



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

**Pablo Lourenço Ribeiro de Almeida**

**Técnicas de Análise de Sobrevivência aplicado a dados de  
pacientes com Mieloma Múltiplo**

CAMPINA GRANDE - PB  
DEZEMBRO/2015

Pablo Lourenço Ribeiro de Almeida

**Técnicas de Análise de Sobrevivência aplicado a dados de  
pacientes com Mieloma Múltiplo**

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.

CAMPINA GRANDE - PB

DEZEMBRO/2015

É expressamente proibida a comercialização deste documento, tanto na forma impressa como eletrônica. Sua reprodução total ou parcial é permitida exclusivamente para fins acadêmicos e científicos, desde que na reprodução figure a identificação do autor, título, instituição e ano da dissertação.

A447t Almeida, Pablo Lourenço Ribeiro de.  
Técnicas de análise de sobrevivência aplicado a dados de pacientes com mieloma múltiplo [manuscrito] / Pablo Lourenço Ribeiro Almeida. - 2015.  
53 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 Estatística".

1. Análise de sobrevivência. 2. Mieloma múltiplo. 3. Estimador de Kaplan-Meier. 4. Modelos de regressão. I. Título.  
21. ed. CDD 519.544

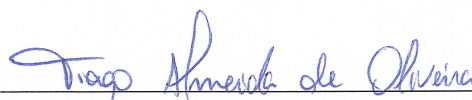
Pablo Lourenço Ribeiro de Almeida

# Técnicas de Análise de Sobrevida aplicado a dados de pacientes com Mieloma Múltiplo

Trabalho de Conclusão de Curso apresentado  
ao Curso de Bacharelado em Estatística da  
Universidade Estadual da Paraíba em cum-  
primento às exigências legais para obtenção  
do título de Bacharel em Estatística.

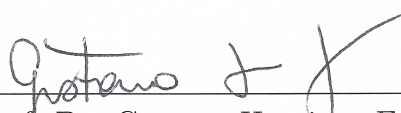
Aprovado em: 11 / 12 / 2015

## Banca Examinadora:



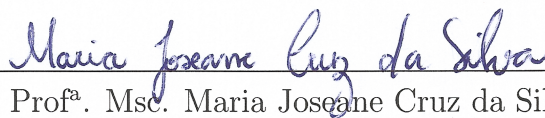
---

Prof<sup>ª</sup>. Dr. Tiago Almeida de Oliveira  
Orientador



---

Prof. Dr. Gustavo Henrique Esteves  
Universidade Estadual da Paraíba - UEPB



---

Prof<sup>ª</sup>. Msc. Maria Joseane Cruz da Silva  
Universidade Estadual da Paraíba - UEPB

# Dedicatória

Aos meus pais,  
Maria José Lourenço Ribeiro e  
Ladimir de Almeida Silva

A minha irmã,  
Poliana Lourenço Ribeiro de Almeida

Com todo amor, DEDICO.

# Agradecimentos

Primeiramente gostaria de agradecer a Deus, por nos ter concedido a dádiva da vida e por estar presente em todos os momentos dela. Através Dele obtive a coragem e sabedoria necessária para vencer os obstáculos da vida e nunca desistir dos meus sonhos.

A minha mãe, por suprir minhas necessidades, me dando vida, amor, cuidado, enxugando minhas lágrimas, corrigindo meus erros, me dando conselhos, dizendo que não era a hora certa, sendo pai e mãe ao mesmo tempo, me ajudando a levantar nos meus tombos e comemorando comigo minhas conquistas. Mãe, te levarei no meu coração por toda a eternidade.

Ao meu pai, em memória, pois mesmo não estando mais entre nós, me ensinou durante os 13 anos que passamos juntos a ser uma pessoa boa, honesta e altruísta. Levarei esses ensinamentos para toda a vida.

A minha irmã, pelo seu amor, conselhos, incentivos e por sempre apoiar as minhas decisões. Ao meu tio José Lourenço Ribeiro, por estar presente em minha vida, sempre disposto a ajudar em todos os momentos.

A minha noiva Dayene Nunes Ribeiro, pelo seu carinho, apoio e companheirismo em todos os momentos.

Aos meus amigos de infância, Cléber Reinaldo, Érica Reis, Felipe Mesquita, Júlio César Viana, Mércio Aurélio, Rafael Henrique, Tancredo Xavier e Wagner Rodrigues pela amizade de toda a vida.

Ao professor Tiago Almeida de Oliveira, meu orientador, pela amizade, paciência, incentivos e apoio que foram fundamentais para o desenvolvimento desse trabalho. A professora Ana Patrícia Bastos, pelos seus ensinamentos, amizade e por sempre me receber tão bem em sua casa nos dias que fui tirar as dúvidas desse trabalho.

A todos os professores do departamento de Estatística da UEPB.

Ao professor Pedro Cesar Pereira Coelho e o Grupo 6 Sigma, pela a oportunidade de estágio, ensinamentos e amizade.

A todos o meu sincero agradecimento.

# Resumo

A Análise de Sobrevivência pode ser caracterizada por um conjunto de técnicas estatísticas que têm como objetivo principal a análise de tempos até a ocorrência de um determinado evento de interesse, onde as observações são acompanhadas ao longo de períodos de tempo. Para este fim, foi utilizado a Análise de Sobrevivência aplicado a dados de pacientes com Mieloma Múltiplo, onde esta análise de sobrevivência é realizada para responder o tempo médio de vida desses pacientes e possíveis causas que alteram esse tempo de vida, assim como, escolher uma distribuição e obter o modelo que melhor se ajustem aos dados. Para essas análises foi utilizado técnicas paramétricas por meio de modelos de tempo de vida acelerado, técnicas não-paramétricas, tais como Kaplan-Maier e modelos de regressão de sobrevivência. Para tanto, foi possível identificar que a presença de uma covariável chamada RENAL teve influência direta no tempo de sobrevivência dos pacientes, além de obter o modelo Log-Normal como sendo o modelo que melhor descreve o comportamento de vida dos pacientes no estudo. Para a aplicação de tais técnicas, contamos com a ajuda do software R 3.2.2.

Palavras-Chave: Análise de sobrevivência; Mieloma múltiplo; Estimador kaplan-meier; Modelos de regressão.

# Abstract

The Survival Analysis can be characterized by a set of statistical techniques that have as main objective the analysis of time until the occurrence of an event of interest, where the observations are monitored over periods of time. In this work we have proposed to make a survival analysis applied to data from patients with multiple myeloma, where this survival analysis is carried out to meet the average lifespan of these patients and possible causes that alter this time. Thereby how to choose a distribution and model that best fits the data. For this analysis is intended to use parametric techniques by accelerated life time models and regression models survival, non-parametric techniques such as Kaplan-Maier. Thus, it was possible to identify the presence of a covariate called RENAL had a direct influence on patient survival time, besides getting the Log Normal model as the model that best describes the life behavior of patients in the study. For application of such techniques, we rely on the help of software R 3.2.2.

Key-Words: Survival analysis; Multiple myeloma; Kaplan-Meier Estimator; Regression models.



# Sumário

Lista de Figuras

Lista de Tabelas

<b>1</b>	<b>Introdução</b>	p. 11
<b>2</b>	<b>Objetivos</b>	p. 13
2.1	Objetivo Geral . . . . .	p. 13
2.2	Objetivos Específicos . . . . .	p. 13
<b>3</b>	<b>Fundamentação Teórica</b>	p. 14
3.1	Análise de Sobrevivência . . . . .	p. 14
3.1.1	Distribuição do Tempo de Sobrevivência . . . . .	p. 14
3.1.2	Censura . . . . .	p. 17
3.2	Modelos Não-Paramétricos . . . . .	p. 18
3.2.1	Estimador de Kaplan-Meier . . . . .	p. 19
3.3	Modelos paramétricos . . . . .	p. 20
3.3.1	Distribuição Gama Generalizada . . . . .	p. 20
3.3.2	Distribuição Exponencial . . . . .	p. 22
3.3.3	Distribuição Weibull . . . . .	p. 23
3.3.4	Distribuição Log-Normal . . . . .	p. 25
3.4	Estimação dos parâmetros . . . . .	p. 26
3.4.1	Máxima Verossimilhança . . . . .	p. 26

3.4.2	Teste Log-rank . . . . .	p. 28
3.4.3	Teste de Hipótese . . . . .	p. 30
3.4.4	Modelos de Regressão . . . . .	p. 31
3.4.4.1	Modelo de Regressão Exponencial . . . . .	p. 32
3.4.4.2	Modelo de Regressão Weibull . . . . .	p. 32
<b>4</b>	<b>Material e Métodos</b>	<b>p. 33</b>
4.1	Material em Estudo . . . . .	p. 33
4.2	Métodos Estatísticos . . . . .	p. 34
<b>5</b>	<b>Resultados e Discussão</b>	<b>p. 35</b>
<b>6</b>	<b>Considerações Finais</b>	<b>p. 46</b>
6.1	Pesquisas Futuras . . . . .	p. 46
	<b>Referências</b>	<b>p. 47</b>
	<b>Apêndice</b>	<b>p. 49</b>

# Lista de Figuras

1	Representação gráfica de censura, em que $\bullet$ representa falha e $\circ$ censura.	p. 18
2	Gráfico de Kaplan-Meier para o tempo de sobrevivência de pacientes com Mieloma Múltiplo . . . . .	p. 36
3	Gráfico de Kaplan-Meier para os dois tratamentos analisados . . . . .	p. 37
4	Gráfico de Kaplan-Meier para pacientes com doença renal e sem doença renal . . . . .	p. 39
5	Curvas ajustadas por Kaplan - Meier versus as distribuições ajustadas (a) e linearização dos modelos exponencial, Weibull e Log-Normal(b) .	p. 41
6	$\hat{S}(t)$ de kaplan-Meier e das distribuições Weibull (a) e Log-Normal (b) .	p. 42
7	Ajuste da distribuição Gama-Generalizada, curva vermelha, com as curvas de Kaplan-Meier e as distribuições Log-Normal em azul (a), Weibull em verde (b) e Exponencial em laranja (c), com respectivos intervalos de confiança . . . . .	p. 42
8	Curvas de sobrevivência estimadas pelo método de regressão Log-Normal para os dois grupos de pacientes com e sem doença renal . . . . .	p. 44

# Lista de Tabelas

1	Estimativas de sobrevivência obtidas pelo método de Kaplan-Meier . . .	p. 35
2	Estimativas de Sobrevivência de pacientes com Mieloma Múltiplo sobre dois diferentes tratamentos . . . . .	p. 37
3	Estimativas de Sobrevivência para pacientes com Mieloma Múltiplo que apresentam e não apresentam doença renal . . . . .	p. 38
4	Teste Log-rank para os dois tratamentos utilizados nos pacientes com Mieloma Múltiplo . . . . .	p. 39
5	Teste Log-rank para os pacientes com Mieloma Múltiplo que possuem ou não possuem doença renal . . . . .	p. 40
6	Sobrevivência estimada segundo as diferentes distribuições estudadas . . .	p. 40
7	Logaritmo da função $L(\theta)$ e resultados dos TRV e AIC . . . . .	p. 42
8	Estimativas dos parâmetros e logaritmo da funções de verossimilhança dos modelos de regressão Log-Normal ajustados para os dados de Mieloma Múltiplo . . . . .	p. 43
9	Resultados dos testes da Razão de Verossimilhança (TRV) . . . . .	p. 43
10	Resultados dos testes da Razão de Verossimilhança (TRV) . . . . .	p. 44

# 1 Introdução

O Mieloma Múltiplo (MM) é a segunda doença onco-hematológica mais comum no mundo, perdendo apenas para os linfomas e chegando a representar 10% dos casos. É uma doença caracterizada pela proliferação descontrolada de células plasmáticas na medula óssea com frequente produção de imunoglobulinas anômalas monoclonais (Proteína M). No Brasil, o MM representa 1% de todos os tipos de câncer, sendo o segundo mais comum entre os hematológicos, ficando atrás dos linfomas Não-Hodgkin, em adultos.

Segundo a Associação Brasileira de Linfoma e Leucemia (ABRALE), o MM tem maior incidência em pessoas idosas, em geral, maiores de 65 anos, sendo mais rara em indivíduos com menos de 35 anos (menos de 1% dos casos). Grandes avanços no tratamento desta enfermidade ocorreram desde a introdução do primeiro tratamento com melfalano e prednisona, ainda na década de 60.

