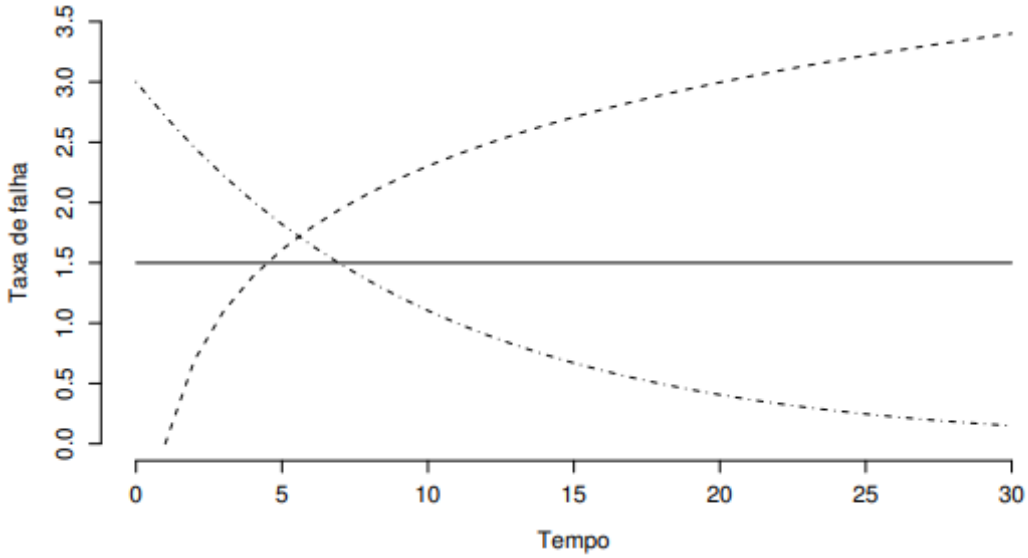
Desta forma, $\lambda(t)$ representa a taxa de falha instantânea no tempo t condicional a sobrevivência até o tempo f quando Δt assume valores pequenos. Logo, a função taxa de falha do T é definida como:

$$\lambda(t) = \frac{P(t \leq T < t+\Delta t | T \geq t)}{\Delta t} \tag{2.4}$$

Existem algumas classificações básicas para a função de risco, sendo três delas: (i) função de risco crescente (FRC), em que a incidência de risco cresce com o tempo; (ii) função de risco decrescente (FRD), em que a incidência de risco decresce com o tempo; e (iii) função de risco constante ou estacionária (FRE), em que a unidade está exposta a uma mesma quantidade de risco em qualquer momento do tempo (Fogliatto e Ribeiro,2011).

Devido ao aumento de aplicações nessas áreas, diferentes funções de risco são necessárias para modelar de forma mais consistente os dados obtidos, com isso, a criação de novas distribuições se torna essencial. Atualmente, a distribuição gama exponenciada, assim como Weibull e Log-Normal, é comumente utilizada para a análise de sobrevivência, devido a sua fácil modelagem.

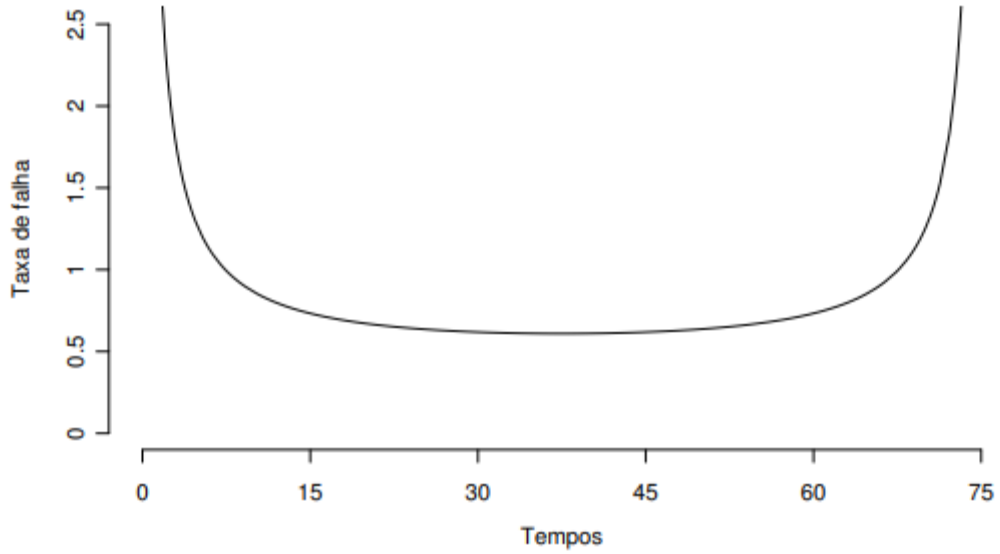
Figura 3 - Tipos de função de risco



Fonte: Colosimo e Giolo (2006)

Outra importante função de risco é denominada “curva da banheira”, muito utilizada para descrever o tempo de vida de seres humanos. No início possui uma taxa de falha decrescente, associado a mortalidade infantil, em sequência se mantem constante na fase intermediaria e termina crescente.

