UNIVERSIDADE ESTADUAL DA PARAÍBA
CENTRO DE CIÊNCIAS E TECNOLOGIA
DEPARTAMENTO DE ESTATÍSTICA

MANOEL JOAQUIM ISIDRO

INTRODUÇÃO AOS MODELOS DE FRAGILIDADES APLICADOS A DADOS DE LEUCEMIA LINFOBLÁSTICA

CAMPINA GRANDE
DEZEMBRO DE 2015

MANOEL JOAQUIM ISIDRO

INTRODUÇÃO AOS MODELOS DE FRAGILIDADES APLICADOS A DADOS DE LEUCEMIA LINFOBLÁSTICA

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Orientador: Prof. Dr Tiago Almeida de Oliveira

CAMPINA GRANDE

DEZEMBRO DE 2015

I81i Isidro, Manoel Joaquim.
Introdução aos modelos de fragilidades aplicados a dados de leucemia linfoblástica [manuscrito] / Manoel Joaquim Isidro. - 2015.
73 p. : il. color.

Digitado.
Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2015.
"Orientação: Prof. Dr. Tiago Almeida de Oliveira, Departamento de Estatistíca".

1. Análise de sobrevivência. 2. Modelos de fragilidades. 3. Kaplan-Meier. 4. Leucemia linfoblástica. I. Título.
21. ed. CDD 519.544

MANOEL JOAQUIM ISIDRO

INTRODUÇÃO AOS MODELOS DE FRAGILIDADES APLICADOS A DADOS DE LEUCEMIA LINFOBLÁSTICA

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Aprovado em: 09 / 12 / 2015

Banca Examinadora:

Tiago Almeida de Oliveira

Prof. Dr Tiago Almeida de Oliveira
Universidade Estadual da Paraíba - DE/CCT
Orientador

DMEstves

Prof. Dra Divanilda Maia Esteves
Universidade Estadual da Paraíba - DE/CCT
Examinador

Michelli Karinne Barros da Silva

Prof. Dra Michelli Karinne Barros da Silva
Universidade Federal de Campina Grande - UAEST
Examinador

Dedicatória

Em memória do meu querido pai Anísio Joaquim.

MINHA HOMENAGEM

A minha noiva Roseane de Alcântara Costa, por me possibilitar,
com apoio e amor, concluir este trabalho.
E a toda minha família.
OFEREÇO E DEDICO.

Agradecimentos

Agradeço de forma muito especial...

A DEUS, pelo dom da vida, pelo dom do conhecimento infinito. Por ter me dado uma excelente família, grandes amigos e pela força diária.

Aos meus pais por terem, da melhor forma, me oferecido até agora uma excelente formação, tanto com relação a valores pessoais quanto à educação intelectual.

Aos meus irmãos, Joacil, Maria, Lucielma e Aucicleide, por mostrar que é possível mudar de situação por meio dos estudos.

A meu grande amigo Jailson Xavier por toda a força e exemplo ao longo do curso.

A todos os professores e funcionários do Departamento de Estatística da UEPB, principalmente ao professor Tiago (orientador).

Aos membros da banca examinadora, as professoras Divanilda Maia Esteves e Michelli Karinne Barros da Silva, pela disponibilidade e pelas sugestões para o enriquecimento deste trabalho.

Finalmente à residência universitária.

“A essência do conhecimento consiste
em aplicá-lo, uma vez possuído.”
- Confúcio.

Resumo

Este trabalho de conclusão de curso tem como objetivo produzir um material pratico e claro sobre a análise de dados de sobrevivência, que possa vir a auxiliar aos que desejarem utilizar essas técnicas, especialmente quando aplicado aos modelos de fragilidades. Compreende-se por dados de sobrevivência, dados provenientes de estudos longitudinais em que, os indivíduos são acompanhado até a ocorrência do evento de interesse. A análise de sobrevivência modela o tempo até a ocorrência do evento do interesse e incorpora a informação das censuras, ou seja, utiliza o tempo até a censura dos indivíduos que participaram do estudo e não falharam. Os dados foram obtidos do institute for Helth & Society da universidade de Wisconsin (Medical College of Wisconsin), o banco de dados é constituído de um total de 137 pacientes (38 LLA, 99 LMA) os quais apresentam uma distinção entre os tipos de câncer, leucemia mieloide aguda (LMA) e leucemia linfoblástica aguda (LLA), os pacientes foram tratados em quatro hospitais. Na realização deste trabalho será estimado a curva de sobrevivência não-paramétrica de Kaplan-Meier para os 3 grupos de pacientes com diferentes tipos de leucemia, em seguido utilizaremos o teste log-rank para investigar se existi diferença significativa entre as cuvas de sobrevivência, após ajustaremos as distribuições paramétricas e, por fim, aplicaremos o modelo de Cox e os modelos de fragilidades Gama e Log-Normal por indivíduos. A análise será realizada utilizando o software R.

Palavras-chave: *Leucemia Linfoblástica, Modelos de Fragilidades, Análise de Sobrevivência.*

Abstract

This course conclusion work aims to produce a practical and clear material on the analysis of survival data, which may help those who wish to use these techniques, especially when applied to the models of weaknesses. It is understood by survival data, data from longitudinal studies in which subjects are followed until the occurrence of the event of interest. The survival analysis models the time to occurrence of the interest of the event and incorporates information from censure, or use the time to censorship of individuals participating in the study and not failed. Data were obtained from the institute for Helth & Society of the University of Wisconsin (Medical College of Wisconsin), the database consists of a total of 137 patients (38 ALL, 99 AML) who present a distinction between the types of cancer, acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL) patients were treated in four hospitals. In this work it will be estimated nonparametric survival curve Kaplan-Meier curves for the 3 groups of patients with different types of leukemia, in a row will use the log-rank test to investigate whether existed significant difference between the cuvas survival after We will adjust the parametric distributions, and finally, we will apply the Cox model and models of Gamma frailty and Log-Normal individuals. The analysis will be performed using the software R.

Keywords: *Lymphoblastic Leukemia, Frailty Models, Survival Analysis.*

Sumário

Lista de Tabelas

Lista de Figuras

Lista de abreviaturas

***t*:** Tempo de Sobrevivência

***EM*:** Esperança Maximização

***MCMC*:** Monte Carlos via Cadeias de Markov

***km*:** Kaplan-Meier

***GVHD*:** Doença enxerto versus hospedeiro

***GG*:** Distribuição Gama Generalizada

***CMV*:** Citomegalovírus

***LLA*:** Leucemia Linfoblástica Aguda

***LMA*:** Leucemia Miloide Aguda

***TMO*:** Transplante de Medula Óssea

***TRV*:** Teste da Razão de Verossimilhança

***MTX*:** Metrotexano

***AIC*:** Critério de Informação de Akaike

***FAB*:** Sistema de classificação que divide a LMA em estágios

***REML*:** Estimação Máxima Verossimilhança Restrita

1 Introdução

A análise de sobrevivência é formada por um conjunto de técnicas estatísticas para analisar dados, que consistem nos tempos até a ocorrência de um evento de interesse, comumente conhecido como tempo de sobrevivência. São exemplos de casos de análise de sobrevivência o tempo até que um aluno possa concluir sua graduação e a morte após o diagnóstico de certa doença. Um fato que caracteriza dados de sobrevivência é a possibilidade da presença de censura, que é a ocorrência da observação parcial da resposta de interesse. A variável resposta foi o tempo em dias até a morte do paciente ou até o término do acompanhamento.

Leucemia Linfoblástica Aguda (LLA) é um câncer que se origina de um grupo de células precursores dos linfócitos. Os linfócitos são glóbulos brancos que defendem o corpo contra infecções. A medula óssea cria inúmeras células subdesenvolvidas conhecidas como blastos. A LLA é resultado de um dano genético adquirido (não herdado) no DNA de um grupo de células (glóbulos brancos) na medula óssea. As células doentes substituem a medula óssea normal e os efeitos são o crescimento incontrolável e o acúmulo de linfoblasto (linfócitos imaturos) que perdem a capacidade de funcionar como células sanguíneas normais, levando a um bloqueio ou diminuição na produção de glóbulos vermelhos, plaquetas e glóbulos brancos na medula óssea. Após o diagnóstico, o tratamento da LLA é dividido em duas partes: terapia de indução e terapia pós-indução ou transplante de medula óssea. A quimioterapia utiliza medicamentos anticancerígenos para destruir as células tumorais, dependendo do tipo e do estado da leucemia, e pode ser utilizada sozinha ou combinada com a radioterapia (PEDROSA; LINS, 2002).

2 Objetivos

2.1 Geral

O objetivo deste estudo é de utilizar o modelo Fragilidade de Cox nos dados de leucemia linfoblástica.

2.1.1 Específicos

- Ajustar os métodos clássicos e introdutórios da análise de sobrevivência;
- Determinar a fragilidade de pacientes com leucemia linfoblástica.

3 Revisão Literária

3.1 Análise de Sobrevivência

A análise de sobrevivência reúne um conjunto de técnicas e métodos estatísticos úteis na análise do tempo de vida de indivíduos em que, geralmente, a variável resposta é o tempo até a ocorrência de um evento de interesse. Estes eventos são chamados de falhas, tendo como principal característica a presença de censura nos dados. A análise de sobrevivência é um conjunto de técnicas estatísticas que mais cresceram nas últimas décadas. Este crescimento se deve ao desenvolvimento e aprimoramento de técnicas estatísticas combinadas com computadores cada vez mais eficientes (COLOSIMO; GIOLO, 2006).

3.1.1 Dados Necessários Para Análise de Sobrevivência

Segundo Colosimo (2009), para realizar uma análise de sobrevivência, é preciso possuir informações sobre o tempo de falha e censura. Estes dois componentes constituem a resposta na análise de sobrevivência. Em estudos clínicos, um conjunto de covariáveis são medidas em cada indivíduo.

3.1.1.1 Tempo de Falha

De acordo com Herrmann (2011), tempo de falha é o tempo decorrido a partir de um instante inicial até a ocorrência do evento de interesse, uma falha pode ser a morte de um indivíduo no estudo ou uma remissão da doença, também pode ser considerado como a melhora do quadro clínico do paciente.

O tempo de início do estudo deve ser bem definido, permitindo comparar os indivíduos na origem do estudo, com exceção de diferenças medidas pela covariáveis. Nos estudos clínicos aleatorizados, a data da aleatorização é a seleção natural para a origem do estudo. Outras escolhas possíveis são as datas do início do tratamento da doenças ou do diagnóstico (FERREIRA, 2007).

De acordo com Prentice et al., (1978), a falha pode ser definida como a ocorrência de um único ou mais de um evento de interesse, quando causas de falhas competem entre si, temos o risco competitivo.

3.1.1.2 Censura

Segundo Colosimo e Giolo (2006), censura é a observação parcial da resposta que foi interrompida por algum motivo, impedindo a observação do tempo de falha do indivíduo. Uma censura pode acontecer por várias razões, tais como: o término do experimento, o paciente pode ter deixado o tratamento, o paciente ter se mudado para uma outra localidade desconhecida, entre outras. Existem três tipos de censuras. Censura à direita, censura à esquerda e censura intervalar.

- Censura à direita acontece quando o evento de interesse ocorre após o término do estudo. É a mais comum, não se observa o desfecho (evento de interesse) do paciente (HERRMANN, 2011).

  De acordo com Lawless (2011), existem três tipos de mecanismo de censura que são utilizadas com mais frequência no caso de censura à direita: Censura tipo I ocorre quando o estudo é encerrado após um período preestabelecido de tempo; Censura Tipo II acontece quando o estudo é encerrado após ter ocorrido o evento de interesse em um número preestabelecido de indivíduos; Censura Aleatória ocorre se a observação for retirada no decorrer do estudo sem ter ocorrido o evento de interesse ou se o evento de interesse ocorrer por uma razão diferente da estudada.

  Segundo Colosimo (2009), pode-se fazer uma representação simples do mecanismo de censura aleatória por meio de duas variáveis aleatórias. Suponha T uma variável aleatória representando o tempo de falha de um paciente, e C uma outra variável aleatória independe de T, representando o tempo de censura associado a este paciente. Observa-se, portanto,

  $$t = min(T,\ C)$$

  $$\delta_i = \begin{cases} 1, & \text{quando } T \leq C, \\ 0, & \text{quando } T > C. \end{cases}$$

  Onde δ_i representa a variável indicadora de falha ou censura.

  A Figura 1 mostra os mecanismos de cesura à direita, onde o tempo de ocorrência do evento de interesse está à direita do tempo observado. Este fato acontece com

frequência em estudos de sobrevivência.

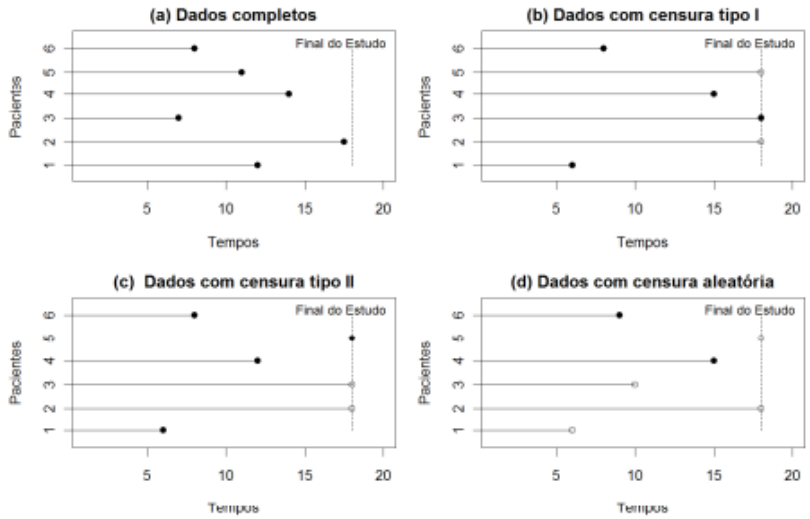

Figura 1: Representação gráfica dos mecanismo de censura, em que • representa falha e ◦ censura. Figura retirada de Colosimo e Giolo (2006).

- De acordo com Colosimo (2009), censura à esquerda acontece quando não conhecemos o momento da ocorrência do evento de interesse, mas sabemos que ele ocorreu antes do tempo observado, isto é, o evento já acorreu quando o paciente foi observado.

- Censura Intervalar acontece quando não se sabe o tempo exato de ocorrência do evento de interesse, sabe-se que ele ocorreu dentro de um intervalo especificado, é o tipo mais geral de censura que ocorre. Pode-se observar em estudos onde os pacientes são acompanhados em visitas periódicas (LAWLESS, 2011).

De acordo com Ferreira (2007), todos os resultados provenientes de um estudo de sobrevivência devem ser utilizado na análise estatística, pois mesmo incompletas, as observações fornecem informações sobre o tempo de vida do paciente. A omissão destas observações no cálculo das estatísticas de interesse provavelmente resultarão em conclusões tendenciosa na análise.

3.1.2 Função de Sobrevivência

A função de sobrevivência é uma função cronológica habitualmente denotada por $S(t)$, onde t representa o tempo de interesse, isto é, permite calcular as porcentagens sobre o

tempo de vida dos indivíduos no estudo, tais como, o tempo médio e o tempo mediano (HERMETO, 2014).

Segundo Louzada (2012), a função de densidade de probabilidade de sobrevivência é escrita como,

$$f(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}, \quad t \geq 0.$$

Em que $f(t)$ é definida como a probabilidade de um indivíduo falhar em um curto período do espaço de tempo. A área entre a curva de densidade e o eixo t é igual a 1.

