



Trabalho de Conclusão de Curso

**Um estudo dos Modelos de Fragilidades Gama e  
Log-Normal via inferência Bayesiana**

Rayssa Miczewski de Araujo

20 de dezembro de 2018

Rayssa Miczewski de Araujo

**Um estudo dos Modelos de Fragilidades Gama e  
Log-Normal via inferência Bayesiana**

Trabalho de Conclusão apresentado à comissão de Graduação do Departamento de Estatística da Universidade Federal do Rio Grande do Sul, como parte dos requisitos para obtenção do título de Bacharel em Estatística.

Orientador(a): Profa. Dra. Silvana Schneider

Porto Alegre  
Dezembro de 2018

Rayssa Miczewski de Araujo

**Um estudo dos Modelos de Fragilidades Gama e  
Log-Normal via inferência Bayesiana**

Este Trabalho foi julgado adequado para obtenção dos créditos da disciplina Trabalho de Conclusão de Curso em Estatística e aprovado em sua forma final pela Orientador(a) e pela Banca Examinadora.

Orientador(a): \_\_\_\_\_  
Profa. Dra. Silvana Schneider, UFRGS  
Doutora pela Universidade Federal de Minas Gerais, Belo Horizonte, MG

Banca Examinadora:

Prof. Dr. Rodrigo Citton Padilha dos Reis, UFRGS  
Doutor pela Universidade Federal de Minas Gerais, Belo Horizonte, MG

Porto Alegre  
Dezembro de 2018

*"Se manter de pé, contra o que vier, vencer os medos, mostrar ao que veio, ter o foco ali, e sempre seguir rumo a vitória!" (Dead Fish)*

# Agradecimentos

Agradeço a minha professora orientadora Silvana Schneider, por ter me passado conhecimento e sabedoria, por me aconselhar e me apoiar durante esse período. Muito obrigada por acreditar e confiar em mim.

Agradeço aos professores da graduação que tive o prazer de conhecer, sem eles esse momento não seria alcançado.

Agradeço aos meus colegas de trabalho da SPGG e principalmente as meninas do NUMA, que sempre estão dispostas a ajudar e tornar o ambiente de trabalho mais leve.

Aos amigos que fiz durante a graduação e principalmente a Nicole, Aline e Helena, amigas que vou levar para a vida inteira.

As minhas amigas de infância que estiveram comigo em praticamente todos os momentos da minha vida, Kristine, Caroline, Giselly, Pauxiana, Daniela, Stephenie e Priscila. Sempre me aconselharam e torceram por mim, sou muito grata por ter vocês na minha vida.

Por fim, agradeço a minha família, que sempre me deu suporte e apoio nas minhas decisões. Agradeço aos meus pais, Felisberto e Bernardete, que sempre me incentivam a estudar e buscar conhecimento, e ao meu irmão Augusto que sempre esteve ao meu lado quando precisei, sem eles eu não teria chegado até aqui. Sou eternamente grata pela família que tenho.

# Resumo

Em análise de sobrevivência o objetivo é analisar o tempo até a ocorrência de algum evento de interesse, conhecido como tempo de falha. Com o passar do tempo foram aprimorando-se os métodos de análises, por exemplo, incluindo-se covariáveis explicativas que pudessem influenciar o evento de interesse. Existem também outros fatores que ajudam a explicar algumas informações que não foram consideradas no estudo, com o modelo de fragilidade que por exemplo, ajuda a explicar a não inclusão de algumas covariáveis importantes para o estudo ou alguma correlação que não foi considerada no modelo. A fragilidade pode ser interpretada como uma componente aleatório que é introduzida na função taxa de falha que descreve possíveis associações entre os tempos de cada indivíduo, ou entre os tempos de cada grupos de indivíduos. O objetivo deste trabalho é entender e aplicar, usando bancos de dados simulados via software estatístico *R* (R Core Team, 2018), e bancos de dados reais, duas das fragilidades que são mais utilizadas em análise de sobrevivência, a fragilidade Gama e a fragilidade Log-Normal. A inferência Bayesiana foi aplicada para a estimação dos parâmetros e os modelos foram ajustados utilizando o software Bayesiano *JAGS* (Plummer, 2015). Os modelos de fragilidades Gama e Log-Normal foram testados em dados simulados para que fosse possível analisar a adequação e o funcionamento dos mesmos. Para isso definiu-se distribuições *a priori* para os parâmetros dos modelos, e depois de aplicar os modelos com fragilidades Gama e Log-Normal nos bancos simulados observou-se as médias *a posteriori*, a partir disto, como os valores estimados estavam bem próximos dos valores reais, conclui-se que os modelos estavam se ajustando bem aos dados. Com isso, os modelos de fragilidade Gama e Log-Normal foram aplicados em dois bancos de dados diferentes, um pertencendo ao pacote Survival do *R* e outro retirado do site da *Fundação Fio Cruz* (<http://sobrevida.fiocruz.br/dados.html>), que também faz parte do livro Carvalho *et al.* (2011), para que fosse observado o comportamento desses modelos em dados reais. Por fim, foram utilizados os critérios de seleção de modelos Bayesianos DIC e WAIC, a partir deles decidiu-se qual modelo de fragilidade adequou-se melhor aos bancos de dados escolhidos.

**Palavras-Chave:** Análise de Sobrevivência, Fragilidade, Estatística Bayesiana.

# Abstract

In survival analysis, the objective is to analyze the time until the occurrence of an event of interest, known as failure time. Over time, the analysis methods were improved, for example, including explanatory covariates that could affect the event of interest. There are other factors which help to explain information that was not considered in the study, such as the frailty model, that helps to explain the non-inclusion of important covariates for the study or explain correlations that were not considered in the model. Frailty can be interpreted as a random component that is introduced in the hazard rate function and describes possible associations between times of each individual or between times of each group of individuals. The purpose of this work is to understand and apply, using  $R$  simulated data and real data, two of the fragilities models that are most used in survival analysis, Gama frailty and Log-Normal frailty. Bayesian inference was applied for parameter estimation and the models were adjusted using a Bayesian software, *JAGS*. Gama and Log-Normal frailty models were tested in simulated data to analyze their adjustment and behavior. *A priori* distributions were defined for the parameters of the models and, after applying Gama and Log-Normal frailty models in simulated data, a posteriori means were observed. Since estimated values were very close to real values, it was concluded that the models were well adjusted to data. Thus, Gama and Log-Normal frailty models were applied to two different databases, one belongs to Survival R package and another from the FundaçãoFio Cruz website (<http://sobrevida.fiocruz.br/dados.html>), which is also part of the book [Carvalho et al. \(2011\)](#), in order to observe the behavior of the models in real data. Finally, DIC and WAIC, Bayesian model selection criteria, were used to decide which frailty model were better adjusted to databases.

**Keywords:** Survival Analysis, Frailty, Bayesian Statistics.

# Sumário

<b>1</b>	<b>Introdução</b>	<b>12</b>
1.1	Contextualização da Análise de Sobrevidência . . . . .	12
1.1.1	Censura e truncamento . . . . .	12
<b>2</b>	<b>Definições e construção da Função de Verossimilhança</b>	<b>15</b>
2.1	Definições das funções de sobrevivência . . . . .	15
2.2	Construção da função de verossimilhança . . . . .	16
2.3	Modelo de Cox . . . . .	18
2.4	Verossimilhança com tempos Weibull . . . . .	19
<b>3</b>	<b>Modelos de Fragilidade</b>	<b>21</b>
3.1	Modelo de fragilidade Gama . . . . .	22
3.2	Modelo de fragilidade Log-Normal . . . . .	23
<b>4</b>	<b>Inferência Bayesiana</b>	<b>25</b>
4.1	Teorema de Bayes . . . . .	25
4.2	Construção da distribuição a posteriori e densidades condi- cionais completas . . . . .	26
4.3	Métodos de estimação - MCMC . . . . .	28
4.4	Crerérios de avaliação - DIC E WAIC . . . . .	30
<b>5</b>	<b>Simulação e Aplicação</b>	<b>32</b>
5.1	Aplicação em bancos de dados reais . . . . .	42
5.1.1	Aplicação dos modelos em dados sobre c4ncer de pulm4o . . . . .	42
5.1.2	Aplicação dos modelos em dados sobre infarto . . . . .	48
<b>6</b>	<b>Conclus4es</b>	<b>55</b>
	<b>Refer4ncias Bibliogr4ficas</b>	<b>55</b>
.1	C4lculo dos tempos em sobreviv4ncia . . . . .	59
.2	Gera4o dos dados . . . . .	59
.3	Fragilidades Log-Normal n4o padronizadas das simula4es . . . . .	60
.4	Fragilidades Log-Normal n4o padronizadas do banco sobre c4n- cer de pulm4o . . . . .	61
.5	Fragilidades Log-Normal n4o padronizadas do banco de dados sobre infarto . . . . .	62

## Lista de Figuras

Figura 5.1: Média, mediana e moda <i>a posteriori</i> dos parâmetros $\beta_1$ e $\beta_2$ com ajuste do modelo com fragilidade Log-Normal. . . . .	33
Figura 5.2: Média, mediana e moda <i>a posteriori</i> dos parâmetros $\lambda$ e $\alpha$ com ajuste do modelo com fragilidade Log-Normal. . . . .	34
Figura 5.3: Viés das médias <i>a posteriori</i> dos parâmetros $\beta_1, \beta_2, \lambda$ e $\alpha$ com ajuste do modelo com fragilidade Log-Normal. . . . .	35
Figura 5.4: Histograma dos 500 valores de DIC e WAIC gerados da fragilidade Log-Normal . . . . .	35
Figura 5.5: Média, mediana e moda <i>a posteriori</i> dos parâmetros $\beta_1$ e $\beta_2$ com ajuste do modelo com fragilidade Gama. . . . .	38
Figura 5.6: Média, mediana e moda <i>a posteriori</i> dos parâmetros $\lambda$ e $\alpha$ com ajuste do modelo com fragilidade Gama. . . . .	39
Figura 5.7: Viés das médias dos parâmetros $\beta_1, \beta_2, \lambda$ e $\lambda$ com ajuste do modelo com fragilidade Gama. . . . .	39
Figura 5.8: Histograma dos 500 valores de DIC e WAIC gerados da fragilidade Gama. . . . .	40
Figura 5.9: Gráfico das distribuições <i>a posteriori</i> das Fragilidades Gama e Log-Normal para os 40 grupos das simulações. . . . .	42
Figura 5.10: Distribuição dos tempos de falha e censura por sexo para os dados sobre Câncer de pulmão. . . . .	44
Figura 5.11: Médias <i>a posterior</i> das fragilidades Gama e Log-Normal do banco sobre dados de câncer de pulmão . . . . .	48
Figura 5.12: Distribuição dos tempos de falha e censura por sexo para os dados sobre infarto . . . . .	49
Figura 5.13: Médias <i>a posterior</i> das fragilidades Gama e Log-Normal do banco sobre dados de infarto . . . . .	54

## Lista de Tabelas

Tabela 5.1:	Resultado das simulações para o modelo de fragilidade Log-Normal.	33
Tabela 5.2:	DIC e WAIC das simulações para o modelo de fragilidade Log-Normal. . . . .	35
Tabela 5.3:	Fragilidades com ajuste do modelo de fragilidade Log-Normal das simulações . . . . .	37
Tabela 5.4:	Resultado das simulações para o modelo de fragilidade Gama. . .	38
Tabela 5.5:	DIC e WAIC das simulações com modelo de fragilidade Gama . .	40
Tabela 5.6:	Fragilidades com ajuste do modelo de fragilidade Gama . . . . .	41
Tabela 5.7:	Análise descritiva das variáveis contínuas do banco sobre câncer de pulmão . . . . .	43
Tabela 5.8:	Resultados para o modelo de fragilidade Log-Normal para os dados sobre câncer de pulmão . . . . .	45
Tabela 5.9:	Resultados para o modelo de fragilidade Gama para os dados sobre câncer de pulmão . . . . .	45
Tabela 5.10:	WAIC e DIC do modelo com fragilidade Log-Normal para os dados sobre câncer de pulmão . . . . .	46
Tabela 5.11:	WAIC e DIC do modelo com fragilidade Gama para os dados sobre câncer de pulmão . . . . .	46
Tabela 5.12:	Fragilidades com ajuste do modelo de fragilidade Log-Normal dados sobre câncer de pulmão . . . . .	47
Tabela 5.13:	Fragilidades com ajuste do modelo de fragilidade Gama para dados sobre câncer de pulmão . . . . .	47
Tabela 5.14:	Análise descritiva das variáveis contínuas do banco infarto . . . .	49
Tabela 5.15:	Resultados para o modelo de fragilidade Log-Normal para os dados sobre infarto . . . . .	50
Tabela 5.16:	Resultados para o modelo de fragilidade Gama para os dados sobre infarto . . . . .	51
Tabela 5.17:	WAIC e DIC do modelo de fragilidade Log-Normal sobre os dados de infarto . . . . .	51
Tabela 5.18:	WAIC e DIC do modelo de fragilidade Gama sobre os dados de infarto . . . . .	52
Tabela 5.19:	Fragilidades com ajuste do modelo de fragilidade Log-Normal para dados sobre infarto . . . . .	52
Tabela 5.20:	Fragilidades com ajuste do modelo de fragilidade Gama para dados sobre infarto . . . . .	53
Tabela 1:	Fragilidades Log-Normal das simulações . . . . .	60

Tabela 2:	Fragilidades Log-Normal das simulações . . . . .	61
Tabela 3:	Fragilidades Log-Normal do banco sobre câncer de pulmão . . . . .	61
Tabela 4:	Fragilidades Log Normal do banco de dados sobre Infarto . . . . .	62

# 1 Introdução

## 1.1 Contextualização da Análise de Sobrevivência

A análise de sobrevivência é uma das áreas da estatística formada por um conjunto de técnicas de análises em que, geralmente a variável resposta estudada é o tempo até um evento de interesse, e esse tempo é chamado de tempo de falha ou tempo de vida (Giolo & Colosimo, 2006). Esse evento de interesse é conhecido com desfecho e pode ser definido como a transição de um estado para o outro. A morte pode ser considerada é a transição do estado vivo para o estado morto. A ocorrência de uma doença pode ser a transição do estado saudável para o estado doente (Hanagal, 2011).

Outra característica dos dados de sobrevivência é a presença de censura, que ocorre quando o evento de interesse não é observado para algum indivíduo (Hanagal, 2011). A censura ocorre quando por alguma razão o acompanhamento de algum indivíduo estudado foi interrompido, seja porque ele mudou de cidade, ou o estudo terminou para a análise dos dados, ou porque o paciente veio a óbito por causas diferentes das estudadas (Giolo & Colosimo, 2006).

Os métodos de análise de sobrevivência têm sido aplicados em várias áreas, na engenharia são comuns os estudos em que produtos ou componentes de um equipamento são colocados sob testes para estimar características relacionadas aos seus tempos de vida, nas ciências sociais várias situações de interesse têm como resposta tempo entre eventos, que pode ser o tempo até uma criança se adaptar a uma nova escola, ou também criminalistas que estudam o tempo entre a liberação de presos e a ocorrência de crimes (Giolo & Colosimo, 2006).

### 1.1.1 Censura e truncamento

Em estudos que tem como objetivo avaliar o tempo até a ocorrência de um evento de interesse, é comum a presença de observações incompletas ou parciais. Essas observações, que são denominadas de Censuras, e que podem ocorrer por diversas situações, como por exemplo, a perda de acompanhamento de um paciente durante um estudo ou a não ocorrência do evento de interesse até o término de um experimento. Lembrando que as observações censuradas devem entrar na análise, pois mesmo incompletas elas fornecem informações sobre o tempo de vida do indivíduo e a omissão das mesmas pode acarretar conclusões viciadas (Giolo & Colosimo, 2006).

A notação usada para representar os dados censurados é feita usando duas va-

riáveis aleatória  $(T_i, C_i)$ , em que  $T_i$  representa o tempo de falha observado para o indivíduo  $i$  e  $C_i$  representa o tempo de censura do indivíduo  $i$  (Giolo & Colosimo, 2006). Os tempos observados são definidos na forma,  $t = \min(T_i, C_i)$ , que define o tempo como sendo o mínimo entre o evento de falha ou censura. Se o tempo de falha for menor que o tempo de censura, significa que ocorreu uma falha, e quando o tempo de censura for menor que o tempo de falha, significa que ocorreu uma censura (Giolo & Colosimo, 2006). A indicadora de falha ou censura é definida como

$$\delta_i = \begin{cases} 1, T_i \leq C_i \\ 0, T_i > C_i, \end{cases} \quad (1.1)$$

em que, quando  $\delta_i = 1$  significa que o indivíduo sofreu uma falha, e quando  $\delta_i = 0$  significa que o indivíduo sofreu uma censura (Carvalho *et al.*, 2011).

Os três tipos de censuras em estudos de sobrevivência são definidas aqui com base nos livros Carvalho *et al.* (2011) e Giolo & Colosimo (2006):

- Censura à direita: ocorre quando o tempo de ocorrência do evento de interesse está à direita do tempo registrado, em outras palavras, quando um indivíduo por algum motivo deixa de ser observado durante o período de tempo estabelecido no estudo. Por exemplo, em um estudo que quer avaliar o tempo até a ocorrência de alguma doença, alguns indivíduos podem ter saído do estudo sem vivenciar o evento. Além disso, esse tipo de censura pode ser ainda dividida em três tipos.
  - Tipo I: Ocorre quando o estudo é encerrado após um período de tempo pré-estabelecido.
  - Tipo II: Ocorre quando o estudo é encerrado após ter ocorrido o evento de interesse em um número preestabelecido de indivíduos.
  - Aleatório: Ocorre se a observação for retirada no decorrer do estudo sem ter ocorrido o evento de interesse, ou se o evento ocorreu por alguma razão diferente.
- Censura à esquerda: ocorre quando o tempo registrado está à esquerda do tempo de falha. Isto é, o evento de interesse já ocorreu quando o indivíduo foi observado. Por exemplo, um estudo sobre a determinação do tempo que crianças de um certo local levaram para aprender a escrever, pode ser que antes do início do estudo, algumas dessas crianças já tivessem aprendido, e portanto, não havia como identificar quando isso ocorreu.
- Censura intervalar: Ocorre quando não se sabe o tempo exato de ocorrência do evento, sabe-se que ele ocorreu em um intervalo especificado. Um exemplo pode ser o tempo entre exames feitos semestralmente em animais de uma fazenda para prevenir uma certa doença e a ocorrência da doença sobre eles, assim não se sabe o tempo exato do evento apenas o intervalo entre o exame e a ocorrência da doença.

As censuras também podem ser classificadas como não informativas que é quando as censuras ocorrem ao acaso e não estão relacionadas com o desfecho, e censura

informativas que é quando elas estão relacionadas com o desfecho (Giolo & Colosimo, 2006).

Uma outra forma de dados de sobrevivência são as observações truncadas, que muitas vezes são confundidos com as censuras (Giolo & Colosimo, 2006). A sua característica é a de que por algum motivo alguns indivíduos são inteiramente excluídos do estudo (Carvalho *et al.*, 2011). Dois tipos de truncamento são definidos na literatura, truncamento à esquerda e o truncamento à direita.

- Truncamento à esquerda: nesse tipo de truncamento os indivíduos só são acompanhados após a ocorrência do evento. Por exemplo em um estudo que quer determinar o tempo de vida após a aposentadoria de moradores de uma cidade, se a amostra usada for de dados previdenciários, os indivíduos que não atingiram a aposentadoria não farão parte do estudo (Giolo & Colosimo, 2006).
- Truncamento à direita: é o critério de seleção dos indivíduos que inclui somente os indivíduos que sofreram o evento de interesse. Por exemplo, em um estudo sobre AIDS, alguns dos indivíduos podem ter o vírus, mas não terem desenvolvido a doença, portanto eles não são incluídos no estudo (Giolo & Colosimo, 2006).