Figura 4 - Função de risco em formato de banheira



Fonte: Colosimo e Giolo (2006)

2.3 Técnicas não paramétricas

2.3.1 Estimador de Kaplan-Meier

O estimador de Kaplan-Meier é atualmente o mais utilizado em estudos científicos na área de análise de sobrevivência. O estimador não paramétrico é uma adaptação da função de sobrevivência empírica e foi proposto por Kaplan e Meier (1958) para estimar a função de sobrevivência. Quando não possui censura, ele é definido como:

$$\hat{S}(t) = \frac{no.\ de\ observações\ que\ não\ falharam\ ate\ o\ tempo\ t}{no.\ total\ de\ observações\ no\ estudo} \quad (2.5)$$

De acordo com Colosimo e Giolo (2006, apud Kaplan e Meier), a expressão geral do estimador de Kaplan-Meier pode então ser apresentada após estas considerações preliminares. Formalmente, considere:

- $t_1 < t_2 \cdots < t_k$, os k tempos distintos e ordenados de falha,
- d_j o número de falhas em t_j , $j = 1, \cdots, k$, e
- n_j o número de indivíduos sob risco em t_j, ou seja, os indivíduos que não falharam e não foram censurados até o instante imediatamente anterior a t_j. O estimador de Kaplan-Meier é, então, definido como:

$$\hat{S}(t) = \prod_{j:\, t_j<t} \left(\frac{n_j - d_j}{n_j}\right) = \prod_{j:\, t_j<t} \left(1 - \frac{d_j}{n_j}\right) \quad (2.6)$$

2.3.2 Gráfico do Tempo Total em Teste

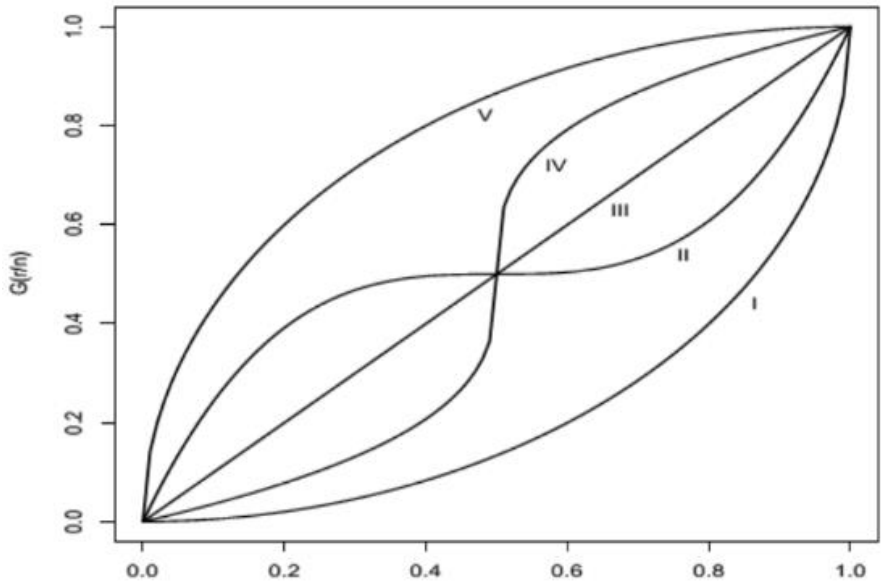
Diversos métodos tem sido proposto para identificar as formas da função de risco e podem ser encontradas em Glaser (1980). Dentre estes métodos destaca-se o método gráfico conhecido como TTT-plot (Gráfico do Tempo Total em Teste) proposto por Barlow e Campo (1975), que é utilizado em situações onde há informações qualitativas sobre a curva de risco.

A versão empírica proposto por Aaset (1985) do TTT-plot é obtida a partir da função dada por:

$$G\left(\frac{r}{n}\right) = \frac{\sum_{i=1}^{r} T_i + (n-r)T_r}{\sum_{i=1}^{n} T_i} \quad (2.7)$$

Em que $r = 1, \ldots, n$ e $T_i,\ i = 1, \ldots, n$ são estatísticas de ordem de amostra (AARSET,1985)

Figura 5 - Diferentes formas de identificar a função de risco



Fonte: www.lume.ufrgs.com.br

A partir de cada curva ilustrada na figura a cima é possível identificar as formas da função de risco $\lambda(t)$;

Curva I – (convexa) indica que $\lambda(t)$ é decrescente.

Curva II – (côncava e convexa) indica que $\lambda(t)$ é unimodal

Curva III – (reta diagonal) indica que $\lambda(t)$ é constante

Curva IV – (convexa e côncava) indica que $\lambda(t)$ é uma forma de banheira

Curva V – (côncava) indica que $\lambda(t)$ é crescente.