De acordo com Hermeto (2014), a função de sobrevivência é uma das principais funções probabilísticas usadas para descrever dados de tempo de sobrevivência, é denotada por

$$S(t) = P(T > t)$$

$S(t)$ é definida como a probabilidade de um indivíduo sobreviver até um certo tempo t, sem o evento. A função de distribuição acumulada $F(t)$ é definida como a probabilidade do indivíduo falhar até o tempo t. É representada por

$$F(t) = P(T \leq t).$$

Escrevendo $S(t)$ em função de $F(t)$, temos

$$S(t) = 1 - F(t)$$

3.1.3 Função de Risco

A função de risco, ou taxa de falha é definida como a probabilidade de um indivíduo falhar entre o tempo t e $t + \Delta t$, dado que ele sobreviveu até o tempo t. É denotada por

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}. \tag{3.1}$$

$\lambda(t)$ expressa o risco instantâneo de ocorrência de uma falha em um pequeno intervalo de tempo, dado que até então a falha não tenha ocorrido (LOUZADA, 2012).

De acordo com Colosimo (2009), a probabilidade da falha ocorrer em um intervalo de tempo $[t_1, t_2)$ pode ser expressa em termos da função de sobrevivência como

$$P[t_1 \leq T < t_2] = P\{T < t_2\} - P\{T < t_1\}$$

$$= 1 - P\{T \geq t_2\} - [1 - P\{T \geq t_1\}]$$

$$= P\{T \geq t_1\} - P\{T \geq t_2\}$$

$$S(t_1) - S(t_2).$$

A função de risco no intervalo $[t_1, t_2)$ é definida como a probabilidade de que a falha ocorra neste intervalo, dado que não ocorreu antes de t_1, e dividida pelo comprimento do intervalo. Assim a taxa de falha no intervalo $[t_1, t_2)$ é expressa por

$$\frac{P\{t_1 \leq T < t_2 | T > t_1\}}{(t_2 - t_1)} = \frac{P\{t_1 \leq T < t_2\}}{(t_2 - t_1)P\{T > t_1\}} = \frac{S(t_1) - S(t_2)}{(t_2 - t_1)S(t_1)}, \tag{3.2}$$

considerando o intervalo $[t, t + \Delta t]$, a expressão (3.2) pode ser reescrita como

$$\frac{S(t) - S(t + \Delta t)}{\Delta t S(t)}. \tag{3.3}$$

Logo à função de risco instantânea no tempo t condicional à sobrevivência até o tempo t, pela a expressão (3.3), fazendo Δt tender a zero, temos

$$\lambda(t) = \lim_{\Delta t \to 0^+} \frac{S(t) - S(t + \Delta t)}{\Delta t S(t)}.$$

Sabendo-se que $S(t) = 1 - F(t)$ e que a função densidade de probabilidade $f(t)$ é igual à derivada da função de distribuição acumulada $F(t)$, então temos

$$f(t) = \lim_{\Delta t \to 0^+} \frac{F(t + \Delta t) - F(t)}{\Delta t}$$

portanto a função de risco pode ser expressa como,

$$= \lim_{\Delta t \to 0^+} \frac{1 - F(t) - \{1 - F(t + \Delta t)\}}{\Delta t S(t)}$$

$$= \lim_{\Delta t \to 0^+} \frac{F(t + \Delta t) - F(t)}{\Delta t S(t)}$$

$$= \frac{1}{S(t)} \lim_{\Delta t \to 0^+} \frac{F(t + \Delta t) - F(t)}{\Delta t}$$

$$\lambda(t) = \frac{f(t)}{S(t)}.$$

Segundo Garcia (2013), quando $\lambda(t)$ é difícil de ser estimada, pode-se usar a função de risco acumulado, pois ela apresenta um estimador com excelentes propriedades.

De acordo com Colosimo (2010), a função de risco é útil para descrever a distribuição do tempo de vida de indivíduos. Ela descreve a forma em que a taxa instantânea de falha muda com o tempo.

Segundo Ramires (2013), a função de risco indica como um indivíduo envelhece. Em

três categorias básicas:

- (i) crescente, IFR *(increasing failure rate)*. A função crescente indica que o risco do paciente aumenta com transcorrer do tempo;
- (ii) decrescente, DFR *(decreasing failure rate)*. A função decrescente mostra que o risco do paciente diminui à medida que o tempo passa; e
- (iii) constante, DFC *(constant failure rate)*. A função constante indica que o risco do paciente não se altera com o passar do tempo.

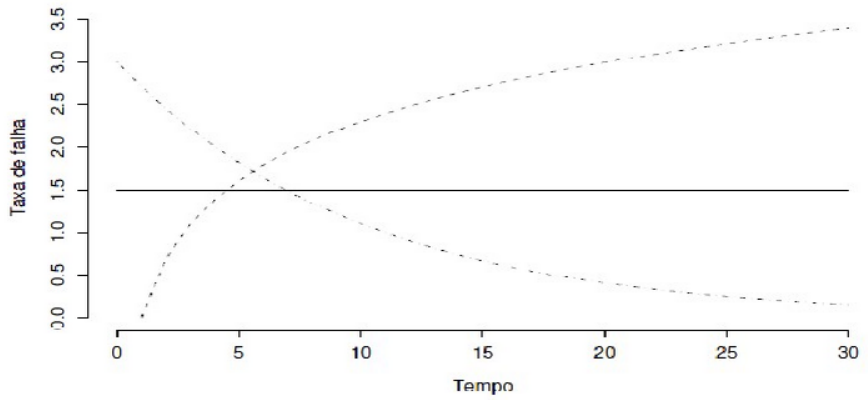

Figura 2: Representação gráfica de três funções de risco. Figura retirada de Colosimo e Giolo (2006).

A função de risco acumulado é denotada por

$$\Lambda(t) = \int_0^t \lambda(u)du \tag{3.4}$$

esta função não tem interpretação direta, mas é útil na avaliação da função de maior interesse que é função de risco, $\lambda(t)$. Existe uma relação entre $S(t)$ e $\lambda(t)$, fazendo a expressão

$$\begin{aligned}
\Lambda(t) &= -\log(S(t)) \\
\Lambda(t) &= \int_0^t \lambda(u)du \\
&= \int_0^t \frac{f(u)}{S(u)}du \\
&= \int_o^t \frac{-d}{du}[\log S(u)]du \\
&= -logS(t) + \log S(0) \\
\Lambda(t) &= -\log S(t).
\end{aligned}$$

3.1.4 Estimação das Funções de Sobrevivência e Risco

Em estudos de sobrevivência deseja-se analisar tempos de vida dos indivíduos. Uma função se destaca neste tipo de análise, que é a função de sobrevivência $S(t)$. Estes estudos frequentemente apresentam censuras. Neste caso utiliza-se os métodos não-paramétricos.

Os estimadores não-paramétricos, usam os próprios dados para estimar as quantidades necessárias da análise, sem fazer suposições à respeito da forma da distribuição do tempo de sobrevivência (HERMETO, 2014).

Uma das utilizações dos métodos não-paramétricos é em situações onde os modelos paramétricos (probabilístico) não estão se adequando aos dados, existem técnicas não-paramétricas para estimar parâmetros em análise de sobrevivência, tendo como principal característica, a opção de ajustar os dados (COLOSIMO; GIOLO, 2006).

Os estimadores de probabilidade de sobrevida, $\hat{S}(t)$ utilizados nos testes não-paramétricos se resumem em três que são: o teste de Kaplan-Meier, a tabela de vida ou atuarial, que é uma das mais antigas técnicas estatística para estimar o tempo de falha, sendo utilizada apenas em amostras homogêneas, e o estimador de Nelson-Aalen que apresenta propriedades similares ao de Kaplan-Meier (HERRMANN, 2011).

Hermeto (2014) ressalta que entre os estimadores não-paramétricos, existe superioridade do estimador de Kaplan-Meier, justificando assim a preferência na utilização deste estimador.

3.1.5 Estimador de Kaplan-Meier

Este estimador é o mais utilizado em estudos estatísticos e atuariais e foi proposto por Kaplan e Meier (1958, apud COLOSIMO; GIOLO, 2006, p.35) e é também conhecido como estimador produto-limite. Esta expressão refere-se ao fato de que a probabilidade de sobrevida até a data especificada é estimada considerando-se que a sobrevivência de cada tempo é independente da sobrevivência até outros tempos, e em consequência, a probabilidade de se chegar até o tempo t é o produto da probabilidade de chegar até a cada um dos tempos anteriores (COLOSIMO; GIOLO, 2006).

O estimador de Kaplan-Meier estima uma curva de sobrevivência incorporando a informação da censura. Este estimador é calculado na forma adaptada da função de

sobrevivência empírica, como:

$$\hat{S}(t) = \frac{\text{número total de observações que não falharam até o tempo } t}{\text{número total de observações no estudo}},$$

$\hat{S}(t)$ é uma função escada com degraus nos tempos observados de falha de tamanho $\frac{1}{n}$, onde n é o tamanho da amostra (BASTOS; ROCHA, 2006).

De acordo com Ferreira (2007), a construção do estimador de Kaplan-Meier considera o número de intervalos iguais ao número de falhas distintas e os limites dos intervalos são os próprios tempos de falhas da amostra. Sejam $t_1, t_2, ..., t_n$ os tempos de falhas de maneira que $t_1 \leq t_2 \leq ... \leq t_k$. O estimador de Kaplan-Meier é então definido como,

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

d_j: é o número de falhas no tempo t_j,
n_j: é o número de itens em risco no tempo t_j (não falhou e não foi censurado antes de t_j), para $j = 1, ..., k$.

Este estimador apresenta-se como não-viciado para grandes amostras, fracamente consistente, e é estimador de máxima verossimilhança para a função de sobrevivência $S(t)$. A estimativa não muda nos tempos censurados, o efeito dos tempos censurados é, entretanto, sentido nos valores de n_j, portanto, nos tamanhos dos degraus em $\hat{S}(t)$ (CARVALHO, 2011).

De acordo com Kaplan e Meier (1958), o estimador Kaplan-Meier de $S(t)$ é um estimador de máxima verossimilhança de $S(t)$. Seja

$$q_j = 1 - p_j = \ P[T < T_j | T \geq T_{j-1}],$$

isto é, a probabilidade do indivíduo falhar dado que sobreviveu a t_{j-1}. Considerando n_j fixo, Beslow e Crowley (1974) afirmam que

$$d_j \sim Bin(n_j, q_j)$$

$$P(d_j < d_j) = \binom{n_j}{d_j} q^{d_j} (1 - q_j)^{n_j - d_j}.$$

A função de verossimilhança é dada por

$$L(p_i) = \prod_{j=1}^{k} \binom{n_j}{d_j} q^{d_j} (1 - q_j)^{n_j - d_j}.$$

$$= \prod_{j=1}^{k} \binom{n_j}{d_j} (1-p_j)^{d_j} p^{n_j - d_j},$$

aplicando o log a função verossimilhança temos

$$l(p_j) = \sum_{j=1}^{k} \log \binom{n_j}{d_j} + \sum_{j=1}^{k} d_j \log(1-p_j) + \sum_{j=1}^{k} (n_j - d_j) \log p \ ;$$

derivando e igualando a zero tem-se

$$\frac{\partial l}{\partial p_j} = 0$$

$$\sum d_j \frac{1}{(1-\hat{p}_j)}(-1) + \sum (n_j - d_j)\frac{1}{\hat{p}_j} = 0$$

$$-\sum d_j \hat{p}_j + \sum (n_j - d_j)(1-\hat{p}_j) = 0$$

$$-\sum d_j \hat{p}_j + \sum (n_j - n_j\hat{p}_j - d_j + d_j\hat{p}_j) = 0$$

$$-\sum d_j \hat{p}_j + \sum n_j - \sum n_j \hat{p}_j - \sum d_j + \sum d_j \hat{p}_j = 0$$

$$\sum n_j \hat{p}_j = \sum n_j - \sum d_j$$

$$\Leftrightarrow \hat{p}_j = \frac{\sum n_j - \sum d_j}{\sum n_j}$$

$$= 1 - \frac{d_j}{n_j}$$

como

$$\hat{S}(t) = \prod_{j:t_j<t} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j:t_j<t} \left(1 - \frac{d_j}{n_j} \right) = \prod_{j:t_j \leq t} \hat{p}_j = \prod_{j:t_j \leq t} (1 - \hat{q}_j)$$

portanto, pelo o Teorema da Invariância, $\hat{S}(t)$ é um estimador de máxima verossimilhança.

3.1.6 Estimador Nelson-Aalen

O estimador de Nelson-Aalen é utilizado para estimar a função de risco acumulado e é denotado por

$$\hat{\Lambda}(t) = \sum_{j:t_j<t} \left(\frac{d_j}{n_j} \right).$$

Obtêm-se uma estimativa não paramétrica da função de sobrevivência, utilizando a função

$$\hat{S}(t) = \exp\left\{ -\hat{\Lambda}(t) \right\},$$

em que d_j e n_j são definidos como no estimador de Kaplan-Meier (CARVALHO, 2011).

O estimador de $\hat{\Lambda}(t)$ foi proposto inicialmente por Nelson (1972, apud COLOSIMO; GIOLO, 2006, p.35) e retomado por Aalen (1978, apud COLOSIMO; GIOLO, 2006, p.35), que provou suas propriedades assintóticas usando processos de contagem. As estimativas obtidas via estimador de Nelson-Aalen são maiores ou iguais às obtidas por meio do estimador de Kaplan-Meier (BOHORIS, 1994).

O estimador de Nelson-Aalen possui características muito semelhantes às do estimador Kaplan-Meier. A utilização do método de Nelson-Aalen é mais indicado quando se analisa dados de amostra pequena. Quando a amostra é grande os resultados dos dois métodos são equivalentes. A função $\Lambda(t)$ não tem interpretação probabilística mas tem utilidade na seleção de modelos (HOSME; LEMESHOW, 1999).

3.1.7 Estimador da Tabela de Vida ou Atuarial

A tabela de vida, que também é conhecida como método atuarial, é um dos instrumentos estatísticos mais antigos. Ela calcula as probabilidades de sobrevivência em intervalos previamente fixados. Por meio da tábua de vida é possível estimar o tempo de sobrevivência dos indivíduos de amostras homogêneas. Usa-se geralmente uma amostra de no mínimo 30 indivíduos para que se possa organizar os tempos de vida em intervalos (FERREIRA, 2007).

Para construir uma tabela de vida primeiramente divide-se o período total de observação em certo número de intervalos e para cada intervalo estima-se o valor da taxa de falha e a partir da obtenção desses valores estima-se a função de sobrevivência. O estimador da tabela de vida apresenta a mesma forma do Kaplan-Meier, mas utiliza q_j uma vez que se tenha d_j e n_j, onde

- d_j = número de falhas no intervalo $[t_{j-1}, t_j)$ e
- $n_j = [\text{número sob risco em } t_{j-1}] - \left[\frac{1}{2} \times \text{número de censuras em} [t_{j-1}, t_j)\right]$,

então a tábua de vida é denotada por

$$\hat{S}(t_j) = \prod_{i=1}^{j} (1 - \hat{q}_{i-1}), \quad j = 1, ..., k$$

(HERRMANN, 2011).

A representação gráfica da função de sobrevivência é uma escada, com valor constante em cada intervalo de tempo. A função de sobrevivência é interpretada como a

probabilidade de um indivíduo não falhar até o tempo t_j (COLOSIMO, 2009).

3.1.8 Comparação de curvas de sobrevivência

Segundo Colosimo e Giolo (2006), um dos testes mais utilizados para investigar diferenças entre grupos, curvas de sobrevivência é o Teste log-rank, que compara os valores observados e esperados de cada estrato sob a hipótese de que o risco é o mesmo em todos os grupos.

O Teste log-rank utiliza distribuição esperada de eventos igual em todos os estratos:

$$ek(t) = N(t)\frac{S_k(t)}{S(t)} \tag{3.5}$$

a estatística de teste log-rank para dois estratos ($k = 2$):

$$Log - rank = \frac{(N1 - E1)^2}{Var(N1 - E1)} \tag{3.6}$$

com $N_1 =$ ao total de eventos observados no estrato 1 e $E_1 =$ ao total de eventos esperados no estrato 1. A variância, que entra no cálculo como um fator de padronização, tem a fórmula (para $k = 2$), em que Var $(N_1 - E_1) = \nu i$, em que

$$\nu i = \sum_{t_i} \frac{S_1(t_i)[S(t_i) - S_1(t_i)]N(t_i)[S(t_i) - N(t_i)]}{S(t_i)^2[S(t_i) - 1]}.$$

A estatística log-rank, sob a hipótese nula, segue uma distribuição χ^2, com $k - 1$ graus de liberdade (COLOSIMO; GIOLO, 2006).