O MM é uma doença sem cura definitiva até o momento. Atualmente, após mais de 30 anos da introdução do transplante na prática médica, sabe-se que o Transplante de Células-Tronco Hematopoéticas (TCTH), apesar de sua eficácia, não é um procedimento curativo do MM. O tempo de resposta ao transplante é uma variável ainda pouco previsível, sendo particular a cada paciente. No momento ainda não estão completamente elucidados os motivos que levam o indivíduo a adquirir esta doença. Entretanto, acredita-se que exista uma combinação de fatores de predisposição genética e exposição ambiental, levando a uma maior suscetibilidade para o aparecimento do mieloma. No entanto, com a moderna abordagem terapêutica, é possível controlar de forma eficaz a doença com um grande benefício nos sintomas e na qualidade de vida dos pacientes.

O uso de técnicas estatísticas tais como a análise de sobrevivência é de grande valia para melhor entender o Mieloma Múltiplo. Em Análise de Sobrevivência, a variável resposta é, geralmente, o tempo até a ocorrência de um evento de interesse. Este tempo é denominado tempo de falha. Neste contexto, podemos considerar uma falha, o tempo de reincidência da doença após o TCTH, ou o tempo de tratamento do paciente até que

ocorra o evento de interesse do estudo (morte do paciente).

A principal característica do conjunto de dados de sobrevivência é a presença de censura. A censura ocorre quando algum sujeito do estudo não experimentou o evento de interesse até o final do estudo. Por exemplo, alguns pacientes podem ainda estar vivos ou livres da doença no final do estudo, o tratamento pode ser interrompido por motivo de mudança de cidade, ou ainda, pode ocorrer a morte do paciente por motivo diferente do de interesse (doença em estudo).

Diante do exposto, o uso da estatística, mais especificamente da análise de sobrevivência pode ajudar a entender os fatores de risco para o tempo de sobrevivência de pacientes com Mieloma Múltiplo.

## 2 Objetivos

### 2.1 Objetivo Geral

O objetivo deste estudo é utilizar modelos paramétricos, não-paramétricos e de regressão aplicados a dados de pacientes com Mieloma Múltiplo.

### 2.2 Objetivos Específicos

- i) Ajustar modelo não paramétrico de Kaplan-Meier;
- ii) Utilizar métodos de análise de sobrevivência paramétricas para investigar a distribuição que melhor descreve os dados;
- iii) Aplicar os modelos de regressão paramétrica para investigar a relação entre co-variáveis e o tempo até a ocorrência do evento de interesse;
- iv) Comparar os resultados obtidos pelos modelos paramétrico e não-paramétrico.

## 3 Fundamentação Teórica

Encontram-se nesta seção as principais metodologias que servirão de base para este trabalho, no que se refere ao tempo de sobrevivência de pacientes portadores de Mieloma Múltiplo, utilizando modelos não paramétricos, paramétricos e de regressão.

### 3.1 Análise de Sobrevivência

Análise de Sobrevivência (ou Teoria da Confiabilidade) é um método estatístico usado para análise de dados de sobrevivência derivados de estudos de laboratórios ou de clínicas relacionadas a doenças agudas ou fatais.

A análise de sobrevivência utiliza dados que envolvem tempo para um certo evento (como morrer, recair, recuperar), ou seja, ela estuda o tempo em que um indivíduo sobrevive a um determinado tratamento, o tempo de resposta de um tratamento, o tempo em que um indivíduo desenvolveu uma doença, etc. Podemos exemplificar, considerando a análise da sobrevivência de pacientes infectados com um vírus letal, considerando que após ser aplicada uma determinada medicação e sabendo que o tempo de sobrevivência está sendo registrado em dias, o método pode responder: Qual o número médio de dias completados até a morte de um indivíduo? Qual a porcentagem de mortes esperada conferindo dois anos de medição? Qual o número de dias para qual 10% dos indivíduos terão morrido?

Por meio dessa análise é possível verificar a eficiência dos tratamentos, desenvolver novos produtos farmacêuticos, selecionar o tratamento mais adequado para cada situação, de acordo com os resultados nas pesquisas.

#### 3.1.1 Distribuição do Tempo de Sobrevivência

Seja  $T$  uma variável aleatória, não-negativa, que representa o tempo de vida de um indivíduo proveniente de uma população homogênea. A distribuição de  $T$  pode ser ca-



racterizada por meio da função densidade de probabilidade,  $f(t)$ , função de sobrevivência,  $S(t)$ , ou função de risco,  $h(t)$ .

A função de densidade de probabilidade é caracterizada pelo evento de interesse ao observar um indivíduo no intervalo de tempo  $[t, t+\Delta t]$  por unidade de tempo que é definida como limite da probabilidade, (LOUZADA; DINIZ, 2012). Expressa por,

$$f(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t} \quad (3.1)$$

em que,  $f(t) \geq 0$  para todo  $t$ , e tem a área abaixo da curva igual a 1.

A função de sobrevivência, denotada por  $S(t)$ , é definida como a probabilidade de um indivíduo sobreviver até um certo tempo  $t$ , sem o evento. Sendo uma das principais funções probabilísticas usadas para descrever dados de tempo de sobrevivência, definida por

$$S(t) = P(T > t) = 1 - F(t) = \int_0^t f(u) du \quad (3.2)$$

sendo que  $S(t) = 1$  quando  $t = 0$  e  $S(t) = 0$  quando  $t \rightarrow \infty$  e  $F(t) = \int_0^t f(u) du$  representa a função de distribuição acumulada.

A função de risco, ou taxa de falha, descreve a forma com que a taxa de falha muda com o tempo, ou seja, demonstra o risco do indivíduo falhar no tempo. É definida como o risco instantâneo de um indivíduo sofrer o evento entre o tempo  $t$  e  $t + \Delta t$ , dado que ele sobreviveu até o tempo  $t$ , uma definição formal é apresentada por Louzada e Diniz (2012),

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T \leq t + \Delta t \mid T \geq t)}{\Delta t}. \quad (3.3)$$

A função de risco pode ser definida, em termos da função de distribuição  $F(t)$  e da função de densidade de probabilidade  $f(t)$ , da seguinte forma:

$$h(t) = \frac{f(t)}{1 - F(t)} \quad (3.4)$$

A função de risco fornece a taxa instantânea de falha, por unidade de tempo, isto é, pode-se caracterizar classes especiais de distribuições de tempo de sobrevivência de acordo com o comportamento em relação ao tempo. Conhecida como força de mortalidade ou taxa de mortalidade condicional.

A função densidade de probabilidade é definida como a derivada da função densidade

de probabilidade acumulada,

$$f(t) = \frac{\partial F(t)}{\partial t}. \quad (3.5)$$

Como  $F(t) = 1 - S(t)$  pode-se escrever

$$f(t) = \frac{\partial[1 - S(t)]}{\partial t} = -S'(t), \quad (3.6)$$

substituindo (3.6) em (3.4) obtêm-se

$$h(t) = -\frac{S'(t)}{S(t)} = -\frac{\partial[\log S(t)]}{\partial t}. \quad (3.7)$$

Dessa forma tem-se

$$\log S(t) = -\int_0^t h(u) du \quad (3.8)$$

ou seja,

$$S(t) = \exp\left(-\int_0^t h(u) du\right). \quad (3.9)$$

Uma outra função importante é a função risco acumulada, definida como

$$H(t) = \int_0^t h(u) du \quad (3.10)$$

Substituindo-se (3.10) em (3.9) tem-se que

$$S(t) = \exp[-H(t)]. \quad (3.11)$$

Como,  $\lim_{t \rightarrow \infty} S(t) = 0$  então

$$\lim_{t \rightarrow \infty} H(t) = \infty.$$

Além disso, de (3.4) e seleção de variáveis

$$f(t) = h(t)S(t) \quad (3.12)$$

Substituindo-se (3.11) em (3.12) tem-se

$$f(t) = h(t) \exp\left(-\int_0^t h(u)du\right). \quad (3.13)$$

A expressão (3.13) é muito importante quando desenvolve-se os procedimentos de estimação somente sobre a função de risco.

### 3.1.2 Censura

A censura ocorre quando um indivíduo participante do estudo não falha. Isto habitualmente acontece nos estudos de acompanhamentos dos indivíduos ao longo do tempo. Em estudos clínicos, por exemplo, pode-se perder contato com alguns pacientes que por algum motivo passaram a residir em outra cidade, morra em função de fatores externos (que não estejam relacionados a doença em estudo), ou porque o estudo acabou sem que este paciente tenha falhado. Segundo Strapasson (2007), para análise de sobrevivência é necessário que as observações sejam representadas por um vetor  $(t_i, \delta_i, x_i)$  em que,  $t_i$  é o tempo observado de falha ou censura e  $\delta_i$  uma variável indicadora de censura, em que  $\delta_i = 1$ , o tempo observado corresponde a uma falha, ou  $\delta_i = 0$ , corresponde a uma censura. Para cada indivíduo observado tem-se uma covariável  $x_i$ , em que  $i, i=1, \dots, n$  são observações representadas pelo um par  $(t_i, \delta_i)$ . Seja  $\delta_i$  uma variável indicadora de cesnsura, temos que:

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

Pode-se ainda ocorrer outros dois tipos de censuras: censura à esquerda e censura intervalar.

Segundo Strapasson (2007), censura à esquerda ocorre quando o evento de interesse já aconteceu, quando o indivíduo foi observado: ou seja, o tempo de vida é menor que o observado.

Censura intervalar é quando não se sabe o tempo exato de ocorrência do evento de interesse, sabe-se que ele ocorreu dentro de um intervalo especificado, por exemplo, ocorre quando não se conhece o exato momento da morte, mas sabe-se que ocorreu no intervalo de tempo. Segundo Colosimo e Giolo (2006) o mecanismo de censura aleatória é aquele em que os tempos de censura são variáveis aleatórias mutuamente independentes e ainda independentes dos tempos de vida. A censura do tipo I é um caso particular da

aleatória, cuja variável aleatória  $t$ , tem uma probabilidade maior do que zero, ou seja,  $t$  é uma variável aleatória mista com um componente contínuo e outro discreto. Dados censurados são representados por sinal “+”. A representação gráfica dos exemplos de tipos de censuras são encontradas na Figura 1.

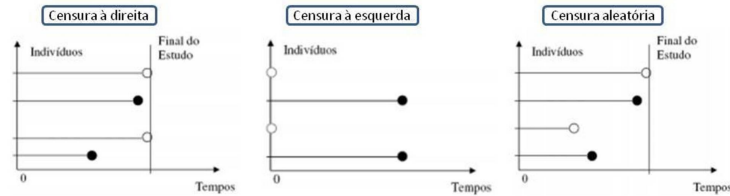


Figura 1: Representação gráfica de censura, em que  $\bullet$  representa falha e  $\circ$  censura.

## 3.2 Modelos Não-Paramétricos

As funções de análise de sobrevivência podem ser estimadas segundo três abordagens: utilizando métodos não-paramétricos, semi-paramétricos e paramétricos. Os modelos não-paramétricos são assim definidos porque não há suposição de distribuição de probabilidade ao tempo de sobrevivência. Contudo, estes modelos tem caráter basicamente descritivo, e podem ser utilizados para auxiliar na escolha do modelo adequado. Segundo Colosimo e Giolo (2006) existem técnicas não-paramétricas para estimar parâmetros em análise de sobrevivência, obtendo a opção de ajustar os dados utilizando-se os modelos paramétricos probabilísticos para tempo de falha.

Os procedimentos não-paramétricos são usados para estimação das funções de densidade de probabilidade,  $f(t)$ , da função de sobrevivência,  $S(t)$ , e da função de risco,  $h(t)$ . A função de densidade de probabilidade,  $\hat{f}(t)$ , pode ser estimada a partir dos dados amostrais se não existirem observações censuradas, a função de sobrevivência,  $\hat{S}(t)$ , é estimada a partir dos dados, como a proporção de pacientes que sobreviveram após um certo período de tempo,  $t$ , e a função de risco,  $\hat{h}(t)$ , é estimada a partir dos dados amostrais quando não existirem observações censuradas.

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 actuarial, que é uma das mais antigas técnicas estatística para estimar o tempo de falha, sendo utilizada apenas em grandes amostras. E o estimador de Nelson-Aalen que apresenta propriedades similares a Kaplan-Meier.

