ANÁLISE DO USO DAS ÁGUAS DE UMA EMPRESA DO RAMO PETROQUÍMICO VIA ANÁLISE DE SOBREVIVÊNCIA

Edimar Izidoro Novaes

Dissertação apresentada à Universidade Estadual Paulista "Júlio de Mesquita Filho" para a obtenção do título de Mestre em Biometria.

BOTUCATU
São Paulo Brasil
Novembro – 2014

ANÁLISE DO USO DAS ÁGUAS DE UMA EMPRESA DO RAMO PETROQUÍMICO VIA ANÁLISE DE SOBREVIVÊNCIA

Edimar Izidoro Novaes

Orientadora: Profa. Dra. **Miriam Harumi Tsunemi**

Dissertação apresentada à Universidade Estadual Paulista "Júlio de Mesquita Filho" para a obtenção do título de Mestre em Biometria.

BOTUCATU
São Paulo Brasil
Novembro – 2014

Ficha Catalográfica

Dedicatória

Dedico esse trabalho aos professores, amigos e familiares que de uma maneira ou outra contribuíram para que eu pudesse conquistar esse título, adquirir e aperfeiçoar conhecimentos estatísticos e matemáticos, assim como amadurecimento para a vida em sociedade.

Agradecimentos

Agradeço primeiramente a Deus pela oportunidade cedida para essa conquista.

Agradeço aos meus pais, Aparecido Fernandes Novaes e Ivone Maria Izidoro Novaes, pelos conselhos, apoio e incentivo.

A professora orientadora Dra. Miriam Harumi Tsunemi pela compreensão, paciência e dedicação na orientação deste trabalho.

Aos professores da banca julgadora deste trabalho, pela disponibilidade de lerem e assim contribuírem para o fechamento do mesmo.

Aos professores e funcionários do departamento de Bioestatística da UNESP de Botucatu.

Aos professores da Faculdade Estadual de Ciências Econômica de Apucarana e da Faculdade de Apucarana, pessoas quem tive o prazer de ter como professores e também o prazer de ter tido como colegas de serviço, a quem sempre me incentivaram, aconselharam, apoiaram e que tenho como espelho de profissionais e seres humanos.

Aos amigos da república Google que passaram a fazer parte da minha família e onde tivemos ótimos momentos de festas, conversas, estudos, conselhos e acima de tudo de perspectivas para um futuro melhor, tanto como ser humano, como profissionais.

Aos amigos da turma do mestrado pela compreensão, ajuda, apoio e companheirismo.

Enfim, agradeço a todos que de uma maneira ou de outra fizeram parte do caminho percorrido por mim para chegar até aqui nesta conquista.

Sumário

Lista de Figuras

Página

Lista de Tabelas

Página

ANÁLISE DO USO DAS ÁGUAS DE UMA EMPRESA DO RAMO PETROQUÍMICO VIA ANÁLISE DE SOBREVIVÊNCIA

Autor: EDIMAR IZIDORO NOVAES
Orientadora: Profa. Dra. MIRIAM HARUMI TSUNEMI

RESUMO

Na literatura é discutido o uso de planárias como bioindicadores de qualidade das águas, pois diferentes dos ensaios de toxicologia clássica, os animais conseguem captar efeitos de exposição a longo prazo. Atualmente, uma das técnicas de se verificar a qualidade das águas com o uso de planárias é realizando o teste cometa. Propomos uma forma alternativa para verificar a qualidade das águas utilizando planárias, a qual consiste em acompanhar o tempo de vida das mesmas inseridas em diferentes ambientes aquáticos. No estudo presente foram utilizados os dados obtidos por uma empresa que atua no ramo petroquímico e que utiliza as águas do rio Jaguari em seu processo de produção, nos quais foram aplicadas as técnicas estatísticas de análise de sobrevivência para comparar os níveis de poluentes tóxicos nos diferentes ambientes aquáticos. Também propomos um modelo de regressão para realizar inferências nos tempos de vida das planárias para os diferentes locais

dos quais foram coletadas as amostras de água. O novo método proposto apresentou uma forma eficiente e de baixo custo no uso das planárias da espécie Dugesia Tigrina como bioindicadores de qualidade de águas.

Palavras-Chaves: Planária Dugesia Tigrina; análise de sobrevivência; modelos paramétricos.

ANALYSIS OF INFLUENCE OF PETROCHEMICAL INDUSTRY COMPANIES IN THE USE OF JAGUARI RIVER'S WATERS USING SURVIVAL ANALYSIS

Author: EDIMAR IZIDORO NOVAES
Adviser: Profa. Dra. MIRIAM HARUMI TSUNEMI

SUMMARY

In literature the use of planarians as bioindicators of water quality is discussed, because differently from classic toxicology tests, the animals are able to capture long term exposition effects. Currently, one of techniques to verify the water quality with the use of planarians is performing the comet assay. We propose a new alternative to verify the water quality by accompany lifetime planarians on different aquatic environments. In the present study we used data obtained from a petrochemical company that uses Jaguari river on its production process. Survival Analysis was used to compare the levels of toxic pollutants on different aquatic environments. We also propose a regression model to make inference in the lifetimes of planarians for the different locations from which water samples were collected. The new proposed method presented an efficient and low cost way to use planarians of Dugesia Tigrina specie as water quality bioindicators.

Keywords: Dugesia Tigrina, Survival analysis, bioindicators.

1 INTRODUÇÃO

Presente em praticamente todas as áreas do conhecimento, a estatística desenvolve um papel de destaque no que diz respeito à análise de dados. Dentre essas áreas, uma das que mais cresceu nos últimos anos foi a de análise de sobrevivência, fato este evidenciado por sua quantidade de aplicações nas mais diversas áreas de pesquisa, como medicina, agronomia, engenharia, biologia e demais áreas relacionadas á saúde.

Um exemplo do uso da análise de sobrevivência na área da saúde é o trabalho realizado por Mota et al. (2012), em que foi avaliado o efeito de covariáveis medidas no tempo até a ocorrência de acidentes vasculares cerebrais recorrentes em pacientes sob diálise. Outro exemplo é o trabalho de Junior et al. (2011) em que foi usado a análise de sobrevivência para descrever e compreender o fluxo escolar de alunos do curso de graduação em Física da Universidade Federal do Rio Grande do Sul, assim como o trabalho de Suzi et al. (2010) para estimar os tempos médios de permanência de usuários em produtos sem que seja necessário observá-los até que finalizem as tarefas.

Um dos principais fatores que contribuiu para o crescimento da análise de sobrevivência nos últimos anos, foi o aprimoramento e desenvolvimento de técnicas e métodos estatísticos, cada vez mais eficientes, aliado ao grande avanço tecnológico e computacional das últimas décadas.