3.2 Métodos Paramétricos

Os métodos paramétricos são ferramentas estatísticas que são trabalhadas através dos parâmetros estimados das amostras em estudo, e geralmente possuem um número de suposições a serem seguidas que dão uma maior confiança na hora da interpretação dos resultados obtidos (COLOSIMO, 2009).

Os modelos paramétricos baseiam-se na suposição de uma distribuição de probabilidade para o tempo de sobrevivência. A utilização das distribuições de probabilidades têm se mostrado bastante adequado na análise estatística de dados de sobrevida. Estes modelos podem também ser chamados de modelos probabilísticos, devido o uso da distribuição probabilística (HERRMANN, 2011).

Os principais modelos probabilísticos utilizados na análise de sobrevivência são Exponencial, Weibull e Log-Normal pois as variáveis tratam do tempo até a falha, por outro lado, a Gaussiana (normal) e a binomial são adequadas para variáveis clínicas e industriais (MIRANDA, 2012).

3.2.1 Modelo Exponencial

O modelo Exponencial é adequado para situações em que, o tempo de falha é bem detalhado através de uma distribuição de probabilidade Exponencial. Este modelo paramétrico é apontado como o método mais simples em termos matemáticos, e é também tido como um dos mais importantes entre os modelos paramétricos. Sua importância na análise de sobrevivência é comparada com a importância de uma distribuição normal nas diversas análises da área estatística (LEE; WANG, 2003).

De acordo com Colosimo (2010), a função de densidade de probabilidade para a variável aleatória tempo de falha T, com distribuição Exponencial é dada por:

$$f(t) = \frac{1}{\alpha} \exp\left\{-\left(\frac{t}{\alpha}\right)\right\}, \qquad t \geq 0. \tag{3.7}$$

O modelo Exponencial possui apenas um parâmetro α. Este parâmetro representa o inverso do tempo médio de sobrevivência, isto é, o tempo médio de sobrevivência é dado por α. A função de sobrevivência do modelo Exponencial é dado por

$$S(t) = \exp\left\{-\left(\frac{t}{\alpha}\right)\right\} \tag{3.8}$$

e a função de risco por

$$\lambda(t) = \frac{1}{\alpha}, \qquad para \;\; t \geq 0. \tag{3.9}$$

Uma característica importante do modelo Exponencial é que ele apresenta a função de risco constante ao longo do tempo, ou seja, o risco de falha é sempre o mesmo para qualquer tempo t. O valor da função de risco é igual ao valor do parâmetro da distribuição. Se o modelo Exponencial for adequado para analisar os dados, sabe-se automaticamente, que o risco de morte para os pacientes que estão com a doença há pouco tempo ou há muito tempo, dado que ainda não tenham morrido é o mesmo (LEE; WANG, 2003).

Outras características de interesse são a média, a variância e os percentis. A média da distribuição Exponencial é α, e a variância α^2. O percentil 100p% corresponde ao tempo em que 100p% do pacientes falharam. Os percentis são importantes para obtenção

de informações sobre falhas prematuras. Eles podem ser obtidos por meio da função de densidade ou da função de sobrevivência. No caso da distribuição Exponencial, o percentil 100p%, t_p, pode ser obtido por

$$t_p = -\alpha \ \ log(1-p)$$

conhecendo o valor de α, o percentil corresponde à mediana é facilmente obtido (COLOSIMO, 2009).

A forma típica dessa três funções para diferentes valores de α podem ser observado na Figura 3, a titulo de ilustração.

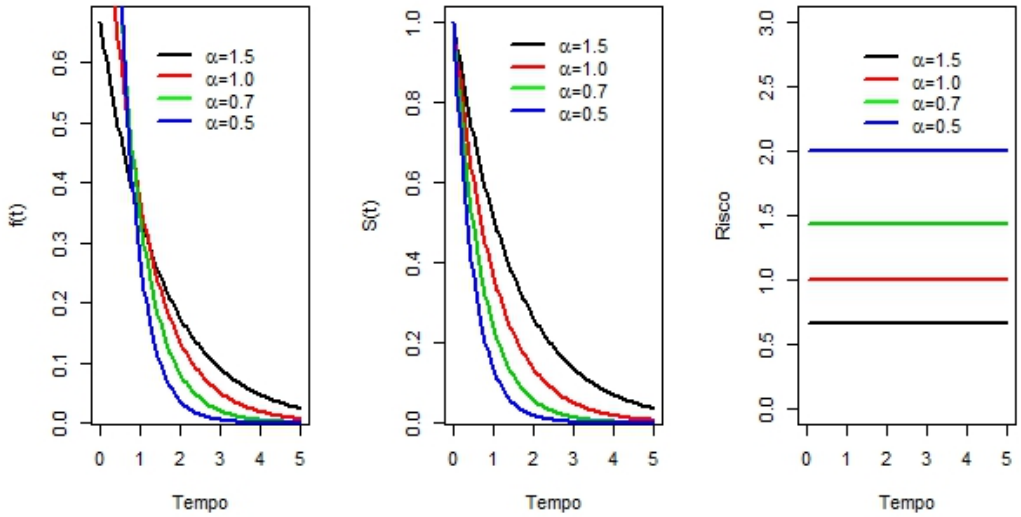

Figura 3: Forma típica das funções de densidade de probabilidade, de sobrevivência e de risco da distribuição Exponencial.

3.2.2 Modelo Weibull

O modelo Weibull é indicado para situações em que o evento de interesse (tempo de falha) é bem definido por meio da distribuição de probabilidade. Este modelo paramétrico tem se mostrado muito útil, pois ele possui uma grande variedade de formas e, por isso, consegue se adaptar a diversas situações práticas (HERRMANN, 2011).

A distribuição Weibull é usada frequentemente para estudos biomédicos, pois apresenta uma grande variedade de formas devido à sua simplicidade, todas com propriedades básicas: função de taxa de falha é monótona, isto é, crescente, decrescente ou constante

(HERRMANN, 2011).

Segundo Colosimo e Giolo (2006), a sua densidade é dada por,

$$f(t) = \frac{\gamma}{\alpha\gamma} t^{\gamma-1} \exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\}, \qquad t \geq 0 \tag{3.10}$$

em que $\gamma > 0$ e $\alpha > 0$ são parâmetros de forma e escala, respectivamente, Pode-se observar, que para $\gamma < 1$, tem-se função de taxa de falha decrescente, enquanto $\gamma > 1$ as funções de taxa de falha são crescente, e $\gamma = 1$ a função de taxa de falha é constante. O parâmetro α tem mesma unidade de medida de t, γ não tem unidade. As funções de risco e de sobrevivência do modelo Weibull são, respectivamente,

$$S(t) = \exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\} \tag{3.11}$$

e

$$\lambda(t) = \frac{\gamma}{\alpha\gamma} t^{\gamma-1} \tag{3.12}$$

para $t \geq 0$, α e $\gamma > 0$. Quando $\gamma = 1$, obtêm-se a distribuição Exponencial como caso particular da distribuição Weibull. Algumas formas das funções de densidade de sobrevivência e de risco de uma variável T com distribuição de Weibull são mostradas a titulo de ilustração na Figura 4.

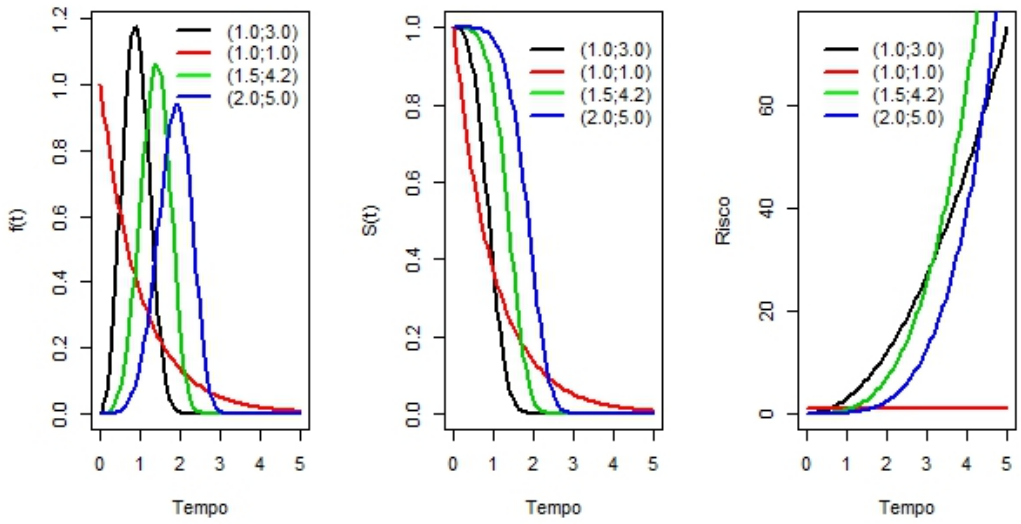

Figura 4: Forma típica das funções de densidade de probabilidade, de sobrevivência e de risco da distribuição de Weibull para alguns valores dos parâmetros (α, γ).

Os percentis da distribuição Weibull são obtidos por

$$t_p = \alpha \left[-\log(1-p)\right]^{\frac{1}{\gamma}}.$$

O modelo Weibull é um caso particular do modelo de Cox para dados de sobrevivência intervalar. Pode-se ainda mostrar que se $T \sim$ Weibull (α, γ), então, $Y = \log(\mathrm{T}) \sim$ Gumbel, ou seja, Y tem distribuição do valor extremo, com função de densidade de probabilidade, função de sobrevivência e função da taxa dado respectivamente, por

$$f(y) = \frac{1}{\sigma} \exp\left\{\left(\frac{y-\mu}{\sigma}\right) - \exp\left\{\frac{y-\mu}{\sigma}\right\}\right\}, \tag{3.13}$$

$$S(y) = \exp\left\{-\exp\left\{\frac{y-\mu}{\sigma}\right\}\right\} \tag{3.14}$$

e

$$\lambda(y) = \frac{1}{\sigma} \exp\left\{\frac{y-\mu}{\sigma}\right\} \tag{3.15}$$

em que Y e $\mu \in \mathbb{R}$ e $\sigma > 0$. Se $\mu = 0$ e $\sigma = 1$, tem-se a distribuição do valor extremo padrão. Os parâmetros μ e σ são denominados parâmetros de locação e escala, respectivamente, e relacionam-se com os parâmetros da distribuição de Weibull da seguinte forma, $\gamma = \frac{1}{\sigma}$ e $\alpha = \exp\{\mu\}$ (STRAPASSON, 2007).

3.2.3 Modelo Log-Normal

O modelo Log-Normal uma característica interessante que é, o logaritmo de uma variável com distribuição Log-Normal com parâmetros μ e σ, tem distribuição normal com média μ e desvio padrão σ (LEE; WANG, 2003).

A distribuição Log-Normal é muito utilizada para caracterizar tempos de vida de produtos e indivíduos. Ela é bastante utilizada para descrever situações clínicas (STRAPASSON, 2007).

A função densidade de uma variável aleatória T com distribuição Log-Normal é denotada por

$$f(t) = \frac{1}{\sqrt{2\pi}t\sigma} \exp\left\{-\frac{1}{2}\left(\frac{\log(t)-\mu}{\sigma}\right)^2\right\}, \qquad t > 0, \tag{3.16}$$

em que μ, é a média do logaritmo do tempo de falha, assim como σ é o desvio padrão

(COLOSIMO, 2010).

As funções de sobrevivência e risco de uma variável Log-Normal não possuem uma forma analítica explícita e são, deste modo, denotada respectivamente, por

$$S(t) = \Phi\left(\frac{-\log(t) + \mu}{\sigma}\right) \tag{3.17}$$

e

$$\lambda(t) = \frac{f(t)}{S(t)}, \tag{3.18}$$

em que $\Phi(\cdot)$, é a função de distribuição acumulada de uma normal padrão (LEE; WANG, 2003). A titulo de ilustração a Figura 5 mostra a forma típica da função de densidade de probabilidade, da função de sobrevivência e da função de risco da distribuição Log-Normal.

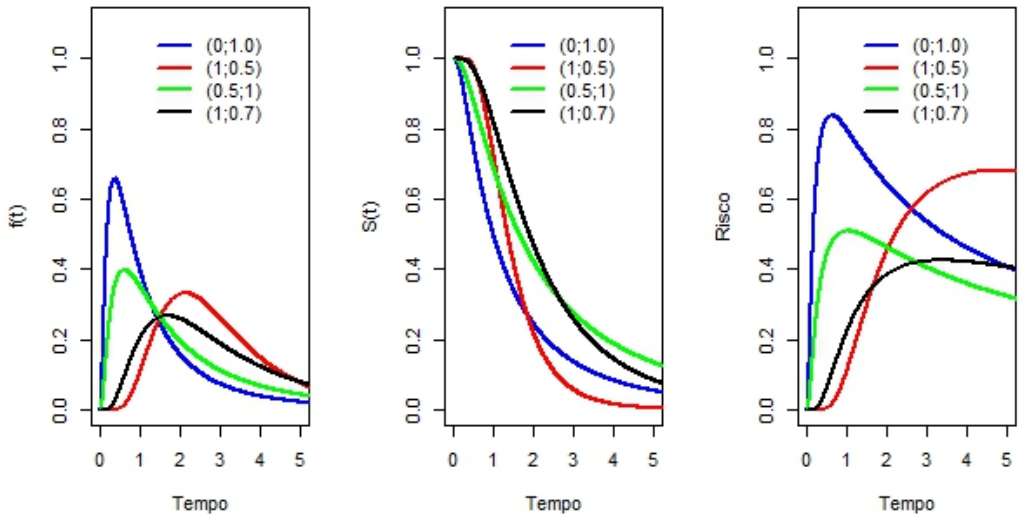

Figura 5: Forma típica das funções de densidade de probabilidade, de sobrevivência e de risco da distribuição de Log-Normal para alguns valores dos parâmetros (μ, σ).

As funções de risco não são monótonas como as da distribuição de Weibull. Elas crescem, atingem um valor máximo e depois decrescem. Os percentis para distribuição Log-Normal podem ser obtidos através da tabela da normal padrão, usando a seguinte expressão

$$t_p = \exp\{z_p\sigma + \mu\},$$

com z_p o 100p% percentil da distribuição normal padrão (COLOSIMO, 2010).

3.2.4 Modelo Gama

A distribuição Gama vem sendo utilizada em problemas de confiabilidade, por se ajustar adequadamente a uma variedade de fenômenos na análise de sobrevivência. Na área médica, recentemente está sendo utilizada para descrever o tempo de vida dos pacientes e, em outras situações que envolvem efeitos aleatórios, como é o caso dos modelos de fragilidade, esta distribuição é assumida com maior frequência para modelar estes componentes (COLOSIMO; GIOLO, 2006).

A função de densidade da distribuição Gama é caracterizada por dois parâmetros, k e α, em que $k > 0$ é chamado parâmetro de forma e $\alpha > 0$ de escala e é dada por:

$$f(t|\alpha, k) = \frac{\alpha^k}{\Gamma(k)} t^{k-1} \exp\left\{-\left(\frac{t}{\alpha}\right)\right\}, \ t > 0, \tag{3.19}$$

com $\Gamma(k)$ sendo a função Gama. Para $k > 0$, esta função densidade apresenta um único pico em $t = (k-1)/\alpha$. A respectiva função de sobrevivência desta distribuição é dada por:

$$S(t) = \int_t^\infty \frac{\alpha^k}{\Gamma(k)} u^{k-1} \exp\left\{-\left(\frac{u}{\alpha}\right)\right\} du. \tag{3.20}$$

A função taxa de falha é obtida da relação $\lambda(t) = f(t)/S(t)$, apresenta um padrão crescente ou decrescente convergindo, no entanto, para um valor constante quando t cresce de 0 a ∞ (CARVALHO, 2011). Na Figura 6 apresentamos os gráficos da função de densidade, função de sobrevivência e função de risco da distribuição Gama a titulo ilustrativo.