### 3.2.1 Estimador de Kaplan-Meier

O estimador Kaplan-Meier, também conhecido como estimador limite-produto, é provavelmente a técnica de análise de sobrevivência mais conhecida e utilizada. Este estimador, criado por Kaplan e Meier, é uma adaptação da função de sobrevivência empírica, ou seja, a função de sobrevivência estimada na ausência de censura. Como o objetivo de uma análise estatística envolvendo dados de sobrevivência está relacionado com a identificação de fatores de prognóstico para uma certa doença ou à comparação de tratamentos em estudos clínicos. Quando a amostra não contiver observações censuradas, há ocorrências de falha em um certo número no intervalo, para a função de sobrevivência é utilizado o estimador não-paramétrico de Kaplan-Meier, conforme apresentado por Colosimo e Giolo (2006).

Portanto, como a função de densidade de probabilidade e a função de risco de modelos paramétricos, a partir dos dados amostrais, não permitem a presença de observações censuradas, as quais são comuns em dados de sobrevivência e confiabilidade. As estimativas podem ser obtidas a partir de métodos não-paramétricos, que não supõem nenhuma distribuição conhecida, como a utilização da função de Kaplan-Meier no qual permite a presença de observações censuradas.

O estimador de Kaplan-Meier permite realizar testes de hipóteses que não requerem pressupostos sobre a forma da distribuição subjacentes aos dados, é usado para analisar dados medidos apenas numa escala ordinal, podendo ocorrer para dados categorizados que são medidos em escala nominal. É adequada para amostras provenientes de diversas populações, sendo usado se todas as observações falharam, ou seja, não existiram censuras. As observações censuradas informam que o tempo até a falha é maior do que aquele que foi registrado. O estimador não-paramétrico de Kaplan-Meier considera a ocorrência de falhas distintas em intervalos de tempo, onde os tempos de sobrevivência são ordenados, isto é,  $t_1 \leq t_2 \leq t_3 \leq \dots \leq t_k$ , podendo ocorrer mais de uma falha no mesmo tempo, expressado por Colosimo e Giolo (2006) por,

- i)  $t_1 \leq t_2 \leq t_3 \leq \dots \leq t_k$  tempos distintos e ordenados de falha,
- ii)  $d_j$ : número de falhas até o tempo  $t_j$ ,  $j= 1, 2, \dots, k$  e
- iii)  $n_j$ : número de itens sob risco, ou seja, os indivíduos não falharam e não censurados até  $t_j$ .

Segundo Colosimo e Giolo (2006) o estimador  $\widehat{S}(t)$  de Kaplan-Meier, é definido por,

$$\widehat{S}(t) = \left( \frac{n_1 - d_1}{n_1} \right) \cdot \left( \frac{n_2 - d_2}{n_2} \right) \cdot \dots \cdot \left( \frac{n_{t_0} - d_{t_0}}{n_{t_0}} \right) = \prod_{i, t_i < t} \frac{n_i - d_i}{n_i}$$

em que  $t_o$  é o maior tempo de falha menor que  $t$ .

As principais propriedades do estimador são: ele é não viciado para amostras grandes, é fracamente consistente, converge assintoticamente para um processo gaussiano e é estimador de máxima verossimilhança de  $S(t)$ .

### 3.3 Modelos paramétricos

Embora existam vários modelos paramétricos (probabilísticos), alguns ocupam maior destaque por sua comprovada adequação a várias situações reais, ou seja, por modelar os tempos de sobrevivência. Os principais modelos probabilísticos utilizados na análise de sobrevivência são a Gama Generalizada, o Exponencial, o Weibull e o Log-Normal, pois as variáveis tratam do tempo até a falha sendo positivos, por outro lado, a Gaussiana (normal) e a binomial são adequadas para variáveis clínicas e industriais.

A distribuição Gama Generalizada despertou o interesse de diversos pesquisadores pelo fato de representar uma família paramétrica que possui como casos particulares a distribuição Exponencial, quando  $\tau = K = 1$ . Se  $\tau = 1$  obtém-se a distribuição Gama e para  $K = 1$  tem-se a distribuição Weibull. A distribuição Exponencial é das mais simples e importantes distribuições de probabilidade utilizadas para modelagem de dados que representam o tempo até a ocorrência do evento em interesse, apresentando a função de risco constante. A distribuição Weibull é a generalização da distribuição exponencial, sendo bastante utilizada no ajuste de dados de confiabilidade em diversas áreas do conhecimento, apresenta função de risco crescente, decrescente ou ainda constante. A distribuição Log-Normal é usada para ajustar dados referentes à confiabilidade, como a distribuição Weibull, sendo que a Weibull e Log-Normal são caracterizados por dois parâmetros e a Exponencial por apenas um.

#### 3.3.1 Distribuição Gama Generalizada

A distribuição Gama Generalizada (GG) foi introduzida por Stacy (1962). Nos últimos anos diversos trabalhos envolvendo a distribuição GG foram propostos, entre os quais destacam-se, por exemplo, Nadarajah e Gupta (2007) que usaram a distribuição com

aplicações em dados de seca. Cox (2008) discutiu e comparou a família F-Generalizada com o modelo GG. Recentemente, Cordeiro *et al.* (2011) propuseram a distribuição Gama Generalizada Exponenciada.

Dessa forma, a função densidade de probabilidade da distribuição GG proposta por Stacy (1962) é dada por:

$$g(t) = \frac{\tau}{\alpha\Gamma(k)} \left(\frac{t}{\alpha}\right)^{\tau k-1} \exp\left[-\left(\frac{t}{\alpha}\right)^\tau\right], t > 0,$$

que  $\alpha > 0$  é o parâmetro de escala,  $\tau > 0$  e  $\kappa > 0$  são os parâmetros de forma e  $\Gamma(\kappa)$  é a função gama, definida por:

$$\Gamma(k) = \int_0^\infty t^{k-1} \exp(-t) dt$$

Se  $T$  é uma variável aleatória positiva com distribuição GG com parâmetros  $\alpha$ ,  $\tau$  e  $\kappa$ , então denota-se que  $T \sim GG(\alpha, \tau, \kappa)$ .

A média e a variância da distribuição GG são dadas por:

$$E(T) = \frac{\alpha\Gamma\left(\frac{\tau k+1}{\tau}\right)}{\Gamma(k)}$$

e

$$V(T) = \frac{\alpha^2}{\Gamma(k)} \left\{ \Gamma\left(\frac{\tau k+2}{\tau}\right) - \frac{\left[\Gamma\left(\frac{\tau k+1}{\tau}\right)\right]^2}{\Gamma(k)} \right\}.$$

A função da distribuição acumulada  $G(t)$ , função de sobrevivência  $S(t)$  e função de risco  $h(t)$  são expressas, respectivamente, por:

$$G(t) = P[T \leq t] = \frac{\gamma(k, (t/\alpha)^\tau)}{\Gamma(k)} = \frac{1}{\Gamma(k)} \int_0^{(t/\alpha)^\tau} w^{k-1} \exp(-w) dw = \gamma_1 \left[ k, \left(\frac{t}{\alpha}\right)^\tau \right],$$

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

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

em que  $\gamma(\kappa, x) = \int_0^x w^{\kappa-1} e^{-w} dw$  é a razão da função gama incompleta, definida por  $\gamma_1(\kappa, x) = \gamma(\kappa, x) \Gamma(\kappa)$ , que é facilmente implementada em vários pacotes estatísticos (R, SAS, Ox).

Algumas outras propriedades da distribuição GG podem ser encontradas em Lawless (1980).

### 3.3.2 Distribuição Exponencial

O modelo exponencial é adequado para situações em que o tempo de falha é bem descrito através de uma distribuição de probabilidade exponencial. Este modelo paramétrico é apontado como o modelo mais simples em termos matemáticos, e é também considerado o modelo paramétrico mais importante. Lee e Wang (2003) comparam a sua importância na análise de sobrevivência à importância de uma distribuição normal nas diversas análises da área da estatística.

A função densidade de probabilidade para a variável aleatória tempo de falha  $T$  é dada pela expressão:

$$f(t) = \lambda \exp \{ -\lambda t \}, t \geq 0 \quad \text{e} \quad \lambda > 0$$

em que o modelo exponencial apresenta apenas um parâmetro,  $\lambda$ . Este parâmetro representa o inverso do tempo médio de sobrevivência, ou seja, o tempo médio de sobrevivência é obtido por  $1/\lambda$ . A função de sobrevivência do modelo exponencial é dada por:



$$S(t) = \exp\{-\lambda t\}, t \geq 0 \quad \text{e} \quad \lambda > 0$$

e a função de falha por:

$$h(t) = \lambda, t \geq 0 \quad \text{e} \quad \lambda > 0.$$

Uma característica marcante do modelo exponencial é que ele possui a função taxa de falha constante ao longo do tempo, ou seja, o risco de falha é sempre o mesmo para qualquer tempo  $t$ . O valor da função taxa de falha é igual ao valor do parâmetro da distribuição, como pode ser visto na expressão acima. Essa propriedade é conhecida como falta de memória da distribuição exponencial.

Por exemplo, considere que um estudo está sendo realizado para investigar o tempo até a morte de pacientes com determinada doença. Se o modelo exponencial for adequado para analisar esses dados, sabe-se, automaticamente, que o risco de morte para os pacientes que estão com a doença há pouco ou há muito tempo, dado que estes pacientes ainda não tenham morrido, é o mesmo.

### 3.3.3 Distribuição Weibull

O modelo Weibull é adequado para situações em que o tempo de falha é bem descrito através de uma distribuição de probabilidade Weibull. Este modelo paramétrico tem se mostrado bastante útil porque ele apresenta uma grande variedade de formas devido à sua simplicidade e, por isso, consegue se adaptar a várias situações práticas.

A função densidade de probabilidade para a variável tempo de falha  $T$  é dada pela expressão:

$$f(t) = \lambda p t^{p-1} \exp\{-\lambda t^p\}, \quad t \geq 0$$

A função de sobrevivência do modelo Weibull é dada por:

$$S(t) = \exp\{-\lambda t^p\}$$

e a função taxa de falha por:

$$h(t) = \lambda p t^{p-1}$$

Note que o modelo exponencial é um caso particular do modelo Weibull, quando  $p = 1$ .

A função taxa de falha da distribuição Weibull tem como propriedades ser uma função monótona, ou seja, é crescente, decrescente ou constante. Ela será crescente quando  $p > 1$ , constante quando  $p = 1$  e decrescente quando  $p < 1$ .

Na função de sobrevivência, o parâmetro  $\lambda$  influencia na rapidez com que a curva decresce: valores altos para este parâmetro fazem a curva de sobrevivência decair mais rapidamente do que valores baixos. Note também que à medida que  $p$  aumenta, o decaimento da curva de sobrevivência ocorre mais rapidamente, indicando que os parâmetros  $p$  e  $\lambda$  influenciam conjuntamente a forma da curva de sobrevivência.

Para a função taxa de falha, no caso onde  $p \neq 1$ , o parâmetro  $\lambda$  influencia na rapidez com que a função taxa de falha cresce ou decresce: o crescimento será mais rápido ( $p > 1$ ) ou o decrescimento será mais lento ( $p < 1$ ), para valores maiores do parâmetro  $\lambda$ . Para  $p = 1$ , a função taxa de falha é contante com taxa de falha maior para valores maiores de  $\lambda$ .

Há situações em que é conveniente utilizar o logaritmo do tempo de falha  $T$ . Quando o tempo de falha segue distribuição Weibull, o logaritmo do tempo de falha segue uma distribuição de Gambel, também chamada de distribuição do valor extremo. Ou seja, se  $T$  segue distribuição Weibull,  $Y = \log(T)$  segue distribuição do valor extremo.

A função densidade de probabilidade, a função de sobrevivência e a função taxa de falha de  $Y$  são dadas por:

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

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

e

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

A relação entre as distribuições Weibull e Gumbel também pode ser expressa por uma relação entre os parâmetros dessas duas distribuições. Para tanto, tem-se que  $p = \frac{1}{\sigma}$  e  $\lambda = \left(\frac{1}{\exp(\mu)}\right)^p$ .

### 3.3.4 Distribuição Log-Normal

O modelo log-normal é adequado para situações em que o tempo de falha é bem descrito através de uma distribuição de probabilidade log-normal. Uma característica interessante dessa distribuição é que o logaritmo de uma variável com distribuição log-normal, com parâmetros  $\mu$  e  $\sigma$ , tem distribuição normal, com média  $\mu$  e desvio padrão  $\sigma$ .

A função densidade de probabilidade para a variável aleatória tempo de falha  $T$  é dada pela expressão:

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi i}} \exp \left\{ -\frac{1}{2} \left( \frac{\log(t) - \mu}{\sigma} \right)^2 \right\}, t > 0, \mu \in \mathfrak{R}, \sigma > 0,$$

em que  $\mu$  e  $\sigma$  representam, respectivamente, a média e o desvio padrão do logaritmo do tempo de falha.