Neste trabalho utiliza-se também a modelagem de análise de sobrevivência na área de bioindicadores de qualidade da água. Na literatura é discutido o uso de planárias como bioindicadores de qualidade das águas, pois diferente dos ensaios de toxicologia clássica, esses animais conseguem captar efeitos de exposição a

longo prazo. “Outros motivos para se tornarem indicadores de amostras ambientais é por serem de fácil cultivo, terem custo baixo e, por serem organismos considerados simples do ponto de vista filogenético, esses animais podem ser utilizados como indicativo de risco de exposição a poluentes para organismos mais complexos”(Lau, 2002).

Diversos estudos sobre a sensibilidade das planárias vem sendo apresentados na literatura. Sáfadi (1993) relata a sensibilidade das planárias Dugesia Tigrina ao entrarem em contato com produtos tóxicos e compostos metálicos (mercúrio, cobre, cádmio, cromo e zinco). Lau (1998) testa a sensibilidade genotóxica de planárias ao metil metanossulfonato (MMS) e Alvarado & Newmark (1998) relatam sobre a capacidade de regeneração das planárias de água doce quando entram em contado com poluentes.

Estudos mais recentes, como o de Ribeiro (2012), relatam que as planárias de água doce possuem vasta distribuição geográfica, plasticidade biológica, facilidade de cultivo em laboratório, sensibilidade a contaminantes diversos e capacidade regenerativa, por isso vem sendo utilizadas como organismos-teste em ensaios ecotoxicologicos.

Em especial destaca-se a planária Dugesia Tigrina, a qual é relativamente resistente à poluição, sendo encontrada em águas paradas, pantanosas e às vezes poluídas, porém possuem alta sensibilidade genotóxica (Lau, 2002). Já Lau (1998) apresenta resultados de experimentos que garantem que esses organismos são adequados para avaliações de genotoxicidade ambiental, especialmente para amostras de águas ou misturas complexas.

Atualmente, uma das técnicas de se verificar a qualidade das águas utilizando planárias é através da realização do teste cometa. Esse teste detecta danos primários causados no DNA, ou seja, ele relaciona a quantidade de mutações genéticas à qualidade da água. Outros mecanismos para verificar mutação gênica são: teste de Ames, ensaio de Salmonella/microssoma, teste de avaliação gênica em pelos estaminais de *Tradescantia* (Trad-SHM), entre outros. Lau (2002) apresenta

avaliação da genotoxidade das águas da bacia do rio Guaíba-RS utilizando planárias, no qual foi possível detectar mutações genéticas, por meio do teste cometa, em planárias que ficaram expostas a um novo ambiente aquático por 12 dias.

Existe também outra possibilidade para verificar a qualidade das águas utilizando planárias, e tal possibilidade é o que esse trabalho aborda. É comum empresas do ramo petroquímico usarem águas de rios em suas atividades de produção, sendo devolvidas aos rios após um tratamento específico. Nos processos industriais, as águas são destinadas, na maior parte dos casos, para a limpeza de materiais, refrigeração de sistemas e geração de vapor. "Estima-se que a cada ano acumulam-se nas águas de rios cerca de 300 a 500 mil toneladas de dejetos provenientes dos efluentes industriais, que muitas vezes podem transportar resíduos tóxicos"(Lora, 2002).

Desta forma, as empresas tratam as águas utilizadas e avaliam posteriormente se o tratamento realizado está sendo eficiente. Como as planárias Dugesia Tigrina são sensíveis aos poluentes tóxicos e possuem um longo ciclo de vida, comparar os tempos de vida das mesmas quando inseridas a um novo ambiente aquático, possivelmente contaminado com produtos tóxicos, pode ser uma alternativa para avaliar a qualidade da água que retorna aos rios. Diferente das técnicas usuais, a nova metodologia proposta neste estudo para avaliação toxicológica das águas possui um custo relativamente mais baixo por dispensar as técnicas laboratoriais.

No estudo presente foram utilizados os dados obtidos por uma empresa que atua no ramo petroquímico e que usa as águas do rio Jaguari em seu processo de produção. "O rio Jaguari possui suas nascentes no estado de Minas Gerais, nos municípios de Sapucaí-Mirim, Camanducaia e Itapeva e ao entrar em território paulista, o rio é represado, sendo este um dos reservatórios integrantes do sistema produtor de água chamado Cantareira, construído para permitir a reversão de água da bacia do Piracicaba para a bacia do Alto Tietê, como reforço ao abastecimento público da Região Metropolitana de São Paulo"(Sabesp, 1989).

Resumidamente, a empresa faz a captação da água no rio Jaguari, após uma análise do produto recolhido utiliza em suas atividades e por fim faz a devolução

da água ao rio. Desta forma, os pontos dos quais foram obtidas as amostras de água deste estudo são: ponto 1 que refere-se, ao Rio Jaguari onde ocorre a captação da água para uso na empresa; ponto 3 é a saída da água tratada da empresa; ponto 4 referente ao montante do rio Atibaia, ou seja, lugar onde nasce o rio Atibaia e por fim ponto 8, referente ao jusante do rio Atibaia, ou seja, lado para onde se dirige a corrente de água do rio. A empresa não usa a água do rio Atibaia, porém o mesmo serve neste trabalho,como parâmetro para analisar a qualidade da água que está sendo devolvida ao rio Jaguari. A seguir na Figura (1), está uma ilustração da localização da empresa Replan e os rios Jaguari e Atibaia.

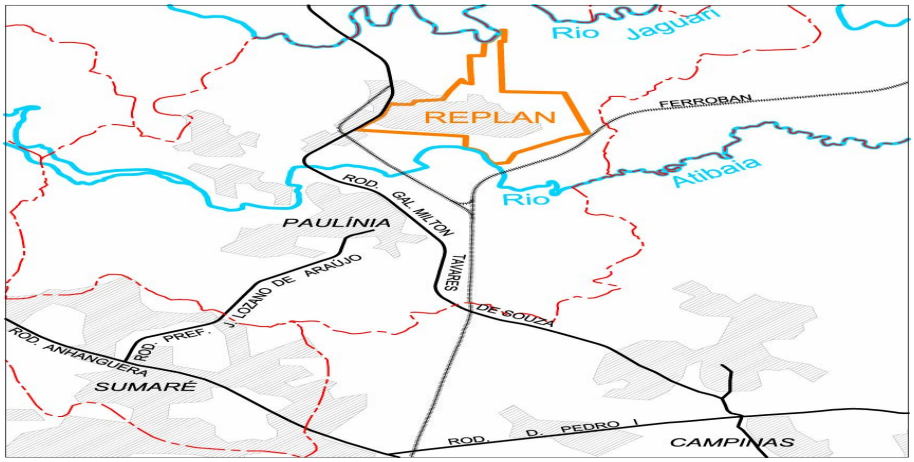

Figura 1: Mapa com sistema viário. Fonte: Relatório de Impacto ao Meio Ambiente. Retirado em http://www.comitepcj.sp.gov.br/download/Replan-RIMA, acesso em 10/08/2014.