A média e variância da distribuição Gama são dadas, respectivamente, por $k\ \alpha$ e $k\ \alpha^2$. A distribuição Gama com parâmetro k restrito a valores inteiros $(1, 2, ...)$ é conhecida como distribuição de Erlang (LEE; WANG, 2003). Segundo Colosimo e Giolo (2006), outra distribuição que merece destaque em análise de sobrevivência é a distribuição Gama generalizada.

3.2.5 Modelo Gama Generalizada

A distribuição Gama Generalizada (GG) foi introduzida por Stacy (1962, apud COLOSIMO; GIOLO, 2006, p.80-81) e é caracterizada por três parâmetros, γ, k e α, todos

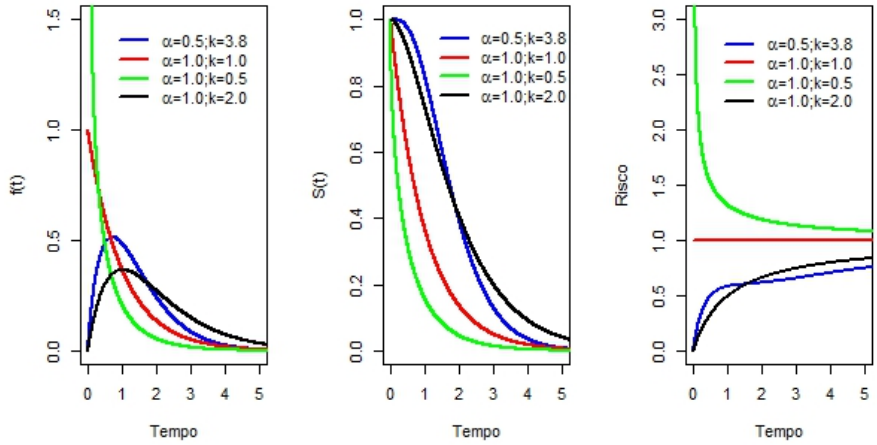

Figura 6: Forma típica das funções de densidade de probabilidade, de sobrevivência e de risco da distribuição Gama para alguns valores dos parâmetros (α, k).

positivos. Sua função de densidade é dada por:

$$f(t) = \frac{\gamma}{\Gamma(k)\alpha^{\gamma k}} t^{\gamma k - 1} \exp\left\{ -\left(\frac{t}{\alpha}\right)^{\gamma} \right\}, \ t > 0, \tag{3.21}$$

em que $\Gamma(k)$ é a função Gama. Para esta distribuição tem-se um parâmetro de escala, α, e dois de forma, γ e k, o que a torna bastante flexível. Um fato interessante sobre a distribuição GG é que ela representa uma família paramétrica e, a partir de sua função de densidade, pode obter-se os seguintes casos particulares:

- i) para $\gamma = k = 1$ tem-se $T \sim Exp(\alpha)$,
- ii) para $k = 1$ tem-se $T \sim Weibull(\gamma, \alpha)$,
- iii) e para $\gamma = 1$ tem-se $T \sim Gama(k, \alpha)$.

Pode-se, ainda, mostrar que a distribuição Log-Normal aparece como um caso limite da distribuição Gama generalizada quando $k \to \infty$ (LAWLESS, 2011).

Do que foi exposto, tem-se que a distribuição Gama Generalizada inclui, como casos especiais, as distribuições: Exponencial, de Weibull, Gama e Log-Normal. Esta propriedade da distribuição Gama Generalizada faz com que a mesma seja de grande utilidade, tipo no caso da discriminação entre modelos probabilísticos alternativos (COLOSIMO; GIOLO, 2006).

De acordo com Lawless (2011), a função de distribuição acumulada G(t), a função de sobrevivência S(t) e a função de risco $\lambda(t)$, são denotadas, respectivamente por

$$G(t) = P[T \leq t] = \frac{\gamma(k, (t/\alpha)^{\gamma})}{\Gamma(k)}$$

$$= \frac{1}{\Gamma(k)} \int_0^{(\frac{t}{\alpha})^{\gamma}} w^{k-1} exp(-w) dw$$

$$G(t) = \gamma_1 \left[k, \left(\frac{t}{\alpha} \right)^{\gamma} \right],$$

$$S(t) = 1 - G(t) = 1 - \gamma_1 \left[k, \left(\frac{t}{\alpha} \right)^{\gamma} \right]$$

a função de risco

$$\lambda(t) = \frac{f(t)}{S(t)} = \frac{t^{\gamma k - 1} \exp\left[- \left(\frac{t}{\alpha} \right)^{\gamma} \right]}{\int_0^{\infty} x^{\gamma k - 1} \exp\left[- \left(\frac{x}{\alpha} \right)^{\gamma} \right] dx}$$

onde $\gamma(k, x) = \int_0^x w^{k-1} e^{-w} dw$ é a função Gama incompleta e $\gamma_1(k, x)$ é a razão da função Gama incompleta, definida por $\gamma_1(k, x)$ =$\gamma(k, x)/\Gamma(k)$, que é facilmente implementando em vários pacotes estatísticos. Representações gráficas das funções de densidades, de sobrevivência e de risco da distribuição (GG), para α =(1.5, 1.1, 2.7, 3.9), γ =(1.2, 2.0, 4.3, 4.5) e k =(0.5, 1.3, 2.4, 0.8), podem ser observados na Figura 7 A titulo ilustrativo

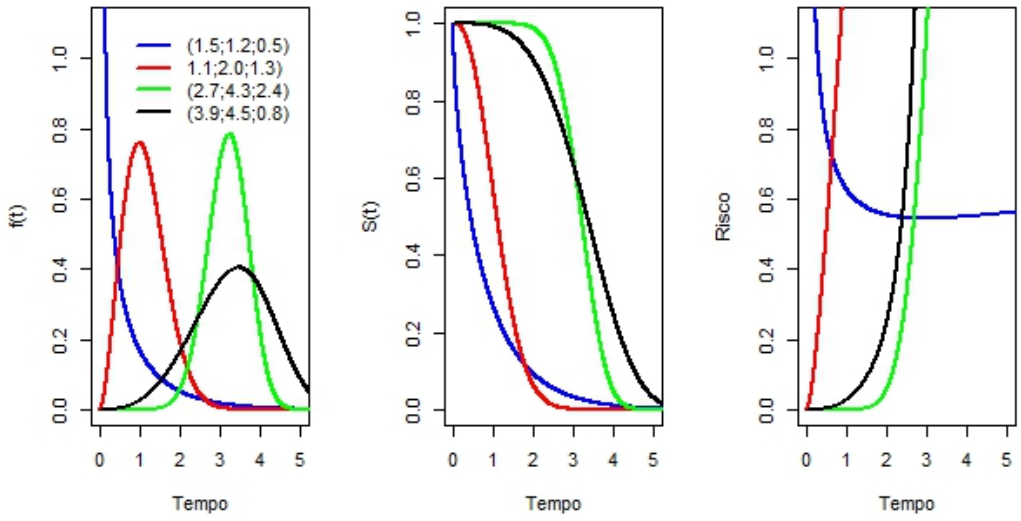

Figura 7: Forma típica das funções de densidade de probabilidade, de sobrevivência e de risco da distribuição Gama Generalizada para alguns valores dos parâmetros (α, γ, k).

3.2.6 Estimação dos parâmetros

Segundo Colosimo e Giolo (2006), os parâmetros são características dos modelos de probabilidade para estudos de tempo de vida. Estes são quantidades desconhecida e que precisam ser estimados a partir das observações amostrais. O método de máxima verossimilhança é uma opção apropriada para dados censurados, incorporando as censuras é relativamente simples, e possui propriedades para grandes amostras.

O uso de qualquer uma das distribuições probabilísticas citadas implica na estimação de seus respectivos parâmetros e como o modelo é paramétrico, será utilizado o método da máxima verossimilhança (COLOSIMO, 2009).

3.2.6.1 Método da Máxima Verossimilhança

O método de máxima verossimilhança trata o problema de estimação baseado nos resultados obtidos pela amostra e qual é a distribuição, entre todas as definidas pelos possíveis valores de seus parâmetros, com maior probabilidade de ter gerado tal amostra. Em uma amostra de n indivíduos, onde γ não são censurados (isto é, possuem tempos de falha completos) e $n - \gamma$ são censurados, e denominado θ o vetor de parâmetros desconhecidos e dados observados, respectivamente, e expressa por:

$$L(\theta) = \prod_{i=1}^{\gamma} f(t_i|\theta) \prod_{i=\gamma+1}^{n} S(t_i|\theta), \tag{3.22}$$

(BASTOS; ROCHA, 2006).

A dependência de f em θ é preciso ser mostrado, pois L é função de θ. Na expressão (3.22), θ pode estar representando um único parâmetro ou um conjunto de parâmetros (COLOSIMO; GIOLO, 2006).

A função de verossimilhança $L(\theta)$ mostra que a contribuição dos indivíduos que falharam será dada pela função densidade, enquanto que a contribuição dos censurados é dada pela função de sobrevivência. Dada a variável indicadora de falha relativa ao i-ésimo indivíduo, δ_i, a função de verossimilhança pode reescrita como

$$L(\theta) = \prod_{i=1}^{n} [f(t_i|\theta)]^{\delta_i} [S(t_i|\theta)]^{1-\delta_i} = \prod_{i=1}^{n} [\lambda(t_i|\theta)]^{\delta_i} S(t_i|\theta). \tag{3.23}$$

Em que a expressão (3.23) vale para os três tipos de mecanismo de censura. Os estimadores de máxima verossimilhança são os valores de θ que maximizam $L(\theta)$, ou

equivalentemente o log $L(\theta)$ e são encontrados resolvendo o sistema de equações

$$U(\theta) = \frac{\partial \log L(\theta)}{\partial \theta} = 0,$$

da qual solução é obtida na maioria das vezes por meio de algoritmos numéricos, como método de Newton-Rhapson. Os intervalos de confiança para estes parâmetros são obtidos através das propriedades assintóticas destes estimadores (LOUZADA, 2012).

3.2.6.2 Teste da Razão de Verossimilhança (TRV)

Segundo Lehmann (1993), o teste da razão de verossimilhança é um teste de hipóteses que compara a qualidade do ajuste de dois ou mais modelos, um modelo irrestrito com todos os parâmetros livres, e seu modelo correspondente restrito pela hipótese nula para menos parâmetros, para determinar qual modelo explica melhor os dados.

De acordo com Lehmann (1993), o TRV é baseado no log da razão entre as duas verossimilhanças, ou seja, na diferença entre o $\log L(\tilde{\theta})$ e $\log L(\hat{\theta})$. Se H_0 for verdadeiro a estatística é dada por

$$TRV = -2[\log L(\tilde{\theta}) - \log L(\hat{\theta})] \sim \chi^2_g \tag{3.24}$$

onde g é o número de restrições. Portanto, distribuído assintoticamente como uma $qui-quadrado$ com g graus de liberdade. Se o valor da estatística de teste for maior que o valor crítico ao nível de significância recomenda-se rejeitar a hipótese nula (H_0).

3.2.6.3 Critério de Informação de Akaike (AIC)

De acordo com Akaike (1973), o AIC procura uma solução satisfatória entre o bom ajuste e o princípio da parcimônia, isto é, o modelo que envolva o mínimo de parâmetros possíveis a serem estimados e que explique bem o comportamento da variável resposta.

O Critério de Informação de Akaike (AIC) admite a existência de um modelo real que descreve os dados que é desconhecido, e tenta escolher dentre um grupo de modelos avaliados, o que minimiza a divergência de Kullback-Leibler (K-L). O valor de K-L para um modelo f com parâmetros θ, em relação ao modelo "real"representado por g é

$$l(g, f\theta) = \int g \log\left(\frac{g(y)}{f(y|\theta)}\right) dy$$

Esta divergência está relacionada à informação perdida por se usar um modelo apro-

ximado e não o real. A estimativa do AIC para um determinado modelo é dada por:

$$AIC = -2L + 2k \tag{3.25}$$

em que, L é o logaritmo da função verossimilhança do modelo e k é vetor de parâmetros. O modelo com menor valor e AIC é considerado o modelo de melhor ajuste (WOLFINGE, 1993).

Segundo Bozdongan (1987), utilizando-se o AIC admite-se que dentre os modelos avaliados nenhum é considerado o que realmente descreve a relação entre a variável dependente e as variáveis explanatórias, tenta-se escolher o modelo que minimize a divergência (K-L).

3.3 Modelo de Cox

O modelo apresentado por Cox (1972) é o mais utilizado em estudo clínicos, devido a sua versatilidade. Isso se deve ao fato de que a estrutura deste modelo conta com um componente não-paramétrico e outro paramétrico, justificando sua denominação de modelo semi paramétrico.

De acordo com Walters (1999) o propósito do modelo de Cox é simultaneamente explorar os efeitos de várias variáveis sobre o tempo de sobrevivência. O modelo de Cox quando é usado na análise de sobrevivência de pacientes em ensaios clínicos e permite isolar os efeitos dos tratamentos de outras variáveis.

O modelo de Cox é expresso por

$$\lambda(t|X) = \lambda_0(t)g(X'\beta), \tag{3.26}$$

sendo g uma função não-negativa com $g(0) = 1$, $\lambda_0(t)$ é a função de risco base que representa o componente não-paramétrico e $g(X'\beta)$ o componente paramétrico do modelo. O componente paramétrico pode ser reescrito como:

$$g(X'\beta) = \exp\{X'\beta\}$$

em que, β é o vetor de parâmetros associados às p covariáveis (COX, 1972).

Segundo Taconeli (2013), o modelo de Cox trata a função de risco como uma função fatorável em duas, em que a primeira contempla a parte dinâmica do risco e a segunda é associada ao efeito que as covariáveis exercem na função de risco. Normalmente usa-se

uma função exponencial, para garantir que seu resultado seja positivo.

O modelo de Cox, também é denominado de modelo de riscos proporcionais, pois a razão entre o risco de dois indivíduos é constante no tempo, dado que a função de risco base é igual para todos os elementos da amostra. Logo a diferença entre eles será explicada apenas pelas covariáveis:

$$\frac{\lambda_i}{\lambda_j} = g(X_i'\beta - X_j'\beta). \tag{3.27}$$

Mesmo com toda essas características que envolvem sua flexibilidade, o modelo de Cox não se ajusta necessariamente a qualquer problema envolvendo dados de sobrevivência, o que torna necessário estudar sua adequação (WALTERS, 1999).

3.3.1 Modelos de Fragilidades

Na análise de sobrevivência, a forma de inclusão de efeitos aleatórios parte do modelo de riscos proporcionais com o efeito aleatório atuando multiplicativamente sobre o risco de base, da mesma forma que as covariáveis observadas. De acordo com Hanagal (2015) foi Vaupel et al., (1979) que introduziu o termo fragilidade para indicar que diferentes indivíduos estão em risco. Eles usaram o termo de fragilidade para representar um efeito aleatório não observável compartilhado por indivíduos com semelhante riscos na análise das taxas de mortalidade. O efeito aleatório descreve o excesso de risco ou fragilidade para categorias distintas, como indivíduos ou famílias, e sobre as covariáveis medidas.