A função de sobrevivência do modelo log-normal é dada por:

$$S(t) = \Phi \left( \frac{-\log(t) + \mu}{\sigma} \right), t > 0, \mu \in \mathfrak{R}, \sigma > 0,$$

e a função taxa de falha por:

$$h(t) = \frac{f(t)}{S(t)}, t > 0, \mu \in \mathfrak{R}, \sigma > 0,$$

em que  $\Phi(\cdot)$  é a função de distribuição acumulada de uma distribuição normal padrão. Note que essas funções não apresentam forma analítica explícita.

A característica da função taxa de falha do modelo paramétrico log-normal é inicialmente crescente e, quando atingir o ponto de máximo passar a decrescer. Valores grandes para  $\mu$  produzem curvas de sobrevivência com sobrevida maior que valores pequenos para

$\mu$ . Ou seja, quanto maior for o parâmetro  $\mu$ , maior será o tempo de sobrevivência dos pacientes. Conseqüentemente, quando menor for o valor de  $\mu$ , mais acentuado será o pico da função taxa de falha. O parâmetro influencia na variabilidade das curvas, ou seja, curvas de sobrevivência que tem valores mais altos para  $\sigma$  terão probabilidade de sobrevivência maior para tempos maiores do que curvas com valores baixos para  $\sigma$ . Uma grande variabilidade acarreta em uma função taxa de falha menor do que seria se a variabilidade fosse pequena.

### 3.4 Estimação dos parâmetros

Segundo Colosimo e Giolo (2006), os parâmetros são características dos modelos de probabilidade para estudos de tempo de vida, existindo-se alguns métodos de estimação. O método de máxima verossimilhança é uma opção apropriada para dados censurados, incorporando-se as censuras relativamente simples por possuir propriedades para grandes amostras.

#### 3.4.1 Máxima Verossimilhança

O método de máxima verossimilhança apresenta os procedimentos de estimação para os parâmetros dos modelos de sobrevivência. Utilizando esse método é possível incorporar as censuras, presentes em muitos dados de tempo de vida. São pressupostos básicos da função de verossimilhança, que as observações sejam independentes e que os tempos de sobrevivência e de censura também sejam independentes.

Segundo Carvalho *et al.* (2011), no contexto de análise de sobrevivência só conhecemos o tempo exato de sobrevivência para os que sofreram o evento, e por isso somente para esses é possível calcular a probabilidade exata do evento ocorrer através da função de densidade  $f(t)$ . Assim, no caso de dados sem censura, a função de verossimilhança  $\mathcal{L}$ , para  $i$  indivíduos na amostra, torna a seguinte forma:

$$\mathcal{L} \propto \prod_i f(t_i).$$

Quando há censuras à direita, sabemos apenas que o tempo de sobrevivência exato é maior que o observado. Logo, a contribuição da observação censurada à direita é dada pela função de sobrevivência e a função de verossimilhança é proporcional à:

$$\mathcal{L} \propto \prod_{i \in O} f(t_i) \prod_{i \in D} S(t_i),$$

em que  $O$  é o conjunto de observações que sofreram o evento e  $D$  é o conjunto de observações que foram censuradas à direita.

Quando existe censura à esquerda sabemos somente que o tempo de sobrevivência exato é menor que o observado e conseqüentemente a contribuição dessas observações é dada pela função de distribuição acumulada  $F(t) = 1 - S(t)$ . Nesse caso, a função de verossimilhança fica assim definida:

$$\mathcal{L} \propto \prod_{i \in O} f(t_i) \prod_{i \in D} S(t_i) \prod_{i \in E} [1 - S(t_i)],$$

em que  $E$  é o conjunto de observações que foram censuradas à esquerda. Para censura intervalar maiores detalhes podem ser vistos em (CARVALHO *et al.*, 2011).

Segundo Carvalho *et al.* (2011), a construção da verossimilhança é feita da mesma forma para a distribuição Weibull e Log-Normal, substituindo-se apenas as funções de densidade e sobrevivência, no entanto neste caso as equações não podem ser resolvidas analiticamente para todos os parâmetros e processos iterativos de maximização são necessários, como o método de Newton-Raphson, por exemplo.

A distribuição do tempo de falha é a Weibull, para cada combinação diferente  $\gamma$  e  $\alpha$ , tendo diferentes distribuições de Weibull. O estimador de máxima verossimilhança escolhe o par de  $\gamma$  e  $\alpha$  que melhor explique a amostra observada.

Segundo Colosimo e Giolo (2006), considera-se uma amostra de observações aleatórias  $t_1, \dots, t_n$  de uma variável aleatória  $T$  com tempos de sobrevivência e de confiabilidade de uma certa população de interesse com  $n$  observações independentes de  $t_i$ , em que  $t_i$ ,  $i = 1, \dots, n$ , indica o tempo de falha ou censura, onde todas são não-censuradas. Com um vetor de parâmetros  $\theta = (\alpha, \beta, \gamma)$ , tem-se a função de verossimilhança para um parâmetro genérico  $\theta$  da população.

O método de máxima verossimilhança é baseado geralmente para modelo em inferência paramétrica e sua teoria assintótica, onde a função de verossimilhança para o vetor de

parâmetros  $\boldsymbol{\theta}$  é expressa por (GUSMÃO *et al.*, 2011)

$$\mathcal{L}(\boldsymbol{\theta}) = \prod_{i=1}^n f(t_i; \boldsymbol{\theta})$$

cujo logaritmo é

$$\begin{aligned} l(\boldsymbol{\theta}) &= r[\log(\gamma) + \log(\beta) + \log(\alpha)] - (\beta + 1) \sum_{i \in F} \log(ti) - \gamma \alpha^\beta \sum_{i \in F} \log(ti)^{-\beta} \quad (3.14) \\ &+ \sum_{i \in C} \log \left\{ 1 - \exp \left[ -\gamma \left( \frac{\alpha}{ti} \right)^\beta \right] \right\} \end{aligned}$$

As funções de pontuação para os parâmetros  $\alpha, \beta, \gamma$  são expressadas por (GUSMÃO *et al.*, 2011)

$$U_\alpha(\boldsymbol{\theta}) = \frac{r\beta}{\alpha} - \gamma\beta\alpha^{\beta-1} \sum_{i \in F} ti^{-\beta} + \gamma\beta\alpha^{\beta-1} \sum_{i \in C} ti^{-\beta} \left( \frac{1 - ui}{ui} \right),$$

$$U_\beta(\boldsymbol{\theta}) = \frac{r}{\beta} + r \log(\alpha) - \sum_{i \in F} \log(ti) - \gamma \alpha^\beta \sum_{i \in F} ti^{-\beta} \log \left( \frac{\alpha}{ti} \right) + \gamma \alpha^\beta \sum_{i \in C} ti^{-\beta} \log \left( \frac{\alpha}{ti} \right) \left( \frac{1 - ui}{ui} \right)$$

e

$$U_\gamma(\boldsymbol{\theta}) = \frac{r}{\gamma} - \alpha^\beta \sum_{i \in F} ti^{-\beta} + \alpha^\beta \sum_{i \in C} ti^{-\beta} \left( \frac{1 - ui}{ui} \right),$$

em que  $ui = 1 - \exp \left[ -\gamma \left( \frac{\alpha}{ti} \right)^\beta \right]$  é a  $i$ -ésima observação transformada.

A estimativa do logaritmo da função de verossimilhança é obtida por meio das probabilidades de equações não-lineares,  $U_\alpha(\boldsymbol{\theta})=0$   $U_\beta(\boldsymbol{\theta})=0$  e  $U_\gamma(\boldsymbol{\theta})=0$  funções escores usando-se o algoritmo de Newton-Raphson.

Segundo Strapasson (2007), as propriedades assintóticas dos estimadores de máxima verossimilhança parcial são necessárias para construção de intervalos de confiança e testes de hipóteses sobre os parâmetros do modelo sob condições de regularidade com média  $\boldsymbol{\theta}$ , matriz de variância e covariância dada pelo inverso da matriz de Fisher ( $I(\boldsymbol{\theta})^{-1}$ ).

### 3.4.2 Teste Log-rank

Para comparar as curvas de sobrevivência mais formalmente, deve-se recorrer a testes de hipóteses. O mais simples é o teste de Mantel-Haenzel, ou log-rank, que compara os valores observados e esperados de cada estrato sob a hipótese de que o risco é o mesmo todos os grupos. Testar se curvas de sobrevivência são iguais equivale a testar se a incidência de

eventos é semelhante em cada estrato. Se for semelhante, a curva de sobrevivência será a mesma. Assim, em Carvalho *et al.* (2011), a hipótese nula é definida como:

$$H_0 : \lambda_1(t) = \lambda_2(t) = \dots = \lambda_k(t)$$

em que  $k$  é o número de estratos. Note que esta é uma hipótese que se refere à função de risco em todo o tempo de observação e não às diferenças/semelhanças em trechos da curva. Rejeitar a hipótese nula significa que pelo menos uma curva difere das outras, significativamente, em algum momento do tempo.

Segundo Carvalho *et al.* (2011), para realizar o teste, calcula-se a estatística em duas etapas: primeiro, estima-se o número de eventos esperados para cada estrato  $k$ , segundo a hipótese nula de incidência igual em todos os estratos. Chamamos esse número esperado de  $E_k(t)$ . Em seguida, calcula-se a estatística do teste. Esta estatística segue uma distribuição  $\chi^2$ , com  $k - 1$  graus de liberdade, quando a hipótese nula é verdadeira.

Formalizando, para calcular a distribuição esperada de eventos, o total de eventos precisamente no tempo  $t$ ,  $\Delta N(t)$ , é redistribuído pelos  $k$  estratos, proporcionalmente ao número de pessoas presentes em cada estrato. Assim, para cada estrato  $k$ , temos:

$$E_k(t) = N(t) \frac{R_k(t)}{R(t)}$$

em que  $\Delta N(t)$  é o número total de eventos observados em  $t$ , e  $R_k(t)$  é o número de pessoas em risco no estrato  $k$  no tempo  $t$ , e  $R(t)$  é o número total de pessoas em risco no estudo no tempo  $t$ .

Quando apenas dois estratos estão sendo comparados, a estatística log-rank é calculada utilizando-se os dados de um dos estratos somente, por exemplo, o estrato 1. O resultado do teste para um estrato se estende ao outro estrato por simetria. Sendo  $E_1$  o total de eventos esperado no estrato 1 e  $O_1$  o total de eventos observados no estrato 1, a estatística log-rank é calculada a partir da diferença entre o número total de eventos observados e o número total de eventos esperados:

$$\text{Log - rank} = \frac{(O_1 - E_1)^2}{\text{Var}(O_1 - E_1)}$$

que segue uma distribuição  $\chi^2$  com um grau de liberdade.

A variância, que entra no cálculo como um fator de padronização, tem a fórmula (para  $k = 2$ ).

$$\text{Var}(O_1 - E_1) = \sum_t \frac{R_1(t)R_2(t)\Delta N(t) [R(t) - \Delta N(t)]}{R(t)^2 [R(t) - 1]}.$$

### 3.4.3 Teste de Hipótese

Teste de hipótese é utilizado para modelos relacionados com um vetor  $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_n)'$  de parâmetro ou um subconjunto dele. E a comparação entre ajustes de modelos de sobrevivência aos dados, quando os mesmos são hierárquicos, é facilmente avaliada por meio de testes formais de ajuste, como o de Wald, o Razão de Verossimilhança e o Escore.

Segundo Colosimo e Giolo (2006), o teste Wald é baseado na distribuição assintótica de  $\hat{\boldsymbol{\theta}}$  e é uma generalização do teste t de Student (WALD, 1943). É conhecida por testar um único parâmetro  $\boldsymbol{\theta}_j$ , tendo aproximadamente uma distribuição qui-quadrado com  $p$  graus de liberdade ( $\chi_p^2$ ).

O teste da razão de máxima verossimilhança é a comparação entre o modelo paramétrico e o sub-modelo do modelo Weibull. Segundo Strapasson (2007), pelo teste da razão verossimilhança, Gomes (2005) faz uma discriminação entre o modelo proposto por Freitas *et al.* (2003) que é baseado no modelo Weibull e o modelo de risco proporcionais, proposto por ela.