Sendo assim, o objetivo principal deste estudo é propor uma forma alternativa de verificar a qualidade das águas, em termos de toxicidade, utilizando métodos estatísticos de análise de sobrevivência. Essa técnica consiste em comparar os tempos de vida das planárias Dugesia Tigrina em ambientes aquáticos, sendo um deles considerado como controle e os outros após sofrer alguma pertubação. Para exemplificar a nova técnica, foi realizado um estudo comparando os tempos de vida das planárias de diferentes pontos de coleta de água. Também, propomos um modelo

de regressão com o intuito de realizar inferências nos tempos de vida dos animais para cada local onde foi coletada a amostra de água.

2 REVISÃO DE LITERATURA

"Assim como para as pessoas, a água é de fundamental importância para a indústria, devido à sua grande utilidade em vários setores como na higiene de pessoas, na limpeza dos ambientes, dos equipamentos e dos instrumentos. Por esses motivos a mesma requer especial atenção nas fontes de abastecimento, no seu tratamento, desinfecção, depósito e distribuição"(SEAP/PR, 2007).

Dependendo da finalidade de utilização, a água deve ter certas características como potabilidade, dureza, teor de metais tóxicos e contagem de patógenos dentro dos padrões estabelecidos, além de ausência de odor e sabor indesejáveis. "Em função da fonte fornecedora, ou seja, água de subsolo, rios, lagos, reservatórios, água já tratada do município, e do uso final da água (limpeza, processamento) é recomendável que a indústria, sempre que possível, tenha o seu próprio sistema de tratamento de água"(Gava, 1984).

No Brasil, as empresas responsáveis pela captação, tratamento e distribuição da água utilizam diversos métodos de purificação de água, que a torna adequada ao consumo humano. Entretanto a qualidade da água tratada por estas empresas, pode influenciar de modo negativo em produtos ou máquinas da indústria devido a componentes químicos para seu tratamento.

"Os tipos de tratamentos da água são diferenciados e consequentemente, os tipos de análises requeridos são determinados de acordo com a constituição vigente. Os processos empregados variam conforme as características da água bruta e da qualidade que se deseja ter, podendo incluir a clarificação, a desinfecção ou a eliminação de impurezas específicas"(Barros et al., 1995).

Desta forma, as indústrias que utilizam a água dos rios devem realizar

tratamentos adequados para que a água devolvida não agrida ao meio ambiente. Com esse objetivo, a seguir será detalhada uma técnica de monitoramento da qualidade da água.

2.1 Controle do Tratamento da Água

Em virtude da importância do controle da água a ser utilizada na indústria, tanto na questão da qualidade da água que adentra a indústria, assim como naquela devolvida para o meio ambiente, empresas que na sua matéria prima possuem a presença de produtos tóxicos, podem usar planárias no processo de verificação da qualidade da água. Sendo assim, é comum empresas usarem para tal processo a planária que é uma espécie de platelminto conhecido como Dugesia Tigrina.

As planárias pertencem a classe de animais classificados como invertebrados. "Animais com tal classificação são animais que não possuem espinha dorsal e também cujo organismo é composto por diferentes grupos de células, ou seja, são animais com formação multicelular onde normalmente sua formação é com célula animal"(Brusca & Brusca, 2007). Na Figura (2) a seguir, há a representação de uma planária da espécie Dugesia Tigrina.

Figura 2: Dugesia Tigrina. Retirado de https://www.google.com.br/dugesiatigrina, acesso em 19/08/2014.

As planárias também são vermes que vivem na terra a milhões de anos e

são segundo Brusca & Brusca (2007) cerca de 20.000 espécies. Já Moore (2003) relata que as planárias vivem principalmente em ambientes aquáticos, medem desde alguns milímetros até metros de comprimento. Como pode ser visto na Figura (2), são animais de corpo achatado, revestido de muco, com epiderme geralmente ciliada, são carnívoros e alimentam de animais vivos ou mortos, tendo assim, grande importâcia na degradação da matéria orgânica. Alguns exemplos de planárias são as tênias e os esquistossomos.

Em relação à àgua doce, a planária é um dos platelmintos mais conhecido, medem cerca de 1,5 centímetro de comprimento e situam-se em locais limpos em cima de pedras e folhas. São animais hermafroditas e se reproduzem sexuada e assexuadamente, alimentam-se de outros vermes e moluscos. Outra característica das planárias de água doce, é a capacidade de regeneração, que é bastante estudada Alvarado & Newmark (1998).

Segundo estudos de bioindicadores de qualidade de água realizado no Instituto de Ciências Gerais da Universidade Federal de Minas Gerais, esses animais são viáveis para a realização de indicadores da qualidade de água e saúde de ecossistemas, pois eles apresentam longo ciclo de vida, são organismos grandes, pouca mobilidade tem elevada diversidade taxônomica, ou seja, são de diversos grupos de seres vivos e são sensíveis a diferentes concentrações de poluentes.

Na literatura é discutido o uso de planárias como indicadores biológicos de qualidade das águas. Por exemplo, Alvarado & Newmark (1998) afirma a sensibilidade das planárias Dugesia Tigrina ao entrarem em contato com produtos tóxicos e compostos metálicos, assim como a capacidade de regeneração que as planárias de água doce possuem quando entram em contato com poluentes.

Estudos mais recentes, como o de Ribeiro (2012), relatam que as planárias de água doce possuem vasta distribuição geográfica, plasticidade biológica, facilidade de cultivo em laboratório, sensibilidade a contaminantes diversos e capacidade regenerativa, por isso vem sendo utilizadas como organismos-teste em ensaios ecotoxicologicos.

Desta forma, as planárias podem ser bons indicadores de toxicidade de água e a técnica de análise de sobrevivência, detalhada a seguir, é instrumento importante no estudo estatístico do problema.

2.2 Análise de Sobrevivência

"A análise de sobrevivência é um conjunto de técnicas estatísticas que estuda o tempo até a ocorrência de um evento de interesse, tendo como principal característica a presença de censura"(Klein & Kleinbaum, 2005).

Nos últimos anos, a análise de sobrevivência foi uma das áreas da estatística que mais se desenvolveu. Segundo Colosimo (2001) devido a fatores como aprimoramento de técnicas estatísticas, combinados com computadores cada vez mais rápidos e com maior capacidade de armazenamento de dados.

Outro fator que contribui para o uso da análise de sobrevivência é a sua aplicabilidade. Na engenharia, usa-se a análise de sobrevivência, que é denominada confiabilidade, para determinar, por exemplo, o tempo de garantia de uma peça ou produto. Na área financeira, por sua vez para verificar o tempo de adesão dos clientes em um pacote, assim como na área da saúde, para analisar o tempo até a cura de um paciente com uma doença.