O objetivo deste estudo é entender e aplicar, usando bancos de dados simulados via software estatístico *R* (R Core Team, 2018), e bancos de dados reais, dois dos modelos de fragilidades mais utilizadas em análise de sobrevivência, o modelo de fragilidade Gama e o modelo de fragilidade Log-Normal.

O trabalho está organizado da seguinte maneira, o segundo capítulo apresenta a construção da função de verossimilhança. No terceiro capítulo é onde são apresentadas as fragilidades Gama e Log-Normal. Como a implementação destes modelos foram feitos via Software *JAGS* e a estimação dos parâmetros via inferência Bayesiana, no quarto capítulo foi apresentado o teorema de Bayes, assim como a construção da distribuição *a posteriori* e as densidades condicionais completas dos parâmetros dos modelos. Neste capítulo também foram estudados os métodos de estimação via cadeias de Markov: os amostradores *Gibbs* e o algoritmo *Metropolis Hastings* e por fim dois critérios de avaliação de modelos conhecidos como *Deviance Information Criterion (DIC)* (Spiegelhalter *et al.*, 2002) e *Watanabe-Akaike Information Criterion (WAIC)* (Watanabe, 2010).

No último capítulo se encontram as simulações dos modelos com bancos de dados simulados e também a aplicação dos modelos de fragilidades Gama e Log-Normal nos dois bancos de dados reais. O primeiro banco real que foi escolhido é o *Lung*, que foi retirado do pacote *Survival* do Software *R* (R Core Team, 2018), e contém dados de indivíduos em estado avançado de câncer de pulmão. O segundo banco escolhido foi o *Infarto*, retirado do site da *Fundação Fio Cruz*, e contém dados de indivíduos que sofreram infartos.

## 2 Definições e construção da Função de Verossimilhança

### 2.1 Definições das funções de sobrevivência

Nesta seção serão apresentadas algumas funções da análise de sobrevivência e como elas se relacionam entre si.

Suponha  $T$  uma variável contínua, positiva, com função densidade de probabilidade  $f(t)$ , representando o tempo até a ocorrência de um evento de interesse. Em sobrevivência essa função pode ser representada como a probabilidade de um indivíduo sofrer um evento em um intervalo imediato de tempo ([Carvalho et al., 2011](#)),

$$f(t) = \lim_{\epsilon \rightarrow 0^+} \frac{Pr(t \leq T < t + \epsilon)}{\epsilon}, \quad (2.1)$$

em que  $\epsilon$  é o incremento de tempo infinitamente pequeno.

Segundo [Carvalho et al. \(2011\)](#), a função de sobrevivência  $S(t)$  é a probabilidade de um indivíduo não falhar, ou seja, de sobreviver a um certo tempo  $t$ , estabelecido no estudo e representada por

$$S(t) = Pr(T > t). \quad (2.2)$$

A função de distribuição acumulada  $F(t)$  de uma variável aleatória,  $T$ , é a probabilidade de um evento ocorrer até o tempo  $t$ ,

$$F(t) = Pr(T \leq t). \quad (2.3)$$

Com essa duas funções definidas (2.2) e (2.3), podemos ver que a função de sobrevivência  $S(t)$  é o complementar da função de distribuição acumulada  $F(t)$ ,

$$S(t) = Pr(T > t) = 1 - Pr(T \leq t) = 1 - F(t). \quad (2.4)$$

A função taxa de falha  $h(t)$  é definida como a taxa instantânea de um indivíduo sofrer o certo evento de interesse, entre o tempo  $t$  e  $t + \epsilon$ , dado que ele sobreviveu até o tempo  $t$ . É definida como ([Carvalho et al., 2011](#))

$$h(t) = \lim_{\epsilon \rightarrow 0^+} \frac{Pr(t \leq T < t + \epsilon | T \geq t)}{\epsilon}. \quad (2.5)$$

A função taxa de falha pode ser relacionada com a função de sobrevivência da seguinte forma (Carvalho *et al.*, 2011)

$$h(t) = \frac{f(t)}{S(t)}. \quad (2.6)$$

Notamos que a função taxa de falha e a função de sobrevivência são inversamente proporcionais. Quando a função taxa de falha aumenta a probabilidade de sobrevivência diminui e vice-versa.

A função de taxa de falha acumulada pode ser útil quando queremos avaliar mais a fundo a função taxa de falha (Giolo & Colosimo, 2006). Essa função serve para medir o risco até um tempo  $t$  determinado, na forma matemática ela é a soma de todos os riscos em todos os tempos até o tempo (Carvalho *et al.*, 2011). Ela pode ser calculada a partir da integral da função de taxa de falha, dada por

$$H(t) = \int_0^t h(u)d(u). \quad (2.7)$$

As funções de sobrevivência, taxa de falha e taxa de falha acumulada podem ser relacionadas da seguinte maneira (Giolo & Colosimo, 2006)

$$h(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}(\log S(t)), \quad (2.8)$$

$$H(t) = \int_0^t h(u)d(u) = -\ln(S(t)), \quad (2.9)$$

$$S(t) = \exp\{-H(t)\} = \exp\left\{-\int_0^t h(u)d(u)\right\}. \quad (2.10)$$

## 2.2 Construção da função de verossimilhança

A função de verossimilhança é utilizada para que possam ser feitas inferências acerca de possíveis valores de parâmetros, a partir de um conjunto de dados quaisquer (Giolo & Colosimo, 2006). Supondo tempos de censura aleatórios em um estudo onde  $n$  indivíduos são observados, e associado ao  $i$ -ésimo indivíduo estão os pares de variáveis aleatórias contínuas  $T_i$  (tempo de falha) com função de distribuição  $F_\theta(\cdot)$  e  $C_i$  (tempo de censura), com função de distribuição  $G_\phi(\cdot)$ . As derivadas das funções de distribuição dos tempos de falha e tempos de censura, respectivamente, nos fornecem as funções de densidade  $f_\theta(\cdot)$  e  $g_\phi(\cdot)$ . Nelas temos também dois parâmetros distintos  $\theta$  e  $\phi$ .

Temos, então, que o tempo  $t_i$  é o mínimo entre o tempo de falha e o tempo de censura, expresso da maneira  $t_i = \min(T_i, C_i)$ , como visto em (1.1), (Lawless, 2011), (Teruel *et al.*, 1993). Assim a função de probabilidade conjunta do tempo do indivíduo  $i$  e a indicadora de falha ou censura,  $(t_i, \delta_i)$ , apresentada por Lawless (2011) fica

$$Pr[t_i = t, \delta_i = 0] = Pr[C_i = t, T_i > C_i] = f_\theta(t_i)[1 - G_\phi(t_i)], \quad (2.11)$$

quando for observada uma censura, e

$$Pr[t_i = t, \delta_i = 1] = Pr[T_i = t, T_i \leq C_i] = g_\phi(t_i)[1 - F_\theta(t_i)], \quad (2.12)$$

quando for observada uma falha.

Com isso, a função de verossimilhança fica dada por

$$L(\theta, \phi) = \prod_{n=1}^n [f_\theta(t_i)[1 - G_\phi(t_i)]^{\delta_i} [g_\phi(t_i)[1 - F_\theta(t_i)]]^{1-\delta_i}. \quad (2.13)$$

Reescrevendo a função (2.13), temos

$$L(\theta, \phi) = \prod_{n=1}^n [f_\theta(t_i)^{\delta_i} [1 - F_\theta(t_i)]^{1-\delta_i} [g_\phi(t_i)^{1-\delta_i} [1 - G_\phi(t_i)]^{\delta_i}]. \quad (2.14)$$

Supondo então há independência entre os tempos de falha ( $T_i$ ) e censura ( $C_i$ ) e que a distribuição de censura não envolve nenhum parâmetro de interesse, a função de verossimilhança fica (Lawless, 2011) dada por

$$L(\theta) = \prod_{n=1}^n [f_\theta(t_i)^{\delta_i} [1 - F_\theta(t_i)]^{1-\delta_i}]. \quad (2.15)$$

Como vimos na Seção 2.1, a função de sobrevivência pode ser escrita da maneira (2.4). Logo a função de verossimilhança é definida como

$$L(\theta) = \prod_{n=1}^n f_\theta(t_i)^{\delta_i} [S_\theta(t_i)]^{1-\delta_i}. \quad (2.16)$$

Na Seção 2.1 vimos também que podemos escrever a função taxa de falha  $h(t)$  como (2.6). Logo,

$$f(t) = h(t)S(t). \quad (2.17)$$

Substituindo (2.17) na função de verossimilhança, temos

$$L(\theta) = \prod_{n=1}^n [h_\theta(t_i)S_\theta(t_i)]^{\delta_i} [S_\theta(t_i)]^{1-\delta_i},$$

$$L(\theta) = \prod_{n=1}^n [h_\theta(t_i)^{\delta_i} S_\theta(t_i)]. \quad (2.18)$$

Em sobrevivência uma das técnicas mais utilizadas para a estimação de parâmetros é o método de máxima verossimilhança. Esse método implica em estimar o parâmetro  $\theta$ , maximizando a equação  $L(\theta)$ . Isto é equivalente a maximizar o logaritmo da função de verossimilhança (Giolo & Colosimo, 2006)

$$Log(L) = \sum_{n=1}^n [\delta_i \log h_\theta(t_i) + \log(S_\theta(t_i))]. \quad (2.19)$$

## 2.3 Modelo de Cox

O modelo de Cox foi apresentado por David Cox em 1972, no Journal of the Royal Statistical Society B (Cox, 1972). O modelo ficou tão conhecido que teve mais de 600 citações entre os anos de 1987 a 1989 (Giolo & Colosimo, 2006). Cox propôs a inclusão de covariáveis na função taxa de falha do modelo de sobrevivência, com isso permitiu um novo tipo de análise para os estudos que tem como resposta o tempo de vida, que é o tempo até que um evento de interesse ocorra (Giolo & Colosimo, 2006). Esse modelo é dito semiparamétrico, pois ele não assume nenhuma distribuição para a função taxa de falha basal  $h_0(t)$ , ele assume apenas que as covariáveis agem multiplicativamente sobre a taxa de falha, que se torna a parte paramétrica do modelo.

Supondo um estudo de caso controle em que os indivíduos são selecionados de forma aleatória, e que tem como objetivo de comparar tempos de falha de dois grupos. Temos um grupo que irá receber um tratamento padrão (grupo 0) e outro grupo que irá receber um novo tratamento (grupo 1). A função taxa de falha do primeiro grupo é dada por  $h_0(t)$  e a do segundo por  $h_1(t)$ , assumindo que as taxas são proporcionais, temos (Giolo & Colosimo, 2006)

$$\frac{h_1(t)}{h_0(t)} = K, \quad (2.20)$$

em que  $K$  representa a razão de taxas de falha, constante para todo tempo  $t$ .

Seja  $x$  variável indicadora de grupos dado por

$$x = \begin{cases} 1, & \text{se grupo 1} \\ 0, & \text{se grupo 0.} \end{cases}$$

Seja  $K = \exp\{\beta x\}$ , logo

$$h(t) = h_0(t) \exp\{\beta x\}, \quad (2.21)$$

ou seja,

$$h(t) = \begin{cases} h_1(t) = h_0(t) \exp\{\beta\}, & x = 1 \\ h_0(t), & x = 0. \end{cases}$$

A expressão (2.21) para o modelo de Cox, é com uma única covariável. Agora considere  $p$  covariáveis, de modo que  $\mathbf{x}$  seja o vetor com os componentes  $x = (x_1, \dots, x_p)'$ . Com isso temos que a expressão geral do modelo de regressão de Cox é

$$h(t) = h_0(t)g(\mathbf{x}'\boldsymbol{\beta}), \quad (2.22)$$

em que  $g(\cdot)$  é uma função não-negativa, que deve ser especificada de modo que  $g(0)=1$  (Giolo & Colosimo, 2006).

Portanto, temos um modelo que é composto pelo produto de uma parte paramétrica e outra não-paramétrica. A parte não-paramétrica é usualmente chamada de função base ou basal e ela é uma função não negativa do tempo dada por  $h(t) = h_0(t)$ , quando  $x=0$  (Giolo & Colosimo, 2006). A parte paramétrica pode ser expressada da seguinte forma

$$g(\mathbf{x}'\boldsymbol{\beta}) = \exp(\mathbf{x}'\boldsymbol{\beta}) = \exp(\beta_1 x_1 + \dots + \beta_p x_p),$$

em que  $\beta$  é o vetor de parâmetros associados as covaráveis.

Agora que vimos o modelo de regressão de Cox, a função taxa de falha para o  $i$ -ésimo indivíduo fica dada por

$$h(t_i) = h_0(t_i) \exp(x'_i \beta). \quad (2.23)$$

Podemos substituir a função (2.23) em (2.18) e encontrar a função de verossimilhança. Como vimos na Seção 2.1, temos da função de sobrevivência e função taxa de falha acumulada relacionadas na forma (2.10).

Com isso temos, pelo modelo de regressão de Cox,

$$\begin{aligned} S(t_i|x_i) &= \exp\left\{-\int_0^{t_i} h_0(u) \exp\{x'_i \beta\} du\right\} \\ &= \exp\left\{-\int_0^{t_i} h_0(u) du \exp\{x'_i \beta\}\right\} \\ &= \left[\exp\left\{-\int_0^{t_i} h_0(u) du\right\}\right]^{\exp\{x'_i \beta\}} \\ &= [S_0(t_i)]^{\exp\{x'_i \beta\}}. \end{aligned} \quad (2.24)$$

Com essas relações vistas, agora podemos montar a função de verossimilhança com a inclusão de covariáveis, ela é apresentada da seguinte maneira (Giolo & Colosimo, 2006)

$$L(\theta) = \prod_{n=1}^n (h_0(t_i) \exp\{x'_i \beta\})^{\delta_i} [S_0(t_i)]^{\exp\{x'_i \beta\}}. \quad (2.25)$$

## 2.4 Verossimilhança com tempos Weibull

A distribuição de Weibull foi proposta em 1939 por Waloddi Weibull (Weibull, 1939), desde então ela é muito utilizada principalmente nas áreas biomédicas e industriais (Giolo & Colosimo, 2006). Por exemplo, em estudos sobre o tempo até a ocorrência de câncer em animais ((Pike, 1966) e (Peto, 1972)). Na medicina, por exemplo, em estudos sobre o tempo até a morte por câncer de pulmão (Whittemore & Altshuler, 1976), (Wienke, 2010). Ela tornou-se muito popular pois a sua função de taxa de falha é uma função monótona, isto é, ela pode ser crescente, decrescente ou contante no tempo. A distribuição Weibull tem dois parâmetros,  $\lambda$  é definido como parâmetro de forma e  $\alpha$  definido como parâmetro de escala. O parâmetro  $\alpha$  é mais flexível e é ele quem torna a função monótona (Giolo & Colosimo, 2006).

A função taxa de falha com tempos Weibull diminui monotonamente quando  $\alpha < 1$ , é constante quando  $\alpha = 1$  e aumenta quando  $\alpha > 1$ . Um caso especial da distribuição Weibull é quando o valor do parâmetro  $\lambda$  é igual a 1. Caso isso aconteça, a distribuição Weibull equivale a uma distribuição Exponencial. Para as funções de sobrevivência temos as seguintes funções com tempos Weibull (Wienke, 2010). As funções de análise de sobrevivência com tempos Weibull, estão definidas a seguir.

Função densidade de probabilidade

$$f(t) = \lambda \alpha t^{\alpha-1} e^{-\lambda t^\alpha}. \quad (2.26)$$

Função de sobrevivência

$$S(t) = e^{-\lambda t^\alpha}. \quad (2.27)$$

Função taxa de falha

$$h(t) = \lambda \alpha t^{\alpha-1}. \quad (2.28)$$

Função taxa de falha acumulada

$$H(t) = \lambda t^\alpha. \quad (2.29)$$

Como vimos na Seção 2.3, a construção da função de sobrevivência com a inclusão de covariáveis via modelo de Cox, agora com as funções citadas acima, incluiremos tempos Weibull a ela. A função de verossimilhança fica

$$L(\theta) = \prod_{i=1}^n \left[ (\lambda \alpha t^{\alpha-1}) \exp(x'_i \beta)^{\delta_i} (e^{-\lambda t^\alpha})^{\exp(x'_i \beta)} \right]. \quad (2.30)$$

### 3 Modelos de Fragilidade

Até agora vimos métodos em que a suposição dos tempos de sobrevivência de indivíduos distintos eram independente, mas essas suposições podem ser válidas para alguns estudos e para outros pode ser inválida. Por exemplo, o tempo de sobrevivência, quando observado em indivíduos de uma mesma família ou em uma ninhada de alguma espécie de animal, caracterizam situações em que a suposição de independência dos tempos pode não ser válida. Isso ocorre porque indivíduos de uma mesma família tendem a ter características muito semelhantes, e por esse motivo é razoável supor que existe alguma associação entre os tempos desse grupo. Existem também os eventos conhecidos como recorrentes, são aqueles em que cada indivíduo de um estudo está sujeito à múltiplos eventos do mesmo tipo, como por exemplo, ataques cardíacos. Nesse caso mais de um tempo de sobrevivência é observado para cada indivíduo, por isso podemos supor que existam associações entre os tempos de um mesmo indivíduo (Giolo & Colosimo, 2006).

A fragilidade foi introduzida por Vaupel *et al.* (1979). Resumidamente, Vaupel *et al.* observou que uma tábua de vida, publicada na época, ignorava o fato de que alguns indivíduos eram mais suscetíveis ou propensos a ter um tempo de vida menor que outros indivíduos. Ele observou que os indivíduos mais “frágeis” morriam antes, com isso a estimativa da taxa de risco seria viesada. Para resolver esse problema de heterogeneidade, ele propôs a inclusão de um fator aleatório (fragilidade) na taxa de falha. Esse fator modificaria o risco de forma multiplicativa, por exemplo um indivíduo com fragilidade dois, tem o dobro do risco de um indivíduo dito “padrão” com fragilidade um (Amrhein, 2014).

Portanto, a inclusão da fragilidade em um estudo é a inclusão de um efeito aleatório na função taxa de falha para descrever uma possível associação entre indivíduos, ou entre um grupo de indivíduos (Giolo & Colosimo, 2006).

Partindo do modelo de Cox, foi introduzido um efeito aleatório  $Z$ , que é uma variável aleatória independente, identicamente distribuída e latente (não observada) (Giolo & Colosimo, 2006). Essa variável tem média 1 e variância  $\zeta$ , que reflete a fragilidade de um indivíduo ou grupo. Quando temos altos valores de  $\zeta$ , isso nos indica uma grande heterogeneidade entre os indivíduos, quando os valores são baixos pouca heterogeneidade, o que indica que é desnecessário a inclusão do efeito no modelo. Portanto com o efeito  $Z=z$ , o modelo de Cox passa a ser (Carvalho *et al.*, 2011)

$$h(t) = zh_0(t) \exp(\mathbf{x}'\boldsymbol{\beta}). \quad (3.1)$$

Os modelos de fragilidade são usados tanto em casos de sobrevivência univariada, que é quando cada indivíduo tem a sua própria fragilidade, quanto em sobrevivência multivariada, que é quando um grupo tem a mesma fragilidade (Giolo & Colosimo, 2006).

Como vimos na Seção (2.3), a função de verossimilhança para o modelo de Cox é dada por (2.25), agora com a inclusão da fragilidade na função temos

$$L(\beta, \theta | Z_1, \dots, Z_n) = \prod_{i=1}^n \left( Z_i h_0(t_i; \theta) e^{\beta' X_i} \right)^{\delta_i} e^{-Z_i H_0(t_i; \theta) e^{\beta' X_i}}. \quad (3.2)$$

Nas próximas seções serão apresentadas duas das fragilidades mais usadas na literatura, a Fragilidade Gama e a Fragilidade Log-Normal (Carvalho *et al.*, 2011).