As estatísticas da razão de máxima verossimilhança, TRV e do teste escore,  $S$  é obtida pelas hipóteses de interesse definidas por

$$\begin{cases} H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0, & \text{o modelo Weibull esta adequado aos dados} \\ H_1 : \boldsymbol{\theta} \neq \boldsymbol{\theta}_0, & \text{o modelo Weibull não esta adequado aos dados} \end{cases}$$

Para testar  $H_0$  pode-se utilizar a estatística de razão de verossimilhança, definida por Colosimo e Giolo (2006),

$$TRV = -2 \log \left[ \frac{L(\hat{\boldsymbol{\theta}}_0)}{L(\hat{\boldsymbol{\theta}})} \right] = 2[\log L(\hat{\boldsymbol{\theta}}) - \log L(\hat{\boldsymbol{\theta}}_0)]$$

em que, sob  $H_0 : \boldsymbol{\theta} = \boldsymbol{\theta}_0$ , dado  $\hat{\boldsymbol{\theta}}$  é o modelo geral e  $\hat{\boldsymbol{\theta}}_0$  modelo de interesse, seguem aproximadamente uma distribuição qui-quadrado com  $p$  graus de liberdade. Como critério



de decisão, a hipótese  $H_0$  é rejeitada, a um nível de  $100\alpha\%$  de significância se  $S > \chi_{p,1-\alpha}^2$ .

Na presença de covariáveis é interessante avaliar a significância das mesmas para o ajuste do modelo. A partir das estatísticas da razão de verossimilhança e escore, em que temos  $\{\boldsymbol{\theta} \mid TRV(\boldsymbol{\theta}) < \chi_{p,1-\alpha}^2\}$  é intervalo de  $(1 - \alpha) 100\%$  de confiança para  $\boldsymbol{\theta}$ .

E por meio da estatística escore em que pode-se construir intervalo de confiança, expressado por (COLOSIMO e GIOLO, 2006).

$$S = U'(\boldsymbol{\theta}_0)[\mathfrak{F}(\boldsymbol{\theta}_0)]^{-1}U(\boldsymbol{\theta}_0)$$

em que  $U(\boldsymbol{\theta}_0)$  é a função escore  $U(\boldsymbol{\theta}) = \frac{\text{partial}L(\boldsymbol{\theta})}{\text{partial}\boldsymbol{\theta}}$  avaliada em  $\boldsymbol{\theta}_0$ , e  $\mathfrak{F}(\boldsymbol{\theta}_0)$  a matriz de variância-covariância observada de  $\hat{\boldsymbol{\theta}}$  também avaliada por  $\boldsymbol{\theta}_0$ . Segundo Colosimo e Giolo (2006), as três estatísticas de teste podem ser adaptadas para o caso em que se tenha interesse somente em um subconjunto de  $\boldsymbol{\theta}$ .

Para Cordeiro (1992), as estatísticas da razão de verossimilhança (TRV) e do teste de escore ( $S$ ) são assintoticamente equivalentes, sob a hipótese nula  $H_0$ , à distribuição qui-quadrado. O problema de escolha entre elas surge quando a estimação segundo ambas as hipóteses apresentar o mesmo grau de dificuldade. As estatísticas (TRV) e ( $S$ ) são invariantes em relação à parametrização da distribuição dos dados.

### 3.4.4 Modelos de Regressão

De acordo com Carvalho *et al.* (2011) o efeito de covariáveis sobre o tempo de sobrevivência é estimado através de um modelo de regressão, no qual o tempo de sobrevivência é a variável resposta e  $\mathbf{x} = (x_1, \dots, x_p)$  é o vetor de covariáveis (variáveis independentes) possivelmente associadas ao tempo. Ainda segundo Carvalho *et al.* (2011) o modelo de regressão paramétrico será constituído de um componente aleatório, responsável por descrever o comportamento probabilístico do tempo de sobrevivência e um componente sistemático ou estrutural, que descreve a relação entre os parâmetros da distribuição de probabilidade e as covariáveis. Assim o modelo matemático que descreve este relacionamento é,

$$\lambda(t|\mathbf{x}) = \lambda_0(t)g(\mathbf{x}\beta),$$

sendo  $\beta$ 's os parâmetros;  $g(\cdot)$  uma função positiva e contínua das covariáveis;  $\lambda_0(t)$  o risco basal de um indivíduo que possui  $g(\mathbf{x}\beta) = 1$ .

Carvalho *et al.* (2011) afirmam também que a razão de riscos de dois indivíduos diferentes é função das covariáveis e não depende do tempo e que para ajustar um modelo de regressão paramétrico é necessário uma distribuição de probabilidade que é utilizada para descrever o tempo de sobrevivência  $T$ , que define a forma da função de sobrevivência, de risco e de risco acumulado.

#### 3.4.4.1 Modelo de Regressão Exponencial

No modelo de regressão exponencial, usado quando se assume que o risco é constante ao longo do tempo, o parâmetro  $\alpha$  depende das covariáveis  $x$  da seguinte forma:

$$\alpha(\mathbf{x}) = \exp(\mathbf{x}\beta)$$

sendo  $\beta$ 's as estimativas dos efeitos das covariáveis e  $\alpha$  o parâmetro que define o risco exponencial.

As funções de risco de sobrevivência para o modelo de regressão exponencial ficam assim definidas (CARVALHO *et al.*, 2011):

$$\lambda(t|\mathbf{x}) = \lambda(\mathbf{x}) = \exp(\mathbf{x}\beta)$$

e

$$S(t|\mathbf{x}) = \exp(-\alpha(\mathbf{x})t) = \exp(-\exp(\mathbf{x}\beta)t)$$

#### 3.4.4.2 Modelo de Regressão Weibull

Utilizando a distribuição Weibull no contexto da modelagem de sobrevivência significa que o tempo  $T$  segue uma distribuição Weibull e o parâmetro  $\alpha$  é modelado pelas covariáveis. Existem outras formas de incluir covariáveis na distribuição Weibull, mas segundo Carvalho *et al.* (2011) esta é mais utilizada. As funções são dadas abaixo:

$$\lambda(t|\mathbf{x}) = \gamma t^{\gamma-1} \alpha(\mathbf{x})^\gamma, \quad e \quad S(t|\mathbf{x}) = \exp(-(\exp(\mathbf{x}\beta)t)^\gamma).$$

## 4 Material e Métodos

Nesse trabalho foi feita uma revisão teórica sobre a análise de sobrevivência e suas principais ferramentas estatísticas, utilizando os métodos paramétricos, não-paramétricos e modelos de regressão paramétrica para estimar as funções de sobrevivência e risco. Para aplicação prática dessa teoria foi usado um banco de dados de pacientes com Mieloma Múltiplo (MM), onde toda metodologia desenvolvida das análises para obtenção dos resultados são provenientes do software R 3.2.2 (TEAM, 2015), por meio do pacote Survival versão 2.36 (THERNEAU; LUMLEY, 2014). Os dados da aplicação prática estão descritos detalhadamente em Allison (2010).

### 4.1 Material em Estudo

O banco de dados utilizado nesse estudo é proveniente do livro Survival Analysis Using SAS, constituído de um total de 25 pacientes portadores da doença Mieloma Múltiplo (MM). Através de uma variável chamada de Tratamento ( $x_1$ ), esses pacientes foram divididos aleatoriamente em dois grupos de tratamentos com drogas distintas, de tal forma que 12 pacientes experimentaram o tratamento tipo 1 e 13 pacientes experimentaram o tratamento tipo 2. A variável Duração ( $x_2$ ) dá o tempo, em dias, que se inicia o tratamento até que ocorra a morte (falha) ou censura do paciente. O acompanhamento máximo foi 2240 dias, em que o paciente foi censurado a partir deste tempo. Analogamente, o acompanhamento mínimo foi 8 dias, em que o paciente morreu. A variável STATUS( $x_3$ ) classifica o estado em que o paciente se encontra no estudo, onde ele pode se encontrar no estado de morte ou censura. Ainda foi adicionado no estudo uma covariável chamada de RENAL ( $x_4$ ), onde indica se o paciente possui ou não possui doença renal.

Classificação das variáveis:

$$x_1 = \begin{cases} 1, & \text{para o tratamento 1,} \\ 2, & \text{para o tratamento 2.} \end{cases}$$

$$x_3 = \begin{cases} 1, & \text{quando o paciente morreu,} \\ 0, & \text{quando o paciente censurou.} \end{cases}$$

Classificação da covariável:

$$x_4 = \begin{cases} 1, & \text{quando o paciente possui doença renal,} \\ 0, & \text{quando o paciente não possui doença renal.} \end{cases}$$

## 4.2 Métodos Estatísticos

Para a análise dos dados em estudo, aplicou-se os métodos não-paramétricos de Kaplan-Meier e o teste log-rank para se saber se existiam diferenças de tempo de vida por grupos. Aplicou-se também os métodos paramétricos, com o intuito de estimar a função de sobrevivência. Por fim, foi aplicado os modelos de regressão paramétricos para se obter o modelo que melhor se ajusta aos dados. O presente estudo buscou analisar a sobrevivência de 25 pacientes com a doença Mieloma Múltiplo, utilizando ferramentas estatísticas citadas acima e seguindo os seguintes passos:

1. Construção da tabela e gráfico de Kaplan-Meier, permitindo observar detalhadamente as probabilidades de sobrevivência dos pacientes e seus respectivos comportamentos de falhas e censuras através do gráfico.
2. Construção da tabela e gráfico de Kaplan-Meier, para os dois tipos de tratamentos, onde observamos o comportamento de cada tratamento isoladamente.
3. Construção da tabela e gráfico de Kaplan-Meier para a covariável  $x_4$  (RENAL), onde também observamos o comportamento isolado dos pacientes que possuem e os que não possuem doença renal.
4. Aplicação do teste Log-rank para variável  $x_1$  (Tratamento) e covariável  $x_4$  (RENAL), com o interesse de se verificar estatisticamente se as diferenças apontadas nas análises anteriores eram de fato verdadeiras.
5. Construção da tabela e gráficos de sobrevivência para as diferentes distribuições paramétricas, e comparação a curva de Kaplan-Meier.
6. Ajuste e validação do melhor modelo de regressão pelo teste de Razão de Verossimilhança (TRV).

## 5 Resultados e Discussão

A seguir demonstraremos os principais resultados obtidos a partir da análise realizada com os dados de pacientes com Mieloma Múltiplo. Primeiramente foram obtidas as estimativas de Sobrevivência utilizando-se o estimador de Kaplan-Meier aos quais são apresentadas na Tabela 1, em que pode-se observar os tempos de falha dos pacientes com suas respectivas probabilidades de sobrevivência.

Tabela 1: Estimativas de sobrevivência obtidas pelo método de Kaplan-Meier

Tempo	Nº de risco	Nº de falha	Sobrevivência	Erro Padrão	Limite Inf. I.C.	Limite Sup. I.C.
8	25	2	0,920	0,0543	0,820	1,000
13	23	1	0,880	0,0650	0,761	1,000
18	22	1	0,840	0,0733	0,708	0,997
23	21	1	0,800	0,0800	0,658	0,973
52	20	1	0,760	0,0854	0,610	0,947
63	19	2	0,680	0,0933	0,520	0,890
70	17	1	0,640	0,0960	0,477	0,859
76	16	1	0,600	0,0980	0,436	0,826
180	15	1	0,560	0,0993	0,396	0,793
195	14	1	0,520	0,0999	0,357	0,758
210	13	1	0,480	0,0999	0,319	0,722
220	12	1	0,440	0,0993	0,283	0,685
632	10	1	0,396	0,0986	0,243	0,645
700	9	1	0,352	0,0970	0,205	0,604
1296	7	1	0,302	0,0953	0,162	0,560

De acordo com a Tabela 1, 60% dos pacientes falharam até o dia 76 do estudo, no qual ocorreu a presença de duas falhas simultâneas no 8º e 63º dia. No dia 210, a probabilidade que o paciente sobrevivesse a doença até o final do estudo já era menor do que 50%.

Uma estimativa gráfica para uma função de sobrevivência é uma função escada, com valor constante para cada intervalo de tempo. Por definição, a função de sobrevivência estimada no primeiro intervalo  $[0, t_1)$  é igual a 1. Por outro lado, a função de probabilidade estimada para o último intervalo,  $[t_k, \infty)$ , é zero, se o maior tempo observado for uma falha, e não atingirá o zero se for uma censura. Na Figura 2, podemos observar o gráfico de Kaplan-Meier para o tempo de sobrevivência de pacientes com Mieloma Múltiplo.