2.4 Estimação para dados sem censura

2.4.1 Método da máxima Verossimilhança

O princípio de Verossimilhança afirma que toda a informação contida em um determinado experimento está contida na Função de Verossimilhança. O método da Máxima Verossimilhança consiste em estimar os parâmetros de um modelo utilizando as estimativas que tornam máximo o valor da função de verossimilhança, Bolfarini e Sandoval (2010) define a função de verossimilhança como:

A função de verossimilhança de θ para amostra aleatória independente e identicamente distribuída é dada por:

$$L(\theta; x) = \prod_{i=1}^{n} f(x_i/\theta) \tag{2.8}$$

no qual o estimador de máxima verossimilhança de θ é o valor $\theta \in \Theta$ que maximiza a função de verossimilhança. Para verificar se a solução da equação (2.8) é um ponto de máximo, é necessário verificar se

$$l''(\theta; x) = \frac{\partial^2 \log L(\theta;x)}{\partial \theta^2} < 0 \tag{2.9}$$

2.5 Estimação para dados censurados

A verossimilhança é um ótimo estimador para o parâmetro de estudo, contudo é necessário se ter cuidado quando se trata de dados censurados. A independência do tempo de censura e o tempo de vida é a principal suposição para aplicação dessa técnica. Caso essas suposições não forem cumpridas, técnicas especificas devem ser utilizadas.

Uma observação do tempo exato da ocorrência do evento nos fornece a probabilidade de o evento ocorrer nesse exato tempo, que é aproximadamente a função densidade da variável X.

Para um conjunto de dados com censura a esquerda, tudo que é possível inferir sobre ele é que o evento já ocorreu, então para adicionar a verossimilhança é utilizada a função de distribuição acumulada avaliada no tempo de início de observação. Já para um dado com censura a direta, a única informação que temos é a ocorrência do evento sendo maior que o tempo limite de estudo, logo utiliza-se a função de sobrevivência avaliada neste tempo.

Dados com censura intervalar, nos fornece apenas o intervalo de tempo em que o evento ocorreu, sendo possível incorporar essa informação utilizando a probabilidade do tempo de ocorrência do evento que está nesse intervalo.

Para cada tipo de censura existe um método de incorporar a informação na verossimilhança. Esse método consiste em utilizar diferentes funções para a estimação, como mostra a seguir:

Tempo de vida exato - $f(t)$

Dados com censura a direita - $S(C_r)$

Dados com censura a esquerda - $1 - S(C_r)$

Dados com censura intervalar - $[S(L) - S(R)]$

Desse modo, é possível construir a verossimilhança como:

$$L = \prod_{i\in D} f_i(x_i) \prod_{i\in R} S_i(C_r) \prod_{i\in L}(1 - S_i(C_l)) \prod_{i\in I}[S_i(L_i) - S_i(R_i)] \quad (2.11)$$

Em que D é o conjunto dos tempos de falha, R é o conjunto de dados com censura a direita, L é o conjunto de dados com censura a esquerda, e I é o conjunto de observações com censura intervalar.

2.5.1 Censura a esquerda

Um tempo de vida associado a um indivíduo em estudo é considerado censurado a esquerda se ele é menor do que o tempo de censura C_l. Os dados observados de um indivíduo podem ser registrados como (T, δ), onde $T = (X, C_l)$ e indica se o dado é censurado $(\delta = 0)$ ou não$(\delta = 1)$. Se tivermos uma amostra aleatória de pares (T_i, δ_i) a função de verossimilhança é dada por

$$L = \prod_{i=1}^{n}[f(x_i)]^{\delta_i}[1 - S(C_l)]^{1-\delta_i} \quad (2.12)$$

3 DISTRIBUIÇÃO GAMA EXPONENCIADA

A distribuição Gama Exponenciada foi proposta por Gupta et al. (1998), onde o modelo é obtido através do método $F(x) = [G(x)]^{\theta}$, sendo $G(x)$ a distribuição base, no caso em estudo, a distribuição Gama e θ (parâmetro forma) é um número real e positivo. Tal distribuição possui a flexibilidade suficiente para modelar taxas de falha monótonas e não monótonas (SHAWKY; BAKOBAN, 2011).

KUMAR, SINGH e YADAV (2015) propuseram a estimação bayesiana para o parâmetro e função de confiabilidade da distribuição gama exponenciada sob amostras com censura tipo 2 progressivas. Nadarajah e Gupta (2007) obtiveram interessantes resultados ao utilizar à distribuição para o tratamento de dados de seca no estado do Nebraska.