### 3.1 Modelo de fragilidade Gama

A distribuição Gama é a distribuição mais usada para as fragilidades, isso se deve ao fato de ela ser muito flexível e de que a sua função densidade é facilmente derivada pela transformada de Laplace, resultando em um expressão fechada (Wienke, 2010). Os parâmetros da distribuição Gama são  $k$  e  $\lambda$ , quando  $k = 1$ , ela se iguala a uma distribuição exponencial, quando  $k$  é um valor grande, lembra a forma de sino da distribuição normal (Wienke, 2010).

Abbring & Van Den Berg publicaram um artigo (Abbring & Van Den Berg, 2007) que racionaliza o uso da distribuição Gama para fragilidades em análises de dados em que a variável de desfecho é o tempo até um evento de interesse. Os autores mostram que em uma grande classe de modelos de fragilidade univariada a distribuição de fragilidade entre os indivíduos sobreviventes, convergiam para uma distribuição Gama (Wienke, 2010). A função densidade da fragilidade Gama é dada por

$$f(z) = \frac{1}{\Gamma(k)} \lambda^k z^{k-1} e^{-\lambda z}. \quad (3.3)$$

A transformada de Laplace desta distribuição fica

$$\begin{aligned} \mathbf{L}(u) &= \frac{1}{\Gamma(k)} \lambda^k \int e^{-uz} z^{k-1} e^{-\lambda z} dz \\ &= \frac{\lambda^k}{(\lambda + u)^k \Gamma(k)} (\lambda + u)^k \int z^{k-1} \exp^{-(\lambda+u)z} dz \\ &= \left( 1 + \frac{u}{\lambda} \right)^{-k}. \end{aligned} \quad (3.4)$$

A primeira e a segunda derivada ficam

$$\mathbf{L}'(u) = -\frac{k}{\lambda} \left( 1 + \frac{u}{\lambda} \right)^{-k-1}, \quad (3.5)$$

$$\mathbf{L}''(u) = \frac{k(k+1)}{\lambda^2} \left( 1 + \frac{u}{\lambda} \right)^{-k-2}. \quad (3.6)$$

Quando nessas derivadas, supormos que  $u = 0$ , isso implica que,

$$\mathbf{E}(Z) = \frac{k}{\lambda}, \quad (3.7)$$

e

$$\mathbf{V}(Z) = \frac{k(k+1)}{\lambda^2} - \frac{k^2}{\lambda^2} = \frac{k}{\lambda^2}. \quad (3.8)$$

Para o modelo ser identificável, a restrição  $k = \lambda$  é usada para a distribuição Gama, que resulta em  $\mathbf{E}(Z) = 1$ , denotando a média e  $\sigma^2 = \frac{1}{\lambda}$  denotando a variância da variável fragilidade.

Se denotarmos a variância  $V(Z) = \frac{1}{\lambda}$  por  $\sigma_2$ , então a função densidade de uma variável  $Z \sim \Gamma\left(\frac{1}{\sigma^2}, \frac{1}{\sigma^2}\right)$  tem distribuição Gama é dada por

$$f(z) = \frac{1}{\Gamma\left(\frac{1}{\sigma^2}\right)} \left(\frac{1}{\sigma^2}\right)^{\frac{1}{\sigma^2}} z^{\frac{1}{\sigma^2}-1} \exp\left(-\frac{z}{\sigma^2}\right). \quad (3.9)$$

No contexto de análise de sobrevivência serão apresentadas as funções de sobrevivência, função de probabilidade e função taxa de falha, com fragilidade gama.

$$S(t) = \frac{1}{(1 + \sigma^2 H_0(t))^{\frac{1}{\sigma^2}}}, \quad (3.10)$$

$$f(t) = \frac{h_0(t)}{(1 + \sigma^2 H_0(t))^{\frac{1}{\sigma^2} + 1}}, \quad (3.11)$$

$$h(t) = \frac{h_0(t)}{1 + \sigma^2 H_0(t)}. \quad (3.12)$$

Utilizando as funções vista acima, a função de verossimilhança com fragilidade Gama resulta em

$$L(\beta, \theta, \sigma^2) = \prod_{n=1}^n \left( \frac{h_0(t; \theta) e^{\beta' x_i}}{1 + \sigma^2 H_0(t; \theta) e^{\beta' x_i}} \right)^{\delta_i} (1 + \sigma^2 H_0(t; \theta) e^{\beta' x_i})^{-\frac{1}{\sigma^2}}. \quad (3.13)$$

## 3.2 Modelo de fragilidade Log-Normal

A distribuição Log-Normal é uma das mais utilizadas para fragilidade, pelo fato de ser muito útil quando se quer modelar dados que tem uma distribuição Normal Multivariada. Assumimos uma variável aleatória com distribuição normal  $W \sim N(m, s^2)$ , onde  $m$  é a média e  $s^2$  é a variância. Se  $W \sim Normal$ , então assumimos que  $Z = e^W$  para gerar fragilidades Log-Normal (Wienke, 2010).

O cálculo dos parâmetros  $\mu$ , média e  $\sigma^2$ , variância dessa distribuição são,

$$\mu = \mathbf{E}(Z) = e^{m + \frac{s^2}{2}}, \quad (3.14)$$

$$\sigma^2 = \mathbf{V}(Z) = e^{2m + s^2} (e^{s^2} - 1). \quad (3.15)$$

Cálculos simples mostram que, sob a restrição  $\mu = 1$ , isso resulta em

$$m = \mathbf{E}(\log(z)) = -\frac{1}{2}s^2, \quad (3.16)$$

$$s^2 = \mathbf{V}(\log(Z)) = \log(1 + \sigma^2). \quad (3.17)$$

É utilizado o *log* para se obter a fragilidade Log-Normal, a partir de uma distribuição Normal.

A função de densidade da fragilidade Log-Normal é

$$f(z) = \frac{1}{\sqrt{2\pi sz}} e^{-\frac{(\log(z))^2}{2s^2}}. \quad (3.18)$$

A interpretação do modelo de fragilidade Log-Normal como modelo de efeitos aleatórios normalmente distribuídos torna esta classe de modelo mais conveniente para aplicação de pesquisadores no campo de análise de sobrevivência ([Wienke, 2010](#)).

A função de verossimilhança quando a distribuição das fragilidades segue uma desatribuição Log-Normal é dada por

$$L(\beta, \theta, s^2) = \prod_{n=1}^n \int (h_0(t_i; \theta) e^{\beta' \mathbf{x}_i + w_i})^{\delta_i} e^{-H_0(t_i; \theta) e^{\beta' \mathbf{x}_i + w_i}} d\Phi(w_i). \quad (3.19)$$

Por ser uma função muito complexa, a integral dessa distribuição não tem uma forma conhecida, ou seja, não é possível fazer os cálculos na forma convencional, apenas computacionalmente. Para termos estimativas dessa função, a abordagem deve ser feita com integrações numéricas, que são métodos onde se busca um resultado mais próximo possível do que seriam os valores reais ([Hanagal, 2011](#)).

## 4 Inferência Bayesiana

Na inferência clássica a amostra de dados é considerada aleatória e seus parâmetros são fixos. Já na inferência Bayesiana, são os próprios parâmetros que seguem uma distribuição de probabilidade, e mesmo antes de conhecer os dados da amostra, se tem um conhecimento prévio da distribuição dos parâmetros (Congdon, 2014).

Segundo Gelman *et al.* (2013), o processo de inferência Bayesiana se divide em 3 etapas, a primeira etapa é estabelecer um modelo de probabilidade conjunta, a segunda etapa é calcular e interpretar a distribuição *a posteriori*, que é a distribuição de probabilidade condicional aos dados do modelo, e é de interesse no final, dado os dados observados. A última etapa é avaliar o ajuste do modelo, ver até onde o modelo se adéqua aos dados, e caso ele não se adéque, fazer as modificações necessárias para que isso ocorra. Neste trabalho a inferência Bayesiana foi utilizada para a estimação dos parâmetros dos modelos e foi feita utilizando os métodos de estimação *Gibbs Sampling* (Gelfand *et al.*, 1990) e o algoritmo *Metropolis Hastings* (Metropolis *et al.*, 1953), baseado nas cadeias de Markov. Outro método que também vem sendo utilizado para esse tipo de estimação é o *Integrated Nested Laplace Approximations - INLA*, publicado em Rue *et al.* (2009). O *INLA* utiliza métodos de estimação diferentes dos métodos *MCMC* e ele se aplica a uma classe de modelos conhecidos como *modelos latentes de Gauss*. Mais informações sobre o *INLA* são encontradas no livro Faraway *et al.* (2018).

### 4.1 Teorema de Bayes

Quando queremos calcular a probabilidade de  $\theta$ , que é um parâmetro desconhecido e aleatório no qual queremos estimar a partir de uma amostra  $y$ , começamos calculando a probabilidade conjunta de  $\theta$  dado  $y$ . Podemos dizer que a inferência Bayesiana segue o princípio da verossimilhança, que afirma que, para uma dada amostra de dados quaisquer, as probabilidade  $p(y|\theta)$  que possuem a mesma função de verossimilhança, produzem a mesma inferência para  $\theta$  (Gelman *et al.*, 2013).

A distribuição *a posteriori* é construída com base no teorema de Bayes, desta maneira

$$\pi(\theta|y) = \frac{\pi(\theta, y)}{p(y)} = \frac{\pi(\theta)p(y|\theta)}{p(y)}, \quad (4.1)$$

onde  $\pi(\theta|y)$  é a distribuição *a posteriori* condicional aos dados  $y$ ,  $p(y|\theta)$ , é a distribuição amostral que define a relação entre  $\theta$  e  $y$ , conhecida como função de verossimilhança e  $\pi(\theta)$  é conhecida como a distribuição *priori* de  $\theta$ . A distribuição  $p(y)$ , é

conhecida como distribuição preditiva de  $y$ , essa função não depende do parâmetro  $\theta$  e fornece previsões para  $y$  baseando-se no conhecimento que se tem sobre  $\theta$  (Gelman *et al.*, 2013), dada por

$$p(y) = \int p(y|\theta)\pi(\theta)d\theta. \quad (4.2)$$

A distribuição  $p(y)$  pode ser omitida do cálculo da distribuição *a posteriori*, pois ela não depende de  $\theta$ . Essa simplificação transforma a igualdade em proporcionalidade e o cálculo da distribuição *a posteriori* fica (Congdon, 2014)

$$\pi(\theta|y) \propto \pi(\theta)p(y|\theta). \quad (4.3)$$

Portanto, podemos dizer que a distribuição *a posteriori* é proporcional a combinação da função de verossimilhança dos parâmetros, multiplicado pelas distribuições *a priori* de cada um dos parâmetros do modelo.

## 4.2 Construção da distribuição *a posteriori* e densidades condicionais completas

Nessa seção será usada inferência Bayesiana para fazer a estimação dos parâmetros dos modelos apresentados no Capítulo 3. Considerando agora a definição da construção da distribuição *a posteriori* (4.3), serão calculadas *as posterioris* conjuntas dos modelos com fragilidade Gama e Log-Normal.

Como vimos nos capítulos anteriores já temos a função de verossimilhança com tempos Weibull e fragilidades Gama e Log-Normal. Agora supondo que os parâmetros dos modelos seguem as seguintes distribuições *a priori*

$$\beta \sim N(\mu_\beta, \sigma_\beta^2),$$

$$\lambda \sim Gama(a_\lambda, b_\lambda),$$

$$\alpha \sim Gama(a_\alpha, b_\alpha),$$

em que  $N(., .)$  denota a densidade da distribuição normal com parâmetros  $\mu_\beta$  e  $\sigma_\beta^2$ , que são respectivamente denotados para a média e o desvio padrão.  $Gama(., .)$ , denota tanto para  $\lambda$  como para  $\alpha$ , a densidade da distribuição Gama com parâmetros  $a_\lambda, b_\lambda, a_\alpha, b_\alpha$ , onde  $a$ , representa o parâmetro de forma e  $b$  o parâmetro de escala.

Com isso podemos construir distribuição *a posteriori* conjunta, que é dada genericamente da seguinte maneira

$$\pi(\theta|y) \propto L(\theta)\pi(\beta)\pi(\lambda)\pi(\alpha)f(z_i, \sigma^2), \quad (4.4)$$

em que  $L(\theta)$  é a função de verossimilhança vista em (3.2),  $\pi(\beta)$  é a distribuição *a priori* de  $\beta$ ,  $\pi(\lambda)$  é a distribuição *a priori* de  $\lambda$ ,  $\pi(\alpha)$  é a distribuição *a priori* de  $\alpha$  e  $f(z_i, \sigma^2)$  denota a função densidade de probabilidade das fragilidades. Consideramos independência entre os parâmetros da distribuição Weibull e os coeficientes de regressão. Logo, utilizamos distribuições *a priori* independentes para os betas,  $\lambda$  e  $\alpha$ .

Fazendo substituições na equação (4.4), usando a função de verossimilhança e as funções de densidade definidas *a priori* de cada um dos parâmetros, a distribuição *a posteriori* conjunta com fragilidade Gama fica dada por

$$\begin{aligned}
L(\theta|d) &= \prod_{i=1}^n [\lambda \alpha t^{\alpha-1} z_i e^{\beta' x_i}]^{\delta_i} \cdot \exp\{-z_i e^{\beta' x_i} \lambda t^\alpha\} \\
&\times \frac{1}{\sqrt{2\pi\sigma_\beta^2}} \exp\left\{\frac{-1}{2} \left(\frac{\beta - \mu_\beta}{\sigma_\beta}\right)^2\right\} \\
&\times \frac{b_\lambda^{a_\lambda}}{\Gamma(a_\lambda)} \lambda^{a_\lambda-1} \exp(-b_\lambda \lambda) \quad , \quad (4.5) \\
&\times \frac{b_\alpha^{a_\alpha}}{\Gamma(a_\alpha)} \alpha^{a_\alpha-1} \exp(-b_\alpha \alpha) \\
&\times \frac{1}{\Gamma\left(\frac{1}{\sigma^2}\right)} \left(\frac{1}{\sigma^2}\right)^{\frac{1}{\sigma^2}} z_i^{\frac{1}{\sigma^2}-1} \exp\left(\frac{-z_i}{\sigma^2}\right).
\end{aligned}$$

onde  $d$ , é o componente de dados na amostra e é representado por  $d = (x, \delta, t_i)$ .

A distribuição *a posteriori* conjunta com fragilidade Log-Normal

$$\begin{aligned}
L(\theta|d) &= \prod_{i=1}^n [\lambda \alpha t^{\alpha-1} e^{\beta' x_i + w_i}]^{\delta_i} \exp\{e^{\beta' x_i + w_i} \lambda t^\alpha\} \\
&\times \frac{1}{\sqrt{2\pi\sigma_\beta^2}} \exp\left\{\frac{-1}{2} \left(\frac{\beta - \mu_\beta}{\sigma_\beta}\right)^2\right\} \\
&\times \frac{b_\lambda^{a_\lambda}}{\Gamma(a_\lambda)} \lambda^{a_\lambda-1} \exp(-b_\lambda \lambda) \quad , \quad (4.6) \\
&\times \frac{b_\alpha^{a_\alpha}}{\Gamma(a_\alpha)} \alpha^{a_\alpha-1} \exp(-b_\alpha \alpha) \\
&\times \frac{1}{\sqrt{2\pi s w_1}} \exp\left(\frac{-(\ln(w_i))^2}{2s^2}\right).
\end{aligned}$$

onde  $d$ , é o componente de dados na amostra e é representado por  $d = (x, \delta, t_i)$ .

Após a construção das distribuições *a posteriori* conjuntas dos modelos, para podermos entender como funciona o amostrador de Gibbs, serão obtidas as densidades condicionais completas de cada um dos parâmetros do modelo, aplicando o logaritmo nas funções (4.5) e (4.6).

Portanto, se as fragilidades  $Z_i$  seguem uma distribuição Gama, então temos as seguintes densidades condicionais completas

$$\begin{aligned}
\pi(\beta|\dots) &\propto \sum_{i=1}^n \delta_i (\beta' x_i) - \lambda \sum_{i=1}^n (z_i \lambda t^\alpha e^{\beta' x_i}) - \frac{1}{2} \left(\frac{\beta - \mu_\beta}{\sigma_\beta}\right)^2, \quad (4.7) \\
\pi(\lambda|\dots) &\propto \text{Gama}\left(\sum_{i=1}^n \delta_i + a_\lambda; \sum_{i=1}^n (-z_i t^\alpha e^{\beta' x_i}) - b_\lambda\right), \\
\pi(\alpha|\dots) &\propto \text{Gama}\left(\sum_{i=1}^n \delta_i + a_\alpha; \sum_{i=1}^n (-z_i t^\alpha e^{\beta' x_i}) - b_\alpha\right),
\end{aligned}$$

$$\pi(z_i|\dots) \propto \text{Gama}\left(\sum_{i=1}^n \delta_i + \frac{1}{\sigma}^2; \sum_{i=1}^n (\lambda t^\alpha e^{\beta' x_i})\right).$$

Se as fragilidades  $w_i \sim \text{Normal}$ , então  $Z_i \sim \text{Log} - \text{Normal}$ , e portanto as densidades condicionais completas são,

$$\pi(\beta|\dots) \propto \sum_{i=1}^n \delta_i(\beta' x_i) - \lambda \sum_{i=1}^n (\exp(\beta' x_i + w_i) t_i^\alpha) - \frac{1}{2} \left( \frac{\beta - \mu_\beta}{\sigma_\beta} \right)^2, \quad (4.8)$$

$$\pi(\lambda|\dots) \propto \text{Gama}\left(\sum_{i=1}^n \delta_i + a_\lambda; \sum_{i=1}^n (\exp(\beta' x_i + w_i) t_i^\alpha) - b_\lambda\right),$$

$$\pi(\alpha|\dots) \propto \text{Gama}\left(\sum_{i=1}^n \delta_i + a_\alpha; \sum_{i=1}^n (\exp(\beta' x_i + w_i) t_i^\alpha) - b_\alpha\right),$$

$$\begin{aligned} \pi(z_i|\dots) &\propto \sum_{i=1}^n \delta_i(w_i) \\ &+ \sum_{i=1}^n \left( -\lambda t^\alpha \exp(\beta' x_i + w_i) + \log(\sqrt{2\pi s w_i}) + \left( \frac{-(\ln(w_i))^2}{2s^2} \right) \right). \end{aligned} \quad (4.9)$$

Podemos notar que tanto nas densidades condicionais completas com fragilidade Gama, quanto nas condicionais completas com fragilidade Log-Normal, temos equações que não tem uma forma fechada (4.7), (4.8) e (4.9), e outras com forma fechada. Para esses casos o método Metropolis Hastings serve para amostrar os dados. Na próxima seção temos uma breve descrição de como funcionam.

### 4.3 Métodos de estimação - MCMC

Nesta seção será introduzida uma breve explicação sobre cadeias de Markov, sobre o amostrado *Gibbs* e sobre o algoritmo *Metropolis Hastings*. A base teórica para a elaboração desta seção foi retirada das notas de aula [Ehlers \(2011\)](#) e dos livros [Gamerman & Lopes \(2006\)](#) e [Casella & George \(1992\)](#). Uma cadeia de Markov pode ser definida como um tipo especial de processo estocástico, que trabalha com uma sequência de variáveis aleatórias ([Gamerman & Lopes, 2006](#)). Em termos mais simples, a cadeia de Markov é um processo em que temos um estado presente, um estado passado e um estado futuro, a principal característica é que a cada tempo  $t \geq 0$ , o estado seguinte da cadeia,  $\theta_{(t+1)}$ , depende do passado só através do tempo atual,  $\theta_{(t)}$ , que pode ser definido como

$$P(\theta^{(t+1)} = y | \theta^{(t)} = x, \theta^{(t-1)} = x_{t-1}, \dots, \theta^{(0)} = x_0), \quad (4.10)$$

para  $x_0, \dots, x_t, y \in S$ .