De acordo com a Figura 2, podemos observar que o último intervalo da função de sobrevivência não atinge o valor zero, ou seja, o gráfico mostra que o último paciente do estudo censurou. Podemos confirmar essa conclusão observando na tabela do banco de

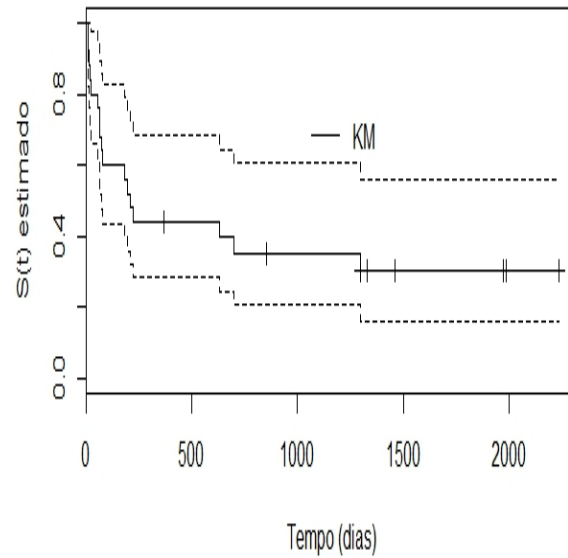


Figura 2: Gráfico de Kaplan-Meier para o tempo de sobrevivência de pacientes com Mieloma Múltiplo

dados a variável  $x_3$  (STATUS), onde no dia 2240 o último paciente foi censurado. Também podemos observar através do gráfico a quantidade de censuras ocorridas no estudo e seu respectivo intervalo de tempo. A forma de identificar a censura no gráfico é através do símbolo (+) na função de sobrevivência, onde podemos contar a presença de 8 censuras no estudo.

É de interesse do nosso estudo fazer as estimativas de sobrevivência para os diferentes tipos de tratamento, em que os pacientes que tomam a droga do tratamento 1 sejam analisados separadamente dos pacientes que tomam a droga do tratamento 2. Dessa forma, foi criada a Tabela 2. Nesta tabela, tem-se as estimativas de sobrevivência dos pacientes com Mieloma Múltiplo sob os dois tratamentos com os respectivos intervalos de confiança.

Tabela 2: Estimativas de Sobrevivência de pacientes com Mieloma Múltiplo sobre dois diferentes tratamentos

Tempo	Nº de risco	Nº de falha	Sobrevivência	Tratamento 1		
				Erro Padrão	Limite Inf. I.C.	Limite Sup. I.C.
8	12	2	0,833	0,108	0,647	1,000
52	10	1	0,750	0,125	0,541	1,000
63	9	2	0,583	0,142	0,362	0,941
220	7	1	0,500	0,144	0,284	0,880
				Tratamento 2		
13	13	1	0,923	0,0739	0,7890	1,000
18	12	1	0,846	0,1001	0,6711	1,000
23	11	1	0,769	0,1169	0,5711	1,000
70	10	1	0,692	0,1280	0,4819	0,995
76	9	1	0,615	0,1349	0,4004	0,946
180	8	1	0,538	0,1383	0,3255	0,891
195	7	1	0,462	0,1383	0,2566	0,830
210	6	1	0,385	0,1349	0,1934	0,765
632	5	1	0,308	0,1280	0,1361	0,695
700	4	1	0,231	0,1169	0,0855	0,623
1296	3	1	0,154	0,1001	0,0430	0,550

De acordo com a Tabela 2, podemos observar que todos os pacientes que participaram do tratamento 1 falharam ou censuraram até o dia 220, e que as falhas simultâneas que ocorreram no 8º e 63º dia estavam nesse grupo. Dessa forma, começamos a pensar que há indícios para se acreditar que o tratamento 2 seja mais eficiente.

Podemos observar melhor esse comportamento através da estimação gráfica da função de sobrevivência para os dois tratamentos, como mostra a Figura 3.

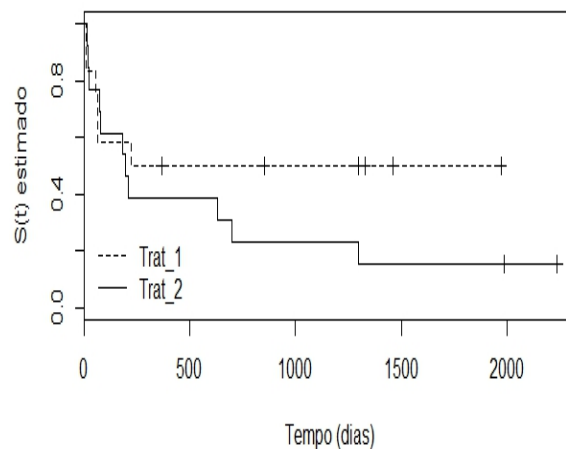


Figura 3: Gráfico de Kaplan-Meier para os dois tratamentos analisados

De acordo com a Figura 3, os tempos de sobrevivência dos pacientes do tratamento 2 foram maiores que dos pacientes do tratamento 1. Por outro lado, o gráfico nos mostra que seis das oito censuras que ocorrem no nosso banco de dados estão contidos no tratamento 1, de tal forma que nos leva a acreditar que a presença dessas censuras estão influenciando nas estimativas do tempo de vida dos pacientes que participam do tratamento 1.

De forma análoga a análise feita para os dois tipos de tratamento, também foi feita para a covariável  $X_4$  (RENAL), onde os pacientes foram divididos em dois novos grupos: os que possuem doença renal e os que não possuem doença renal. Como mostra a Tabela 3.

Tabela 3: Estimativas de Sobrevivência para pacientes com Mieloma Múltiplo que apresentam e não apresentam doença renal

Tempo	Nº de risco	Nº de falha	Não possui doença renal			
			Sobrevivência	Erro Padrão	Limite Inf. I.C.	Limite Sup. I.C.
8	18	1	0,944	0,0540	0,844	1,000
70	17	1	0,889	0,0741	0,755	1,000
76	16	1	0,833	0,0878	0,678	1,000
180	15	1	0,778	0,0980	0,608	0,996
195	14	1	0,722	0,1056	0,542	0,962
210	13	1	0,667	0,1111	0,481	0,924
220	12	1	0,611	0,1149	0,423	0,883
632	10	1	0,550	0,1186	0,360	0,839
700	9	1	0,489	0,1201	0,302	0,791
1296	7	1	0,419	0,1216	0,237	0,740
Possui doença renal						
8	7	1	0,857	0,132	0,6334	1.000
13	6	1	0,714	0,171	0,4471	1.000
18	5	1	0,571	0,187	0,3008	1.000
23	4	1	0,429	0,187	0,1822	1.000
52	3	1	0,286	0,171	0,0886	0,922
63	2	2	0,000	NaN	NA	NA

Pode-se observar na Tabela 3 que os pacientes que possuem doença renal, o máximo que sobreviveram foi até o 63º dia, sendo que neste dia ocorreram 2 falhas simultâneas. Percebe-se também que o tempo de vida máximo dos pacientes com doença renal é 20 vezes menor em relação ao último paciente que falhou no grupo que não possuía doença renal. No grupo com doença renal o tempo de sobrevivência estimado convergiu rapidamente a zero, enquanto que o grupo sem doença renal manteve-se até o final do estudo.



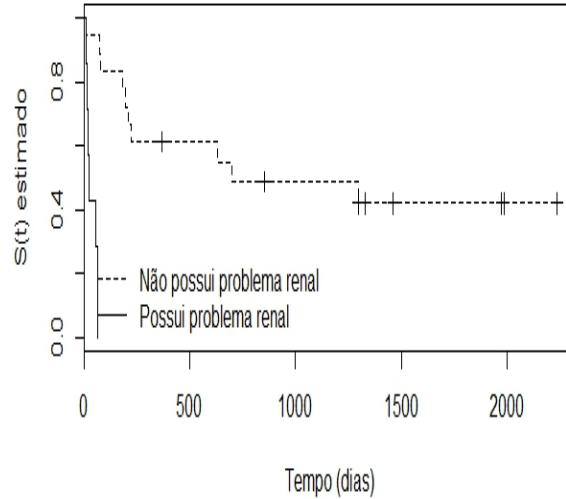


Figura 4: Gráfico de Kaplan-Meier para pacientes com doença renal e sem doença renal

No gráfico da Figura 4, os tempos de sobrevivência para os pacientes que possuem doença renal foram significativamente menores do que os pacientes que possuem função renal normal, resultado este obtido a partir da Tabela 3. Observamos também que todos os pacientes que possuem doença renal chegaram a óbito. Outra observação importante é que todos os dados censurados estão no grupo de pacientes com função renal normal, ou seja, dos oito pacientes censurados no estudo, observamos que sete censuram antes de 2000 dias e apenas um paciente conseguiu chegar ao fim do estudo.

Após aplicação do estimador de Kaplan-Meier aos dados de Mieloma Múltiplo (Tabelas 1, 2 e 3), procedeu-se a análise com o teste de Log-rank. O teste Log-rank foi aplicado com o interesse de se verificar estatisticamente se as diferenças apontadas nas análises das Tabelas 1, 2 e 3 eram de fato verdadeiras. Na Tabelas 4 e 5 tem-se os valores observados e esperados que compõem o teste de  $\chi^2$ . Na Tabela 4 tem-se o teste  $\chi^2$  para os pacientes que receberam o tratamento 1 e 2.

Tabela 4: Teste Log-rank para os dois tratamentos utilizados nos pacientes com Mieloma Múltiplo

Doença	N	Observado	Esperado	$(O - E)^2/E$	$(O - E)^2/V$
Trat. 1	12	6	8,34	0,655	1,31
Trat. 2	13	11	8,66	0,631	1,31

A estatística encontrada foi  $\chi^2 = 1,3$  sobre 1 grau de liberdade, obtendo um valor

$p = 0,252$ , indicando que os tratamentos não obtiveram efeitos distintos, ou seja, não houve diferença entre os tratamentos aplicados em relação ao tempo de sobrevivência.

Tabela 5: Teste Log-rank para os pacientes com Mieloma Múltiplo que possuem ou não possuem doença renal

Doença	N	Observado	Esperado	$(O - E)^2/E$	$(O - E)^2/V$
Não possuem	18	10	15,4	1,89	24
Possuem	7	7	1,6	18,24	24

A estatística encontrada foi  $\chi^2 = 24$  sobre 1 grau de liberdade, com um valor  $p = 9,54 \times 10^{-07}$ , indicando que houve diferença entre as curvas estudadas, ou seja, existe diferença no tempo de vida entre os pacientes que possuem doença renal e os pacientes sem doença renal. Desta forma o teste Log-rank comprova o que foi discutido no gráfico da figura 4, em que os pacientes com doença renal apresentam menor tempo de sobrevivência.

Dando sequência aos resultados aplicou-se diferentes distribuições paramétricas aos dados de Mieloma Múltiplo. O objetivo é encontrar qual distribuição se ajusta melhor aos dados e assim as estatísticas estimadas o serem com maior fidedignidade.

Na Tabela 6, estão descritas todas as estimativas de sobrevivência para as distribuições de Kaplan Meier, Exponencial, Weibull e Log-Normal.

Tabela 6: Sobrevivência estimada segundo as diferentes distribuições estudadas

Tempo	$\hat{S}(t)_{KM}$	$\hat{S}(t)_{exp}$	$\hat{S}(t)_W$	$\hat{S}(t)_{ln}$	
1	8,00	0,92	0,99	0,90	1,00
2	13,00	0,88	0,99	0,87	1,00
3	18,00	0,84	0,98	0,85	1,00
4	23,00	0,80	0,97	0,84	1,00
5	52,00	0,76	0,94	0,77	0,98
6	63,00	0,68	0,93	0,75	0,97
7	70,00	0,64	0,93	0,74	0,96
8	76,00	0,60	0,92	0,73	0,95
9	180,00	0,56	0,82	0,62	0,74
10	195,00	0,52	0,81	0,61	0,71
11	210,00	0,48	0,79	0,60	0,69
12	220,00	0,44	0,78	0,59	0,67
13	365,00	0,44	0,67	0,52	0,45
14	632,00	0,40	0,50	0,42	0,23
15	700,00	0,35	0,46	0,41	0,20
16	852,00	0,35	0,39	0,37	0,14
17	1296,00	0,30	0,24	0,30	0,06
18	1328,00	0,30	0,23	0,29	0,06
19	1460,00	0,30	0,20	0,28	0,05
20	1976,00	0,30	0,11	0,23	0,02
21	1990,00	0,30	0,11	0,23	0,02
22	2240,00	0,30	0,08	0,21	0,02

Na tabela 6, observamos que no dia 180 do estudo, a probabilidade de sobrevivência de um paciente estimado pela distribuição de Kaplan-Meier é menor que 60%, enquanto

que as demais distribuições paramétricas tiveram suas probabilidades de sobrevivência maiores que 60%, com destaque para a distribuição Exponencial que ainda possuía uma probabilidade de 82% de sobrevivência. A partir do dia 1296, observamos uma queda acentuada na sobrevivência dos pacientes estimados pela distribuição Log-Normal em relação as demais distribuições, onde a estimativa de vida Log-Normal cai para 6%, enquanto que as demais estimativas se mantêm ainda acima de 20%.