Em um modelo de fragilidade é assumido que a função de risco pode ser separado em componentes multiplicativos: fragilidade, a função de risco da linha de base, e o preditor linear. Para dados classificado por grupo de genótipo, a função de risco para indivíduo j no grupo I é expressa por

$$z_i \lambda_0(t) \exp(X_{ij}'\beta), \tag{3.28}$$

onde z_i é a fragilidade comum de cada indivíduo no grupo I que é compartilhada por todos os indivíduos do grupo I e é chamado de fragilidade compartilhada, $\lambda_0(t)$ dá a função de risco linha de base e $\exp(y_{ij}'\beta)$ é o Exponencial do preditor linear. A variabilidade de z_i determina o grau de heterogeneidade entre os grupos e a sua distribuição é expressa pela função de densidade de probabilidade $g(z)$, onde $G(z)$ é interpretado como a distribuição de fragilidade genotípica na população (PRENTICE et al., 1978).

Segundo Vaupel et al. (1979), a fragilidade é uma medida de risco relativo, porque quanto maior a fragilidade de um indivíduo, em relação a alguma causa de morte, maior

a susceptibilidade do indivíduo para a causa da morte, ou seja, se um indivíduo com fragilidade 1 é chamado um indivíduo normal, em seguida, um indivíduo com fragilidade 2 está com o dobro em comparação ao perigo do indivíduo normal em um determinado tempo t. Este conceito de fragilidade assume que cada indivíduo nasce com um certo nível de fragilidade relativa e permanece nesse nível toda a sua vida.

De acordo com Taconeli (2013), um modelo de fragilidade é capaz de incluir variáveis explanatórias no modelo. A razão é que a fragilidade descreve a influência de fatores desconhecidos comuns. Se algumas covariáveis comuns estão incluídas no modelo, há variação devido às covariáveis desconhecidas que devem ser reduzida. Covariáveis comuns são comuns a todos os indivíduos do grupo.

3.3.1.1 Modelo de Fragilidade Gama

O modelo de fragilidade Gama, é expresso por:

$$\lambda_{ij} = z_j \lambda_0(t) \exp\{X'_{ij}\beta\}, \tag{3.29}$$

considerando agora as fragilidades z_j (j = 1,..., m), são assumidas serem uma amostra independente de variáveis aleatórias Z_j com distribuição Gama de média igual a 1 e variância desconhecida ξ, isto é, $Z_j \sim \Gamma(1/\xi, 1/\xi)$. A variância ξ pode ser no modelo como uma escolha natural para medir o quanto a heterogeneidade está presente (COLOSIMO, 2009).

A função de sobrevivência, com a incorporação do termo de fragilidade é expressa por:

$$S(t|X_{ij}) = [S_o(t)]^{(z_j \exp\{X'_{ij}\beta\})}, \quad t \geq 0 \tag{3.30}$$

em que $S_0(t)$ corresponde à sobrevivência de base. O modelo de fragilidade Gama utiliza no processo de estimação, a função de verossimilhança parcial penalizada. Para testar a existência de heterogeneidade ou associação entre as observações. Por meio da variância da variável de fragilidade é possível quantificar a fragilidade. Se $\xi = 0$ indica que todas a fragilidades serão iguais a 1, ficando reduzido no Modelo de Cox e valores grandes de ξ indicam um alto grau de heterogeneidade entre os grupos e forte associação dentro dos grupos (TACONELI, 2013).

Klein (2003) apresenta uma contribuição para compará-la à razão de risco em modelos com fragilidade Gama no contexto multivariado, fazendo a razão de risco entre indivíduos

de um mesmo grupo, entre dois indivíduos de grupos diferentes mas com mesmo valor nas covariáveis e finalmente comparando dois indivíduos com covariáveis diferentes e pertencente a grupos distintos.

Segundo Colosimo (2009), a forma clássica de estimar modelos de sobrevivência com fragilidade Gama é por meio do algoritmo Esperança-Maximização, conhecido como algoritmo EM. A ideia é analisar as fragilidades como dados não observáveis, que são então estimado no passo E do algoritmo, no passo M são obtidos os valores do coeficiente de regressão (os βs) que maximizam a verossimilhança parcial. O algoritmo EM utiliza o $MCMC$ (Cadeias de Markov com Método de Monte Carlo) no passo E permitindo a inclusão de fragilidade a nível de indivíduo.

3.3.1.2 Modelo de Fragilidade Log-Normal

O surgimento do modelo Log-Normal foi motivado principalmente pela necessidade de contemplar estruturas de correlação mais complicadas, que poderiam ser mais facilmente modeladas pela matriz de covariâncias de uma distribuição normal multivariada que por extensões do modelo Gama (MCGILCHRIST; AISBETT, 1991).

De acordo com Ripatti e Palmgren (2000), o procedimento assume que w tem distribuição normal p-variada com vetor de médias O_p e a matriz de covariâncias $D(\xi)$. No modelo de fragilidade compartilhado, $w_1, ..., w_p$ são variáveis independentes e identicamente distribuídas o que implica em $D(\xi) = \xi I_{pxp}$. Escrevendo a densidade de w_i como

$$f_w(w_i, \xi) = \frac{1}{\sqrt{2\pi\xi}} \exp\left\{-\frac{w_i^2}{2\xi}\right\}, \tag{3.31}$$

os autores mostraram que, utilizando o método de laplace para aproximação de integrais (COX; BARNDORFF-NIELSEN, 1989), a log-verossimilhança marginal pode ser aproximada por

$$lm(\lambda_0(t), \beta, \xi) \approx logL(\lambda_0(t), \beta, \tilde{W}, \xi) - g(\tilde{W}, \xi) - \frac{q}{2} log\xi - \frac{1}{2} log(det[K(\beta, \tilde{W}, \xi)]) \tag{3.32}$$

onde $\tilde{W}$ é solução de

$$\frac{\partial logL(\lambda_0(t), \beta, \tilde{W}, \xi)}{\partial \tilde{W}} = 0, \tag{3.33}$$

e

$$K(\beta, \tilde{W}, \xi) = \frac{\partial^2 logL(\lambda_0(t), \beta, \tilde{W}, \xi)}{\partial^2 \tilde{W}} = 0. \tag{3.34}$$

A Log-Normal para $z_i = \exp(w_i)$, em que o parâmetro ξ é a variância dos efeitos aleatório. Para a estimação do parâmetro ξ pode ser feita minimizado o critério de informação de Akaike (AIC). Alternativamente quando a fragilidade é Gama, a variância pode ser estimada através de verossimilhança perfilada. Já para fragilidades lognormais, a estimação é feita com base em verossimilhança restrita aproximada.

Os testes de hipótese aplicados aos modelos de fragilidades são os mesmos aplicados aos modelos de Cox: teste de wald e teste da razão de verossimilhança, adaptados para incluir, além dos βs a variância do efeito aleatório (COLOSIMO; GIOLO, 2006).

O teste de wald segue uma distribuição normal multivariada e, na sua forma quadrática, segue assintoticamente uma distribuição qui-quadrado, no caso dos parâmetros fixos, com um grau de liberdade. O modelo Log-Normal usa o algoritmo $REML$ para estimar os modelos de sobrevivência com fragilidades lognormais (COLOSIMO, 2009).

4 Metodologia

Nesse trabalho foi feita uma revisão teórica sobre a análise de sobrevivência e suas principais ferramentas estatísticas, onde foi estudado o tempo até a ocorrência da Morte dos pacientes após o transplante de Medula Óssea . Também foram estudadas as características que constituem o banco de dados, realizando uma aplicação dessas ferramentas estatísticas em um banco de dados. Os resultados obtidos será apresentado neste trabalho. Os dados foram obtidos do *institute for Helth & Society* da universidade de Wisconsin (*Medical College of Wisconsin*), estes dados estão descritos detalhadamente em Klein e Moeschberger (2003).

4.1 Material

O banco de dados é constituído de um total de 137 pacientes (38 LLA, 99 LMA) os quais possuem uma distinção entre os tipos de câncer, leucemia mieloide aguda (LMA) e leucemia linfoblástica aguda (LLA) foram tratados em um de quatro hospitais avaliados: 76 pacientes no Hospital da Universidade Estadual de Ohio (OSU), em Columbus, 21 pacientes de Hahnemann University (HU), na Filadélfia, 23 pacientes no Hospital St. Vincent (SVH) em Sydney, Austrália, e 17 pacientes no Alfred Hospital (AH) em Melbourne.

O estudo consiste de transplantes realizados nessas instituições no período de 1 de março de 1984 a 30 de Junho de 1989. O máximo de acompanhamento foi de 7 anos. Foram 42 pacientes que recaíram e 41 que morreram enquanto em remissão. Vinte e seis pacientes tiveram um episódio de GVHD aguda (doença do hospedeiro) e 17 pacientes tiveram recidiva ou morreram em remissão, sem suas plaquetas voltarem para níveis normais.

Vários fatores potenciais de risco foram medidos no momento do transplante. Para cada doença, os pacientes foram agrupados em categorias de risco com base no seu estado no momento de transplante. Essas categorias foram os seguintes: LLA (38 pacientes),

LMA baixo risco primeira remissão (54 pacientes), LMA e segunda remissão de alto risco ou não tratada primeira recaída (15 pacientes) ou segunda ou maior de recaída ou nunca em remissão (30 pacientes). Foi estudado o tempo até a morte dos pacientes com leucemia linfoblástica. As covariáveis utilizadas foram 10 ao todo e identificadas como:

Z_1: Idade do paciente; Z_2: Idade do doador;

Z_3: Sexo do paciente; Z_4: Sexo do doador;

Z_5: CMV do paciente; Z_6: CMV do doador;

Z_7: Tempo de espera em dias para o transplante de medula óssea; Z_8: FAB;

Z_9: Hospital; Z_{10}: MTX.

$gg1$: Grupo de risco LLA; $gg2$: Grupo LMA baixo risco;

$gg3$: Grupo de risco LMA alto risco;

a = GVHD agudo; c =GVHD crônico;

p = Plaquetas em níveis normais(recuperação das plaquetas);

4.2 Métodos

Aplicou-se o método não-paramétrico de Kaplan-Meier e o teste log-rank por grupos às distribuições paramétricas de Weibull, Exponencial, Log-Normal e Gama, para estimar a função de sobrevivência e por fim aplicou-se o modelo tradicional de Cox e os modelos de Fragilidades Gama e Log-normal por indivíduos. A aplicação das técnicas e métodos de análise de sobrevivência foi possível com auxílio do ambiente R (R CORE TEAM, 2014) e do pacote survival (THERNEAU, 2014).

5 Resultados e Discussões

A seguir demonstram-se os principais resultados obtidos a partir de uma análise realizada com o auxílio do software R.

O banco de dados com 137 pacientes (38 LLA, 54 LMA baixo risco, 45 LMA alto risco) os quais possuem uma distinção entre os tipos de câncer, leucemia mieloide aguda (LMA) e leucemia linfoblástica aguda (LLA) foram tratados em um de quatro hospitais avaliados.

A curva de sobrevivência de Kaplan-Meier foi aplicada ao banco de dados, considerando-se os três grupos constituídos por (grupo 1 - LLA; grupo 2 - LMA baixo risco e o grupo 3 - LMA alto risco). Na tabela 1 tem-se as estatísticas descritivas estimadas para cada um dos grupos pelo método de Kaplan-Meier.

Tabela 1: Kaplan Meier para os dados de Leucemia Linfoblástica.

Grupos	n	eventos	mediana	$(Limite\ inferior)$	$(Limite\ superior)$
grupo=1	38	23	487	276	NA
grupo=2	54	21	NA	1156	NA
grupo=3	45	33	318	164	547

De acordo com a Tabela 1, o grupo 1 (LLA) possui 38 pacientes dos quais 23 vieram a óbito durante o tratamento, o mesmo apresentou um tempo mediano de 478 dias, ou seja, 50% dos pacientes presentes no grupo 1 sobreviveram aproximadamente 1 ano e 4 meses após o transplante de medula óssea (TMO). Já os paciente do grupo 2 (LMA baixo risco) é composto por 54 pacientes dos quais 21 sofreram o evento de interesse (morte) durante o tratamento, o software não consegui calcular a estimativa do tempo mediano do grupo 2 e por fim o grupo 3 (LMA alto risco) é formado por 45 pacientes dos quais 33 vieram a óbito ao longo do tratamento, e apresentou um tempo mediano de 318 dias, isto significa que 50% dos pacientes sobreviveram aproximadamente 10 meses e 15 dias após TMO.

Na Figura 8, é possível observar a função de risco dos pacientes nos três grupos estimada por meio da função cumhaz.

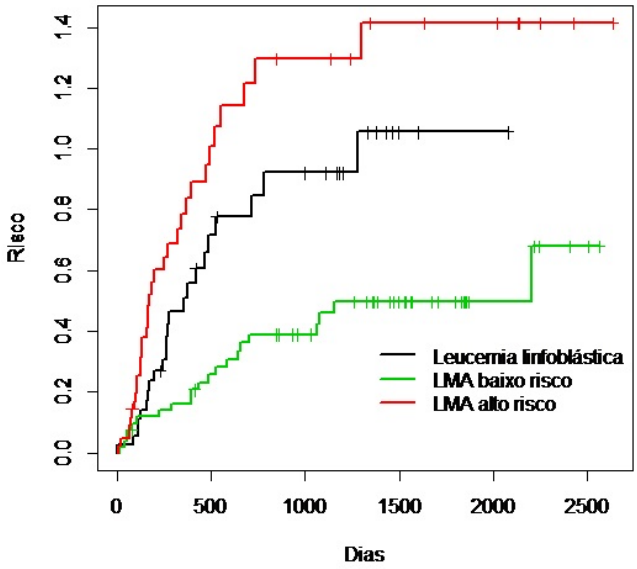

Figura 8: Estimativas da função de risco dos pacientes por grupos.

A Figura 8 mostra que o risco de óbito dos paciente aumenta com o transcorrer do tempo. Este comportamento revela um efeito gradual da gravidade da doença em cada grupo, percebe-se também que os pacientes do grupo 3 (LMA alto risco) possuem um risco maior que os demais grupos de vir a óbito, o que era esperado.

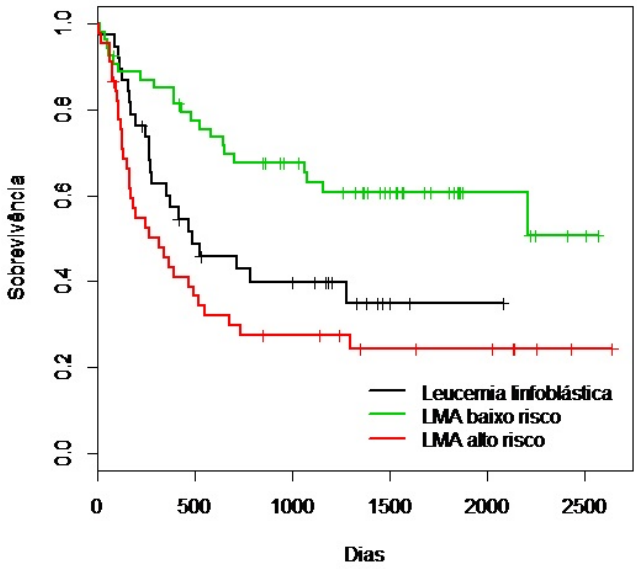

Figura 9: Estimativas da função de sobrevivência dos pacientes por grupos.

De acordo com à Figura 9, observa-se que os pacientes dos três grupos sobreviveram

mais que 1500 dias após TMO, percebe-se também que aproximadamente 60% dos pacientes do grupo (LMA baixo risco) sobreviveram mais de 1500 dias, enquanto os pacientes do grupo (LMA alto risco) aproximadamente 30% sobreviveram 1500 dias. Os pacientes do grupo (LMA alto risco) possuem a menor sobrevivência entre os grupos.

Após a aplicação do método de Kaplan-Meier, efetuou-se a comparação entre as curvas através do teste log-rank, para investigar se existe diferença entre as curvas, resultando numa estatística $\chi^2 = 16{,}3$ para 2 graus de liberdade e um p-valor igual a 0,000294, ou seja, indicando que existe diferença significativa entre as três curvas, ou seja, os pacientes possui tempos de vida diferentes entre os grupos.

Foi realizado o ajuste das distribuições paramétricas e comparou-se por meio das técnicas gráficas e do teste da razão de verossimilhança e do critério de Akaike.