Essa equação é denominada de probabilidade de transição estacionária a um passo e é denotada como  $P_{xy}$ . Normalmente as probabilidades da equação (4.10) dependem de  $x$ ,  $y$  e  $t$ , mas quando essas probabilidades não dependem de  $t$ , diz-se que a cadeia é homogênea ([Gamerman & Lopes, 2006](#)). Podemos também definir  $P(\theta^{(t+1)} | \theta^{(t)})$  como a função de transição da cadeia ou função *kernel* ([Gamerman &](#)

Lopes, 2006).

O amostrador de *Gibbis* ficou conhecido mais amplamente no meio estatístico depois da publicação do trabalho de [Gelfand et al. \(1990\)](#), onde as técnicas que já haviam sido demonstradas no artigo de [Geman & Geman \(1984\)](#), que tratava do processamento de imagens usando algoritmos para simulações de dados de uma distribuição de Gibbs, foram aprimoradas e usadas em outras aplicações estatísticas ([Gamerman & Lopes, 2006](#)).

O amostrador de *Gibbis*, usa a função kernel para fazer as transições das cadeias e as distribuições condicionais completas dos parâmetros ([Gamerman & Lopes, 2006](#)). Supondo que temos uma distribuição  $\pi(\theta)$  em que  $\theta$  pode ser um escalar, uma matriz ou um vetor e é representado por  $\theta = (\theta_1, \dots, \theta_d)'$ , e com isso, também supondo que as condicionais completas são representadas por  $\pi_i(\theta_i) = \pi(\theta_i | \theta_{-i})$  com  $i$  indo do  $i = (1, \dots, p)$  e  $\theta'_{-i} = (\theta_1, \dots, \theta_{i-1}, \theta_{i+1}, \dots, \theta_p)$  ([Gamerman & Lopes, 2006](#)).

O amostrador de *Gibbis* funciona da seguinte maneira:

1. É iniciado o contar da cadeia com  $t = 1$  e os valores iniciais  $\theta^0 = (\theta_1^0, \dots, \theta_p^0)'$ ;
2. Novos valores de  $\theta^t$  são gerados sucessivamente a partir de  $\theta^{t-1}$ , da seguinte maneira:

$$\lambda^{(t)} \sim \pi(\lambda | \alpha^{(t-1)}, \beta_1^{(t-1)}, \beta_2^{(t-1)}, Z_1^{(t-1)}, \dots, Z_{40}^{(t-1)}),$$

$$\alpha^{(t)} \sim \pi(\alpha | \lambda^{(t)}, \beta_1^{(t-1)}, \beta_2^{(t-1)}, Z_1^{(t-1)}, \dots, Z_{40}^{(t-1)}),$$

$$\beta_1^{(t)} \sim \pi(\beta_1 | \lambda^{(t)}, \alpha^{(t)}, \beta_2^{(t-1)}, Z_1^{(t-1)}, \dots, Z_{40}^{(t-1)}),$$

$$\beta_2^{(t)} \sim \pi(\beta_2 | \lambda^{(t)}, \alpha^{(t)}, \beta_1^{(t)}, Z_1^{(t-1)}, \dots, Z_{40}^{(t-1)}),$$

$$Z_1^{(t)} \sim \pi(Z_1 | \lambda^{(t)}, \alpha^{(t)}, \beta_1^{(t)}, \beta_2^{(t)}, \dots, Z_{40}^{(t-1)}),$$

⋮

$$Z_{40}^{(t)} \sim \pi(Z_{40} | \lambda^{(t)}, \alpha^{(t)}, \beta_1^{(t)}, \beta_2^{(t)}, Z_1^{(t)}, \dots, Z_{39}^{(t)}).$$

3. Incrementa-se  $t$  e repete-se o item 2 até um tamanho que garanta a convergência da cadeia.

A probabilidade de aceitação das probabilidades é igual a 1, por isso a cadeia sempre mudará de estado ([Gamerman & Lopes, 2006](#)).

O algoritmo *Metropolis Hastings* foi aprimorado por [Hastings \(1970\)](#), a partir da publicação [Metropolis et al. \(1953\)](#). Quando a distribuição da densidade condicional completa não tem uma forma conhecida, o *Metropolis Hastings* é utilizado para gerar as amostras das distribuições *a posteriori*. Uma das características desse algoritmo

é que precisamos conhecer a distribuição parcialmente para podermos implementá-lo (Gamerman & Lopes, 2006). Basicamente ele funciona da seguinte maneira, a partir de uma distribuição proposta  $p(\cdot)$ , que deve ser muito parecida com a distribuição  $p(\theta|y)$ , um valor  $\theta'$  é gerado de um  $\theta$  inicial em um tempo  $t$ . Esse mecanismo converge as amostras para à distribuição *a posteriori* do modelo com uma probabilidade de aceitação ou rejeição, isto significa que dependendo do valor gerado da amostra a cadeia pode avançar ou se manter no mesmo lugar (Casella & George, 1992).

## 4.4 Critérios de avaliação - DIC E WAIC

Nesta seção serão abordados dois tipos de critérios de avaliação de ajuste de modelos estatísticos Bayesianos, o *Deviance Information Criterion (DIC)*, proposto em Spiegelhalter *et al.* (2002) e, *Watanabe-Akaike Information Criterion (WAIC)*, proposto em Watanabe (2010).

O primeiro critério de avaliação estudado será o Deviance Information Criterion. Nesse critério, valores menores indicam um melhor ajuste do modelo. O DIC se baseia em duas partes: uma que serve para penalizar o modelo e que considera a complexidade das estimativas do número de parâmetros, e outra parte que mede o quão bom é o ajuste do modelo (Gelman *et al.*, 2013). Neste critério a deviance é utilizada para o cálculo e ela é representada por

$$D(\theta|y) = -2 \log p(y|\hat{\theta}), \quad (4.11)$$

em que  $y = (y_1, y_2, \dots, y_n)$  são os dados.

DIC é uma medida de ajuste que pode ser aplicadas em modelos Bayesianos e é formalizado por Dey *et al.*, como (Dey *et al.*, 2000)

$$DIC = \bar{D} + pd, \quad (4.12)$$

onde,

$$\bar{D} = E(D(\theta|y)) = E[-2 \log p(y|\hat{\theta})], \quad (4.13)$$

é o desvio médio *a posteriori* esperado que resume e quantifica o ajuste do modelo e

$$pd = \frac{1}{2} \text{Var}[D(\theta|y)], \quad (4.14)$$

é interpretado como o número efetivo de parâmetros, mensurados no modelo.

O outro critério conhecido como *Watanabe-Akaike Information Criterion (WAIC)*. Nesse critério, valores maiores indicam um melhor ajuste do modelo. O WAIC utiliza as distribuições *a posteriori* e também uma medida de complexidade dada pelo número efetivo de parâmetros (Gelman *et al.*, 2013), e é definido como

$$WAIC = -2 \widehat{elpd}, \quad (4.15)$$

em que,  $elpd$  é o logaritmo do valor esperado da densidade preditiva de um indivíduo, e é representado por

$$\widehat{elpd} = \widehat{lpd} - \hat{p}_{waic}. \quad (4.16)$$

O termo  $\widehat{lpd}$  é dado por

$$\widehat{lpd} = \sum_{i=1}^n \log \left( \frac{1}{S} \sum_{s=1}^S p(y_i | \theta_{(s)}) \right), \quad (4.17)$$

que representa o logaritmo da densidade preditiva do indivíduo  $i$ .

O segundo termo é definido como

$$\hat{p}_{waic} = \sum_{i=1}^n \text{var} \left( \log \left( p(y_i | \theta_{(s)}) \right) \right), \quad (4.18)$$

que é definido como uma estimativa do número de parâmetros do modelo. E  $s = 1, 2, \dots, S$  representa o número de interações do modelo (Gelman *et al.*, 2013).

## 5 Simulação e Aplicação

Neste capítulo serão apresentadas, as simulações de dados para dois modelos com fragilidades Gama e Log-Normal, apresentados no Capítulo 3. Inicialmente gerando dados, posteriormente sendo feita a análise dos resultados e por fim os modelos serão aplicados em dois bancos de dados reais.

Foram gerados dois conjuntos de dados, um com fragilidades que seguiam uma distribuição normal, para ser utilizado com o modelo de fragilidade Log-Normal, e outro em que as fragilidades seguiam uma distribuição Gama, para ser utilizado com a fragilidade Gama.

Cada banco de dados gerado contém 200 observações, elas estão separadas em 40 grupos, cada grupo contém 5 observações. Cada banco também tem duas covariáveis  $x_1$  seguindo uma distribuição de probabilidade  $N(0, 1)$  e a covariável  $x_2$  seguindo uma distribuição de probabilidade  $Bin(1, 0.5)$ . Cada indivíduo foi sorteado de forma aleatória nos grupos. Os tempos seguem distribuição Weibull e foram gerados utilizando a fórmula a seguir, e se encontra de forma mais detalhada no apêndice deste trabalho (Bender *et al.*, 2005)

$$t = \left( \frac{-\log(U)^\alpha}{\lambda e^{x\beta + w_i}} \right)^{1/\alpha}, \quad (5.1)$$

em que  $\lambda$  e  $\alpha$  denotam os parâmetros de forma e escala, respectivamente, de uma distribuição Weibull,  $\beta$  denota nos coeficientes das covariáveis,  $w$  denota as fragilidades de cada um dos grupos e  $U$  denota valores gerados de uma distribuição Uniforme,  $U(0,1)$ . Um passo a passo de como foi feita a geração dos dados se encontra no apêndice deste trabalho.

Foram simulados 500 bancos de dados utilizando a fragilidade Gama e 500 bancos de dados utilizando a fragilidade Log-Normal, afim de comparar as resultantes com os valores reais utilizados para os parâmetros. Este procedimento é necessário para avaliar se o modelo está funcionando da maneira correta.

Para a aplicação de inferência Bayesiana o software JAGS (Plummer, 2015) foi utilizado. Nele foi implementado os dois modelos de fragilidade utilizando as verossimilhanças vistas anteriormente. Para a definição das cadeias foi definido um *burn-in* de 2000, pois observou-se que no gráfico das estimativas *a posteriori*, houve convergência após esse número de interações. O *lag* escolhido foi de 40, baseado na visualização dos gráficos de autocorrelação. Considerando uma amostra *a posteriori* de 5000 interações.

A tabela 5.1 nos mostra os valor reais dos parâmetros  $\beta_1, \beta_2, \lambda$  e  $\alpha$ , a média das

médias à posteriori e o viés das 500 médias à posteriori geradas, a média e o viés das medianas, a média e o viés das modas, a média dos desvios padrões e a média dos intervalos de probabilidade.

Observando os resultados gerados, nota-se que as estimativas *a posteriori* ficaram praticamente iguais aos valores iniciais que foram passados para os parâmetros. O viés dos parâmetros estão todos muito próximos de zero, isso mostra que as estimativas *a posteriori* estão muito pouco viesados. Com isso podemos concluir que o modelo ajustado está sendo executado da maneira correta e desempenhando um bom ajuste.

Tabela 5.1: Resultado das simulações para o modelo de fragilidade Log-Normal.

	Valor real	Média(viés%)	Mediana(viés%)	Moda(viés%)	DP	Hpd
$\beta_1$	1,2	1,218(1,6%)	1,218(1,5%)	1,216(1,3%)	0,126	[0,973;1,465]
$\beta_2$	-1,4	-1,432(-2,3%)	-1,431(-2,2%)	-1,430(-2,2%)	0,217	[-1,856;-1,009]
$\lambda$	1	1,035(3,5%)	1,015(1,5%)	0,977(-2,3%)	0,212	[0,648;1,458]
$\alpha$	1	1,017 (1,8%)	1,017(1,7%)	1,015(1,5%)	0,069	[0,886;1,153]

Viés relativo= $(\hat{\theta} - \theta)/|\theta|$ , onde  $\theta = \beta_1, \beta_2, \lambda, \alpha$ ; DP=desvio padrão, Hpd(Highest Probability Density) = Mais alta densidade de probabilidade.

Os gráficos Box-Plot, da Figura 5.1, nos apresentam o resultado das média, medianas e modas *a posteriori* dos parâmetros  $\beta_1$  e  $\beta_2$ . Visualmente, podemos notar que tanto para o  $\beta_1$  quanto para  $\beta_2$  os valores estimados estão muito próximos dos valores reais que foram informados. Podemos ver também, pela posição da mediana, que os dados são simétricos possuem outliers em todos os gráficos tanto na cauda superior quanto na inferior, mas é no gráfico da moda que eles aparecem em maior quantidade.

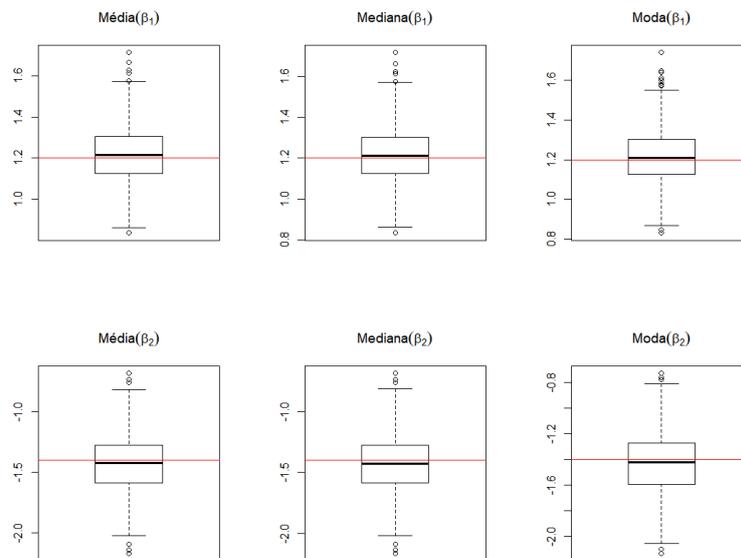


Figura 5.1: Média, mediana e moda *a posteriori* dos parâmetros  $\beta_1$  e  $\beta_2$  com ajuste do modelo com fragilidade Log-Normal.

No caso dos parâmetros  $\lambda$  e  $\alpha$ , os gráficos, da Figura 5.2, mostram que para o parâmetro  $\lambda$ , a média, mediana e moda *a posteriori* ficaram bem próximas do valor real, os dados são simétricos, e notamos que existem em ambos os gráfico outliers na cauda superior. Nos gráficos do  $\alpha$ , as estimativas também ficaram próximas, os dados são simétricos e os outliers estão em maior quantidade na média e na mediana, tanto na cauda inferior quanto na superior.

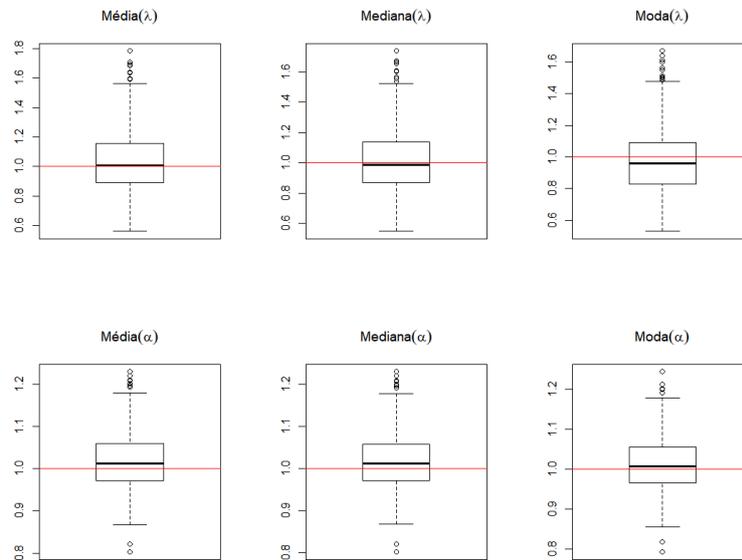


Figura 5.2: Média, mediana e moda *a posteriori* dos parâmetros  $\lambda$  e  $\alpha$  com ajuste do modelo com fragilidade Log-Normal.

Nos gráficos para os vieses, Figura 5.3, também há presença de outliers, mas os vieses estão próximos do valor zero, como vimos anteriormente na Tabela 5.1, o que representa uma boa modelagem.

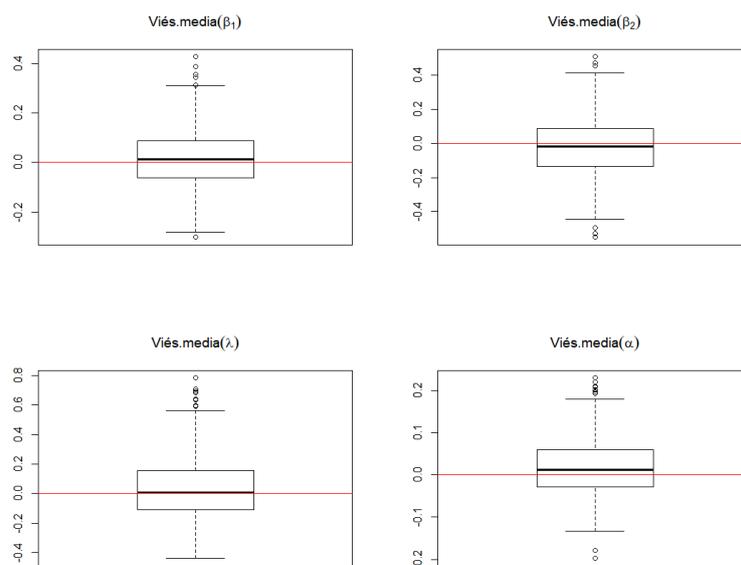


Figura 5.3: Viés das médias *a posteriori* dos parâmetros  $\beta_1$ ,  $\beta_2$ ,  $\lambda$  e  $\alpha$  com ajuste do modelo com fragilidade Log-Normal.

Foram calculados também os dois critérios de avaliação DIC e WAIC, vistos na Seção 4.4, e foram obtidos os seguintes resultados na Tabela 5.2

Tabela 5.2: DIC e WAIC das simulações para o modelo de fragilidade Log-Normal.

Critérios	DIC	WAIC
Médias	341,82	327,84

WAIC foi padronizado para ficar igual ao DIC:  $DIC \equiv (-2WAIC)$

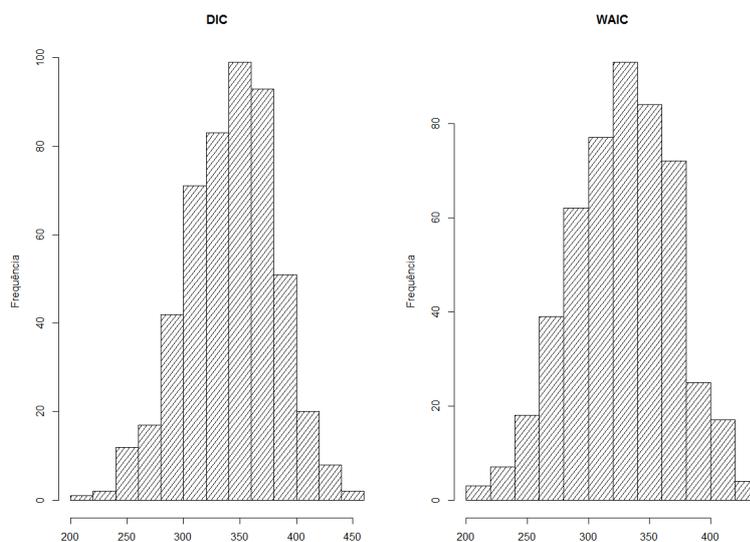


Figura 5.4: Histograma dos 500 valores de DIC e WAIC gerados da fragilidade Log-Normal

Os valores de WAIC foram padronizados para que ficassem na mesma escala dos valores dos WAIC, pela fórmula  $DIC \equiv (-2WAIC)$  (Gelman *et al.*, 2013). Na Figura 5.4, nota-se que os valores dos dois critérios ficaram bem próximos, há também uma assimetria à esquerda nos dois gráficos.