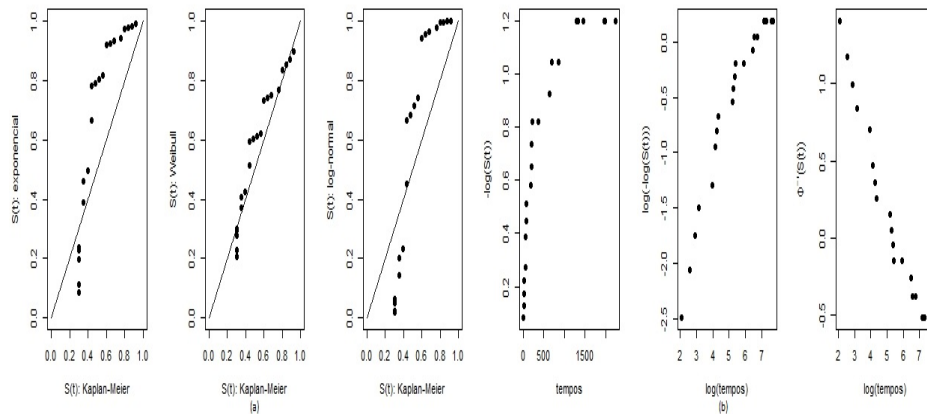


Figura 5: Curvas ajustadas por Kaplan - Meier versus as distribuições ajustadas (a) e linearização dos modelos exponencial, Weibull e Log-Normal(b)

De acordo com a figura 6 (a), é possível ver que o modelo Weibull é o que melhor se ajusta aos dados, pois a curva se apresenta mais próxima da reta  $y = x$ . Por outro lado, os modelos Exponencial e Log-Normal parecem ser menos adequados, já que a curva dos dados se distanciam mais da reta  $y = x$ . Pela figura 6 (b), o modelo Weibull pela linearização é o que mais se aproxima de uma reta e a exponencial é o pior modelo ajustado aos dados.

Após a aplicação de estatísticas de adequação de modelos, realizou-se um ajuste da curva de sobrevivência estimada por Kaplan-Meier em conjunto com as distribuições Weibull e Log-normal. Por meio da Figura 6, percebe-se que a distribuição Weibull teve melhor ajuste sobre a curva de Kaplan-Meier, sendo indicativo de melhor ajuste desta distribuição aos dados de sobrevivência.

Com o Teste da Razão de Verossimilhança (TRV) para  $\alpha = 1\%$  de significância, e para as seguintes hipóteses i) o modelo exponencial é adequado, ii) o modelo Weibull é adequado, iii) o modelo da Log-Normal é adequado, observamos na Tabela 7 que, o p-valor das funções Exponencial e Weibull são muito pequenos, havendo evidências suficientes para rejeitamos a hipóteses i) e ii). Para a distribuição Log-Normal o valor foi limitrofe.

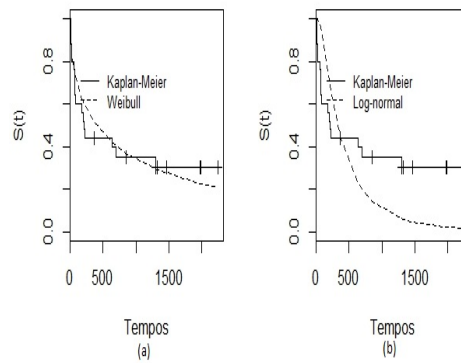


Figura 6:  $\hat{S}(t)$  de kaplan-Meier e das distribuições Weibull (a) e Log-Normal (b)

Ainda de acordo com a Tabela 7, pelo Critério de AIC a distribuição Exponencial foi a que apresentou pior ajuste (267,36) e a distribuição Log-Normal foi a que obteve melhor ajuste (247,47) um pouco superior ao valor da distribuição Weibull (251,29).

Tabela 7: Logaritmo da função  $L(\theta)$  e resultados dos TRV e AIC

MODELO	$\log L(\theta)$	TRV	valor P	AIC
Gama Generalizada	-118,4		—	242,95
Exponencial	-132,7	28,6	$6,16 \times 10^{-07}$	267,36
Weibull	-123,6	10,4	0,0012	251,29
Lognormal	-121,7	6,6	0,01	247,47

Na Figura 7 fica melhor evidenciado o que foi observado na tabela 7, pois a curva da distribuição exponencial (em laranja) foi a que pior se ajustou comparada as curvas da distribuição Gama Generalizada (em vermelho) e de Kaplan-Meier.

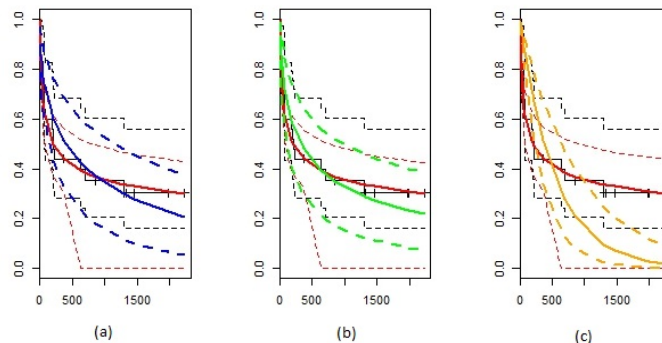


Figura 7: Ajuste da distribuição Gama-Generalizada, curva vermelha, com as curvas de Kaplan-Meier e as distribuições Log-Normal em azul (a), Weibull em verde (b) e Exponencial em laranja (c), com respectivos intervalos de confiança

A curva da distribuição Weibull e Log-Normal, respectivamente gráficos com curvas em verde e azul (Figura 8), foram as que melhor se ajustaram, tendo em vista o ajuste das curvas da distribuição Gama Generalizada (em vermelho) e de Kaplan-Meier. Nesta figura também é possível ver os intervalos de confiança paramétrico de acordo com a distribuição utilizada e do estimador de Kaplan-Meier.

Considerando o modelo de regressão Log-Normal e as covariáveis  $x_1$  (Tratamento) e  $x_4$  (RENAL), foram obtidos os resultados das estimativas dos parâmetros e os valores dos logaritmos das funções de verossimilhança apresentada na Tabela 8, para cinco modelos possíveis, um com a interação de  $x_1$  e  $x_4$ .

Tabela 8: Estimativas dos parâmetros e logaritmo da funções de verossimilhança dos modelos de regressão Log-Normal ajustados para os dados de Mieloma Múltiplo

Modelo	Covariáveis	Estimativas	Log Verossimilhança
M1	Nenhuma	$\hat{\beta}_0 = 5,78$	$l_1 = -121,7$
M2	$x_1$	$\hat{\beta}_0 = 6,85$ $\hat{\beta}_1 = -0,69$	$l_2 = -121,5$
M3	$x_4$	$\hat{\beta}_0 = 6,61$ $\hat{\beta}_1 = -3,34$	$l_3 = -115,1$
M4	$x_1$ e $x_4$	$\hat{\beta}_0 = 8,22$ $\hat{\beta}_1 = -0,99$ $\hat{\beta}_2 = -3,52$	$l_4 = -114,2$
M5	$x_1, x_4$ e $x_1 \times x_4$	$\hat{\beta}_0 = 8,45$ $\hat{\beta}_1 = -1,13$ $\hat{\beta}_2 = -4,16$ $\hat{\beta}_3 = 0,42$	$l_5 = 114,1$

Para se testar a significância da interação (Tabela 9), foi usado o teste da razão de verossimilhança TRV = 0,2 e valor  $p = 0,6547$  com 1 g.l. Deste resultado, pode-se concluir não haver evidências estatísticas que a interação  $x_1$  e  $x_4$  seja significativa. A covariável  $x_4$  na presença de  $x_1$  obteve um resultado de TRV = 14,6 e valor  $p = 0,0001$  indicando que esta covariável é estatisticamente significativa. A variável  $x_1$  com o valor de TRV = 1,8 e valor  $p$  sob 1 g.l. de 0,1797 foi não significativa, o que implica que esta covariável não apresenta significância estatística, devendo ser retirada do modelo de regressão para a explicação do tempo de vida dos pacientes com Mieloma Múltiplo.

Tabela 9: Resultados dos testes da Razão de Verossimilhança (TRV)

Efeito	Hipótese Nula	TRV	Valor p
Interação	$H_0 : \beta_3 = 0$	$2(114,2-114,1)=0,2$	0,6547
de $x_4 x_1$	$H_0 : \beta_2 = 0$	$2(121,5-114,2)=14,6$	0,0001
de $x_1$	$H_0 : \beta_1 = 0$	$2(115,1-114,2)=1,8$	0,1797

No modelo ajustado final apresentado na Tabela 10, a variável  $x_4$  apresenta sinal ne-

gativo ( $\hat{\beta}_1 = -3,344$ ), o que implica que pacientes com doença renal ( $x_4 = 1$ ), apresentam probabilidade de sobrevivência estimada 3,334 vezes menor do que a dos pacientes sem doença renal ( $x_4 = 0$ ). Este fato foi observado claramente na Figura 4, por meio das curvas de Kaplan-Meier.

Tabela 10: Resultados dos testes da Razão de Verossimilhança (TRV)

Modelo	Estimativas	Erro Padrão	z	p
(Intercepto)	6,617	0,462	14,33	$1,35 \times 10^{-46}$
$x_4$ (RENAL)	-3,344	0,806	-4,15	$3,32 \times 10^{-05}$
Log(escala)	0,558	0,184	3,04	$2,40 \times 10^{-03}$

Após todas as análises realizadas, a Figura 8 apresenta as curvas de sobrevivência estimada pela distribuição Log-Normal no modelo de regressão paramétrico de sobrevivência, considerando como covariável a presença ou ausência de doença renal (RENAL ( $x_4 = 1$ ) ou RENAL ( $x_4 = 0$ )), é evidente que o ajuste foi satisfatório, pois a distribuição Log-Normal no modelo de regressão com a covariável RENAL descreve bem o comportamento do tempo de sobrevida de pacientes em estudo.

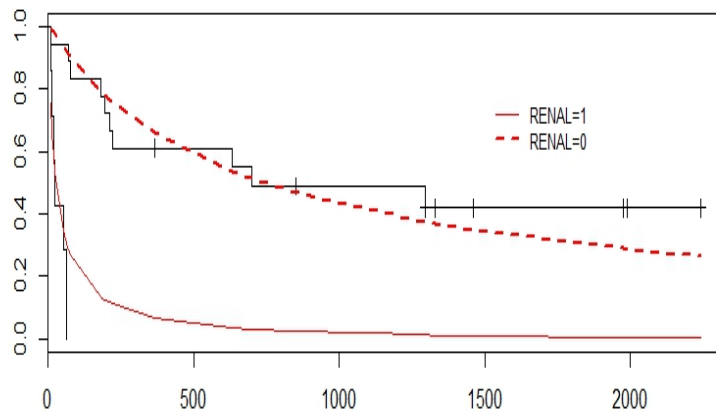


Figura 8: Curvas de sobrevivência estimadas pelo método de regressão Log-Normal para os dois grupos de pacientes com e sem doença renal

Durante todo este estudo foi notável a importância do estimador de Kaplan-Meier no desenvolvimento desse trabalho, pois observamos que ele está presente em quase todos os passos do nosso estudo, seja analisado diretamente, passos (1, 2 e 3), ou servindo como base de comparações entre outros modelos, passos (5 e 6).

No passo (1), analisamos o estudo como um todo, onde retiramos conclusões das estimativas de vida e tempo de falha de cada paciente.

Um exemplo disso, foi quando concluímos que a probabilidade de um paciente a partir do dia 210 do estudo era menor que 50%. Outra análise importante foi a estimativa gráfica da função de sobrevivência em forma de escada, em que cada degrau informava uma falha e cada sinal de (+) informava uma censura.