Em estudos de análise de sobrevivência é de extrema importância estabelecer qual será o evento de interesse, por exemplo, a cura de uma doença nos pacientes, assim como o início e o fim do acompanhamento. Tal importância se dá, por exemplo, ao fato de um indivíduo completar ou não o período de acompanhamento, uma vez que isso vai implicar no que é chamado de falha ou censura. A censura ocorre se por algum motivo, não foi possível observar o evento de interesse ou quando não se conhece o tempo de vida exato.

"É importante ressaltar, mesmo censurados, todos os dados do estudo precisam ser analisados, uma vez que mesmo incompletas, as observações censuradas fornecem informações sobre os tempos de vida do objeto sob estudo"(Strapasson, 2007).

Normalmente, as situações em que há censura é quando o estudo termina e o evento não ocorreu, quando, durante o estudo, perdeu-se o contato com o paciente ou, em se tratando de doença, o paciente morre por outra causa e não pela doença estudada.

É possível segundo, Parreira (2007), relatar a censura à direita de maneira geral como:

• Censura tipo 1: quando os tempos de sobrevivência são maiores que o final do período do experimento.

• Censura tipo 2: quando o estudo termina depois que um número preestabelecido de falhas ocorram.

• Censura aleatória: quando o indivíduo deixa de fazer parte do grupo observado sem ter ocorrido a falha ou pela ocorrência de um evento diferente daquele de interesse.

Já Colosimo & Giolo (2006) classificam à esquerda quando o tempo registrado no acompanhamento é maior do que o tempo de falha, ou seja, quando o indivíduo foi observado, o evento de interesse já havia acontecido. Por exemplo, em um estudo em uma comunidade com o objetivo de determinar a idade com que crianças aprendem a ler os pesquisadores constatam que algumas crianças já aprenderam a ler antes de iniciar o estudo, porém não se lembram da idade exata.

Colosimo & Giolo (2006), definem censura à direita, quando não ocorre o evento de interesse durante o estudo, por exemplo, no caso anterior da pesquisa em uma comunidade sobre crianças que sabem ou não ler, censura a direita seria ao fazer a coleta dos dados as crianças ainda não sabiam ler.

Por fim, na censura intervalar o acompanhamento de um evento pode ser feito em período pré-estabelecido, desse modo o evento de interesse ocorre durante um intervalo de tempo conhecido. Por exemplo, considere um estudo em que o tempo de interesse da reincidência de um particular câncer seguido de uma cirurgia para a remoção do tumor. Suponha que três meses após a cirurgia, um paciente é observado e constata-se que ele não está doente, mas após seis meses, verifica-se a recorrência

da mesma. "O tempo de recorrência da doença neste paciente não é conhecido, sabe-se apenas que está entre três e seis meses, e portanto, um tempo com censura intervalar"(Strapasson, 2007).

Formalmente, nestes estudos, as informações são representadas por t_i tempo observado de falha ou de censura do i-ésimo indivíduo, δ_i que é uma variável indicadora de censura, $\delta_i = 0$ e falha $\delta_i = 1$ e $\mathbf{x_i}$ vetor que representa as covariáveis do indivíduo, por exemplo, idade, sexo, tratamento. Desta forma para cada indivíduo tem-se as informações $(t_i, \delta_i, \mathbf{x_i})$.

A partir das informações levantadas no decorrer do experimento, as principais características de interesse em análise de sobrevivência são a função de sobrevivência, a função de risco e vida média residual que são definidas a seguir. Seja T uma variável aleatória que representa o tempo até o evento de interesse que possui a função densidade de probabilidade $f(t)$ associada, então a função de sobrevivência é dada por

$$S(t) = P(T > t) = \int_0^t f(u)du, t > 0 \tag{1}$$

A função de sobrevivência representa a probabilidade de um indivíduo sobreviver além de certo tempo t, onde T representa o tempo de vida, e f é a função densidade. A função de sobrevivência é muito importante, pois tendo a probabilidade de sobrevivência para diferentes valores de tempos, tem-se um resumo importante dos dados de sobrevivência.

Além disso, considere que um indivíduo tenha sobrevivido até um instante de tempo t. O limite da probabilidade de haver uma falha em um intervalo de tempo t até t+Δt com Δt tendendo a zero, sabendo que o indivíduo sobreviveu até o instante t, é conhecido como função de risco, ou seja,

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t} \tag{2}$$

Essa função de risco da, de modo instantâneo, o risco de um indivíduo morrer considerando que ele sobreviveu até um instante t.

Outra relação importante em análise de sobrevivência é que a função de risco pode ser escrita como sendo a razão entre a função densidade da distribuição pela função de sobrevivência, ou seja,

$$h(t) = \frac{f(t)}{S(t)} \tag{3}$$

Importante ressaltar que, para alguns modelos, a função de risco exposta acima, não apresenta uma forma analítica explícita.

2.3 Estimador de Kaplan-Meier

Para estimar a função de sobrevivência referente aos dados observados, estando os tempos ordenados para cada grupo usa-se, por exemplo, o estimador Kaplan & Meier (1958), também conhecido como estimador limite-produto. "Esse consiste em um método de estimação não paramétrico da função de sobrevivência bastante usado na área médica e que tem ganhado cada vez mais espaço em estudos de análise de sobrevivência"(Colosimo & Giolo, 2006).

A definição do estimador de Kaplan-Meier referente a função de sobrevivência é

$$\hat{S}(t) = \prod_{j:t_j<t} \left(\frac{n_j - d_j}{n_j} \right) = \prod_{j:t_j<t} \left(1 - \frac{d_j}{n_j} \right), t > 0 \tag{4}$$

em que $\hat{S}(t)$ é uma função escada com degraus nos tempos observados de falha com tamanho $1/n$ onde n é o tamanho da amostra, $t_1 < t_2 ... < t_k$ são os k tempos distintos e ordenados de falha, d_j o número de falhas em t_j, $j = 1, ..., k$ e n_j o número de elementos sob risco em t_j, ou seja, os indivíduos que não falharam e não foram censurados até o instante imediatamente anterior a t_j.

Na construção do estimador de Kaplan-Meier, a quantidade de intervalos de tempos e os tempos de falhas distintos da amostra são considerados em quantidades equivalentes, ou seja, os limites dos intervalos dos tempos dos estimadores são os tempos de falha da amostra. Em estudos de Breslow & Crowley (1974) é relatado que o estimador de Kaplan-Meier possui as propriedades de ser não viciado

para grandes amostras, ou seja, a esperança do estimador Kaplan-Meier será igual a função de sobrevivência da população a que se deseja estimar. Ainda, o estimador de Kaplan-Meier converge assintoticamente para um processo gaussiano e ainda é o estimador de máxima verossimilhança para a função de sobrevivência.