As fragilidades Log-Normal de cada um dos 40 grupos transformadas da forma  $z_i = \exp(w_i)$ , estão apresentadas na Tabela 5.3 junto com resultados das médias, modas e medianas *a posteriori* das simulações. Podemos notar que existem grupos que tem médias de fragilidades maiores que 1 e outros com médias de fragilidades menores do que 1.

Giolo & Colosimo (2006, p.287), definiram a fragilidade de um indivíduo da seguinte maneira:

Para valores de  $z_j$  maiores que 1, tem-se que os indivíduos dentro dos grupos correspondentes a esses valores tendem a apresentar sobrevivência mais curta do que a predita pelo modelo de independência em que os  $z_j$ 's são assumidos serem todos iguais a 1. Por outro lado, para valores de  $z_j$  menores que 1, tem-se que os indivíduos deste grupo tendem a apresentar sobrevivência mais longa que a predita pelo modelo que assume independência entre os tempos (Giolo & Colosimo, 2006).

Tabela 5.3: Fragilidades com ajuste do modelo de fragilidade Log-Normal das simulações

Fragilidades	Média	Mediana	Moda	DP	Hpd
Z <sub>1</sub>	0,988	1,012	1,068	1,679	[0,354;2,645]
Z <sub>2</sub>	1,031	1,056	1,104	1,688	[0,365;2,789]
Z <sub>3</sub>	0,994	1,018	1,069	1,683	[0,354;2,670]
Z <sub>4</sub>	0,968	0,992	1,026	1,680	[0,346;2,591]
Z <sub>5</sub>	1,058	1,084	1,141	1,685	[0,376;2,848]
Z <sub>6</sub>	0,982	1,006	1,050	1,684	[0,349;2,643]
Z <sub>7</sub>	0,986	1,010	1,053	1,685	[0,351;2,659]
Z <sub>8</sub>	0,992	1,016	1,065	1,687	[0,352;2,678]
Z <sub>9</sub>	1,015	1,040	1,088	1,681	[0,363;2,723]
Z <sub>10</sub>	0,991	1,014	1,059	1,680	[0,355;2,664]
Z <sub>11</sub>	0,997	1,021	1,065	1,684	[0,355;2,684]
Z <sub>12</sub>	1,044	1,069	1,113	1,687	[0,370;2,82]0
Z <sub>13</sub>	0,957	0,980	1,027	1,680	[0,343;2,570]
Z <sub>14</sub>	1,007	1,031	1,074	1,681	[0,360;2,702]
Z <sub>15</sub>	0,984	1,008	1,055	1,680	[0,352;2,639]
Z <sub>16</sub>	1,039	1,063	1,103	1,677	[0,374;2,785]
Z <sub>17</sub>	1,012	1,036	1,083	1,687	[0,359;2,738]
Z <sub>18</sub>	1,053	1,079	1,127	1,691	[0,371;2,856]
Z <sub>19</sub>	0,965	0,988	1,023	1,682	[0,344;2,596]
Z <sub>20</sub>	0,955	0,978	1,023	1,686	[0,339;2,578]
Z <sub>21</sub>	1,056	1,082	1,139	1,682	[0,376;2,831]
Z <sub>22</sub>	1,000	1,025	1,085	1,685	[0,356;2,696]
Z <sub>23</sub>	1,054	1,080	1,124	1,689	[0,372;2,857]
Z <sub>24</sub>	0,943	0,966	1,016	1,682	[0,336;2,531]
Z <sub>25</sub>	0,990	1,015	1,066	1,688	[0,351;2,679]
Z <sub>26</sub>	0,920	0,942	0,982	1,678	[0,331;2,465]
Z <sub>27</sub>	0,988	1,012	1,049	1,681	[0,354;2,660]
Z <sub>28</sub>	1,015	1,040	1,096	1,680	[0,363;2,721]
Z <sub>29</sub>	0,988	1,012	1,054	1,680	[0,353;2,647]
Z <sub>30</sub>	0,992	1,015	1,060	1,685	[0,351;2,666]
Z <sub>31</sub>	0,966	0,989	1,031	1,686	[0,343;2,603]
Z <sub>32</sub>	1,023	1,047	1,094	1,684	[0,364;2,756]
Z <sub>33</sub>	1,024	1,049	1,110	1,679	[0,368;2,749]
Z <sub>34</sub>	1,042	1,068	1,107	1,685	[0,371;2,816]
Z <sub>35</sub>	1,018	1,044	1,094	1,687	[0,361;2,749]
Z <sub>36</sub>	0,976	1,000	1,043	1,686	[0,347;2,642]
Z <sub>37</sub>	0,969	0,993	1,041	1,684	[0,344;2,605]
Z <sub>38</sub>	1,009	1,033	1,088	1,678	[0,362;2,696]
Z <sub>39</sub>	0,996	1,020	1,064	1,677	[0,357;2,658]
Z <sub>40</sub>	1,032	1,057	1,105	1,678	[0,371;2,769]

DP=Desvio padrão; Hpd(Highest Probability Density) = Mais alta densidade de probabilidade.

Os mesmos procedimentos foram feitos para o ajuste do modelo com fragilidade Gama. A Tabela 5.4 contém os valores reais dos parâmetros, e as médias das médias *a posteriori* e o viés, a média das medianas *a posteriori* e o viés, a média das modas *a posteriori* e o viés. A tabela também contém a média dos desvios padrão e as médias dos intervalos. Como no ajuste do modelo com fragilidade Log-Normal, percebemos que o modelo ajustado também se adequou bem aos dados, pois as estimativas *a posteriori* ficaram muito próximos dos reais.

Tabela 5.4: Resultado das simulações para o modelo de fragilidade Gama.

	Valor real	Média(viés%)	Mediana(viés%)	Moda(viés%)	DP	Hpd
$\beta_1$	1,2	1,223(-1,9%)	1,222(-1,8%)	1,218(1,5%)	0,135	[0,960;1,488]
$\beta_2$	-1,4	-1,403(0,2%)	-1,401(-0,1%)	-1,400(-0,02%)	0,233	[-1,859; -0,951]
$\lambda$	1	1,007(-0,7%)	0,982(-1,8%)	0,938(-6,2%)	0,220	[0,613;1,446]
$\alpha$	1	1,01(-1,1%)	1,009(0,9%)	1,007(0,7%)	0,074	[0,868;1,156]

Viés relativo= $(\hat{\theta} - \theta)/|\theta|$ , onde  $\theta = \beta_1, \beta_2, \lambda, \alpha$ ; DP=desvio padrão, Hpd(Highest Probability Density) = Mais alta densidade de probabilidade.

No caso do modelo com fragilidade Gama os gráficos, na Tabela 5.5, da média, mediana e moda *a posteriori* do  $\beta_1$  ficaram próximos ao valor real, os dados dão simétricos e há poucos outliers. Para o parâmetro  $\beta_2$  nota-se pelos gráficos, que as médias, medianas e modas *a posteriori* ficam praticamente na mesma linha do valor real, nota-se também uma leve assimetria positiva dos dados. Os outliers estão presentes em maior quantidade, quando comparado com os gráficos do  $\beta_1$ .

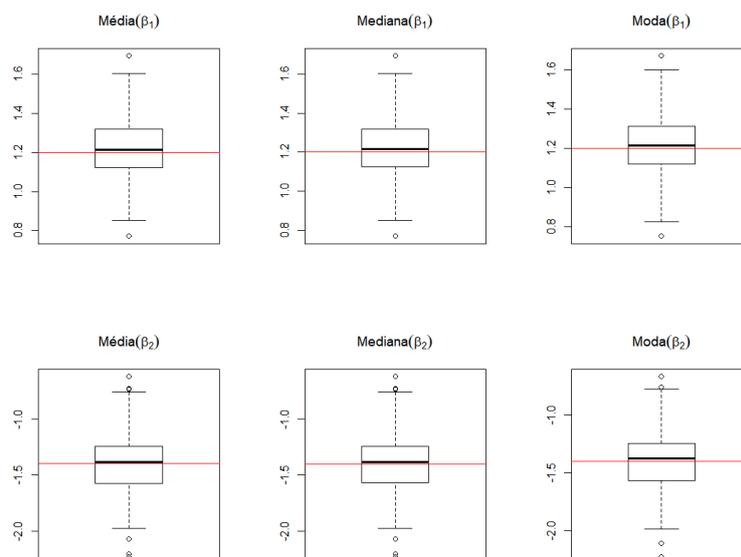


Figura 5.5: Média, mediana e moda *a posteriori* dos parâmetros  $\beta_1$  e  $\beta_2$  com ajuste do modelo com fragilidade Gama.

Nos gráficos da Figura 5.6 do parâmetro  $\lambda$ , a mediana e a moda *a posteriori* estão próximas do valor real, mas no gráfico da moda, percebemos uma leve subes-

timação desse valor. Os dados dos três gráficos parecem ser simétricos e a presença de outliers está mais visível na cauda superior. Já para o parâmetro  $\alpha$ , todos os gráficos mostram valores próximos do valor real, os dados são simétricos e também nota-se que a presença de outliers quase não existe.

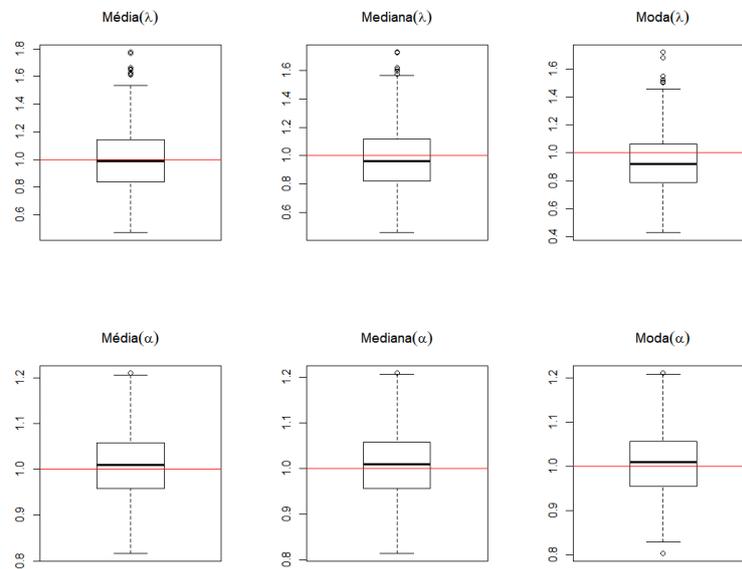


Figura 5.6: Média, mediana e moda *a posteriori* dos parâmetros  $\lambda$  e  $\alpha$  com ajuste do modelo com fragilidade Gama.

A Figura 5.7, com os gráficos das médias *a posteriori* dos vieses também mostram que eles estão centrados próximo de zero, isso é um sinal de que o modelo está bem ajustado aos dados, como no caso do modelo Log-Normal. Também é possível observar que, nos gráficos do parâmetro  $\alpha$  é onde se encontram o menor número de outliers.

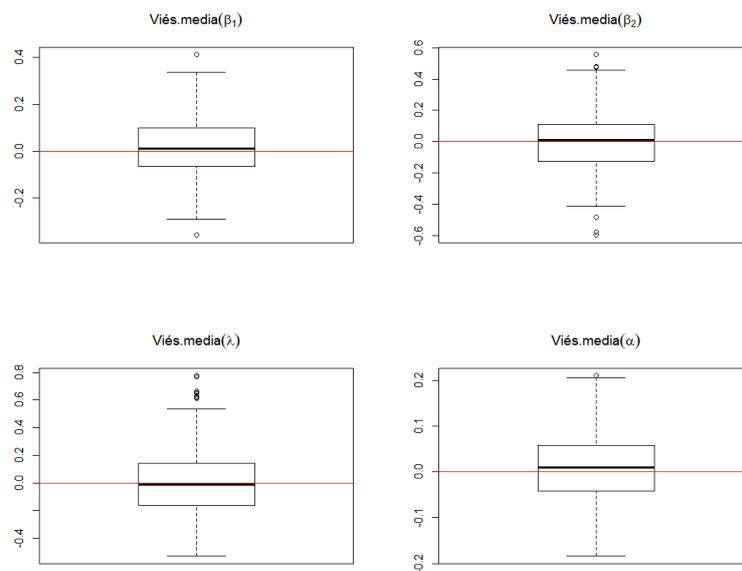


Figura 5.7: Viés das médias dos parâmetros  $\beta_1$ ,  $\beta_2$ ,  $\lambda$  e  $\lambda$  com ajuste do modelo com fragilidade Gama.

Também foram calculados os critérios DIC e WAIC fazendo a padronização pela fórmula  $DIC \equiv (-2WAIC)$ . O histograma dos 500 valores de DIC e WAIC se encontram na Figura (5.8). Os valores das médias dos DIC e WAIC estão bem próximos.

Tabela 5.5: DIC e WAIC das simulações com modelo de fragilidade Gama

Critérios	DIC	WAIC
Médias	375,68	362,44

WAIC foi padronizado para ficar igual ao DIC:  $DIC \equiv (-2WAIC)$

Pelo histograma dos 500 DIC e WAIC calculados das simulação, após padronização, é possível notar que eles são muito parecido, o gráfico do DIC parece ser assimétrico a esquerda e o gráfico do WAIC assimétrico à direita.

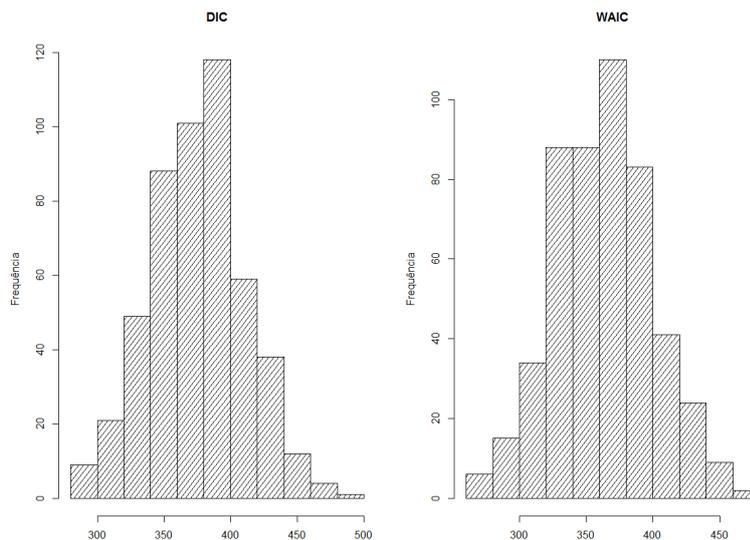


Figura 5.8: Histograma dos 500 valores de DIC e WAIC gerados da fragilidade Gama.

A Tabela 5.6 com as fragilidades do modelos gama dos 40 grupos, apresentam valores bem próximos de 1, sendo que a menor fragilidade é do grupo 12 (0,925) e o maior valor de fragilidade é do grupo 28 (1,070), olhando as médias *a posteriori*.

Tabela 5.6: Fragilidades com ajuste do modelo de fragilidade Gama

Fragilidades	Média	Mediana	Moda	DP	Hpd inf	Hpd sup
$Z_1$	0,933	0,849	0,703	0,475	0,180	1,870
$Z_2$	1,009	0,922	0,762	0,503	0,206	2,002
$Z_3$	1,039	0,950	0,787	0,517	0,216	2,060
$Z_4$	1,014	0,926	0,772	0,505	0,206	2,010
$Z_5$	1,017	0,929	0,764	0,507	0,205	2,016
$Z_6$	1,037	0,947	0,789	0,518	0,206	2,057
$Z_7$	1,001	0,913	0,748	0,503	0,199	1,992
$Z_8$	0,993	0,905	0,746	0,498	0,196	1,973
$Z_9$	0,959	0,873	0,718	0,488	0,186	1,922
$Z_{10}$	1,047	0,955	0,790	0,522	0,212	2,076
$Z_{11}$	1,009	0,922	0,761	0,502	0,206	2,000
$Z_{12}$	0,925	0,842	0,688	0,470	0,181	1,851
$Z_{13}$	1,035	0,946	0,782	0,515	0,210	2,050
$Z_{14}$	0,976	0,889	0,736	0,494	0,190	1,951
$Z_{15}$	0,995	0,907	0,743	0,504	0,195	1,991
$Z_{16}$	0,931	0,847	0,688	0,474	0,179	1,866
$Z_{17}$	1,012	0,924	0,770	0,506	0,203	2,009
$Z_{18}$	1,017	0,930	0,769	0,506	0,207	2,015
$Z_{19}$	0,968	0,882	0,720	0,489	0,191	1,931
$Z_{20}$	0,932	0,850	0,694	0,470	0,183	1,856
$Z_{21}$	1,031	0,943	0,786	0,511	0,213	2,037
$Z_{22}$	0,998	0,911	0,752	0,499	0,202	1,982
$Z_{23}$	1,034	0,945	0,782	0,513	0,210	2,045
$Z_{24}$	1,000	0,913	0,744	0,501	0,201	1,988
$Z_{25}$	1,001	0,914	0,744	0,502	0,200	1,989
$Z_{26}$	0,966	0,880	0,722	0,490	0,187	1,932
$Z_{27}$	0,968	0,883	0,724	0,485	0,194	1,923
$Z_{28}$	1,070	0,977	0,810	0,533	0,218	2,122
$Z_{29}$	1,047	0,957	0,787	0,520	0,214	2,071
$Z_{30}$	1,046	0,956	0,791	0,519	0,215	2,071
$Z_{31}$	1,006	0,918	0,751	0,504	0,202	2,001
$Z_{32}$	1,011	0,922	0,756	0,506	0,203	2,009
$Z_{33}$	1,024	0,936	0,772	0,510	0,208	2,030
$Z_{34}$	0,972	0,888	0,730	0,486	0,194	1,929
$Z_{35}$	1,024	0,935	0,779	0,511	0,209	2,033
$Z_{36}$	1,013	0,925	0,767	0,506	0,206	2,011
$Z_{37}$	1,014	0,925	0,767	0,507	0,204	2,012
$Z_{38}$	0,958	0,874	0,720	0,482	0,190	1,906
$Z_{39}$	1,029	0,941	0,779	0,510	0,211	2,034
$Z_{40}$	1,031	0,943	0,786	0,511	0,211	2,037

DP=Desvio padrão; Hpd(Highest Probability Density) = Mais alta densidade de probabilidade.