Nos passos (2) e (3), analisamos o estudo em função das variáveis. Em (2), o gráfico de Kaplan-Meier nos mostrou uma grande presença de censura no Tratamento 1. No passo (3), o gráfico além de nos mostrar que todas as censuras estão presentes no grupo de pacientes sem doença renal, nos mostra a rápida convergência da curva de Kaplan-Meier dos pacientes com doença renal para zero, indicando que o fato do paciente possuir doença renal influencia negativamente no seu tempo de vida.

No passo (4), temos estatisticamente a confirmação das análises feitas nos passos (2) e (3). Pelo teste Log-rank, concluímos que em (2), o teste foi não significativo, ou seja, não houve diferença entre os tratamentos aplicados em relação ao tempo de sobrevivência dos pacientes. Já em (3), o teste foi significativo, confirmando o que foi dito acima, que os pacientes que possuem doença renal apresentam menor tempo de vida.

No passo (5), ajustamos as curvas das distribuições Exponencial, Weibull e Log-Normal em relação a curva de Kaplan-Meier, onde observamos visualmente que as distribuições Weibull e Log-Normal melhor se adequaram. Pelo critério de AIC, o modelo que melhor se ajustou a os dados foi o Log-Normal.

No passo (6), foi escolhido o modelo de regressão que melhor se ajustou aos dados. Pelo teste da Razão de Verossimilhança, selecionamos o melhor modelo

## 6 Considerações Finais

Observamos que o estimador e o gráfico de Kaplan-Meier são ótimas ferramentas estatísticas para se investigar as estimativas de tempo de sobrevivência com presenças de falhas e censuras em banco de dados epidemiológicos, como por exemplo, o nosso estudo sobre a doença Mieloma Múltiplo.

A distribuição paramétrica que melhor se ajustou aos dados de pacientes com Mieloma Múltiplo foi a Log-Normal, na qual foi selecionada pelas comparações dos ajustes gráficos, TRV e pelo Critério de Akaike.

O modelo de regressão que melhor se ajustou aos dados foi o Log-Normal com a presença da covariável  $x_4$  (RENAL), como foi observado no gráfico da figura 9, onde as curvas do modelo de regressão Log-Normal se ajustaram bem as curvas do modelo de Kaplan-Meier para covariável  $x_4$  (RENAL). Com isso, pôde-se concluir que os pacientes com doença renal apresentam uma estimativa de vida 3,334 vezes menor do que os pacientes sem doença renal.

Portanto, as técnicas de análise de sobrevivência permitem aos especialistas da área de saúde entender os fatores que afetam o tempo de sobrevivência dos pacientes, ajudando no desenvolvimento de melhores estratégias para se conduzir os tratamentos.

### 6.1 Pesquisas Futuras

Considerando que os métodos e modelos descritos nesse trabalho foram feitos como uma revisão teórica para o estudo das técnicas de análise de Sobrevivência, tais métodos e modelos podem ser facilmente aplicados a um banco de dados real. Um bom exemplo de aplicação real aqui em nossa cidade, seria aplicar tais técnicas a dados de pacientes com câncer de próstata do Hospital da FAP. Em vista que a ideia inicial para esse trabalho seria a aplicação desse exemplo, mas por motivos maiores não conseguimos coletar os dados, seria de grande valia uma parceria entre a FAP e o DE-UEPB.



# Referências

- ALLISON, P. D. *Survival Analysis Using SAS: A Pratical Guide, Second Edition*. 2. ed. [S.l.]: SAS Institute Inc., Cary, NC, USA, 2010.
- CARVALHO, M. S.; ANDREOZZI, V. L.; CODEÇO, C. T.; CAMPOS, D. P.; BARBOSA, M. T. S.; SHIMAKURA., S. E. *Análise de Sobrevivência: teoria e aplicações em saúde*. [S.l.]: SciELO-Editora FIOCRUZ, 2011.
- COLOSIMO, E. A.; GIOLO, S. R. *Análise de Sobrevivência Aplicada*. 1. ed. [S.l.]: Edgard Blucher Ltda, 2006.
- CORDEIRO, G. M. Introdução à teoria de verossimilhança. *Simpósio Nacional de Probabilidade e Estatística*, v. 10, p. 174, 1992.
- CORDEIRO, G. M.; ORTEGA, E. M. M.; SILVA, G. O. The exponentiated generalized gamma distribution with application to lifetime data. *Journal of Statistical Computation and Simulation*, v. 81, 2011.
- COX, C. The generalized f distribution: an umbrella for parametric survival analysis. *Statistics in medicine*, v. 27, 2008.
- FREITAS, M. A.; BORGES, W.; HO, L. L. A statistical model for shelf life estimation using sensory evaluations scores. *Communications in Statistics - Theory and Methods*, v. 32, p. 1559–1589, 2003.
- GOMES, R. C. D. *Estimando o tempo de vida de produto em prateleira utilizando modelo de risco proporcionais em dados oriundos de avaliações Sensoriais*. Dissertação (Tese de Mestrado) — Universidade Federal de Minas Gerais, Belo Horizonte, 2005.
- GUSMÃO, F. R. S.; ORTEGA, E. M. M.; CORDEIRO, G. M. The generalized inverse weibull distribution. *Stat Papers*, v. 52, p. 591 – 619, 2011.
- LAWLESS, J. F. Inference in the generalized gamma and log gamma distributions. *Technometrics*, v. 22, p. 409 – 419, 1980.
- LEE, E. T.; WANG, J. *Statistical methods for survival data analysis*. [S.l.]: John Wiley & Sons, 2003.
- LOUZADA, F.; DINIZ, C. *Modelagem Estatística para risco de crédito*. 2. ed. [S.l.]: ABE, 2012.
- NADARAJAH, S.; GUPTA, A. K. The exponentiated gamma distribution with application to drought data. *Bulletin of the Calcutta Statistical Association*, v. 59, p. 233–234, 2007.

- STACY, E. W. A generalization of the gamma distribution. *Institute of Mathematical Statistics*, v. 33, p. 1187–1192, 1962.
- STRAPASSON, E. *Comparação de modelos com censura intervalos em análise de sobrevivência*. Tese (Tese de Doutorado) — ESALQ, Piracicaba, 2007.
- TEAM, R. C. *R: a language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing; 2012. 2015.
- THERNEAU, T.; LUMLEY, T. Survival. r package version 2.36-12. 2014.
- WALD, A. Tests of statistical hypotheses concerning several parameters when the number of observations is large. *American Mathematical Society*, v. 54, p. 426–482, 1943.

# Apêndice

```

setwd(dir)
dados=read.table("dados.txt",h=T)
head(dados)
require(survival)
ekm<- survfit(Surv(DUR,STATUS)~1,data=dados)
summary(ekm)
plot(ekm, xlab="Tempo (dias)",ylab="S(t) estimado")
legend(1000,0.8,c("KM"),lwd=1, bty="n")
ekm1<- survfit(Surv(DUR,STATUS)~TREAT,data=dados,conf.type="plain")
summary(ekm1)
ekm2<- survfit(Surv(DUR,STATUS)~TREAT,data=dados,conf.type="log-log")
summary(ekm2) \#Na rotina do Anexo retirar
ekm3<- survfit(Surv(DUR,STATUS)~TREAT,data=dados,conf.type="log")
summary(ekm3) \#Na rotina do Anexo retirar
plot(ekm1, lty=c(2,1), xlab="Tempo (dias)",ylab="S(t) estimado")
legend(1,0.3,lty=$c(2,1),c("Trat_1","Trat_2"),lwd=1, bty="n")
ekm4<- survfit(Surv(DUR,STATUS)~RENAL, data=dados)
summary(ekm4)
plot(ekm4, lty=c(2,1), xlab="Tempo (dias)",ylab="S(t) estimado")
legend(1,0.3,lty=c(2,1),c("Não possui problema renal",
"Possui problema enal"),lwd=1, bty="n")
\##### Falta Teste Log-rank \#####
survdif(Surv(DUR,STATUS)~RENAL, data=dados)
survdif(Surv(DUR,STATUS)~TREAT, data=dados)
\#####Modelos Paramétricos \#####
require(survival)
ajust1<-survreg(Surv(DUR,STATUS)~1,dist='exponential',data=dados)
ajust1
alpha<-exp(ajust1$coefficients[1])

```

```

alpha
ajust2<-survreg(Surv(DUR,STATUS)~1,dist='weibull',data=dados)
ajust2
alpha2<-exp(ajust2$coefficients[1])
gama<-1/ajust2$scale
cbind(gama, alpha2)
ajust3<-survreg(Surv(DUR,STATUS)~1,dist='lognorm',data=dados)
ajust3
alpha3=ajust3$coefficients
theta=ajust3$icoef[2]
ekm<-survfit(Surv(DUR,STATUS)~1,data=dados)
time<-ekm$time
st<-ekm$surv
ste<- exp(-time/alpha)
stw<- exp(-(time/alpha2)^gama)
stln<- pnorm((-log(time)+ alpha3)/theta)
require(xtable)
xtable(cbind(time,st,ste,stw,stln))
require(flexsurv)
## generalized gamma fit
fitg <- flexsurvreg(formula = Surv(DUR, STATUS)~1, data = dados,
ist="gengamma");fitg
par(mfrow=c(1,3))
plot(st,ste,pch=16,ylim=range(c(0.0,1)), xlim=range(c(0,1)),
xlab = "S(t): Kaplan-Meier",
ylab="S(t): exponencial")
lines(c(0,1), c(0,1), type="l", lty=1)
plot(st,stw,pch=16,ylim=range(c(0.0,1)), xlim=range(c(0,1)),
xlab = "S(t): Kaplan-Meier",
ylab="S(t): Weibull")
lines(c(0,1), c(0,1), type="l", lty=1)
plot(st,stln,pch=16,ylim=range(c(0.0,1)), xlim=range(c(0,1)),
xlab = "S(t): Kaplan-Meier",
ylab="S(t): log-normal")
lines(c(0,1), c(0,1), type="l", lty=1)

```

```

par(mfrow=c(1,3))
invst<-qnorm(st)
plot(time, -log(st),pch=16,xlab="tempos",ylab="-log(S(t))")
plot(log(time),log(-log(st)),pch=16,xlab="log(tempos)",
ylab="log(-log(S(t)))")
plot(log(time),invst,pch=16,xlab="log(tempos)",
ylab=expression(Phi^-1* (S(t))))
ajust1$loglik[2]
ajust2$loglik[2]
ajust3$loglik[2]
par(mfrow=c(1,2))
plot(ekm, conf.int=F, xlab="Tempos", ylab="S(t)")
lines(c(0,time),c(1,stw), lty=2)
legend(25,0.8,lty=c(1,2),c("Kaplan-Meier", "Weibull"),
bty="n",cex=0.8)
plot(ekm, conf.int=F, xlab="Tempos", ylab="S(t)")
lines(c(0,time),c(1,stln), lty=2)
legend(25,0.8,lty=c(1,2),c("Kaplan-Meier", "Log-normal"),
bty="n",cex=0.8)

#####Regressão Paramétrica#####

dados<-as.data.frame(dados)
ajust1<-survreg(Surv(dados$DUR, dados$STATUS)~dados$TREAT +
dados$RENAL, dist='exponential')
ajust1
summary(ajust1)
ajust1$loglik
ajust2<-survreg(Surv(dados$DUR, dados$STATUS)~dados$TREAT,
dist='weibull')
ajust2
ajust2$loglik
gama<-1/ajust2$scale
gama

```

```
#####Seleção do Modelo #####

require(flexsurv)

## Compare generalized gamma fit with Weibull
fitg <- flexsurvreg(formula = Surv(DUR, STATUS)~RENAL + TREAT,
  data = dados, dist="gengamma")
fitg
fitw <- flexsurvreg(formula = Surv(DUR, STATUS)~RENAL + TREAT,
  data = dados, dist="weibull")
fitw
plot(fitg)
lines(fitw, col="blue", lwd.ci=1, lty.ci=1)
fitl <- flexsurvreg(formula = Surv(DUR, STATUS)~RENAL + TREAT,
  data = dados, dist="lognormal")
fitl
plot(fitg)
lines(fitl, col="blue", lwd.ci=1, lty.ci=1)

fite <- flexsurvreg(formula = Surv(DUR, STATUS)~RENAL + TREAT,
  data = dados, dist="exponential")
fite
plot(fitg)
lines(fite, col="blue", lwd.ci=1, lty.ci=1)

par(mfrow=c(1,3))
plot(fitg)
lines(fitw, col="blue", lwd.ci=2, lty.ci=2)
plot(fitg)
lines(fitl, col="green", lwd.ci=2, lty.ci=2)
plot(fitg)
lines(fite, col="orange", lwd.ci=2, lty.ci=2)
```