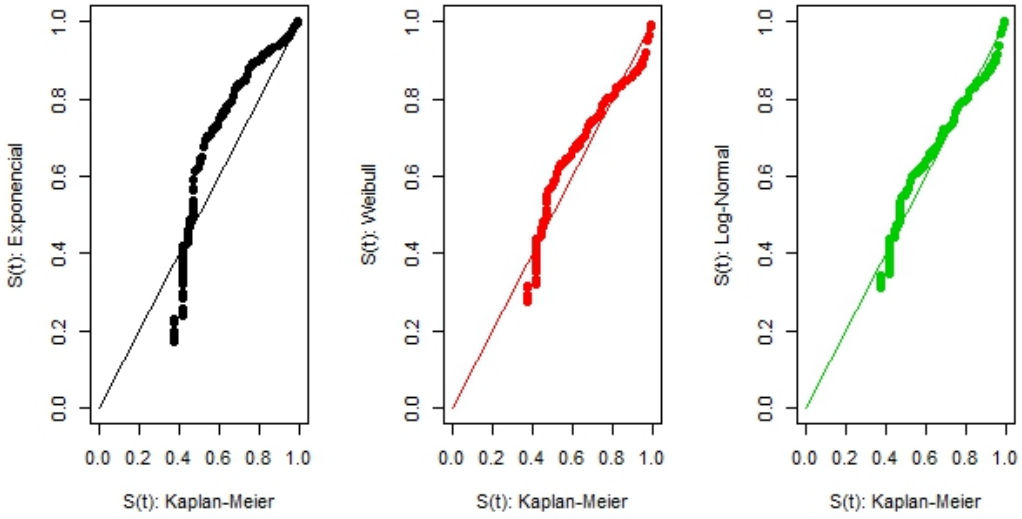

Figura 10: Gráfico das sobrevivência estimada por Kaplan-Meier versus as sobrevivências estimadas pelos modelos Exponencial, de Weibull e Log-Normal.

De acordo com os gráficos da Figura 10, é possível ver que o modelo Exponencial parece não ser adequado para esse dados, pois a curva se apresenta um tanto afastada da reta $y = x$. Por outro lado, os modelos de Weibull e Log-Normal acompanham mais de perto a reta $y = x$, indicando ser um desses modelos, possivelmente, adequado para os dados sob estudo.

A Figura 11 utilizou-se métodos de análise de sobrevivência paramétrica para verificar e determinar curvas de sobrevida dos pacientes transplantados. Foi verificado que os dois modelos apresentam resultados próximos, mas percebe-se um ajuste melhor do modelo

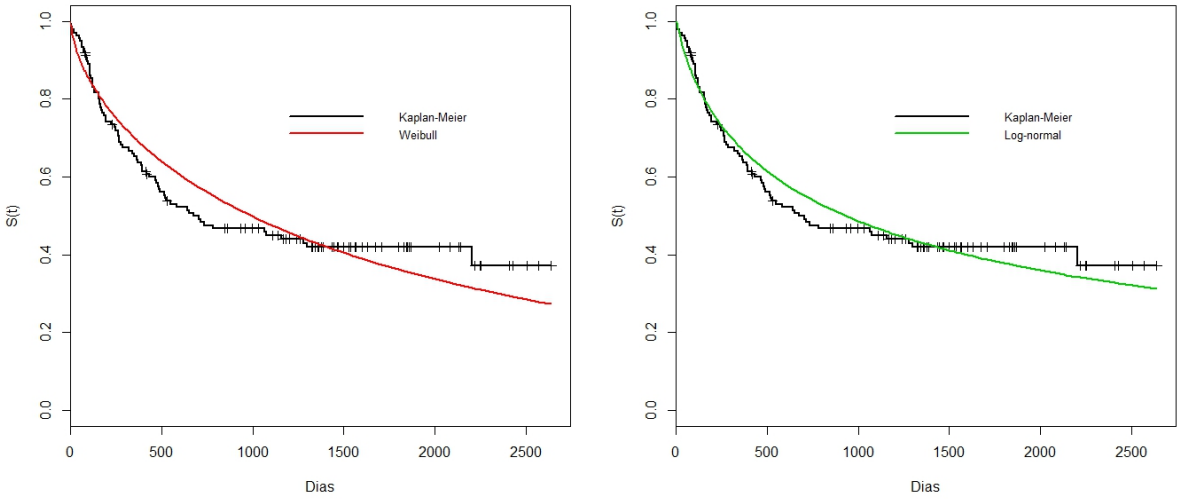

Figura 11: Curvas de sobrevivência estimadas pelos modelos de Weibul e Log-Normal versus a curva de sobrevivência estimada por Kaplan-Meier.

Log-Normal. Para confirmar, fez-se necessário aplicar o teste da razão de verossimilhança e o critério de Akaike.

Os resultados obtidos na tabela 2, indicam que a distribuição Log-Normal foi a que melhor se ajustou aos dados sob estudo.

Tabela 2: Logaritmo da função $L(\theta)$ e resultados dos TRV e AIC.

Modelo	$\log(L(\theta))$	TRV	Valor p	AIC (menor melhor)
Gama generalizado	-632.2	-	-	-
Weibull	-627.6	2(627.6-632.2)=-9.2	1	1275.34
Log-normal	-623.7	2(623.7-632.2)=-17.0	1	1268.42

Ajustou-se também o modelo de Riscos Proporcionais de Cox e os modelos de fragilidade Gama e Log-Normal (Tabelas 3, 4 e 5, respectivamente).

Tabela 3: Modelo de Riscos Proporcionais de Cox aplicado a dados de Leucemia Linfoblástica.

variáveis	exp(coef)	exp(-coef)	Limite inferior	Limite superior
p	0.3786	2.6413	0.2028	0.7067
c	0.4254	2.3506	0.2614	0.6923
z8	1.5954	0.6268	0.8930	2.8502
z9	0.7215	1.3861	0.5300	0.9821
z10	2.1616	0.4626	1.1157	4.1878
gg1	0.9318	1.0732	0.4718	1.8404
gg2	0.4358	2.2945	0.2398	0.7920

De acordo com o modelo ajustado de Riscos proporcionais Cox, observamos que

variável doença do hospedeiro (c) se mostrou como fator protetor: pacientes sem a doença do hospedeiro têm risco maior de ir a óbito do que as com a doença do hospedeiro. Este efeito protetor pode ser interpretado como um efeito indireto. A variável FAB (z8) apresentou-se como um fator de risco, enquanto que a variável hospital (z9) atuou como fator protetor, assim como a variável MTX (z10) se mostro como um fator de risco. O modelo apresentou um poder explicativo absoluto de 32,2%. A probabilidade de concordância estimada pelo modelo teve alto valor discriminatório ou preditivo (76,8%). As variáveis selecionadas foram (p, c, z8, z9, z10, gg1, gg2), que são varáveis relacionadas a volta das plaquetas as condições normais (p), ocorrência de GVHD crônico (c), FAB nível 1 FAB Grade 4 ou 5 e AML 0 caso contrário (z8), Hospital (1: Ohio State University, 2: Alferd , 3: St. Vincent, 4: Hahnemann, z9 e MTX Usado como um tratamento para GraftVersus-Host- Prophylactic (1 para sim, 0 para Não). Estas foram as variáveis que em geral apresentaram em todos os modelos estudados algum resultado significativo para o tempo até a morte do paciente.

As tabelas 4 e 5 mostram uma comparação do modelo de Fragilidade Gama com o modelo de Fragilidade Log-Normal.

Tabela 4: Modelo de Fragilidade Gama (algoritmo EM) para dados de Leucemia Linfoblástica.

Fragilidade-Gama	exp(coef)	exp(-coef)	Limite inferior	Limite superior
p	0.0094116	106.2518	1.263e-03	0.070113
c	0.1257838	7.9501	4.308e-02	0.367223
z8	5.4150324	0.1847	1.642e+00	17.856282
z9	0.5565978	1.7966	3.364e-01	0.920810
z10	5.5644895	0.1797	1.616e+00	19.163442
gg1	0.6688206	1.4952	1.690e-01	2.647367
gg2	0.0953316	10.4897	2.762e-02	0.329098

Na Tabela 4, têm-se o modelo de Fragilidade Gama (algoritmo EM), o qual utilizou 33 interações do algoritmo de Newton-Raphson para estimar a variância dos efeitos aleatórios. O modelo apresentou uma concordância estimada de alto valor discriminatório e/ou preditivo (96,6%), isto é, o modelo foi bem ajustado. A variância dos efeitos aleatórios estimada pelo modelo foi de 3,199897. A inclusão da fragilidade alterou os os efeitos das variáveis.

A Tabela 5, mostra o modelo de Fragilidade Log-Normal (algoritmo REML), ele utilizou 21 interações do algoritmo de Newton-Raphson para estimar a variância dos efeitos aleatórios, o modelo apresentou uma concordância estimada de alto valor discriminatório ou preditivo (95,9%), ou seja, o modelo também foi bem estimado. A variância estimada

Tabela 5: Modelo de Fragilidade Log-normal (algoritmo REML) para dados de Leucemia Linfoblástica.

Fragilidade-LogNormal	exp(coef)	exp(-coef)	Limite inferior	Limite superior
p	0.08425	11.8693	0.026780	0.2651
c	0.23011	4.3457	0.104061	0.5089
z8	2.40000	0.4167	0.967981	5.9505
z9	0.63312	1.5795	0.416213	0.9631
z10	3.36885	0.2968	1.242339	9.1353
gg1	0.75457	1.3253	0.255363	2.2297
gg2	0.23945	4.1763	0.094842	0.6045

dos efeitos aleatórios do modelo foi 2,5249. A fragilidade foi significativa. Os pacientes do grupo gg2 apresentou um risco menor de morte que os pacientes do gg1.

A distribuição Gama utiliza a verossimilhança perfilada para θ (método EM), enquanto a distribuição Log-normal utiliza a máxima verossimilhança restrita (método REML). Ambos os modelos avaliam a consistência usando as mesmas unidades com risco diferenciado e utilizam as estimativas dos efeitos fixos e intervalos de confiança semelhantes.

Os resultados encontrados sugerem que a distribuição Log-Normal foi a que melhor se ajustou aos dados em estudo, com base na variância dos efeitos aleatórios e no número 21 interações do algoritmo de Newton-Raphson. A Figura 12, mostra que Ambas as curvas dos modelos Log-Normal e Gama apresentam assimetria à direita, porém a assimetria do modelo Log-Normal é menor. Percebe-se também que em ambos os modelos é possível destacar três picos, justamente os três grupos considerados no estudo (LLA, LMA baixo risco e alto risco).

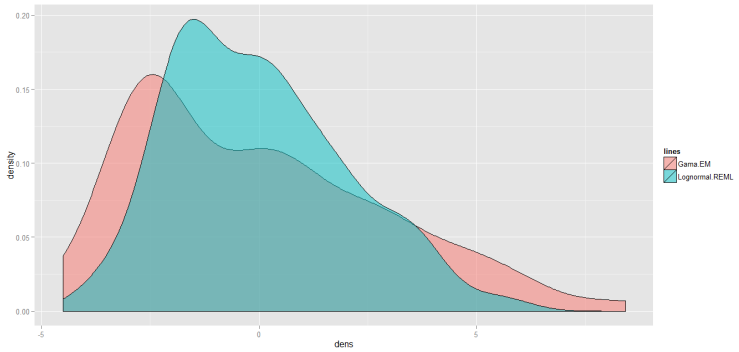

Figura 12: Distribuição das fragilidades estimadas segundo diferentes modelos para os dados de TMO - efeito dos indivíduos

6 Considerações Finais

O objetivo deste estudo foi utilizar o modelo tradicional de Cox e os modelos de Fragilidades Gama e Log-Normal a dados de pacientes transplantados de leucemia óssea (TMO). Ajustou-se os métodos clássicos e introdutórios da análise de sobrevivência.

Inicialmente, foram realizadas algumas estatísticas descritivas para observar o comportamento das covariáveis analisadas no estudo. Foi construída uma tabela de frequência para analisar o tempo mediano de vida dos pacientes nos três grupos. Foram ajustadas curvas de Kaplan-Meier aos dados de pacientes transplantados de medula óssea, as curvas indicaram existir diferença entre os grupos, aplicou-se o teste log-rank e confirmou-se que as curvas apresentaram diferença entre si. Evidenciou-se que os pacientes do grupo LMA alto risco, foram os que apresentaram a maior probabilidade de óbito em relação aos demais grupos.

A distribuição de probabilidade que foi usada para explicar os dados da amostra, foi a Log-Normal selecionada pelas técnicas gráficas e pelos testes da razão de verossimilhança e o critério de Akaike.

O modelo de Riscos Proporcionais de Cox e os modelos de fragilidades Gama e Log-Normal foram adequados para estimar o risco dos pacientes com Leucemia. O modelo de Cox com fragilidade Log-Normal revelou que o risco de morte dos pacientes do grupo LMA de baixo risco é menor que os pacientes do grupo LLA.

A partir das técnicas estatísticas usadas nesse trabalho, evidenciou-se que análise de sobrevivência é uma importante ferramenta na área da saúde desde que todos os critérios que cada técnica estatística possuem sejam seguidos de forma correta, ajudando a melhor entender o comportamento e quais características afetam os pacientes sobre risco.

7 Referências

AKAIKE, H. **Information theory as an extension of the maximum likelihood principle**. Proceedings. Budapest, Akadêmia Kiadó, p.267-281. 1973.

AALEN, O. O. **Nonparametric Inference for a Family of Counting Processes**. Annals of Statistics, 1978, vol.6, p.701-726.

BASTOS,J.; ROCHA, C. **Análise de sobrevivência: Conceitos Básicos**. Arq Med, vol.20, p.185-187. 2006.

BRESLOW, Norman et al. A large sample study of the life table and product limit estimates under random censorship. **The Annals of Statistics**, v. 2, n. 3, p. 437-453, 1974.

BOHORIS, G. A. Comparison of the cumulative-hazard and Kaplan-Meier estimators of the survivor function. **Reliability, IEEE Transactions on**, v. 43, n. 2, p. 230-232, 1994.

BOZDOGAN, Hamparsum. Model selection and Akaike's information criterion (AIC): The general theory and its analytical extensions. **Psychometrika**, v. 52, n. 3, p. 345-370, 1987.

CARVALHO, Marília Sá et al. **Análise de Sobrevivência: teoria e aplicações em saúde**. SciELO-Editora FIOCRUZ, 2011.

COLOSIMO, E.A.; GIOLO, S.R. **Análise de Sobrevivência Aplicada**. 1ª edição. São Paulo: Editora Edgard Blucher, 2006.

COLOSIMO, E.A **A Generalized Log-Normal Model for Grouped Survival Data**. Communications in Statistics. Theory and Methods, v. 39, p. 2659-2666, 2010.

COX, David R. Regression models and life-tables. **Journal of the Royal Statistical Society. Series B (Methodological)**, p. 187-220, 1972.

COX, D. R.; BARNDORFF-NIELSEN, O. E. **Inference and asymptotics**. CRC Press,

1994.

FERREIRA, J. M. **Análise de sobrevivência: uma visão de risco comporta - mental na utilização de cartão de crédito** / Joanne Medeiros Ferreira. 2007.

GARCIA, Priscila Nascimento de Alcântara. Aplicação de técnicas de análise de sobrevivência em pacientes submetidos à intervenção coronária percutânea. 2014.

HANAGAL, D. D. Modeling survival data using frailty models. 2015.

HERMETO, R. T. **Análise de Sobrevivência na Modelagem do Tempo de Vida de Redes de Sensores sem Fio** / Rodrigo Teles Hermeto. 2014.

HERRMANN, Letícia. Estimação de curvas de sobrevivência para estudos de custo-efetividade. 2011.

HOSME, D. W. JR.; LEMESHOW, S. **Applied Survival Analysis**. New York. Editora John Wisley & Sons, 1999.