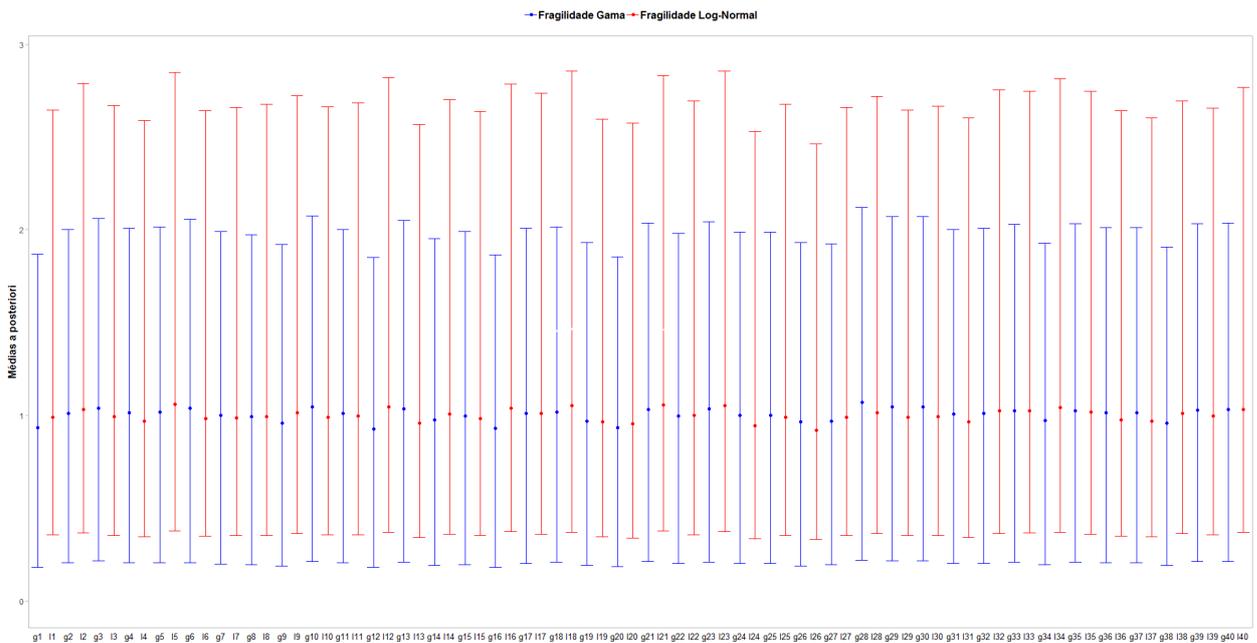


Figura 5.9: Gráfico das distribuições *a posteriori* das Fragilidades Gama e Log-Normal para os 40 grupos das simulações.

A figura (5.9) mostra as médias *a posteriori* e os intervalos HpD (Highest Probability Density), que são denominados em português como a mais alta densidade de probabilidade, das fragilidades. No gráfico podemos notar que a amplitude dos intervalos hpD são maiores nos grupos com fragilidades Log-Normal, quando comparados com os grupos que têm fragilidades Gama.

## 5.1 Aplicação em bancos de dados reais

### 5.1.1 Aplicação dos modelos em dados sobre câncer de pulmão

Um dos bancos de dados escolhidos para a aplicação dos modelos contém observações de indivíduos em estado avançado de câncer de pulmão, denominado *Lung*, foi retirado do pacote *Survival*, do software R. O banco foi publicado no artigo [Loprinzi et al. \(1994\)](#), teve como objetivo avaliar se a informação descritiva de um questionário preenchido pelo paciente poderia fornecer informações prognósticas independentes daquelas já obtidas pelo médico do paciente.

As variáveis do banco são:

1. inst = Código da instituição;
2. tempo = tempo em dias até o óbito;
3. status = indicadora de falha ou censura;
4. idade = idade em anos;

5. sexo = sexo;
6. ph.ecog = pontuação que varia entre boa e óbito, pelo critério do ECOG;
7. ph.karno = pontuação que varia entre ruim e bom que avalia desempenho pelo critério de Karnofsky e é avaliado pelo médico;
8. pat.karno = pontuação que varia entre ruim e bom que avalia desempenho pelo critério de Karnofsky e é avaliado pelo paciente;
9. cal.consu = calorias consumidas durante as refeição;
10. perda.peso = a perda de peso em seis meses.

O banco contém 228 indivíduos e 10 variáveis, dentre elas temos 3 variáveis que são classificadas como critérios de avaliação. A variável *ph.ecog* nos fornece valores de 0 a 5, em que 5 representaria indivíduos que morreram antes de serem avaliados, pelo critério *ECOG* (Oken *et al.*, 1982). As variáveis *ph.karno* e *pat.karno* são classificadas pelo critério de *Karnofsky* (Karnofsky, 1949), que tem uma escala de pontuação que vai de 0 até 100, quanto menor o valor pior é o estado de saúde do indivíduo.

Para ajustar os modelos deste trabalho, optou-se por retirar do banco as informações que continham dados ausentes, que são os dados que não foram informados no banco de dados, contabilizando 61 observações. Com isso o banco final ficou com 167 indivíduos. A variável *inst*, foi recodificada pois as instituições estavam com uma numeração que não seguia uma ordem. A variável *status*, estava como 1 e 2 nas categorias, foi transformada em 0 e 1. A variável *sex* estava com valores 1 e 2, foi recodificada para 0 e 1. Somente a variável *ph.ecog* foi retirada do modelo, todas as outras foram utilizadas.

Podemos perceber 47 (28%) dados censurados e 120 (72%) falhas. Homens representam 103 (62%) dos indivíduos, e entre eles 21 (20%) são censura e 82 (80%) falhas. As mulheres representam 64 (38%) dos indivíduos, e entre elas 26 (41%) são censuras e 38 (59%) falhas. As médias dos tempos de falha foi 288,74 dias e as médias das censuras foi 364,04 dias. A seguir, na Tabela 5.7, será apresentada uma breve análise descritiva dos dados.

Tabela 5.7: Análise descritiva das variáveis contínuas do banco sobre câncer de pulmão

	Mín.	1ºQ	Mediana	3ºQ	Média	Máx.
Idade	39	57	64	62,5	70	82
Cal.consu	90,0	619,0	975,0	929,1	1162,5	26000,0
Tempo	5,0	174,5	268,0	309,9	419,5	1022,0
Perda.peso	-24,0	0,0	7,0	9,7	15,0	68,0

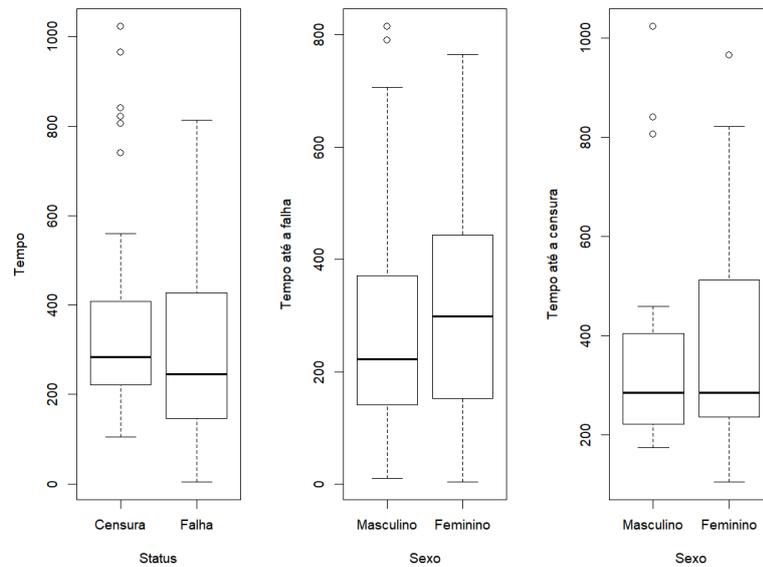


Figura 5.10: Distribuição dos tempos de falha e censura por sexo para os dados sobre Câncer de pulmão.

Nos Box-Plot da Figura 5.10 foi feita uma comparação utilizando os sexos dos indivíduos e o tempo em dias que eles levaram até sofrerem uma falha ou uma censura. No primeiro Box-Plot, nota-se que, de forma geral, os indivíduos levaram mais tempo para sofrer uma falha do que um censura. O segundo Box-Plot compara os sexos, em relação aos tempos de falha e o último gráfico também compara os sexos, mas agora olhando os tempos de censura. Observando o segundo gráfico é possível notar que os indivíduos do sexo feminino tem um tempo de falha maior do que os indivíduos do sexo masculino, mas quando olhamos os tempos de censura nota-se que as medianas estão praticamente iguai.

A fim de compreender melhor os resultados das aplicações dos modelos de fragilidade Log-normal e Gama, os resultados serão apresentados em uma sequência.

As Tabelas 5.8 e 5.9 apresentam os valores estimados dos parâmetros que foram incluídos nos modelos. Elas nos fornecem as médias, medianas, modas, desvios padrão e o intervalo de probabilidade *a posteriori*. As estimações dos parâmetros nos dois modelos ficaram bem parecidas, em ambas as tabelas o parâmetro  $\lambda$  apresenta estimativas relativamente pequenas.

Tabela 5.8: Resultados para o modelo de fragilidade Log-Normal para os dados sobre câncer de pulmão

	Média	Mediana	Moda	DP	Hpd
$\beta_{idade}$	0,076	0,074	0,065	0,117	[-0,141;0,312]
$\beta_{sexo}$	-0,272	-0,270	-0,264	0,105	[-0,475;-0,067]
$\beta_{ph.karno}$	-0,049	-0,050	-0,021	0,113	[-0,266;0,177]
$\beta_{pat.karno}$	-0,334	-0,334	-0,317	0,126	[-0,588;-0,095]
$\beta_{cal.consu}$	-0,017	-0,016	-0,022	0,106	[-0,217;0,197]
$\beta_{perda.peso}$	-0,128	-0,127	-0,154	0,105	[-0,324;0,083]
$\lambda$	$1,9 \cdot 10^{-4}$	$1,5 \cdot 10^{-4}$	$0,8 \cdot 10^{-4}$	$1,4 \cdot 10^{-4}$	$[2 \cdot 10^{-5};4,8 \cdot 10^{-4}]$
$\alpha$	1,440	1,439	1,435	0,109	[1,237;1,655]

Viés relativo= $(\hat{\theta} - \theta)/|\theta|$ , onde  $\theta = \beta, \lambda, \alpha$ ; DP=desvio padrão, Hpd(Highest Probability Density) = Mais alta densidade de probabilidade.

Na Tabela 5.8, observando quais intervalos de confiança *a posteriori* não continham o valor zero, no modelo com fragilidade Log-Normal aplicado no banco de dados sobre câncer de pulmão. As covariáveis que tiveram efeito significativo no modelo foram,  $\beta_{sexo}$  e  $\beta_{pat.karno}$ . Tomando agora o exponencial do  $E(1 - e^{\beta_{sexo}}) = 23,4\%$ , para um indivíduo do sexo feminino, estima-se um tempo até a falha reduz em 23,4% a função taxa de falha. Já a covariável  $\beta_{pat.karno}$ , que é a pontuação de avaliação *Karnofsky* feita pelo paciente, o valor fica  $E(1 - e^{\beta_{pat.karno}}) = 27,8\%$ , o que significa uma redução de 27,6% na função taxa de falha.

Tabela 5.9: Resultados para o modelo de fragilidade Gama para os dados sobre câncer de pulmão

	Média	Mediana	Moda	DP	Hpd
$\beta_{idade}$	0,072	0,070	0,054	0,116	[-0,155;0,301]
$\beta_{sexo}$	-0,269	-0,268	-0,276	0,102	[-0,474;-0,073]
$\beta_{ph.karno}$	-0,048	-0,047	-0,027	0,113	[-0,268;0,170]
$\beta_{pat.karno}$	-0,336	-0,336	-0,299	0,127	[-0,578;-0,085]
$\beta_{cal.consu}$	-0,015	-0,012	-0,003	0,107	[-0,228;0,189]
$\beta_{perda.peso}$	-0,127	-0,125	-0,133	0,103	[-0,335;0,067]
$\lambda$	$2,1 \cdot 10^{-4}$	$1,7 \cdot 10^{-4}$	$1,5 \cdot 10^{-4}$	$1,6 \cdot 10^{-4}$	$[2,3 \cdot 10^{-4};5,3 \cdot 10^{-4}]$
$\alpha$	1,447	1,446	1,417	0,107	[1,247;1,661]

Viés relativo= $(\hat{\theta} - \theta)/|\theta|$ , onde  $\theta = \beta, \lambda, \alpha$ ; DP=desvio padrão, Hpd(Highest Probability Density) = Mais alta densidade de probabilidade.

A Tabela 5.9, que contém os resultados do modelo de fragilidade Gama aplicado no banco de dados sobre câncer de pulmão, obteve significância nas mesmas covariáveis do modelo Log-Normal,  $\beta_{sexo}$  e  $\beta_{pat.karno}$ . Tomando o exponencial do  $E(1 - e^{\beta_{sexo}}) = 23,4\%$ , para um indivíduo do sexo feminino, estima-se uma redução de 23,4% na função taxa de falha. Já a covariável  $\beta_{pat.karno}$ , o valor fica  $E(1 - e^{\beta_{pat.karno}}) = 27,9\%$ , o que significa uma redução de 27,9% na função taxa de falha.

Como visto na Seção 4.4 do Capítulo 4, os critérios de avaliação DIC e WAIC nos ajudam a decidir qual modelo se ajusta melhor ao banco de estudado, os resultados dos dois modelos se encontram nas tabelas 5.10 e 5.11.

Tabela 5.10: WAIC e DIC do modelo com fragilidade Log-Normal para os dados sobre câncer de pulmão

WAIC	DIC
1696,74	1690,22

WAIC foi padronizado para ficar igual ao DIC:  $DIC \equiv (-2WAIC)$ .

Tabela 5.11: WAIC e DIC do modelo com fragilidade Gama para os dados sobre câncer de pulmão

WAIC	DIC
1697,56	1691,41

WAIC foi padronizado para ficar igual ao DIC:  $DIC \equiv (-2WAIC)$ .

Valores menores de WAIC e DIC indicam um melhor ajuste dos modelos aos dados. Os resultados apresentados ficaram bem parecidos, mas tanto o DIC quanto o WAIC ficaram menores para o modelo com fragilidade Log-Normal. Portanto, podemos concluir que para o banco de dados sobre câncer de pulmão, o modelo que se adequou melhor aos dados foi o modelo com fragilidade Log-Normal.

As fragilidades Log-Normal dos 17 grupos, que são as instituições médicas, do banco foram padronizadas e assim elas ficaram na mesma escala que as fragilidades Gama. Os resultados para cada grupo variam bastante, mas com uma visualização não muito criteriosa nas tabelas podemos concluir que os valores das fragilidades Log-Normal e Gama ficaram bem próximas, comparando o mesmo grupo nos dois modelos, salvo algumas exceções. As fragilidades Log-Normal e Gama dos 17 grupos estão apresentadas nas tabelas 5.12 e 5.13.

Tabela 5.12: Fragilidades com ajuste do modelo de fragilidade Log-Normal dados sobre câncer de pulmão

Fragilidades	Média	Mediana	Moda	DP	Hpd
$Z_1$	1,387	1,388	1,245	1,402	[0,727;2,698]
$Z_2$	1,603	1,626	1,544	1,813	[0,467;4,734]
$Z_3$	0,874	0,884	0,952	1,491	[0,416;1,992]
$Z_4$	0,743	0,755	0,703	1,664	[0,282;2,016]
$Z_5$	1,008	1,024	1,053	1,601	[0,398;2,444]
$Z_6$	1,416	1,420	1,349	1,482	[0,682;3,117]
$Z_7$	0,863	0,879	1,069	1,612	[0,327;2,114]
$Z_8$	1,661	1,703	1,699	1,700	[0,578;4,387]
$Z_9$	0,825	0,837	0,918	1,520	[0,373;1,889]
$Z_{10}$	0,959	0,959	0,902	1,462	[0,451;2,014]
$Z_{11}$	0,678	0,685	0,818	1,529	[0,297;1,546]
$Z_{12}$	0,886	0,910	0,908	1,666	[0,335;2,386]
$Z_{13}$	0,568	0,571	0,519	1,560	[0,236;1,368]
$Z_{14}$	2,064	2,070	2,022	1,536	[0,884;4,676]
$Z_{15}$	0,679	0,683	0,757	1,475	[0,307;1,432]
$Z_{16}$	0,960	0,991	0,961	1,898	[0,275;3,432]
$Z_{17}$	0,827	0,846	0,834	1,840	[0,235;2,560]

DP=Desvio padrão; Hpd(Highest Probability Density) = Mais alta densidade de probabilidade.

Tabela 5.13: Fragilidades com ajuste do modelo de fragilidade Gama para dados sobre câncer de pulmão

Fragilidades	Média	Mediana	Moda	DP	Hpd
$Z_1$	1,242	1,188	1,113	0,420	0,474;2,070
$Z_2$	1,531	1,381	1,139	0,818	0,277;3,122
$Z_3$	0,791	0,741	0,583	0,327	0,217;1,391
$Z_4$	0,715	0,644	0,542	0,379	[0,139;1,490]
$Z_5$	0,944	0,868	0,749	0,444	[0,243;1,820]
$Z_6$	1,297	1,225	1,127	0,504	[0,419;2,284]
$Z_7$	0,818	0,741	0,593	0,403	[0,177;1,623]
$Z_8$	1,555	1,414	1,132	0,769	[0,343;3,125]
$Z_9$	0,779	0,719	0,604	0,344	[0,202;1,449]
$Z_{10}$	0,879	0,825	0,641	0,352	[0,297;1,574]
$Z_{11}$	0,624	0,577	0,464	0,282	[0,170;1,189]
$Z_{12}$	0,854	0,774	0,655	0,452	[0,168;1,743]
$Z_{13}$	0,506	0,461	0,443	0,246	[0,102;0,981]
$Z_{14}$	1,873	1,758	1,543	0,761	[0,535;3,318]
$Z_{15}$	0,617	0,576	0,516	0,254	[0,200;1,121]
$Z_{16}$	0,971	0,846	0,442	0,601	[0,100;2,170]
$Z_{17}$	0,835	0,724	0,527	0,530	[0,069;1,902]

DP=Desvio padrão; Hpd(Highest Probability Density) = Mais alta densidade de probabilidade.

Na figura (5.11) é apresentado o gráfico da distribuição *a posteriori* das fragi-

lidades Gama e Log-Normal com os seus respectivos intervalos de probabilidade *a posteriori* dos 17 grupos do banco de dados sobre câncer de pulmão. Nota-se que os intervalos de probabilidade *a posteriori* da fragilidade Log-Normal são bem mais amplos que os intervalos da fragilidade Gama, como visto nas tabelas (5.12) e (5.13).