2.4 Principais Modelos Probabilísticos para dados de Sobrevivência

Para fazer ajustes de modelos aos dados de uma pesquisa, existem na estatística vários modelos para serem testados. Na análise de sobrevivência alguns modelos possuem destaque devido aos bons ajustes a diversas situações práticas tais como modelo exponencial, Weibull, log-normal e o gama. A seguir temos uma explanação sobre estes modelos.

Modelo Exponencial

Na modelagem paramétrica, um dos modelos mais simples para descrever o tempo de falha é o exponencial, pois apresenta um único parâmetro α e tem a função de risco constante. "No final dos anos de 1940, a distribuição exponencial era usada em tempos de vida e em remissão de doenças crônicas e infecciosas, assim como em estudos de confiabilidade de sistemas eletrônicos"(Andreozzi et al., 2011).

A função densidade de probabilidade para a variável aleatória tempo de vida T com distribuição exponencial é dada por

$$f(t;\alpha) = \frac{1}{\alpha} exp\left\{-\left(\frac{t}{\alpha}\right)\right\}, t \geq 0, \tag{5}$$

em que o parâmetro $\alpha > 0$ é o tempo médio de vida e possui a mesma unidade do tempo de falha t.

Já as funções de sobrevivência $S(t;\alpha)$ e de taxa de falha $h(t;\alpha)$ são dadas, respectivamente, por

$$S(t;\alpha) = exp\left\{-\left(\frac{t}{\alpha}\right)\right\} \tag{6}$$

e

$$h(t;\alpha) = \frac{1}{\alpha}, t \geq 0. \tag{7}$$

Segundo Brito & Souza (2010), uma característica do modelo exponencial é que ele possui uma função de risco constante ao longo do tempo, assim se o modelo exponencial se ajustar bem aos dados de um grupo de pacientes com determinada doença, indiferente do tempo que cada paciente está com a doença, o risco de morte será o mesmo. Abaixo, na Figura (3), está a representação gráfica, segundo Colosimo & Giolo (2006) para a distribuição exponencial com diferentes valores de parâmetros.

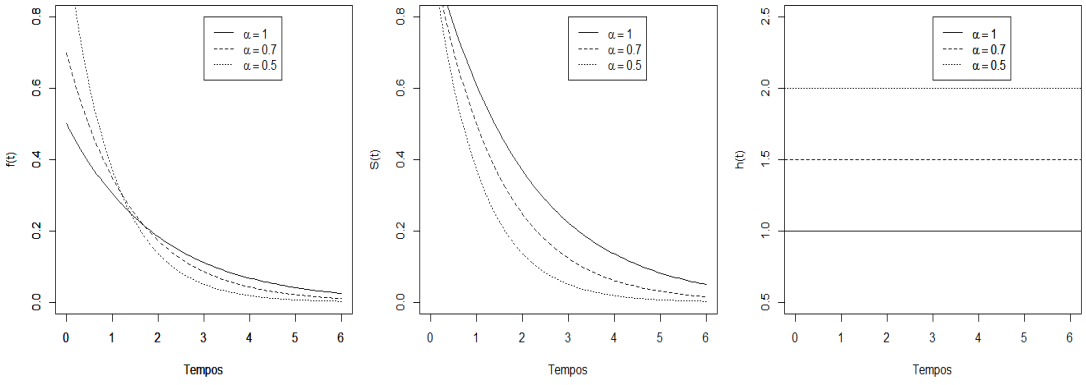

Figura 3: Forma típica das funções de densidade de probabilidade, de sobrevivência e de risco da distribuição exponencial para $\alpha = 1,0; 0,7; 0,5$. Fonte: Colosimo e Giolo (2006).

Modelo Weibull

Outro modelo muito utilizado é a distribuição Weibull, proposta em 1939. Segundo Andreozzi et al. (2011), a distribuição de Weibull é bastante usado em estudos biomédicos e Collet (1994) relata que a distribuição Weibull é tão importante para análise de confiabilidade, quanto a distribuição normal é para os modelos lineares.

A distribuição Weibull possui dois parâmetros, um parâmetro de escala $\alpha > 0$ e um parâmetro de forma $\gamma > 0$. Tem a característica de que a função de risco pode ser decrescente, crescente ou constante e quando o parâmetro $\gamma = 1$, a Weibull tem a distribuição exponencial como um caso particular.

A distribuição de Weibull tem função de densidade de probabilidade, função de sobrevivência e risco, dadas respectivamente por

$$f(t;\alpha,\gamma) = \frac{\gamma}{\alpha^{\gamma}} t^{\gamma-1} exp\left\{-\left(-\frac{t}{\alpha}\right)^{\gamma}\right\}, t \geq 0 \tag{8}$$

$$S(t;\alpha,\gamma) = exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\} \tag{9}$$

$$h(t;\alpha,\gamma) = \frac{\gamma}{\alpha^{\gamma}} t^{\gamma-1} \tag{10}$$

Abaixo na Figura (4) segue uma representação dos gráficos da distribuição Weibull.

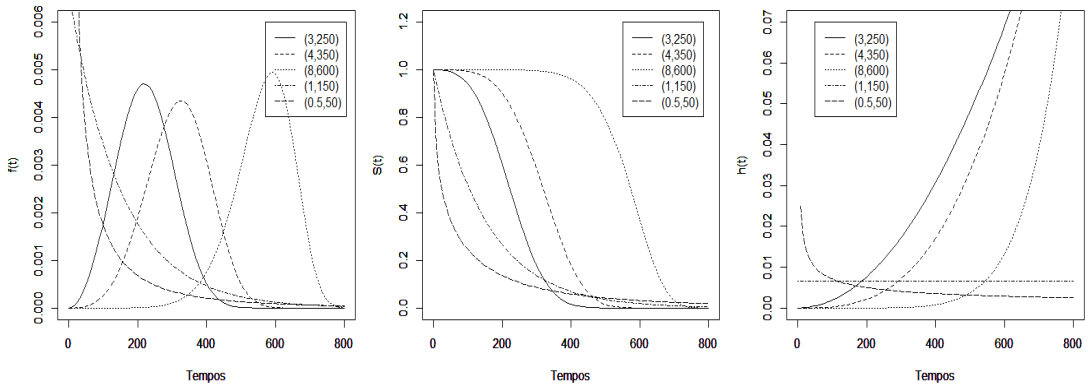

Figura 4: Forma típica das funções de densidade de probabilidade, de sobrevivência e de risco da distribuição Weibull para alguns valores dos parâmetros γ, α respectivamente. Fonte: Colosimo e Giolo (2006).

Modelo Log-normal

Outra distribuição muito utilizada para caracterizar tempos de vida é a distribuição log-normal. O logaritmo de uma variável com distribuição log-normal de parâmetros μ e σ tem distribuição normal, com média μ e desvio padrão σ, assim dados provenientes de uma distribuição log-normal podem ser analisados segundo uma distribuição normal, considerando, ao invés dos dados originais, o seu logaritmo.