KAPLAN, Edward L.; MEIER, Paul. Nonparametric estimation from incomplete observations. Journal of the American statistical association, v. 53, n. 282, p. 457-481, 1958.

KLEIN, John P.; MOESCHBERGER, Melvin L. **Survival analysis: techniques for censored and truncated data**. Springer Science & Business Media, 2003.

LAWLESS, Jerald F. **Statistical models and methods for lifetime data**. John Wiley & Sons, 2011.

LEE, Elisa T.; WANG, John. **Statistical methods for survival data analysis**. John Wiley & Sons, 2003.

LEHMANN, Erich L. The Fisher, Neyman-Pearson theories of testing hypotheses: One theory or two?. **Journal of the American Statistical Association**, v. 88, n. 424, p. 1242-1249, 1993.

LOUZADA, F.; DINIZ, C. Modelagem Estat?stica para risco de crédito. **ABE, Sao Paulo-SP**, 2012.

MCGILCHRIST, C. A.; AISBETT, C. W. Regression with frailty in survival analysis. **Biometrics**, p. 461-466, 1991.

MIRANDA, Marconi Silva et al. Técnicas não-paramétricas e paramétricas usadas na análise de sobrevivência de Chrysoperla externa (Neuroptera: Chrysopidae). 2012.

NELSON, Wayne. Theory and applications of hazard plotting for censored failure data. **Technometrics**, v. 14, n. 4, p. 945-966, 1972.

PEDROSA, F.; LINS, M. **Leucemia linfóide aguda: uma doença curáve**. Rev. bras. saúde matern. infant, v. 2, n. 1, p. 63-68, 2002.

PRENTICE, Ross L. Linear rank tests with right censored data. **Biometrika**, v. 65, n. 1, p. 167-179, 1978.

RAMIRES, Thiago Gentil. **A distribuição beta semi-normal generalizada geométrica**. 2013. Tese de Doutorado. Escola Superior de Agricultura Luiz de Queiroz.

RIPATTI, Samuli; PALMGREN, Juni. Estimation of multivariate frailty models using penalized partial likelihood. **Biometrics**, v. 56, n. 4, p. 1016-1022, 2000.

STACY, Eo W. A generalization of the gamma distribution. **The Annals of Mathematical Statistics**, p. 1187-1192, 1962.

STRAPASSON, Elizabeth. **Comparaçao de modelos com censura intervalar em análise de sobrevivência**. 2007. Tese de Doutorado. Escola Superior de Agricultura ?Luiz de Queiroz.

STRAPASSON, Elizabeth. A SIMULATION STUDY TO COMPARE IMPUTATION METHODS TO HANDLE G ROUPED SURV IV AL DATA. **Rev. Bras. Biom**, v. 27, n. 2, p. 210-224, 2009.

TACONELI, João Paulo. Modelo de Mistura Paramétrico com Fragilidade na Presença de Covariáveis. 2013.

THERNEAU T (2014). survival: A Package for Survival Analysis in S. R package version 2.37-7, URL http://CRAN.R-project.org/package=survival.

VAUPEL, James W.; MANTON, Kenneth G.; STALLARD, Eric. The impact of heterogeneity in individual frailty on the dynamics of mortality. **Demography**, v. 16, n. 3, p. 439-454, 1979.

WALTERS, Stephen John. **What is a Cox model?**. Hayward Medical Communications, 1999.

WEIBULL, W. A Statistical Theory of the Strength of Materials. Ingeniors Vetenskaps Akademien Handlingar, n.151: **The Phenomenon of Rupture in Solid**, P.293-297, 1939.

WOLFINGER, Russ. Covariance structure selection in general mixed models. **Commu-**

nications in statistics-Simulation and computation, v. 22, n. 4, p. 1079-1106, 1993.