A distribuição gama exponenciada tem sua função densidade de probabilidade (f.d.p):

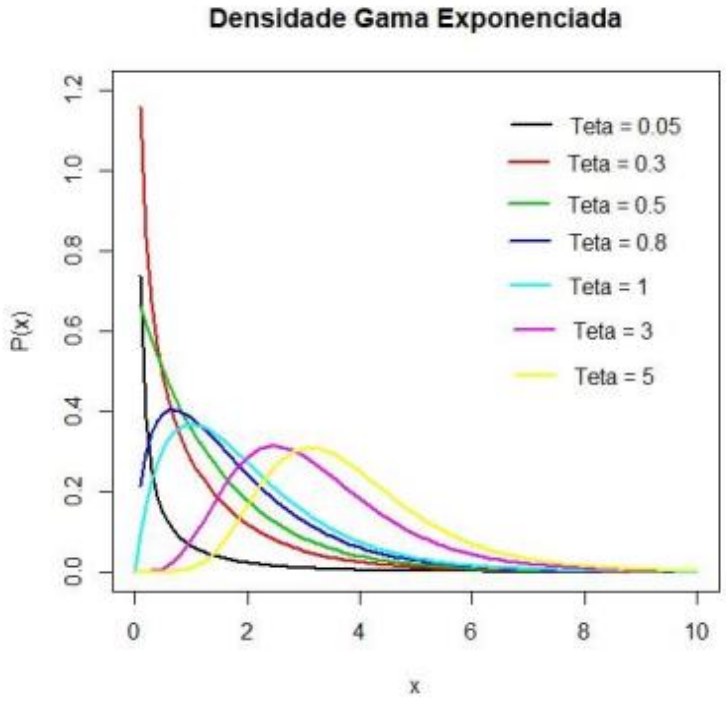
$$f(x) = \theta x e^{-x}\{1 - e^{-x}(1 + x)\}^{\theta-1} \quad (3.1)$$

e sua função densidade acumulada (f.d.a):

$$F(x) = [1 - e^{-x}(x + 1)]^{\theta} \quad (3.2)$$

onde θ é o parâmetro forma. Vale ressaltar que quando $\theta = 1$ a distribuição se torna uma Gama com parâmetro $\alpha = 2$ e $\beta = 1$ i.e. $G(2,1)$.

Figura 6 - distribuição gama exponenciada para diferentes valores de parâmetro



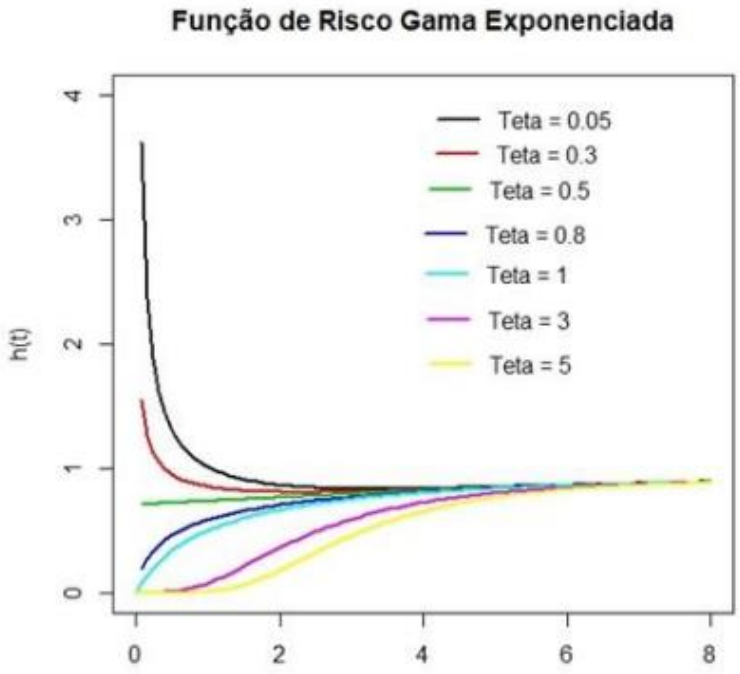
Fonte: imagem do autor

As funções de confiabilidade e taxa de falhas da distribuição GE é dada respectivamente por:

$$S(t) = 1 - \ [1 - e^{-t}(t+1)]^{\theta} \qquad t > 0, \theta > 0, \qquad (3.3)$$

$$h(t) = \ \theta t e^{-t}[1 - e^{-t}(t+1)]^{\theta-1}\left\{1 - [1 - e^{-t}(t+1)]^{\theta}\right\}^{-1}, \ t > 0, \theta > 0 \qquad (3.4)$$

Figura 7 – função de risco da distribuição gama exponenciada para diferentes valores de parâmetro



Fonte: imagem do autor

3.1 Estimação para a distribuição Gama Exponenciada

Seja $X_1, X_2, \ldots, X_n$ uma amostra aleatória independente e identicamente distribuída da distribuição Gama Exponenciada. A função de verossimilhança de θ é dada por:

$$L(\theta; x) = \prod_{i=1}^{n} \theta x_i e^{-x_i} \left\{1 - e^{(1+x_i)}\right\}^{\theta-1}$$

$$L(\theta; x) = \theta^n \prod_{i=1}^{n} x_i e^{-\sum_{i=1}^{n} x_i} \prod_{i=1}^{n} (1 - e^{-x_i} - x_i e^{-x_i})^{\theta-1}$$