Segundo Colosimo & Giolo (2006), a distribuição log-normal é muito usada na caracterização de isolação elétrica assim como para descrever situações clínicas como tempo de vida dos pacientes. A distribuição log-normal também possui dois parâmetros μ e σ e sua função de risco pode ser decrescente, crescente e é unimodal. Andreozzi et al. (2011) relata que a função de risco da distribuição log-normal, é decrescente para grandes valores do tempo de vida.

A distribuição log-normal tem como função densidade e função de sobrevivência, respectivamente

$$f(t;\mu,\sigma) = \frac{1}{\sqrt{2\pi}t\sigma} exp\left\{-\frac{1}{2}\left(\frac{log(t)-\mu}{\sigma}\right)^2\right\}, t > 0 \tag{11}$$

$$S(t;\mu,\sigma) = \Phi\left(\frac{-log(t)+\mu}{\sigma}\right) \tag{12}$$

em que μ é a média do logaritmo do tempo de falha, assim como σ é o desvio-padrão, $\Phi(.)$ é a função de distribuição acumulada de uma normal padrão.

A função de risco da log-normal não possui uma forma analítica e é obtida como a razão entre a função densidade e a função de sobrevivência, como na equação (3).

Abaixo na Figura (5) segue uma representação gráfica da distribuição log-normal.

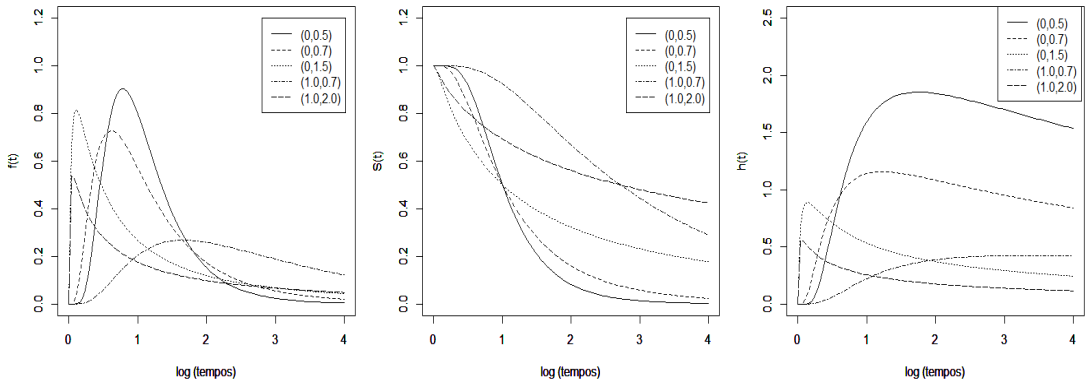

Figura 5: Forma típica das funções de densidade de probabilidade, de sobrevivência e de risco da distribuição log-normal para μ, σ respectivamente. Fonte: Colosimo e Giolo (2006).

Modelo Gama

Outra distribuição importante é a gama, que quando o parâmetro $k = 1$, inclui a exponencial e que tem como utilidade, por exemplo, a descrição do tempo de vida de materiais eletrônicos e efeitos aleatórios. Sua função de densidade é

$$f(t;\alpha,k) = \frac{1}{\Gamma(k)\alpha^k} t^{k-1} exp\left\{-\left(\frac{t}{\alpha}\right)\right\}, t > 0 \tag{13}$$

onde $\Gamma(k)$ é a função gama, $\alpha > 0$ é um parâmetro de escala e $k > 0$ é um parâmetro de forma.

A função de sobrevivência é dada por

$$S(t;\alpha,k) = \int_t^\infty \frac{1}{\Gamma(k)\alpha^k} u^{k-1} exp\left\{-\left(\frac{u}{\alpha}\right)\right\} du. \tag{14}$$

Assim como na log-normal, a função de risco não possui uma forma analítica e fica representada como a razão entre a função densidade e a função de sobrevivência.

Abaixo na figura (6) segue uma representação gráfica para as funções da distribuição gama.

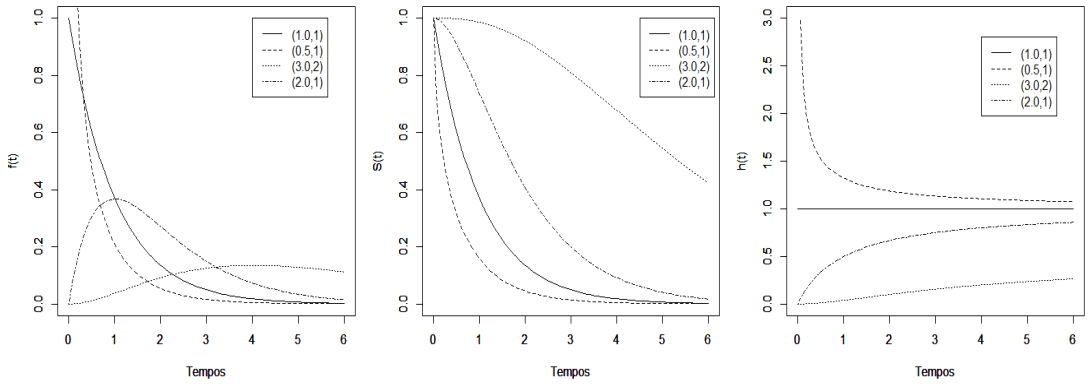

Figura 6: Forma típica das funções de densidade de probabilidade, de sobrevivência e de risco da distribuição gama para alguns valores dos parâmetros k, α respectivamente. Fonte: Colosimo e Giolo (2006).

Modelo Gama Generalizada

Outra distribuição de destaque é a distribuição gama generalizada. "Tal distribuição foi introduzida em 1962 por Stacy e tem a vantagem de possuir as distribuições exponencial, Weibull e gama como modelos encaixados, o que possibilita usar o teste da razão de verossimilhança para testar a hipótese nula e hipótese alternativa para a discriminação de modelos a serem usados"(Colosimo & Giolo, 2006).

Na distribuição gama generalizada, para $\gamma = k = 1$ tem-se uma distribuição exponencial com parâmetro α. Para $k = 1$ tem-se uma distribuição Weibull com parâmetros γ e α e para $\gamma = 1$ tem-se uma distribuição gama com parâmetros k e α. Além disso, segundo Lawlees (1982), a distribuição log-normal aparece como sendo um caso limite quando $k \longrightarrow \infty$ na distribuição gama generalizada.

A função densidade da distribuição gama generalizada é expressa por

$$f(t;\alpha,\gamma,k) = \frac{\gamma}{\Gamma(k)\alpha^{\gamma k}} t^{k-1} exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\}, t > 0 \qquad (15)$$

em que $\Gamma(k)$ é a função gama. Note que nesta distribuição tem-se três parâmetros

γ, k e α onde todos são positivos.