APÊNDICE A – Programa em linguagem R utilizado para a análise

```
## Gráfico com a forma típica das funções de densidade de probabilidade,
## de sobrevivência e de risco da distribuição Exponencial:

par(mfrow=c(1,3))

alfa <- 1.5
curve((1/alfa)*exp(-x/alfa), from = 0, to = 5, col=1, lty=1, lwd=2, ylab
= "f(t)", xlab = "Tempo")
alfa1 <- 1.0
curve((1/alfa1)*exp(-x/alfa1), from = 0, to = 5,col=2, lty=1, lwd=2, add
=T)
alfa2 <- 0.7
curve((1/alfa2)*exp(-x/alfa2), from = 0, to = 5,col=3, lty=1, lwd=2, add
=T)
alfa3 <- 0.5
curve((1/alfa3)*exp(-x/alfa3), from = 0, to = 5,col=4, lty=1, lwd=2, add
=T)
legend(1,0.65, expression(paste(alpha,"=1.5"),
paste(alpha,"=1.0"),
paste(alpha,"=0.7"),
paste(alpha,"=0.5 ")),
lty=c(1,1,1,1), lwd=c(2,2,2,2),col=c('black','red','green','blue'),
bty="n", cex=1)
```

```
alfa <- 1.5
curve(exp(-x/alfa), from = 0, to = 5,col=1, lty=1, lwd=2, ylab = "S(t)",
xlab = "Tempo")
alfa <- 1.0
curve(exp(-x/alfa), from = 0, to = 5, col=2, lty=1, lwd=2, add=T)
alfa <- 0.7
curve(exp(-x/alfa), from = 0, to = 5,  col=3, lty=1,lwd=2, add=T)
alfa <- 0.5
curve(exp(-x/alfa), from = 0, to = 5, col=4, lty=1, lwd=2, add=T)
legend(1,1, expression(paste(alpha,"=1.5"),
paste(alpha,"=1.0"),
paste(alpha,"=0.7"),
paste(alpha,"=0.5 ")),
lty=c(1,1,1,1), lwd=c(2,2,2,2),col=c(‘black’,‘red’,‘green’,‘blue’),
bty="n", cex=1)

alfa <- 1.5
curve((1/alfa)*(x/x), from = 0, to = 5,ylim=c(0,3), col=1, lty=1,lwd=2,
ylab = "Risco", xlab = "Tempo")
alfa <- 1.0
curve((1/alfa)*(x/x), from = 0, to = 5,ylim=c(0,3), col=2, lty=1, lwd=2,
add=T)
alfa <- 0.7
curve((1/alfa)*(x/x), from = 0, to = 5, ylim=c(0,3), col=3, lty=1,lwd=2,
add=T)
alfa <- 0.5
curve((1/alfa)*(x/x), from = 0, to = 5,ylim=c(0,3), col=4, lty=1, lwd=2,
add=T)
legend(1,2.9, expression(paste(alpha,"=1.5"),
paste(alpha,"=1.0"),
paste(alpha,"=0.7"),
paste(alpha,"=0.5 ")),
lty=c(1,1,1,1), lwd=c(2,2,2,2),col=c(‘black’,‘red’,‘green’,‘blue’),
bty="n", cex=1)
```

```
## Gráfico da forma típica das funções de densidade de probabilidade, de
## sobrevivência e de risco da distribuição de Weibull

par(mfrow=c(1,3))

alfa <- 1.0
gama <-3.0
curve((gama/alfa^gama)*(x^(gama-1))*exp(-(x/alfa )^gama),
from = 0, to = 5,col=1, lty=1, lwd=2, ylab = "f(t)", xlab = "Tempo")
alfa <- 1.0
gama <- 1.0
curve((gama/alfa^gama)*(x^(gama-1))*exp(-(x/alfa )^gama),
from = 0, to = 5,col=2, lty=1, lwd=2, add=T)
alfa <- 1.5
gama <- 4.2
curve((gama/alfa^gama)*(x^(gama-1))*exp(-(x/alfa )^gama),
from = 0, to = 5,col=3, lty=1, lwd=2, add=T)
alfa <- 2.0
gama <- 5.0
curve((gama/alfa^gama)*(x^(gama-1))*exp(-(x/alfa )^gama),
from = 0, to = 5,col=4, lty=1, lwd=2, add=T)

legend(1.5,1.23, expression(paste("(1.0;","3.0)"),
paste("(1.0;","1.0)"),
paste("(1.5;","4.2)"),
paste("(2.0;","5.0)")),
lty=c(1,1,1,1), lwd=c(2,2,2,2),col=c('black','red','green','blue'),
bty="n", cex=1)

alfa <- 1.0
gama <- 3.0
curve(exp(-(x/alfa )^gama), from = 0, to = 5, col=1, lty=1, lwd=2,
ylab = "S(t)", xlab = "Tempo")
alfa <- 1.0
gama <- 1.0
```

```
curve(exp(-(x/alfa )^gama), from = 0, to =5,col=2, lty=1, lwd=2, add=T)
alfa <- 1.5
gama <- 4.2
curve(exp(-(x/alfa )^gama), from = 0, to = 5,col=3, lty=1, lwd=2, add=T)
alfa <- 2.0
gama <- 5.0
curve(exp(-(x/alfa )^gama), from = 0, to = 5,col=4, lty=1, lwd=2, add=T)

legend(1.5,1, expression(paste("(1.0;","3.0)"),
paste("(1.0;","1.0)"),
paste("(1.5;","4.2)"),
paste("(2.0;","5.0)")),
lty=c(1,1,1,1), lwd=c(2,2,2,2),col=c('black','red','green','blue'),
bty="n", cex=1)

alfa <- 1.0
gama <- 3.0
curve((gama/alfa^gama)*(x^(gama-1)), from = 0, to = 5,col=1,lty=1,lwd=2,
ylab = "Risco", xlab = "Tempo")
alfa <- 1.0
gama <- 1.0
curve((gama/alfa^gama)*(x^(gama-1)),from = 0, to = 5,col=2,lty=1, lwd=2,
add=T)
alfa <- 1.5
gama <- 4.2
curve((gama/alfa^gama)*(x^(gama-1)),from = 0, to = 5,col=3,lty=1, lwd=2,
add=T)
alfa <- 2.0
gama <- 5.0
curve((gama/alfa^gama)*(x^(gama-1)),from = 0, to = 5,col=4,lty=1, lwd=2,
add=T)

legend(0,75, expression(paste("(1.0;","3.0)"),
paste("(1.0;","1.0)"),
paste("(1.5;","4.2)"),
```

```
paste("(2.0;","5.0)")),
lty=c(1,1,1,1), lwd=c(2,2,2,2),col=c('black','red','green','blue'),
bty="n", cex=1)

## Gráfico da forma típica das funções de densidade de probabilidade, de
## sobrevivência e de risco da distribuição de Log-Normal

par(mfrow=c(1,3))
plot(c(0,5), c(0,1.1), type="n", xlab="Tempo", ylab="f(t)")
t<-seq(0,10,.001)

mu <- 0
sigma <- 1.0
x=(t/sigma)
f=(1/sqrt(2*pi*x^2*sigma^2)*exp((-1/2)*((log(x)-mu)/sigma)^2))

mu1 <- 1.0
sigma1 <- 0.5
f1=(1/sqrt(2*pi*x^2*sigma1^2)*exp((-1/2)*((log(x)-mu1)/sigma1)^2))

mu2 <- 0.5
sigma2 <- 1.0
f2=(1/sqrt(2*pi*x^2*sigma2^2)*exp((-1/2)*((log(x)-mu2)/sigma2)^2))
mu3 <- 1
sigma3 <- 0.7
f3=(1/sqrt(2*pi*x^2*sigma3^2)*exp((-1/2)*((log(x)-mu3)/sigma3)^2))

lines(t,f,col='blue', lty=1, lwd=2)
lines(t,f1,col='red', lty=1, lwd=2)
lines(t,f2,col='green', lty=1, lwd=2)
lines(t,f3,col='black', lty=1, lwd=2)

legend(1,1.1, expression(paste("(0;","1.0)"),
paste("(1;","0.5)"),
paste("(0.5;","1)"),
```

```
paste("(1;","0.7)")),
lty=c(1,1,1,1), lwd=c(2,2,2,2),col=c('blue','red','green','black'),
bty="n", cex=1)

plot(c(0,5), c(0,1.1), type="n", xlab="Tempo", ylab="S(t)")
t<-seq(0,10,.001)

mu <- 0
sigma <- 1.0
x=(t/sigma)
s=(pnorm((-log(x)+ mu)/sigma))

mu1 <- 1
sigma1 <- 0.5
x1=(t/sigma1)
s1=(pnorm((-log(x1)+ mu1)/sigma1))

mu2 <- 0.5
sigma2 <- 1
x2=(t/sigma2)
s2=(pnorm((-log(x2)+ mu2)/sigma2))

mu3 <- 1
sigma3 <- 0.7
x3=(t/sigma3)
s3=(pnorm((-log(x3)+ mu3)/sigma3))

lines(t,s,col='blue', lty=1, lwd=2)
lines(t,s1,col='red', lty=1, lwd=2)
lines(t,s2,col='green', lty=1, lwd=2)
lines(t,s3,col='black', lty=1, lwd=2)

legend(1,1.1, expression(paste("(0;","1.0)"),
paste("(1;","0.5)"),
paste("(0.5;","1)"),
```

```
paste("(1;","0.7)")),
lty=c(1,1,1,1), lwd=c(2,2,2,2),col=c('blue','red','green','black'),
bty="n", cex=1)

plot(c(0,5), c(0,1.1), type="n", xlab="Tempo", ylab="Risco")
t<-seq(0,10,.001)

mu <- 0
sigma <- 1.0
x=(t/sigma)
f=(1/sqrt(2*pi*x^2*sigma^2)*exp((-1/2)*((log(x)-mu)/sigma)^2))/(pnorm((-
log(x)+ mu)/sigma))

mu1 <- 1.0
sigma1 <- 0.5
x1=(t/sigma)
f1=(1/sqrt(2*pi*x^2*sigma1^2)*exp((-1/2)*((log(x1)-mu1)/sigma1)^2))/(pno
rm((-log(x1)+ mu1)/sigma1))

mu2 <- 0.5
sigma2 <- 1.0
x2=(t/sigma)
f2=(1/sqrt(2*pi*x^2*sigma2^2)*exp((-1/2)*((log(x2)-mu2)/sigma2)^2))/(pno
rm((-log(x2)+ mu2)/sigma2))
mu3 <- 1
sigma3 <- 0.7
x3=(t/sigma)
f3=(1/sqrt(2*pi*x^2*sigma3^2)*exp((-1/2)*((log(x3)-mu3)/sigma3)^2))/(pno
rm((-log(x3)+ mu3)/sigma3))

lines(t,f,col='blue', lty=1, lwd=2)
lines(t,f1,col='red', lty=1, lwd=2)
lines(t,f2,col='green', lty=1, lwd=2)
lines(t,f3,col='black', lty=1, lwd=2)
```

```
legend(1,1.1, expression(paste("(0;","1.0)"),
paste("(1;","0.5)"),
paste("(0.5;","1)"),
paste("(1;","0.7)")),
lty=c(1,1,1,1), lwd=c(2,2,2,2),col=c('blue','red','green','black'),
bty="n", cex=1)

## Gráfico da forma típica das funções de densidade de probabilidade, de
## sobrevivência e de risco da distribuição Gama

par(mfrow=c(1,3))

plot(c(0,5), c(0,1.5), type="n", xlab="Tempo", ylab="f(t)")
t<-seq(0,10,0.001)

alfa=0.5
tau=1
k=3.8
lambda=0.5
x=(t/alfa)^(tau)
f=((lambda*tau)/(alfa*gamma(k)))*(((t)/(alfa))^((tau*k)-1))*
exp(-((t/alfa)^(tau)))*((pgamma(x,k))^(lambda-1))

alfa1=1.0
tau1=1
k1=1.0
f1=((tau1)/(alfa1*(gamma(k1))))*((t/alfa1)^((tau1*k1)-1))*
(exp(-((t/alfa1)^tau1)))

alfa2=1.0
tau2=1
k2=0.5

f2=((tau2)/(alfa2*(gamma(k2))))*((t/alfa2)^((tau2*k2)-1))*
(exp(-((t/alfa2)^tau2)))
```

```
alfa4=1.0
tau4=1
k4=2.0

f3=((tau4)/(alfa4*(gamma(k4))))*((t/alfa4)^((tau4*k4)-1))*
(exp(-((t/alfa4)^tau4)))

lines(t,f,col='blue', lty=1, lwd=2)
lines(t,f1,col='red', lty=1, lwd=2)
lines(t,f2,col='green', lty=1, lwd=2)
lines(t,f3,col='black', lty=1, lwd=2)

legend(0.5,1.49, expression(paste(alpha,"=0.5;",k,"=3.8"),
paste(alpha,"=1.0;",k,"=1.0"),
paste(alpha,"=1.0;",k,"=0.5"),
paste(alpha,"=1.0;",k,"=2.0")),
lty=c(1,1,1,1), lwd=c(2,2,2,2),col=c('blue','red','green','black'),
bty="n", cex=1)

plot(c(0,5), c(0,1), type="n", xlab="Tempo", ylab="S(t)")
t<-seq(0,40,.001)

alfa=0.5
tau=1
k=3.8
x=((t/alfa)^tau)
S=(1-(pgamma(x,k)))

alfa1=1.0
tau1=1
k1=1.0
x1=((t/alfa1)^tau1)
S1=(1-(pgamma(x1,k1)))
```

```
alfa2=1.0
tau2=1
k2=0.5
x2=((t/alfa2)^tau2)
S2=(1-(pgamma(x2,k2)))

alfa3=1.0
tau3=1
k3=2.0
x3=((t/alfa3)^tau3)
S3=(1-(pgamma(x3,k3)))

lines(t,S,col='blue', lty=1, lwd=2)
lines(t,S1,col='red', lty=1, lwd=2)
lines(t,S2,col='green', lty=1, lwd=2)
lines(t,S3,col='black', lty=1, lwd=2)

legend(1.0,1, expression(paste(alpha,"=0.5;",k,"=3.8"),
paste(alpha,"=1.0;",k,"=1.0"),
paste(alpha,"=1.0;",k,"=0.5"),
paste(alpha,"=1.0;",k,"=2.0")),
lty=c(1,1,1,1), lwd=c(2,2,2,2),col=c('blue','red','green','black'),
bty="n", cex=1)

plot(c(0,5), c(0,3), type="n", xlab="Tempo", ylab="Risco")
t<-seq(0,10,0.001)

alfa=0.5
tau=1
k=3.8
lambda=0.5
x=(t/alfa)^(tau)
f=((lambda*tau)/(alfa*gamma(k)))*(((t)/(alfa))^((tau*k)-1))*
exp(-((t/alfa)^(tau)))*((pgamma(x,k))^(lambda-1))/(1-(pgamma(x,k)))
```

```
alfa1=1.0
tau1=1
k1=1.0
x1=((t/alfa1)^tau1)
f1=((tau1)/(alfa1*(gamma(k1))))*((t/alfa1)^((tau1*k1)-1))*
(exp(-((t/alfa1)^tau1)))/(1-(pgamma(x1,k1)))

alfa2=1.0
tau2=1
k2=0.5
x2=((t/alfa2)^tau2)
f2=((tau2)/(alfa2*(gamma(k2))))*((t/alfa2)^((tau2*k2)-1))*
(exp(-((t/alfa2)^tau2)))/(1-(pgamma(x2,k2)))

alfa3=1.0
tau3=1
k3=2.0
x3=((t/alfa3)^tau3)
f3=((tau3)/(alfa3*(gamma(k3))))*((t/alfa3)^((tau3*k3)-1))*
(exp(-((t/alfa3)^tau3)))/(1-(pgamma(x3,k3)))

lines(t,f,col='blue', lty=1, lwd=2)
lines(t,f1,col='red', lty=1, lwd=2)
lines(t,f2,col='green', lty=1, lwd=2)
lines(t,f3,col='black', lty=1, lwd=2)

legend(0.5,2.95, expression(paste(alpha,"=0.5;",k,"=3.8"),
paste(alpha,"=1.0;",k,"=1.0"),
paste(alpha,"=1.0;",k,"=0.5"),
paste(alpha,"=1.0;",k,"=2.0")),
lty=c(1,1,1,1), lwd=c(2,2,2,2),col=c('blue','red','green','black'),
bty="n", cex=1)

## Gráfico da forma típica das funções de densidade de probabilidade, de
## sobrevivência e de risco da distribuição Gama Generalizada
```

```
par(mfrow=c(1,3))

plot(c(0,5), c(0,1.1), type="n", xlab="Tempo", ylab="f(t)")
t<-seq(0,10,.001)

alfa=1.5
tau=1.2
k=0.5
lambda=0.5
x=(t/alfa)^(tau)
f=((lambda*tau)/(alfa*gamma(k)))*(((t)/(alfa))^((tau*k)-1))*
exp(-((t/alfa)^(tau)))*((pgamma(x,k))^(lambda-1))

alfa1=1.1
tau1=2
k1=1.3
f1=((tau1)/(alfa1*(gamma(k1))))*((t/alfa1)^((tau1*k1)-1))*
(exp(-((t/alfa1)^tau1)))

alfa2=2.7
tau2=4.3
k2=2.4
f2=((tau2)/(alfa2*(gamma(k2))))*((t/alfa2)^((tau2*k2)-1))*
(exp(-((t/alfa2)^tau2)))

alfa3=3.9
tau3=4.5
k3=0.8
f3=((tau3)/(alfa3*(gamma(k3))))*((t/alfa3)^((tau3*k3)-1))*
(exp(-((t/alfa3)^tau3)))

lines(t,f,col='blue', lty=1, lwd=2)
lines(t,f1,col='red', lty=1, lwd=2)
lines(t,f2,col='green', lty=1, lwd=2)
```

```
lines(t,f3,col='black', lty=1, lwd=2)

legend(0.5,1.1, expression(paste("(1.5;","1.2;","0.5)"),
paste("1.1;","2.0;","1.3)"),
paste("(2.7;","4.3;","2.4)"),
paste("(3.9;","4.5;","0.8)")),
lty=c(1,1,1,1), lwd=c(2,2,2,2),col=c('blue','red','green','black'),
bty="n", cex=1)

plot(c(0,5), c(0,1), type="n", xlab="Tempo", ylab="S(t)")
t<-seq(0,40,.001)

alfa=1.5
tau=1.2
k=0.5
x=((t/alfa)^tau)
S=(1-(pgamma(x,k)))

alfa1=1.1
tau1=2.0
k1=1.3
x1=((t/alfa1)^tau1)
S1=(1-(pgamma(x1,k1)))

alfa2=2.7
tau2=4.3
k2=2.4
x2=((t/alfa2)^tau2)
S2=(1-(pgamma(x2,k2)))

alfa3=3.9
tau3=4.5
k3=0.8
x3=((t/alfa3)^tau3)
S3=(1-(pgamma(x3,k3)))
```

```
lines(t,S,col='blue', lty=1, lwd=2)
lines(t,S1,col='red', lty=1, lwd=2)
lines(t,S2,col='green', lty=1, lwd=2)
lines(t,S3,col='black', lty=1, lwd=2)

plot(c(0,5), c(0,1.1), type="n", xlab="Tempo", ylab="Risco")
t<-seq(0,10,.001)

alfa=1.5
tau=1.2
k=0.5
lambda=0.5
x=(t/alfa)^(tau)
f=((lambda*tau)/(alfa*gamma(k)))*(((t)/(alfa))^((tau*k)-1))*
exp(-((t/alfa)^(tau)))*((pgamma(x,k))^(lambda-1))/(1-(pgamma(x,k)))

alfa1=1.1
tau1=2
k1=1.3
x1=((t/alfa1)^tau1)
f1=((tau1)/(alfa1*(gamma(k1))))*((t/alfa1)^((tau1*k1)-1))*
(exp(-((t/alfa1)^tau1)))/(1-(pgamma(x1,k1)))

alfa2=2.7
tau2=4.3
k2=2.4
x2=((t/alfa2)^tau2)
f2=((tau2)/(alfa2*(gamma(k2))))*((t/alfa2)^((tau2*k2)-1))*
(exp(-((t/alfa2)^tau2)))/(1-(pgamma(x2,k2)))

alfa3=3.9
tau3=4.5
k3=0.8
x3=((t/alfa2)^tau2)
```

```
f3=((tau3)/(alfa3*(gamma(k3))))*((t/alfa3)^((tau3*k3)-1))*
(exp(-((t/alfa3)^tau3)))/(1-(pgamma(x3,k3)))

lines(t,f,col='blue', lty=1, lwd=2)
lines(t,f1,col='red', lty=1, lwd=2)
lines(t,f2,col='green', lty=1, lwd=2)
lines(t,f3,col='black', lty=1, lwd=2)

# Carregando o banco de dados

dd = read.table("dados.txt", header = T, sep = )
head(dd)

# Carregando o pacote survival para análise dos dados.
require(survival)

# usando o estimador de kaplan-Meier por grupos

tempo1=c(2081, 1602, 1496, 1462, 1433, 1377, 1330, 996, 226, 1199, 1111,
         530,  1182, 1167, 418,  417,  276,  156,  781,  172,  487, 716,
         194,  371,  526,  122,  1279, 110,  243,  86,   466,  262, 162,
         262,  1,   107,  269, 350,  2569, 2506, 2409, 2218, 1857, 1829,
         1562, 1470, 1363, 1030, 860, 1258, 2246, 1870, 1799, 1709,1674,
         1568, 1527, 1324, 957,  932,  847,  848, 1850, 1843, 1535,1447,
         1384, 414,  2204, 1063, 481,  105,  641,  390,  288,  522, 79,
         1156, 583,  48,   431,  1074, 393,  10,   53,   80,   35, 1499,
         704,  653,  222,  1356,2640, 2430, 2252, 2140, 2133, 1238,1631,
         2024, 1345, 1136, 845,  491,  162,  1298, 121,  2,    62,  265,
         547,  341,  318,  195,  469,  93,   515,  183,  105,  128, 164,
         129,  122,  80,   677,  73,   168,  74,   16,   248,  732, 105,
         392,  63,   97,  153, 363 )
length(tempo1)

censura=c ( 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1,
            1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 0,
```

```
            0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
            0, 0, 0, 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1,
            1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0,
            0, 0, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1,
            1, 1, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1)
length(censura)

grupos<-c(rep(1, 38), rep(2, 54), rep(3, 45))
length(grupos)

ekm=survfit(Surv(tempo1, censura)~grupos)

# Resumo do Kaplan-Meier
summary(ekm)

# Realizando o teste log-rank
survdiff(Surv(tempo1, censura)~grupos, rho=0)

# Gráfico da função de risco
plot(ekm, lty=c(1,1,1), lwd=c(2, 2, 2), fun = "cumhaz", col = c(1, 3,2),
ylab = "Risco", xlab = "Dias", conf.int= F)
legend(1300, 0.4, c("Leucemia linfoblástica","LMA baixo risco","LMA alto
risco"), lty=c(1,1,1),lwd=c(2,2,2), col = c(1,3,2), bty= "n")

# Gráfico da função de sobrevivência por Kaplan-Meier
plot(ekm, lty=c(1, 1, 1), lwd=c(2, 2, 2), col=c(1, 3, 2), xlab="Dias",
ylab="Sobrevivência")
legend(1300, 0.2,lty=c(1, 1, 1), lwd=c(2,2,2), col=c(1,3,2), c("Leucemia
linfoblástica", "LMA baixo risco","LMA alto risco"),bty="n")

## Ajustando os modelos paramétricos

 ajust1<-survreg(Surv(tempo1,censura)~1,dist='exponential')
 ajust1
 alpha<-exp(ajust1$coefficients[1])
```

```
 alpha
 ajust2<-survreg(Surv(tempo1,censura)~1,dist='weibull')
 ajust2
 alpha<-exp(ajust2$coefficients[1])
 gama<-1/ajust2$scale
 cbind(gama, alpha)
 ajust3<-survreg(Surv(tempo1,censura)~1,dist='lognorm')
 ajust3

 ekm<-survfit(Surv(tempo1,censura)~1)
 time<-ekm$time
 st<-ekm$surv
 ste<- exp(-time/1493.052)
 stw<- exp(-(time/1756.913)^0.640918)
 stln<- pnorm((-log(time)+ 6.835316)/2.131652)
 cbind(time,st,ste,stw,stln)

# Gráfico de sobrevivência estimada por Kaplan-Meier versus a sobrevivên-
# cias estimadas pelos modelos Exponencial, de Weibull  e Log-Normal

par(mfrow=c(1,3))
plot(st, ste, pch=16, ylim=range(c(0.0,1)), col=1, xlim=range(c(0,1)),
xlab = "S(t): Kaplan-Meier", ylab="S(t): Exponencial"")
lines(c(0,1), c(0,1),col=1, type="l", lty=1)
plot(st, stw, pch=16, ylim=range(c(0.0,1)), col=2, xlim=range(c(0,1)),
xlab = "S(t): Kaplan-Meier", ylab="S(t): Weibull")
lines(c(0,1), c(0,1),col=2, type="l", lty=1)
plot(st, stln, pch=16, ylim=range(c(0.0,1)), col=3, xlim=range(c(0,1)),
xlab = "S(t): Kaplan-Meier", ylab="S(t): "Log-Normal")
lines(c(0,1), c(0,1), col=3,type="l", lty=1)

# Curvas de sobrevivência estimadas pelos modelos de weibul e Log-Normal
# versus a curva de sobrevivência estimada por Kaplan-Meier.

par(mfrow=c(1,2))
```

```
plot(ekm, conf.int=F, lty=1,lwd=2,col=1, xlab="Dias", ylab="S(t)")
lines(c(0,time),c(1,stw),lty=1,lwd=2,col=2)
legend(1000,0.8, lty=c(1, 1), lwd=c(2, 2), col=1:2, c("Kaplan-Meier",
"Weibull"), bty="n",cex=0.8)
plot(ekm, conf.int=F, lty=1,lwd=2,col=1, xlab="Dias", ylab="S(t)")
lines(c(0, time), c(1, stln), lty=1,lwd=2,col=3)
legend(1000,0.8, lty=c(1,1), lwd=c(2,2), col=c(1,3),c("Kaplan-Meier",
"Log-normal"), bty="n", cex=0.8)

# Ajustando o modelo uniparamétrico de Cox
id=seq(1:137)
length(id)
ddd=cbind(id,dd)
head(dd)
gg=as.factor(dd$g)
gg=relevel(gg,ref="3")

uni.cox=coxph(Surv(tempo1, censura) ~ p + c + z8 + z9 + z10 + gg,da
ta=dd)

# Resumo do modelo
summary(uni.cox)

# Verossimilhança para fragilidade
uni.cox$loglik

# Ajustando o modelo de Fragilidade-Gama (algoritmo EM)
mult.cox.EM=coxph(Surv(tempo1,censura) ~ p + c  + z8 + z9 + z10 + gg
+ frailty(id, sparse=FALSE), data=dd)

# Resumo do modelo
summary(mult.cox.EM)

# Verossimilhança para fragilidade
mult.cox.EM$loglik
```

```
# Ajustando modelo de Fragilidade Log-normal (algoritmo REML)
mult.coxG=coxph(Surv(tempo1,censura) ~ p + c + z8 + z9 + z10 + gg +
frailty(id, sparse=FALSE, dist="gauss"), data=dd)

# Verossimilhança para fragilidade
mult.coxG$loglik

# Resumo do modelo
summary(mult.coxG)

# Gráfico das distribuição de fragilidades estimadas segundo diferentes
# modelos para os dados de TMO - efeito dos indivíduos

# Carregando o pacote ggplot2 para análise dos dados.
require(ggplot2)
library(ggplot2)

dat <- data.frame(dens = c(mult.cox.EM$linear.
predictors, mult.coxG$linear.predictors),
lines = rep(c("Gama.EM", "Lognormal.REML"), each = 137))

# Plot.|zzz
ggplot(dat, aes(x = dens, fill = lines)) + geom_density(alpha = 0.5)
```