Por ser uma função monótona, o logaritmo da verossimilhança facilita os cálculos, visto que o estimador de θ que maximiza $l(\theta)$ é o mesmo que maximiza $L(\theta)$. Portanto,

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

Derivando a expressão acima e igualando-a a zero afim de encontrar seu ponto de máximo, obtemos a estimativa de máxima verossimilhança $\hat{\theta}$ para o parâmetro de forma da distribuição Gama Exponenciada

$$\frac{\partial l(\theta)}{\partial \theta} = \frac{n}{\theta} + \sum_{i=1}^{n} log(1 - e^{-x_i} - x_i e^{-x_i}) = 0$$

com isso, $\hat{\theta}$ é definido como:

$$\hat{\theta} = -n[\sum_{i=1}^{n} log\,(1 - e^{-x_i} - x_i e^{-x_i})]^{-1} \quad (3.5)$$

3.2 Distribuição Gama Exponenciada na presença de censura a esquerda

Sejam $X_{(r+1)}, X_{(r+2)}, \dots, X_{(n)}$ as últimas $(n-r)$ estatística de ordem de uma amostra aleatória de tamanho "n".Então a verossimilhança é dada por:

$$L(\theta) = \left[F(x_{(r+1)})\right]^r \prod_{i=r+1}^{n} f(C_l)$$

$$= \left[1 - e^{-x_{(r+1)}}(x_{(r+1)} + 1)\right]^{r\theta} \prod_{i=r+1}^{n} \theta x_{(i)} e^{-x_{(i)}} \left[1 - e^{-x_{(i)}}(x_{(i)} + 1)\right]^{\theta - 1}$$

$$= \left[1 - e^{-x_{(r+1)}}(x_{(r+1)} + 1)\right]^{r\theta} \theta^{n-r} \prod_{i=r+1}^{n} x_{(i)} e^{-x_{(i)}} \left[1 - e^{-x_{(i)}}(x_{(i)} + 1)\right]^{\theta - 1}$$

aplicando o logaritmo:

$$\log\big(L(\theta)\big) = (n-r)log(\theta) + r\theta \log\Big(1 - e^{-x_{(r+1)}}\big(-x_{(r+1)} + 1\big)\Big)$$

$$(\theta - 1) \sum_{i=r+1}^{n} log\,\big(1 - e^{-x_i}(x_i + 1)\big)$$

derivando em relação a θ:

$$\frac{\partial log\,(L(\theta))}{\partial \theta} = \frac{n-r}{\theta} + rlog\,\Big(1 - e^{-x_{(r+1)}}\big(x_{(r+1)} + 1\big)\Big) \sum_{i=r+1}^{r} log\,\big(1 - e^{-x_i}(x_i + 1)\big)$$

igualando a derivada a zero para maximizar a função de verossimilhança, temos

$$\hat{\theta} = \frac{n-r}{rlog\left(1-e^{-x_{(r+1)}}(x_{(r+1)}+1)\right)\sum_{i=r+1}^{n} log\left(1-e^{-x_i}(x_i+1)\right)} \quad (3.6)$$

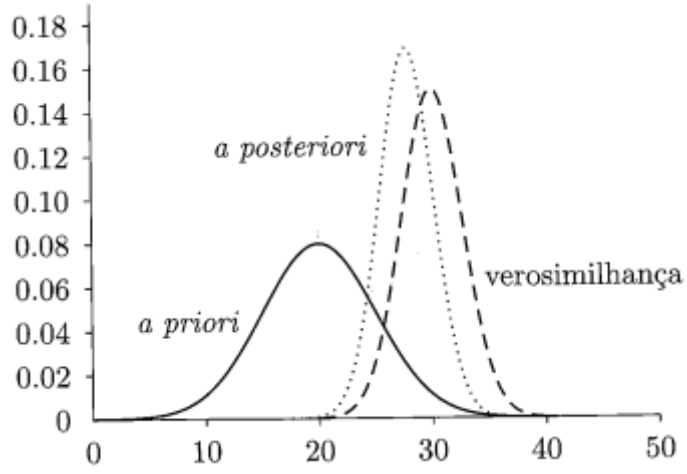
4 INFERÊNCIA BAYESIANA

No modelo clássico o parâmetro $\theta, \theta \in \Theta$, é um vetor desconhecido, mas fixo, i.e., igual ao valor particular que indexa uma distribuição que descreve apropriadamente o processo físico que gera as observações. Já no modelo bayesiano o parâmetro $\theta, \theta \in \Theta$, é tomado como um vetor aleatório, sendo assim, a incerteza do parâmetro é quantificada probabilisticamente para que possa ser inferido.