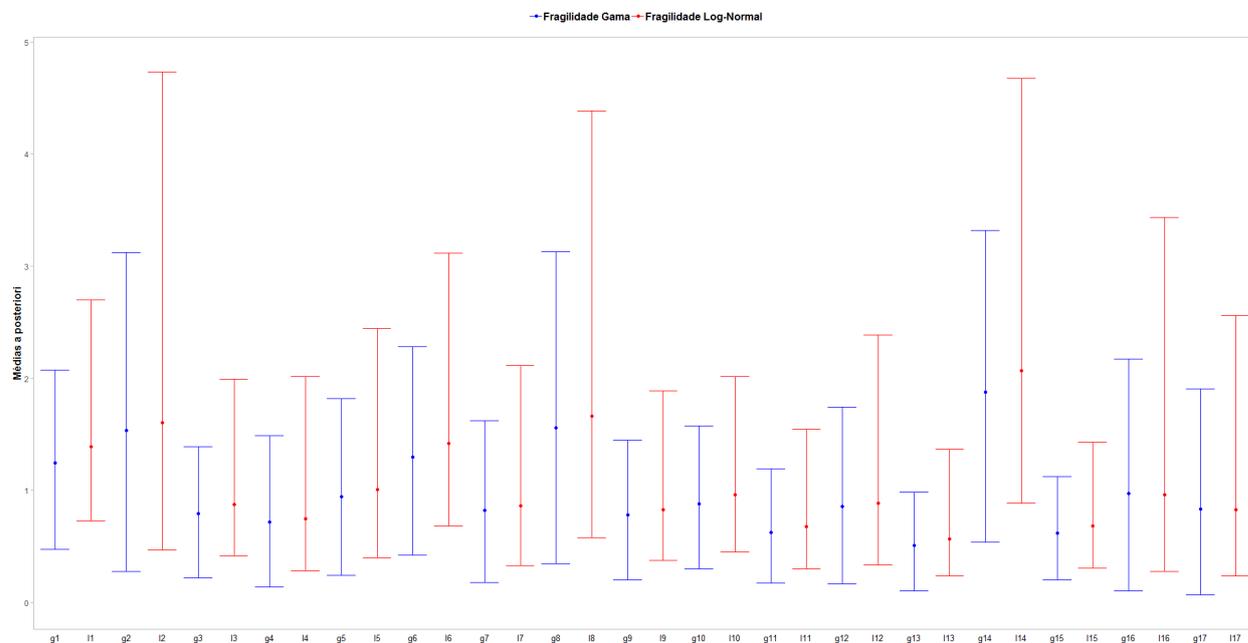


Figura 5.11: Médias *a posteriori* das fragilidades Gama e Log-Normal do banco sobre dados de câncer de pulmão

### 5.1.2 Aplicação dos modelos em dados sobre infarto

O Segundo banco de dados escolhido para a aplicação dos modelos foi um banco com dados de indivíduos que sofreram infarto. O banco denominado *Infarto*, foi retirado do repositório online da Fundação Fio Cruz e faz parte do livro [Carvalho et al. \(2011\)](#). A primeira publicação que utilizou esse banco foi [Melo \(2004\)](#), com o objetivo de avaliar os efeitos das variáveis que eram relacionadas aos pacientes (sexo, idade), com as variáveis que eram relacionadas ao hospital (natureza jurídica, número de internações, número de leitos de UTI), em que elas estavam sendo assistidas. O banco *Infarto* contém 3176 observações, 10 variáveis definidas como:

1. Hospital = Identificador do hospital;
2. Id = Identificador do paciente;
3. Ini = Momento do infarto;
4. Fim = Momento do óbito ou da censura;
5. Status = Falha ou censura;
6. Sexo = Masculino ou Feminino;
7. Idade = Idade na data do infarto;

8. Volume = Número de internações: vp = menor que 25, vg = maior ou igual a 25;
9. Luti = Número de leitos de UTI: n = nenhum, 1a24, 25+;
10. Natureza = Natureza jurídica do hospital: PM = público municipal, C = contratado, PE = público estadual, PFU = público federal e universitário;

Neste banco de dados, 1181 (37%) eram indivíduos do sexo feminino e entre elas 911 (77%) sofrem uma censura e 270 (23%) sofreram uma falha. Já os indivíduos do sexo masculino representam 1995 (63%) indivíduos, dentre eles 1689 (85%) sofreram uma censura e 306 (15%) sofreram uma falha.

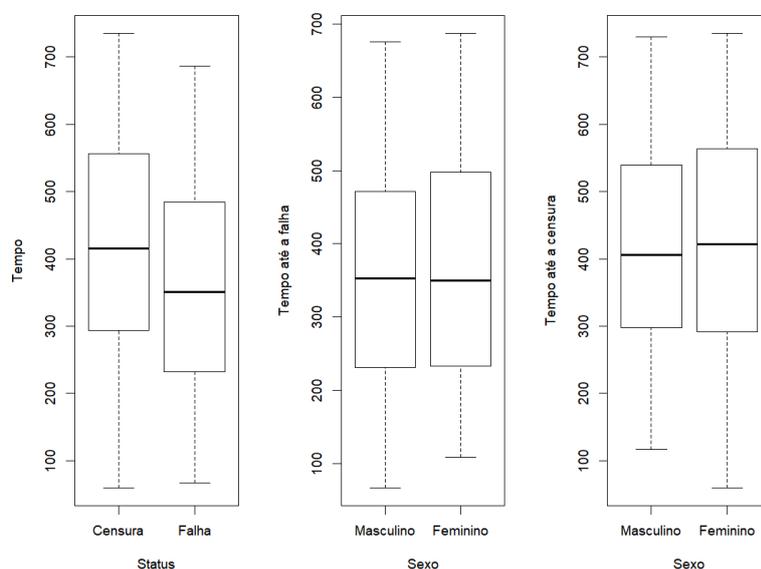


Figura 5.12: Distribuição dos tempos de falha e censura por sexo para os dados sobre infarto

A Figura 5.12 contém os gráficos Box-Plot dos tempos de falha e censura, relacionados com os sexos masculino e feminino. O primeiro gráfico mostra que os tempos de censura foram maiores que os tempos de falha. O segundo gráfico relaciona os sexos com os tempos de falha, os dois sexos estão com tempos de falha muito parecidos, sendo que o sexo feminino aparentemente levou um pouco mais de tempo do que o masculino. O último gráfico mostra uma relação dos sexos com os tempos de censura, os tempos de ambos os sexos também são bem parecidos, sendo que o sexo feminino leva um pouco mais de tempo que o masculino.

A Tabela 5.14 mostra a análise descritiva das variáveis contínuas Ini, Fim e idade.

Tabela 5.14: Análise descritiva das variáveis contínuas do banco infarto

	Min.	1º Q	Median.	3º Q	Média	Máx
Ini	0	232	354	492	363,1	677
Fim	60	283	404	545	413,6	735
Idade	15	51	61	70	60,9	99

As variáveis relacionadas ao hospital são categóricas e estão distribuídas da seguinte maneira, quanto a variável *Natureza*, 1658 (52%) de natureza jurídica público municipal, 927 (29%) são público estadual, 531 (17%) são público federal e universitário e por fim, 60 (2%) são contratados. A variável *Volume*, que representa o número de internações, está dividida em, 3118 (98%) tem número de internações maior ou igual a 25 e 58 (2%) tem um número de internações menor que 25. E por fim, a última variável é a *Luti*, que representa o número de leitos, 2088 (66%) tem de 1 a 24 leitos, 801 (25%) tem mais de 25 leitos, e 287 (9%) não tem nenhum leito.

Para a aplicação dos modelos deste trabalho, foram mantidas todas as observações e foram feitos alguns ajustes no banco, como a recodificação das variáveis, *Hospital*, *Sexo*, *Volume*, *Luti* e *Natureza*. Além disso foram criadas variáveis *dummies* a partir das variáveis categóricas, *Volume*, *Luti* e *Natureza*.

As Tabelas 5.15 e 5.16 apresentam os valores estimados dos parâmetros que foram obtidas nos modelos. Elas nos fornece as médias, medianas, modas, desvios padrão e o intervalo de probabilidade *a posteriori*. No caso do banco de dados sobre *infarto*, as estimações de alguns parâmetros não ficam muito parecidas, nota-se que grande parte das estimativas do modelo Gama são bem maiores quando comparadas com as estimativas dos mesmos parâmetros, mas no modelo Log-Normal. Também em ambas as tabelas o parâmetro  $\lambda$  apresenta um valor relativamente pequeno.

Começando pelo modelo de Fragilidade Log-Normal, foram obtidos os seguintes resultados para os parâmetros, apresentados na tabela 5.15.

Tabela 5.15: Resultados para o modelo de fragilidade Log-Normal para os dados sobre *infarto*

	Média	Mediana	Moda	DP	Hpd
$\beta_{ini}$	-2,409	-2,408	-2,419	0,096	[-2,593;-2,220]
$\beta_{sexo}$	0,506	0,338	0,409	0,896	[-0,774;2,316]
$\beta_{idade}$	0,880	0,639	0,752	1,437	[-1,206;3,731]
$\beta_{nat.pm}$	1,207	0,986	0,613	1,316	[-0,762;3,784]
$\beta_{nat.c}$	0,117	0,117	0,117	0,042	[0,033;0,195]
$\beta_{nat.pe}$	0,534	0,534	0,535	0,045	[0,448;0,624]
$\beta_{nat.pfu}$	-0,155	-0,388	-0,617	0,868	[-1,374;1,844]
$\beta_{vol.vg}$	0,041	0,007	-0,027	0,260	[-0,386;0,630]
$\beta_{vol.vp}$	0,088	-0,076	-0,384	0,782	[-1,086;1,841]
$\beta_{luti.n}$	-0,318	-0,458	-0,675	0,662	[-1,383;1,200]
$\beta_{luti.1a26}$	1,112	1,197	1,454	1,287	[-1,096;3,436]
$\beta_{luti.25+}$	1,182	1,271	1,516	1,283	[-0,979;3,522]
$\lambda$	$2,6 \cdot 10^{-16}$	$1,3 \cdot 10^{-16}$	$-5,8 \cdot 10^{-5}$	$4,4 \cdot 10^{-16}$	$[3,3 \cdot 10^{-16};1,0 \cdot 10^{-16}]$
$\alpha$	5,794	5,794	5,776	0,203	[5,395;6,187]

Viés relativo= $(\hat{\theta} - \theta)/|\theta|$ , onde  $\theta = \beta, \lambda, \alpha$ ; DP=desvio padrão, Hpd(Highest Probability Density) = Mais alta densidade de probabilidade.

Observando os intervalos de probabilidade em que o valor zero não está presente, na tabela 5.15, concluímos que as covariáveis  $\beta_{ini}$ ,  $\beta_{nat.c}$ ,  $\beta_{nat.pe}$ , foram significativas

no modelo com fragilidade Log-Normal. Calculando  $E(1 - e^\beta)$  para as estimativas desses parâmetros que foram significativos, podemos concluir que a covariável  $\beta_{ini}$ , reduz a função taxa de falha e as covariáveis  $\beta_{nat.c}$ ,  $\beta_{nat.pc}$ , aumentam a função taxa falha.

Aplicando o modelos de Fragilidade Gama no banco sobre infarto, temos os seguintes resultados apresentados na tabela 5.16:

Tabela 5.16: Resultados para o modelo de fragilidade Gama para os dados sobre infarto

	Média	Mediana	Moda	DP	Hpd
$\beta_{ini}$	-2,412	-2,409	-2,386	0,103	[-2,612;2,208]
$\beta_{sexo}$	2,580	1,756	0,936	1,776	[0,366;5,809]
$\beta_{idade}$	4,299	3,068	2,112	2,903	[0,547;9,331]
$\beta_{nat.pm}$	4,361	3,226	2,315	2,670	[0,947;9,086]
$\beta_{nat.c}$	0,117	0,117	0,113	0,041	[0,036;0,198]
$\beta_{nat.pe}$	0,534	0,533	0,529	0,045	[0,449;0,624]
$\beta_{nat.pfu}$	-0,149	-0,122	-0,147	0,833	[-2,203;1,321]
$\beta_{vol.vg}$	0,076	0,080	0,098	0,248	[-0,457;0,560]
$\beta_{vol.vp}$	0,067	0,092	-0,176	0,765	[-1,626;1,543]
$\beta_{uti.n}$	-0,344	-0,297	-0,278	0,603	[-1,596;0,792]
$\beta_{uti.1a26}$	1,651	1,509	1,252	1,482	[-0,930;4,539]
$\beta_{uti.25+}$	1,748	1,625	1,876	1,465	[-0,810;4,569]
$\lambda$	$3,7 \cdot 10^{-17}$	$1,6 \cdot 10^{-17}$	$-5,8 \cdot 10^{-6}$	$6,4 \cdot 10^{-16}$	$[2,1 \cdot 10^{-16};1,4 \cdot 10^{-16}]$
$\alpha$	5,801	5,795	5,777	0,226	[5,350;6,17]

Viés relativo= $(\hat{\theta} - \theta)/|\theta|$ , onde  $\theta = \beta, \lambda, \alpha$ ; DP=desvio padrão, Hpd(Highest Probability Density) = Mais alta densidade de probabilidade.

Já observando os intervalos de probabilidade em que o valor zero não está presente, na tabela 5.16, concluímos que as covariáveis  $\beta_{sexo}$ ,  $\beta_{idade}$ ,  $\beta_{nat.pm}$ ,  $\beta_{nat.c}$  e  $\beta_{nat.pc}$ , foram significativas no modelo com fragilidade Gama. Calculando  $E(1 - e^\beta)$  para as estimativas desses parâmetros que foram significativos, conclui-se que todas as covariáveis aumentam a função taxa de falha.

No banco *infarto*, também foi utilizado os critérios de avaliação de modelos, DIC e WAIC, os resultados são apresentados nas tabelas 5.17 e 5.18.

Tabela 5.17: WAIC e DIC do modelo de fragilidade Log-Normal sobre os dados de infarto

WAIC	DIC
8621,2	8620,51

WAIC foi padronizado para ficar igual ao DIC:  $DIC \equiv (-2WAIC)$

Tabela 5.18: WAIC e DIC do modelo de fragilidade Gama sobre os dados de infarto

WAIC	DIC
8620,72	8618,95

WAIC foi padronizado para ficar igual ao DIC:  $DIC \equiv (-2WAIC)$

Como valores menores de DIC e WAIC indicam um bom ajuste do modelos sobre os dados, conclui-se que para o banco de dados *infarto*, o modelo que melhor se ajustou foi o com fragilidade Gama.

Fragilidades Log-Normal também foram padronizadas para os 26 grupos, que são representados pelos hospitais no estudo. Em grande parte dos grupo, em ambos os modelos, notamos que os valores das fragilidades ficaram bem próximas.

Tabela 5.19: Fragilidades com ajuste do modelo de fragilidade Log-Normal para dados sobre infarto

Fragilidades	Média	Mediana	Moda	DP	Hpd
Z <sub>1</sub>	0,792	0,796	0,833	1,633	[0,301;2,031]
Z <sub>2</sub>	0,908	0,912	1,161	1,629	[0,330;2,208]
Z <sub>3</sub>	0,897	0,899	0,837	1,749	[0,286;2,513]
Z <sub>4</sub>	0,889	0,895	0,956	1,533	[0,393;2,083]
Z <sub>5</sub>	0,467	0,481	0,676	2,208	[0,100;2,218]
Z <sub>6</sub>	2,170	2,153	2,194	1,513	[0,993;4,945]
Z <sub>7</sub>	0,847	0,849	0,941	1,558	[0,355;2,007]
Z <sub>8</sub>	0,475	0,479	0,488	1,722	[0,170;1,420]
Z <sub>9</sub>	1,492	1,478	1,419	1,509	[0,685;3,487]
Z <sub>10</sub>	0,855	0,856	1,041	1,638	[0,343;2,352]
Z <sub>11</sub>	0,970	0,959	0,850	1,514	[0,441;2,209]
Z <sub>12</sub>	1,030	1,017	0,980	1,492	[0,468;2,247]
Z <sub>13</sub>	0,687	0,690	0,833	2,334	[0,127;3,613]
Z <sub>14</sub>	1,335	1,327	1,269	1,497	[0,589;2,882]
Z <sub>15</sub>	1,239	1,248	1,290	1,609	[0,487;3,177]
Z <sub>16</sub>	1,121	1,127	1,137	2,409	[0,200;6,258]
Z <sub>17</sub>	1,456	1,471	1,463	1,717	[0,495;4,204]
Z <sub>18</sub>	0,915	0,923	0,977	2,436	[0,164;5,180]
Z <sub>19</sub>	1,791	1,790	1,719	2,316	[0,358;9,296]
Z <sub>20</sub>	0,739	0,727	0,581	2,058	[0,187;3,137]
Z <sub>21</sub>	0,642	0,643	0,759	1,544	[0,285;1,558]
Z <sub>22</sub>	1,323	1,325	1,432	1,496	[0,596;2,865]
Z <sub>23</sub>	1,110	1,111	1,063	1,504	[0,506;2,510]
Z <sub>24</sub>	1,069	1,083	1,123	1,531	[0,456;2,460]
Z <sub>25</sub>	1,407	1,435	1,495	2,107	[0,359;6,360]
Z <sub>26</sub>	1,039	1,050	0,991	1,583	[0,412;2,412]

DP=Desvio padrão; Hpd(Highest Probability Density) = Mais alta densidade de probabilidade.

Tabela 5.20: Fragilidades com ajuste do modelo de fragilidade Gama para dados sobre infarto

Fragilidades	Média	Mediana	Moda	DP	Hpd
$Z_1$	0,750	0,673	0,551	0,386	[0,139;1,536]
$Z_2$	0,855	0,771	0,744	0,433	[0,156;1,653]
$Z_3$	0,882	0,773	0,602	0,515	[0,137;1,904]
$Z_4$	0,839	0,775	0,670	0,372	[0,228;1,572]
$Z_5$	0,403	0,264	0,023	0,440	[0,000;1,286]
$Z_6$	1,895	1,801	1,447	0,735	[0,685;3,417]
$Z_7$	0,758	0,701	0,665	0,335	[0,189;1,402]
$Z_8$	0,426	0,362	0,209	0,275	[0,045;0,965]
$Z_9$	1,324	1,263	1,290	0,519	[0,451 ;2,419]
$Z_{10}$	0,768	0,668	0,423	0,453	[0,150;1,686]
$Z_{11}$	0,861	0,821	0,849	0,342	[0,283;1,523]
$Z_{12}$	0,902	0,855	0,757	0,350	[0,290;1,586]
$Z_{13}$	0,719	0,471	0,168	0,803	[0,000;2,247]
$Z_{14}$	1,168	1,097	0,984	0,450	[0,367;2,060]
$Z_{15}$	1,163	1,062	0,891	0,564	[0,209;2,225]
$Z_{16}$	0,957	0,705	0,142	0,896	[0,004;2,740]
$Z_{17}$	1,449	1,288	1,162	0,787	[0,294;3,049]
$Z_{18}$	0,881	0,647	0,230	0,819	[0,001;2,503]
$Z_{19}$	1,634	1,356	0,888	1,205	[0,013;3,936]
$Z_{20}$	0,710	0,523	0,231	0,623	[0,012;1,920]
$Z_{21}$	0,606	0,552	0,522	0,280	[0,145;1,148]
$Z_{22}$	1,241	1,159	1,049	0,519	[0,392;2,308]
$Z_{23}$	1,048	0,979	0,798	0,447	[0,307;1,922]
$Z_{24}$	1,012	0,931	0,744	0,456	[0,281;1,938]
$Z_{25}$	1,503	1,258	0,842	1,053	[0,108;3,588]
$Z_{26}$	0,997	0,908	0,714	0,470	[0,247;1,929]

DP=Desvio padrão; Hpd(Highest Probability Density) = Mais alta densidade de probabilidade.

Na figura (5.13), temos as médias *a posteriori* e os intervalos de probabilidade *a posteriori* dos 26 grupos do banco de dados sobre infarto. Como pode ser visto nas tabelas (5.19) e (5.20), os grupos com fragilidades Log-Normal tem um intervalo de probabilidade *a posteriori* mais amplos, quando comparado com das fragilidades Gama dos mesmos grupos.

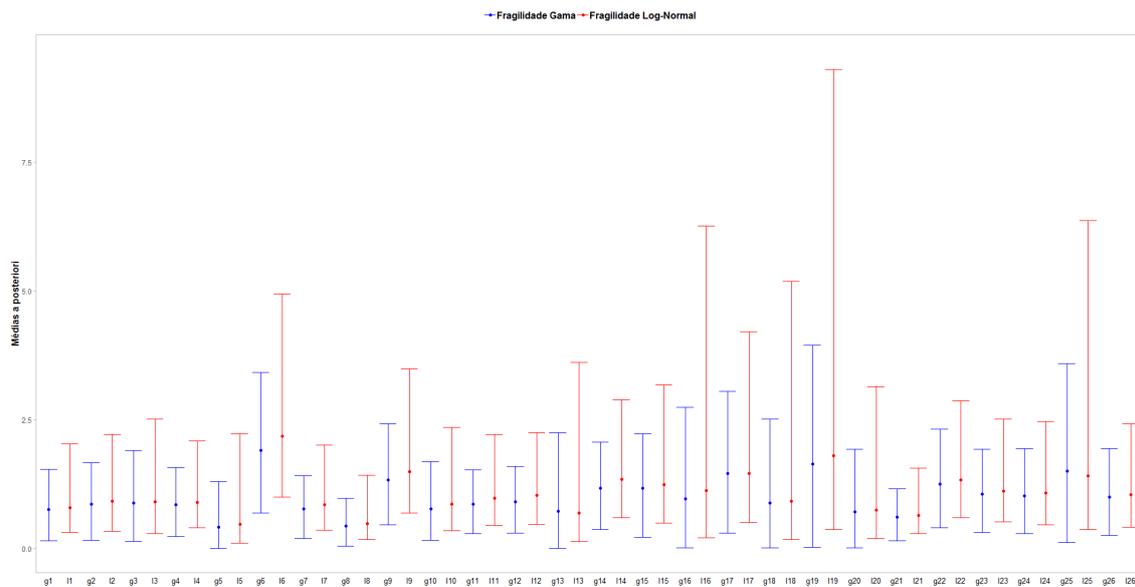


Figura 5.13: Médias *a posteriori* das fragilidades Gama e Log-Normal do banco sobre dados de infarto

## 6 Conclusões

Este trabalho teve como objetivo estudar e aplicar o modelo de fragilidade, considerando duas distribuições muito utilizadas na literatura, a fragilidade com distribuição Gama e a fragilidade com distribuição Log-Normal.