A função de sobrevivência para esta distribuição é dada por

$$S(t;\alpha,k,\gamma) = \int_t^{\infty} \frac{1}{\Gamma(k)\alpha^k} u^{k-1} exp\left\{-\left(\frac{u}{\alpha}\right)\right\} du \tag{16}$$

A função de risco não possui uma forma analítica e fica representada como a razão entre a função densidade e a função de sobrevivência.

2.5 Formas de Estimação dos Parâmetros do Modelo

Os modelos probabilísticos como por exemplo, Weibull, log-normal, exponencial, gama, possuem os parâmetros, que são quantidades desconhecidas e que precisam ser estimadas. Segundo Colosimo & Giolo (2006) em estudo de tempos de falha, os parâmetros devem ser estimados a partir das observações amostrais, para uma boa determinação do modelo.

Dentre os métodos para estimar os parâmetros, segundo Nelson (1990) e Gujarati (2006) um dos mais conhecidos e também mais simples é o método de estimação de mínimos quadrados, porém de acordo com Pinheiro & Bates (2001) para estudo de tempo de vida, este método é inapropriado, uma vez que ele não incorpora a censura no processo de estimação. Com isso, uma boa opção para fazer a estimação dos parâmetros é usando o método de máxima verossimilhança, pois este incorpora a censura de forma relativamente simples de ser entendido e com boas propriedades, principalmente para grandes amostras.

Suponha uma amostra observada $t_1, ..., t_n$ de uma população de interesse em que todas as amostras são não censuradas e que a função de densidade da população seja $f(t;\boldsymbol{\theta})$, sendo $\boldsymbol{\theta}$ parâmetro ou vetor de parâmetro a ser estimado. Nesse caso a função de verossimilhança será

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} f(t_i;\boldsymbol{\theta}) \tag{17}$$

Os estimadores de máxima verossimilhança no caso sem censura é determinando encontrando-se o máximo da função de verossimilhança na equação (17).

Além disso, o parâmetro $\boldsymbol{\theta}$ pode estar representando um ou mais parâmetros. Por exemplo, na distribuição log-normal, $\boldsymbol{\theta}$ representará a média μ e o desvio padrão σ.

Agora quando há a censura à direita a função de verossimilhança será

$$\begin{aligned} L(\boldsymbol{\theta}) &= \prod_{i=1}^{n} \left[f\left(t_i;\boldsymbol{\theta}\right)\right]^{\delta_i} S\left[\left(t_i;\boldsymbol{\theta}\right)\right]^{1-\delta_i} \\ &= \prod_{i=1}^{n} \left[h\left(t_i;\boldsymbol{\theta}\right)\right]^{\delta_i} \left[S\left(t_i;\boldsymbol{\theta}\right)\right] \end{aligned} \tag{18}$$

em que $f\left(t_i;\boldsymbol{\theta}\right)$ contribui com a informação não censurada, a contribuição da informação censurada é dada pela $S\left(t_i;\boldsymbol{\theta}\right)$, $\boldsymbol{\theta}$ é o vetor de parâmetros a ser estimado e δ_i é uma variável indicadora de falha ou censura,ou seja, $\delta_i = 0$ para censura e $\delta_i = 1$ para falha.

Com base em resultados obtidos pela amostra, o estimador de máxima verossimilhança escolhe a melhor combinação dos parâmetros da distribuição e que melhor explique a amostra observada, ou seja, um valor para um parâmetro, por exemplo $\boldsymbol{\theta}$, de maneira que tal valor maximize a função de verossimilhança $L(\boldsymbol{\theta})$.

Por exemplo, usando a equação (18), vamos determinar o estimador de máxima verossimilhança para o parâmetro $\theta = \alpha$ modelo exponencial. Sabemos que a função de risco e a função de sobrevivência, são dadas pelas equações (6) e (7). Substituindo as mesmas na equação (18), temos

$$L(\alpha) = \prod_{i=1}^{n} \left[\frac{1}{\alpha} e^{-\left(\frac{t_i}{\alpha}\right)}\right]^{\delta_i} \left[e^{-\left(\frac{t_i}{\alpha}\right)}\right]^{1-\delta_i}, \tag{19}$$

que é igual

$$\begin{aligned} L(\alpha) &= \prod_{i=1}^{n} \left[\left(\frac{1}{\alpha}\right)^{\delta_i} e^{-\left(\frac{t_i}{\alpha}\right)\delta_i}\right] \left[e^{-\left(\frac{t_i}{\alpha}\right)} \frac{1}{e^{-\left(\frac{t_i}{\alpha}\right)^{\delta_i}}}\right] \\ &= \prod_{i=1}^{n} \left(\frac{1}{\alpha}\right)^{\delta_i} e^{-\left(\frac{t_i}{\alpha}\right)} \end{aligned} \tag{20}$$

Aplicando o logaritmo na equação (20)

$$\begin{aligned} lnL(\alpha) &= \sum_{i=1}^{n} \delta_i ln\left(\frac{1}{\alpha}\right) - \frac{1}{\alpha}\sum_{i=1}^{n} t_i \\ &= \sum_{i=1}^{n} \delta_i \left(ln(1) - ln(\alpha)\right) - \frac{1}{\alpha}\sum_{i=1}^{n} t_i \end{aligned}$$

$$= \sum_{i=1}^{n} \delta_i ln(\alpha) - \frac{1}{\alpha} \sum_{i=1}^{n} t_i. \tag{21}$$

Derivando a equação (21) em relação ao parâmetro α

$$\frac{\partial lnL(\alpha)}{\partial \alpha} = \frac{1}{\alpha} \sum_{i=1}^{n} \delta_i + \frac{1}{\alpha^2} \sum_{i=1}^{n} t_i, \tag{22}$$

igualando a zero

$$\frac{1}{\hat{\alpha}} \sum_{i=1}^{n} \delta_i + \frac{1}{\hat{\alpha}^2} \sum_{i=1}^{n} t_i = 0, \tag{23}$$

isolando o parâmetro $\hat{\alpha}$

$$\hat{\alpha} = \frac{\sum_{i=1}^{n} t_i}{\sum_{i=1}^{n} \delta_i} \tag{24}$$

Assim a equação (24) é o estimador de máxima verossimilhança para o parâmetro do modelo exponencial.

Já para o modelo Weibull, segundo Mantovani & Franco (2004), tem-se os dois parâmetros α e γ. Usando a equação (18) temos

$$L(\alpha;\gamma) = \prod_{i=1}^{n} \left[\frac{\gamma}{\alpha^\gamma} t_i^{\gamma-1} \right]^{\delta_i} e^{-\left(\frac{t_i}{\alpha}\right)\gamma} \tag{25}$$

e

$$lnL(\alpha;\gamma) = rln(\gamma) - r\gamma ln(\alpha) + (\gamma - 1) \sum_{i=1}^{n} \delta_i ln(t_i) - \sum_{i=1}^{n} \left(\frac{t_i}{\alpha} \right)^\gamma, \tag{26}$$

em que $r = \sum_{i=1}^{n} \delta_i$ é o número de observações não censuradas.