Suponha que dispomos de alguma informação sobre o parâmetro θ representado pela distribuição $\pi(\theta)$, conhecido como distribuição a priori sobre θ. A distribuição a priori representa o que sabemos sobre θ sem o conhecimento da informação proveniente dos dados. Acrescentando essa informação a priori com a observação de dados amostrais $x_1, x_2, \ldots, x_n$ relacionado com θ, é possível obter uma distribuição a posteriori $p(\theta|x)$ que represente o que sabemos após a coleta dos dados.

O Teorema de Bayes é a base teórica em que a inferência Bayesiana está inserida. A partir desse teorema é possível criar uma regra de atualização para quantificar o aumento de informação, ou seja, partir da priori $\pi(\theta)$ e chegar na distribuição a posteriori $p(\theta|x)$. A função de verossimilhança tem um importante papel na fórmula de Bayes pois representa o meio através do qual os dados são incorporados junto a priori.

Figura 8 - representação visual da inferência bayesiana



Fonte: www.ime.unicamp.br

Desse modo, a distribuição a posteriori é dada por:

$$p(\theta|x) = \frac{\pi(\theta)L(x|\theta)}{\int \pi(\theta|x)L(x|\theta)\, d\theta}$$

Em outras palavras temos:

$$p(\theta|x) \propto \pi(\theta)L(x|\theta) \tag{4.1}$$

4.1 Escolha da Priori

A informação a priori que se deseja incorporar na análise é a informação que um especialista da área possui sobre o assunto que contém elementos subjetivos. Esses elementos podem ser também fontes objetivas como o uso de dados históricos, problemas análogos e entre outros. Contudo, em muitas situações não se tem o conhecimento proveniente de um expert e com isso é recorrido a outros tipos de priori.

Um outro método para ser abordado é através da utilização de priori não informativas, em que a informação utilizada não se torna tão relevante para a análise da distribuição. Pode-se pensar também como uma informação que considere todos os possíveis valores do parâmetro a ser analisado como igualmente prováveis.

A escolha da priori é subjetiva e foi feita no desenvolver pro projeto para que se tenha o melhor desempenho e que seja feita a melhor análise possível sobre os dados.

4.2 Priori Conjugada

Outro ponto a ser considerado para a escolha da priori é baseado no uso da família conjugada. A ideia é que a distribuição a priori e a posteriori pertençam a mesma classe de distribuições e assim a atualização do conhecimento que se tem de θ envolve apenas a mudança nos parâmetros indexadores. Naturalmente essa priori tem utilidade prática em termos de interpretabilidade.

Utilizando a verossimilhança da distribuição Gama exponenciada:

$$L(\theta) = \theta^{n-r}\left[1 - e^{-x_{(r+1)}}(x_{(r+1)} + 1)\right]^{r\theta} \prod_{i=r+1}^{n} x_{(i)} e^{-x_{(i)}}\left[1 - e^{-x_{(i)}}(x_{(i)} + 1)\right]^{\theta-1}$$

Considerando

$$c = 1 - e^{-x_{(r+1)}}\left(x_{(r+1)} + 1\right) \text{ e } d = \prod_{i=r+1}^{n} 1 - e^{-x_{(i)}}(x_{(i)} + 1)$$

substituindo na equação:

$$L(\theta) \propto \theta^{n-r} c^{r\theta} d^{\theta}$$

simplificando:

$$\propto \theta^{n-r}(c^r d)^{\theta}$$

Temos, então:

$$\propto \theta^{n-r}\exp\{\theta\log(c^r + d)\}$$

Por fim, é possível resumir a forma da distribuição gama:

$$L(\theta) \propto \theta^{n-r}e^{-u\theta} \tag{4.2}$$

onde

$$u = -rlog(c) - \log(d).$$

Portanto, ao utilizar a priori $Gama$ será possível chegar em uma priori conjugada com posteriori $Gama$, no qual verendo isso no capítulo seguinte.

4.3 estimador bayesiano

Dado uma priori $Gama(\alpha, \beta)$ para o parâmetro θ com função densidade de probabilidade igual a $\pi(\theta) \propto \theta^{\alpha-1}e^{-\beta\theta}$,então a correspondente posteriori é dada por uma distribuição gama com parâmetros $\alpha^* = n + \alpha - r$ e $\beta^* = u + \beta$, isto é,

$$P(\theta|x) \propto L(\theta|x)\pi(\theta)$$

utilizando a equação 4.2:

$$P(\theta|x) \propto \theta^{n-r}e^{-u\theta}\theta^{\alpha-1}e^{-\beta\theta}$$