Foram implementados dois modelos, um que continha fragilidade Gama e outro com fragilidade Log-Normal, optou-se por utilizar inferência Bayesiana para as estimativas dos parâmetros dos mesmos. Os modelos foram aplicados em bancos de dados simulados, para a partir dos resultados obtidos, compará-los com os valores iniciais que foram definidos *a priori*. A partir dos resultados *a posteriori*, concluiu-se que os modelos estavam sendo executados da maneira correta e fornecendo um bom ajuste.

Com o interesse de observar como esses modelos se comportavam em bancos de dados reais, buscou-se dados que se adequavam aos estudados em análise de sobrevivência, em que o desfecho é o tempo até a ocorrência de algum evento de interesse. Os bancos escolhidos continham dados de indivíduos com *câncer de pulmão* e outro em que os indivíduos sofreram um *infarto*, após aplicar os modelos nos bancos, definiu-se via critério de avaliação Bayesiano *DIC* e *WAIC*, que para o banco com dados de *câncer de pulmão* o modelo com fragilidade Log-Normal se adaptou melhor aos dados. Já para o banco com dados sobre *infarto*, o modelo que melhor se adequou foi o com fragilidade Gama. Vários fatores podem ter influenciado esses resultados para ambos os modelos. Uma suposição é de que o tamanho do banco e a quantidade de covariáveis podem ter interferido nos resultados, pois o banco com informações sobre *infarto* possuía um tamanho amostral maior e tinha mais covariáveis do que o banco com informações de *câncer de pulmão*.

Em um futuro trabalho pode-se investigar que efeito teríamos se trocássemos a distribuição da fragilidade. Foram gerados dados em que a fragilidade tinha distribuição Gama para serem utilizados com o modelo fragilidade Gama e também foram gerados dados com fragilidade seguindo uma distribuição Normal, para serem modelados com fragilidade Log-Normal. Seria interessante observar como seriam os resultados e como os modelos se comportariam com as fragilidades seguindo uma distribuição diferente da utilizada na geração dos dados. Além dessa troca, também pode-se investigar outros estudos que tenham utilizado os dados reais apresentados nesse trabalho e compará-los com as estimativas obtidas para os parâmetros.

## Referências Bibliográficas

- Abbring, Jaap H, & Van Den Berg, Gerard J. 2007. The unobserved heterogeneity distribution in duration analysis. *Biometrika*, **94**(1), 87–99.
- Amrhein, John. 2014. Introduction to frailty models. *McDougall Scientific Ltd.*, **23**.
- Bender, Ralf, Augustin, Thomas, & Blettner, Maria. 2005. Generating survival times to simulate Cox proportional hazards models. *Statistics in medicine*, **24**(11), 1713–1723.
- Carvalho, Marília Sá, Andreozzi, Valeska Lima, Codeço, Claudia Torres, Campos, Dayse Pereira, Barbosa, Maria Tereza Serrano, & Shimakura, Silvia Emiko. 2011. *Análise de Sobrevivência: teoria e aplicações em saúde*. 1 edn. Scielo-Editora Fio Cruz.
- Casella, George, & George, Edward I. 1992. Explaining the Gibbs sampler. *The American Statistician*, **46**(3), 167–174.
- Congdon, Peter. 2014. *Applied bayesian modelling*. 2 edn. Vol. 595. John Wiley & Sons.
- Cox, David R. 1972. Models and life-tables regression. *JR Stat. Soc. Ser. B*, **34**, 187–220.
- Dey, Dipak K, Ghosh, Sujit K, & Mallick, Bani K. 2000. *Generalized linear models: A Bayesian perspective*. 1 edn. CRC Press.
- Ehlers, Ricardo S. 2011. Inferência bayesiana. *Departamento de Matemática Aplicada e Estatística, ICMC-USP*, 64.
- Faraway, Julian J, Wang, Xiaofeng, & Ryan, Yu Yue. 2018. *Bayesian Regression Modeling with INLA*. Chapman and Hall/CRC.
- Gamerman, Dani, & Lopes, Hedibert F. 2006. *Markov chain Monte Carlo: stochastic simulation for Bayesian inference*. 2 edn. Chapman and Hall/CRC.
- Gelfand, Alan E, Hills, Susan E, Racine-Poon, Amy, & Smith, Adrian FM. 1990. Illustration of Bayesian inference in normal data models using Gibbs sampling. *Journal of the American Statistical Association*, **85**(412), 972–985.
- Gelman, Andrew, Stern, Hal S, Carlin, John B, Dunson, David B, Vehtari, Aki, & Rubin, Donald B. 2013. *Bayesian data analysis*. 3 edn. Chapman and Hall/CRC.

- Geman, Stuart, & Geman, Donald. 1984. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on pattern analysis and machine intelligence*, 721–741.
- Giolo, Suely Ruiz, & Colosimo, Enrico Antônio. 2006. *Análise de sobrevivência aplicada*. 1 edn. Edgard Blucher.
- Hanagal, David D. 2011. *Modeling survival data using frailty models*. 1 edn. Chapman and Hall/CRC.
- Hastings, W Keith. 1970. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**(1), 97–109.
- Karnofsky, David A. 1949. The clinical evaluation of chemotherapeutic agents in cancer. *Evaluation of chemotherapeutic agents*, 191–205.
- Lawless, Jerald F. 2011. *Statistical Models and Methods for Lifetime Data*. 2 edn. John Wiley & Sons.
- Loprinzi, Charles Lawrence, Laurie, John A, Wieand, H Sam, Krook, James E, Novotny, Paul J, Kugler, John W, Bartel, Joan, Law, Marlys, Bateman, Marilyn, & Klatt, Nancy E. 1994. Prospective evaluation of prognostic variables from patient-completed questionnaires. North Central Cancer Treatment Group. *Journal of Clinical Oncology*, **12**(3), 601–607.
- Melo, Enirtes Caetano Prates. 2004. *Infarto agudo do miocárdio no Município do Rio de Janeiro: qualidade dos dados, sobrevida e distribuição espacial*. Ph.D. thesis, Escola Nacional de Saúde Pública da Fundação Oswaldo Cruz.
- Metropolis, Nicholas, Rosenbluth, Arianna W, Rosenbluth, Marshall N, Teller, Augusta H, & Teller, Edward. 1953. Equation of state calculations by fast computing machines. *The journal of chemical physics*, **21**(6), 1087–1092.
- Oken, Martin M, Creech, Richard H, Tormey, Douglass C, Horton, John, Davis, Thomas E, Mcfadden, Eleanor T, & Carbone, Paul P. 1982. Toxicity and response criteria of the Eastern Cooperative Oncology Group. *American journal of clinical oncology*, **5**(6), 649–656.
- Peto, Richard. 1972. Rank tests of maximal power against Lehmann-type alternatives. *Biometrika*, **59**(2), 472–475.
- Pike, MC. 1966. A method of analysis of a certain class of experiments in carcinogenesis. *Biometrics*, **22**(1), 142–161.
- Plummer, Martyn. 2015. *JAGS Version 3.4.0 manual*.
- R Core Team. 2018. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.
- Rue, Håvard, Martino, Sara, & Chopin, Nicolas. 2009. Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, **71**(2), 319–392.

- Spiegelhalter, David J, Best, Nicola G, Carlin, Bradley P, & Van Der Linde, Angelika. 2002. Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **64**(4), 583–639.
- Teruel, Maria Bernadete Sanches, *et al.* 1993. *Modelo de regressão logístico e exponencial por partes em análise de sobrevivência*. M.Phil. thesis, Universidade Estadual de Campinas - UNICAMP.
- Vaupel, James W, Manton, Kenneth G, & Stallard, Eric. 1979. The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, **16**(3), 439–454.
- Watanabe, Sumio. 2010. Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, **11**, 3571–3594.
- Weibull, Waloddi. 1939. A statistical theory of strength of materials. *Ingeniors Vetenskaps Akademien Handlingar*, **3**(151), 45–55.
- Whittemore, Alice, & Altshuler, Bernard. 1976. Lung cancer incidence in cigarette smokers: further analysis of Doll and Hill's data for British physicians. *Biometrics*, 805–816.
- Wienke, Andreas. 2010. *Frailty models in survival analysis*. 1 edn. Chapman and Hall/CRC.

## .1 Cálculo dos tempos em sobrevivência

Por definição, seja  $Y$  uma variável aleatória com função de distribuição  $F$ , então  $U = F(Y)$  segue uma distribuição uniforme no intervalo entre  $[0, 1]$ . Então

$$1 - \exp\{-H_0(t)e^{x\beta+w_i}\} = U, \quad (1)$$

$$\exp\{-H_0(t)e^{x\beta+w_i}\} = 1 - U,$$

$$\exp\{-H_0(t)e^{x\beta+w_i}\} = U. \quad (2)$$

Como a função taxa de falha acumulada da Weibull é dada por (2.29), substituindo na função acima temos,

$$\exp\{-\lambda t^\alpha e^{x\beta+w_i}\} = U, \quad (3)$$

$$-\lambda t^\alpha e^{x\beta+w_i} = \log(U).$$

Isolando o  $t$ ,

$$t^\alpha = \frac{-\log(U)}{\lambda e^{x\beta+w_i}}. \quad (4)$$

Assim o tempo para cada indivíduo do banco de dados é representado por:

$$t = \left( \frac{-\log(U)^\alpha}{\lambda e^{x\beta+w_i}} \right)^{1/\alpha}. \quad (5)$$

## .2 Geração dos dados

Os dados foram gerados utilizando o Software *R* (R Core Team, 2018), e a geração foi feita da seguinte maneira:

1. Definiu-se um tamanho de amostra de 200 observações.
2. Criou-se um vetor com duas covariáveis, uma seguindo uma normal (0,1) e outra uma binomial (1,0.5).
3. Dividiu-se as observações em 40 grupos, cada grupo contém 5 observações.
4. Definiu-se que as fragilidades seguiriam uma distribuição Gama para os dados com fragilidades Gama e para os dados com fragilidades Log-Normal, as fragilidades seguiriam uma distribuição Log-Normal.
5. Definiu-se um vetor que recebe os tempos de cada observação através da fórmula (5) (Bender *et al.*, 2005).

### .3 Fragilidades Log-Normal não padronizadas das simulações

Tabela 1: Fragilidades Log-Normal das simulações

Fragilidades	Média	Mediana	Moda	DP	Hpd
$w_1$	-0,012	0,0119	0,0658	0,518	[-1,0381;0,9727]
$w_2$	0,0306	0,0548	0,0992	0,524	[-1,0078;1,0256]
$w_3$	-0,0058	0,018	0,0664	0,520	[-1,0394;0,9822]
$w_4$	-0,0324	-0,0085	0,0258	0,519	[-1,0619;0,9521]
$w_5$	0,0563	0,0806	0,132	0,522	[-0,9789;1,0466]
$w_6$	-0,0177	0,0062	0,0486	0,521	[-1,0522;0,9718]
$w_7$	-0,0145	0,0096	0,0516	0,522	[-1,0472;0,9778]
$w_8$	-0,0079	0,0162	0,0626	0,523	[-1,0442;0,9851]
$w_9$	0,0149	0,039	0,0847	0,519	[-1,0147;1,0017]
$w_{10}$	-0,0093	0,0143	0,0572	0,519	[-1,0366;0,98]
$w_{11}$	-0,0032	0,0207	0,0634	0,521	[-1,0359;0,9872]
$w_{12}$	0,0427	0,0669	0,1075	0,523	[-0,9945;1,0369]
$w_{13}$	-0,0439	-0,0203	0,0264	0,519	[-1,0708;0,9438]
$w_{14}$	0,0069	0,0309	0,0713	0,519	[-1,0225;0,9941]
$w_{15}$	-0,0159	0,0079	0,0537	0,519	[-1,044;0,9703]
$w_{16}$	0,0379	0,0615	0,0981	0,517	[-0,9833;1,0241]
$w_{17}$	0,0116	0,0351	0,08	0,523	[-1,0241;1,0073]
$w_{18}$	0,0518	0,0757	0,1192	0,526	[-0,9905;1,0494]
$w_{19}$	-0,0353	-0,0117	0,0231	0,520	[-1,0663;0,9539]
$w_{20}$	-0,0458	-0,022	0,0231	0,522	[-1,0813;0,9469]
$w_{21}$	0,0543	0,0784	0,1299	0,520	[-0,9773;1,0408]
$w_{22}$	0,0003	0,0244	0,0814	0,522	[-1,0338;0,9918]
$w_{23}$	0,0522	0,0767	0,1167	0,524	[-0,9884;1,0497]
$w_{24}$	-0,0584	-0,0348	0,0161	0,520	[-1,0912;0,9285]
$w_{25}$	-0,0102	0,0144	0,0637	0,524	[-1,0468;0,9853]
$w_{26}$	-0,0829	-0,0595	-0,0179	0,518	[-1,1066;0,9023]
$w_{27}$	-0,0119	0,0119	0,0475	0,519	[-1,0393;0,9784]
$w_{28}$	0,0149	0,0391	0,0915	0,519	[-1,0126;1,001]
$w_{29}$	-0,0122	0,0117	0,0526	0,519	[-1,0411;0,9736]

DP=Desvio padrão; Hpd(Highest Probability Density) = Mais alta densidade de probabilidade.

Tabela 2: Fragilidades Log-Normal das simulações

Fragilidades	Média	Mediana	Moda	DP	Hpd
$w_{30}$	-0,0083	0,0153	0,0585	0,522	[-1,0464;0,9804]
$w_{31}$	-0,0351	-0,011	0,0301	0,522	[-1,0697;0,9568]
$w_{32}$	0,0223	0,0462	0,0897	0,521	[-1,0098;1,0138]
$w_{33}$	0,0239	0,0477	0,1045	0,518	[-1,0008;1,0113]
$w_{34}$	0,0416	0,0654	0,1014	0,522	[-0,9911;1,0352]
$w_{35}$	0,0181	0,0426	0,0894	0,523	[-1,0192;1,0112]
$w_{36}$	-0,0238	-0,0001	0,0423	0,523	[-1,0577;0,9717]
$w_{37}$	-0,0314	-0,0074	0,0405	0,521	[-1,0661;0,9576]
$w_{38}$	0,009	0,0328	0,0842	0,518	[-1,0168;0,9919]
$w_{39}$	-0,0039	0,0199	0,0617	0,517	[-1,0297;0,9776]
$w_{40}$	0,0316	0,0555	0,1001	0,518	[-0,9928;1,0184]

DP=Desvio padrão; Hpd(Highest Probability Density) = Mais alta densidade de probabilidade.

## .4 Fragilidades Log-Normal não padronizadas do banco sobre câncer de pulmão

Tabela 3: Fragilidades Log-Normal do banco sobre câncer de pulmão

Fragilidades	Média	Mediana	Moda	DP	Hpd
$w_1$	0,327	0,328	0,219	0,338	[-0,319;0,992]
$w_2$	0,472	0,486	0,435	0,595	[-0,762;1,555]
$w_3$	-0,135	-0,124	-0,049	0,400	[-0,877;0,689]
$w_4$	-0,297	-0,281	-0,353	0,509	[-1,266;0,701]
$w_5$	0,008	0,024	0,051	0,470	[-0,921;0,894]
$w_6$	0,348	0,350	0,299	0,393	[-0,382;1,137]
$w_7$	-0,148	-0,129	0,066	0,477	[-1,117;0,749]
$w_8$	0,508	0,533	0,530	0,531	[-0,548;1,479]
$w_9$	-0,192	-0,178	-0,085	0,419	[-0,985;0,636]
$w_{10}$	-0,042	-0,042	-0,104	0,380	[-0,797;0,700]
$w_{11}$	-0,389	-0,379	-0,200	0,425	[-1,214;0,436]
$w_{12}$	-0,121	-0,094	-0,096	0,510	[-1,093;0,870]
$w_{13}$	-0,566	-0,561	-0,656	0,445	[-1,445;0,313]
$w_{14}$	0,725	0,727	0,704	0,429	[-0,124;1,542]
$w_{15}$	-0,386	-0,381	-0,278	0,388	[-1,182;0,359]
$w_{16}$	-0,041	-0,009	-0,040	0,641	[-1,291;1,233]
$w_{17}$	-0,190	-0,167	-0,181	0,610	[-1,448;0,940]

DP=Desvio padrão; Hpd(Highest Probability Density) = Mais alta densidade de probabilidade.

## .5 Fragilidades Log-Normal não padronizadas do banco de dados sobre infarto

Tabela 4: Fragilidades Log Normal do banco de dados sobre Infarto

Fragilidades	Média	Mediana	Moda	DP	Hpd
$w_1$	-0,234	-0,228	-0,183	0,490	[-1,202;0,709]
$w_2$	-0,097	-0,092	0,150	0,488	[-1,109;0,792]
$w_3$	-0,109	-0,107	-0,178	0,559	[-1,252;0,921]
$w_4$	-0,118	-0,111	-0,045	0,427	[-0,935;0,734]
$w_5$	-0,761	-0,732	-0,392	0,792	[-2,304;0,797]
$w_6$	0,775	0,767	0,786	0,414	[-0,007;1,598]
$w_7$	-0,166	-0,164	-0,061	0,443	[-1,037;0,697]
$w_8$	-0,744	-0,736	-0,718	0,543	[-1,771;0,351]
$w_9$	0,400	0,391	0,350	0,411	[-0,379;1,249]
$w_{10}$	-0,156	-0,155	0,040	0,493	[-1,071;0,855]
$w_{11}$	-0,030	-0,042	-0,162	0,415	[-0,819;0,793]
$w_{12}$	0,029	0,017	-0,020	0,400	[-0,759;0,810]
$w_{13}$	-0,376	-0,371	-0,183	0,848	[-2,063;1,285]
$w_{14}$	0,289	0,283	0,238	0,403	[-0,530;1,059]
$w_{15}$	0,215	0,222	0,255	0,476	[-0,720;1,156]
$w_{16}$	0,114	0,120	0,128	0,879	[-1,608;1,834]
$w_{17}$	0,376	0,386	0,381	0,541	[-0,703;1,436]
$w_{18}$	-0,088	-0,080	-0,024	0,890	[-1,806;1,645]
$w_{19}$	0,583	0,582	0,542	0,840	[-1,026;2,230]
$w_{20}$	-0,302	-0,318	-0,543	0,722	[-1,678;1,143]
$w_{21}$	-0,444	-0,441	-0,276	0,435	[-1,257;0,443]
$w_{22}$	0,280	0,282	0,359	0,403	[-0,518;1,053]
$w_{23}$	0,105	0,105	0,061	0,408	[-0,682;0,920]
$w_{24}$	0,067	0,080	0,116	0,426	[-0,785;0,900]
$w_{24}$	0,341	0,361	0,402	0,745	[-1,024;1,850]
$w_{26}$	0,039	0,048	-0,009	0,459	[-0,887;0,881]

DP=Desvio padrão; Hpd(Highest Probability Density) = Mais alta densidade de probabilidade.