Juntando as mesmas bases, chegamos em:

$$P(\theta|x) \propto \theta^{(n+\alpha-r)-1}e^{-(u+\beta)\theta}$$

Ou seja,

$$\theta|x \sim \Gamma(n + \alpha - r, u + \beta)$$

O estimador Bayesiano $\hat{\theta} = E(\theta|x)$ para a distribuição $Gama(\alpha^*, \beta^*)$ é dado por:

$$E(\theta|x) = \frac{\alpha^*}{\beta^*} = \frac{n + \alpha - r}{u + \beta}$$

logo,

$$\hat{\theta} = \frac{n+\alpha-r}{\beta - rlog\left[1-e^{-x_{(r+1)}}(x_{(r+1)}+1)\right] - \sum_{i=r+1}^{n}\ln\left[1-e^{-x_{(i)}}(x_{(i)}+1)\right]} \tag{4.3}$$

5 ESTUDO SIMULADO

5.1 geração de amostra

Para se gerar uma amostra aleatória da distribuição gama exponenciada foi utilizado o método da transformação inversa. O método da transformação inversa consiste em gerar uma amostra de uma distribuição uniforme conhecida sobre o

intervalo [0,1] igualando-a a função acumulada da distribuição gama exponenciada, na qual contempla o mesmo intervalo.

$$U = F(t) = (1 - e^{-t} - te^{-t})^{\theta}$$

Achando a função inversa é possível gerar valores da distribuição estudada. Como não é possível achar a inversa da distribuição Gama Exponenciada foi utilizado o Método de Newton Raphson, no qual é possível achar a raiz da função, ou seja, achar o ponto onde a equação é igual a zero.

5.2 Simulação

Será considerado os seguintes tamanhos de amostras $n = 20, 50, 100$ para diferentes valores do parâmetro $\theta = 0.1, 1, 2$. Para censura a esquerda foi utilizado os primeiros 10% e 20% dos dados com todas as combinações de n e θ.

Para gerar amostra com censura a esquerda foi utilizado os seguintes passos:

1. Escolha um valor arbitrário para θ, que é o parâmetro de forma da distribuição $(\theta > 0)$
2. Usando o parâmetro escolhido no passo 1, gere um valor aleatório de diferentes tamanhos (n = 20,50,100) da distribuição gama exponenciada utilizando o método da transformação inversa através do método de Newton Raphson.
3. Determine os pontos de terminação da censura à esquerda, ou seja, determine os valores de x_r.
4. As observações menores ou iguais a x_r foram consideradas censuradas.
5. Repita do passo 1 ao 5, 1000 vezes

5.3 Dados Completos

A seguir foram geradas 1000 amostras de tamanho 20,50 e 100 e resultados dos parâmetros estimados podem ser encontrados na tabela abaixo, juntamente com o Erro quadrático Médio (EQM) que mede a média das distancias ao quadrado do parâmetro estimado com o verdadeiro parâmetro.

Tabela 1 – Resultado da simulação para dados completos, com o parâmetro fixado em $\theta = 0.5$

n	Clássico		Bayesiano	
	$\hat{\theta}$	EQM	$\hat{\theta}$	EQM
20	0.5378808	0.099216	0.5193356	0.0847143
50	0.5173798	0.068831	0.5073546	0.0556432
100	0.5094662	0.040383	0.5092772	0.0420071

Tabela 2 – Resultado da simulação para dados completos, com o parâmetro fixado em $\theta = 1$

n	Clássico		Bayesiano	
	$\hat{\theta}$	EQM	$\hat{\theta}$	EQM
20	1.055551	0.068852	1.022923	0.058226
50	1.046059	0.051292	1.006676	0.021447
100	1.050692	0.047702	1.002632	0.010764

Tabela 3 – Resultado da simulação para dados completos, com o parâmetro fixado em $\theta = 2$

n	Clássico		Bayesiano	
	$\hat{\theta}$	EQM	$\hat{\theta}$	EQM
20	2.108752	0.114582	2.067225	0.098589
50	2.035705	0.087801	2.010122	0.085318
100	2.021235	0.040725	1.993126	0.041224

5.4 Dados Censurados

A seguir também foram geradas 1000 amostras de tamanho 20,50 e 100, utilizando-se de um nível de censura de 10% e 20%. Os resultados dos parâmetros estimados podem ser encontrados na tabela abaixo, juntamente com o Erro quadrático Médio (EQM).

Tabela 4 – Resultado da simulação para dados com censura, com o parâmetro fixado em $\theta = 0.5$

n	censura	Clássico		Bayesiano	
		$\hat{\theta}$	EQM	$\hat{\theta}$	EQM
20	10%	0.536121	0.019979	0.539083	0.020408
	20%	0.540256	0.0246622	0.543613	0.025228
50	10%	0.515501	0.0069278	0.516640	0.006993
	20%	0.515669	0.0077300	0.516952	0.007808
100	10%	0.509977	0.0030818	0.510541	0.003100
	20%	0.510541	0.0035781	0.510953	0.003600

Tabela 5 – Resultado da simulação para dados com censura, com o parâmetro fixado em $\theta = 1$

n	censura	Clássico		Bayesiano	
		$\hat{\theta}$	EQM	$\hat{\theta}$	EQM
20	10%	1.056259	0.0769063	1.062061	0.078394
	20%	1.057813	0.0832452	1.064349	0.085020
50	10%	1.017365	0.0237608	1.019602	0.023945
	20%	1.022544	0.0281774	1.025075	0.028433
100	10%	1.021616	0.0130429	1.022739	0.013120
	20%	1.023686	0.0149072	1.024952	0.015003

Tabela 6 – Resultado da simulação para dados com censura, com o parâmetro fixado em $\theta = 2$

n	censura	Clássico		Bayesiano	
		$\hat{\theta}$	EQM	$\hat{\theta}$	EQM
20	10%	2.091301	0.2621318	2.102661	0.267032
	20%	2.104272	0.3017942	2.117121	0.308112
50	10%	2.042172	0.0993893	2.086585	0.109298
	20%	2.043712	0.1085495	2.093712	0.120576
100	10%	2.022069	0.0483362	2.044072	2.044072
	20%	2.019505	0.0549293	2.044221	0.057818

6 CONCLUSÃO

A partir do resultado das simulações, observamos que para um nível fixo de censura à esquerda temos que o aumento no tamanho da amostra diminui o erro quadrático médio e também se tem uma melhor estimativa do parâmetro. Também é visível que a diminuição do θ fixado influência na diminuição do erro. Por exemplo, para um nível de 10% de censura à esquerda e tamanho amostral n = 20, temos uma média de EQM para θ = 2 de aproximadamente 0.26 que é reduzido para 0.1 para n = 50 e 0.05 para n= 100.Essa tendencia é observada também para outros valores de parâmetros e níveis de censura (20%).

Por fim, é visto que com o uso da inferência bayesiana é possível receber resultados muito próximos comparado ao uso da inferência clássica. Essa proximidade

de resultado pode ser justificada pelo uso de priori não informativos, no qual a maior parte do conhecimento é provindo da amostra utilizada.

Referências

AARSET, M.V. “The Null Distribution for a Test of Constant versus ‘Bathtub’ Failure Rate.” *Scandinavian Journal of Statistics*, vol. 12, no. 1, 1985, pp. 55–61. *JSTOR*, http://www.jstor.org/stable/4615972. Accessed 7 Feb. 2023.

BOLFARINE, H.; SANDOVAL, M.C. Introdução à inferência estatística, SMB, 2. Ed., [S.I.], p. 159, 2010

COLOSIMO, E. A.; GIOLO, S. R. Análise de sobrevivência aplicada. In: ABE-Projeto Fisher. [S.l.]: Edgard Blücher, 2006.

DONGARRA. J.; SULLIVAN F.;Guest Editors’ Introduction: The Top 10 Algorithms, Computing in Science and Engineering, v.2, 22-23, 2000, Disponivel em: https://ieeexplore.ieee.org/document/814652/authors#authors. Acesso em: 18 de Jan. 2022.

FERREIRA, D. F. Estatística Computacional Utilizando R, Lavras, Universidade Federal de lavras p.122, 2009. Disponível em: https://www.cin.ufpe.br/~maod/ESAP/R/apeco.pdf . Acesso em 19 de jan. 2022.

FOGLIATTO. F.S.; RIBEIRO, J.LD. Confiabilidade e Manutenção Industrial. Editora: Elsevier., [S.I.], p. 229,1. Ed.,2009

GLASER, R.E. Bathtub and Related Failure Rate Characterizations. Journal of the American Statistical Association,75,667-672,1980.

GUPTA, R. C.; GUPTA, P. L.; GUPTA, R. D. Modeling failure time data by lehman alternatives. Communications in Statistics-Theory and methods, Taylor & Francis, v. 27, n. 4, p. 887–904, 1998.

HASTINGS, W. K. Monte carlo sampling methods using markov chains and their applications. Oxford University Press, Oxford, Vol. 57, p.13; 1970

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations.Journal of the American statistical association, Taylor & Francis, v. 53, n. 282, 1958.

MITRA, SHARMISHTHA & KUNDU, DEBASIS. (2008). Analysis of left censored data from the generalized exponential distribution. Journal of Statistical Computation and Simulation.

OLIVEIRA, A. S.; Bayesian analysis of the Exponentiated Gamma distribution under left censorship, Presidente Prudente: [s.n.];2020

PAULINO,C.D.,AMARALTURKMAN,M.A.,MUTEIRA,B.,SILVA,G.L.(2018).Estatistica Bayesiana, 2. e.d. Fundação Calouste Hulbenkian, Lisboa.

SHAWKY, A.; BAKOBAN, R. Bayesian and non-bayesian estimations on the exponentiated gamma distribution. Applied Mathematical Sciences, v. 2, n. 51, 2008.

SHAWKY, A. I.; BAKOBAN, R. A.; Exponentiated Gamma Distribution: Different Methods of Estimations, C. Conca, Arabia Saudita,2012. Disponivel em: https://www.hindawi.com/journals/jam/2012/284296/ . Acesso em 18 de Jan. 2022.