Derivando a equação (26) em relação aos parâmetros α e γ, tem-se

$$\frac{\partial}{\partial \alpha} lnL(\alpha;\gamma) = \frac{r\gamma}{\alpha} + \frac{\gamma}{\alpha} \sum_{i=1}^{n} \left(\frac{t_i}{\alpha} \right)^\gamma \tag{27}$$

e

$$\frac{\partial}{\partial \gamma} lnL(\theta;\alpha/\gamma) = \frac{r}{\gamma} rln(\alpha) + \sum_{i=1}^{n} \delta_i ln(t_i) \sum_{i=1}^{n} \left(\frac{t_i}{\alpha} \right) \gamma ln \left(\frac{t_i}{\alpha} \right) \tag{28}$$

Os estimadores de máxima verossimilhança de α e γ podem ser obtidos, de acordo com Mantovani & Franco (2004), usando métodos computacionais, por exemplo, método de Newton-Raphson nas equações (27) e (28).

Outra distribuição citada neste trabalho é a gama generalizada. Para obter seus estimadores de máxima verossimilhança, a função densidade que está na equação (15) deste trabalho, será escrita com base em Ramos et al. (2014), como.

$$f(t;\boldsymbol{\theta}) = \frac{\alpha}{\Gamma(\Phi)}\mu^{\alpha\Phi}t^{\alpha\Phi-1}e^{(-(\mu t)^{\alpha})} \tag{29}$$

em que $\alpha, \Phi > 0$, α, Φ são dois parâmetros de forma e μ parâmetro de escala.

A função de verossimilhança é dada por

$$L(\boldsymbol{\theta}) = \frac{\alpha^n}{\Gamma(\Phi)^n}\mu^{n\alpha\Phi}\left\{\prod_{i=1}^{n} t_i^{\alpha\Phi-1}\right\} e^{\left\{-\mu^{\alpha}\sum_{i=1}^{n} t_i^{\alpha}\right\}}, \tag{30}$$

em que $\boldsymbol{\theta} = (\Phi, \mu, \alpha)$.

Fazendo as derivadas $\frac{\partial}{\partial\alpha}ln(L(\boldsymbol{\theta}))$, $\frac{\partial}{\partial\mu}ln(L(\boldsymbol{\theta}))$ e $\frac{\partial}{\partial\Phi}ln(L(\boldsymbol{\theta}))$ e igualando a zero, tem-se as equações de verossimilhança dadas por

$$n\psi(\hat{\phi}) = n\hat{\alpha}ln(\hat{\mu}) + \hat{\alpha}\sum_{i=0}^{n} ln(t_i) \tag{31}$$

$$n\hat{\phi} = \hat{\mu}^{\hat{\alpha}}\sum_{i=1}^{n} t_i^{\hat{\alpha}} \tag{32}$$

$$\frac{n}{\hat{\alpha}} + n\hat{\phi}ln(\mu) + \phi\sum_{i=1}^{n} ln(t_i) = \hat{\mu}^{\hat{\alpha}}\sum_{i=1}^{n} t_i^{\hat{\alpha}}ln(\hat{\mu}t_i), \tag{33}$$

em que $\psi(k) = \frac{\partial}{\partial k}ln\Gamma(k) = \frac{\Gamma'(k)}{\Gamma(k)}$. As soluções para as equações (31) a (33), segundo Ramos et al. (2014), necessitam de auxílio de métodos numéricos, por exemplo, método de Newton-Raphson, e tais soluções serão os estimadores de máxima verossimilhança dos parâmetros do modelo gama generalizada.

Em seguida, para o modelo log-normal, assim como ocorre para Weibull e gama generalizada, a estimação dos parâmetros requer métodos computacionais. Para a obtenção da estimação dos parâmetros do modelo log-normal, será feito com base em Lawlees (2003). Para isso é feita uma representação da sobrevivência como

$$S(y;u,b) = S_0\left(\frac{y-u}{b}\right), -\infty < y < \infty, \tag{34}$$

em que $-\infty < u < \infty$ sendo parâmetro de localização, $b > 0$ o parâmetro de escala e $S_0()$ sendo função de sobrevivência especificada e definida nos números reais.

Sendo T uma variável que indica o tempo de vida e fazendo $Y = lnT$, a equação (34) fica

$$S_0^*(t;\alpha,\beta) = S_0\left(\frac{lnt-u}{b}\right) = S_0^*\left[\left(\frac{t}{\alpha}\right)^{\beta}\right], \tag{35}$$

em que $\alpha = e^u$, $\beta = b^{-1}$ e $0 < \omega < \infty$, $S_0^*(\omega) = S_0(ln\omega)$.

Assim a verossimilhança para o caso de censura a direita com r observações censuradas será

$$L(u,b) = \prod_{i=1}^{n}\left[\frac{1}{b}f_0\left(\frac{y_i-u}{b}\right)\right]^{\delta_i} S_0\left(\frac{y_i-u}{b}\right)^{1-\delta_i}, \tag{36}$$

em que $y_i = ln(t_i)$ e $f_0(z) = -d\frac{S_0(z)}{dz}$ é a função densidade de probabilidade que corresponde a $S_0(z)$. Tomando $z_i = \frac{y_i-u}{b}$ e $r = \sum_{i=1}^{n}\delta_i$, o logaritmo da função de verossimilhança fica

$$lnL(u,b) = -rln(b) + \sum_{i=1}^{n}\left[\delta_i lnf_0(z_i) + (1-\delta_i)lnS_0(z_i)\right] \tag{37}$$

Derivando $\frac{\partial z_i}{\partial u} = -b^{-1}$ e $\frac{\partial z_i}{\partial b} = -z_i b^{-1}$, tem-se

$$\frac{\partial logL(u,b)}{\partial u} = -\frac{1}{b}\sum_{i=1}^{n}\left[\delta_i\frac{\partial lnf_0(z_i)}{\partial z_i} + (1-\delta_i)\frac{\partial lnS_0(z_i)}{\partial z_i}\right] \tag{38}$$

$$\frac{\partial logL(u,b)}{\partial b} = -\frac{r}{b} - \frac{1}{b}\sum_{i=1}^{n}\left[\delta_i z_i\frac{\partial lnf_0(z_i)}{\partial z_i} + (1-\delta_i)z_i\frac{\partial lnS_0(z_i)}{\partial z_i}\right], \tag{39}$$

e a segunda derivada

$$\frac{\partial^2 logL(u,b)}{\partial u^2} = \frac{1}{b^2}\sum_{i=1}^{n}\left[\delta_i\frac{\partial^2 lnf_0(z_i)}{\partial z_i^2} + (1-\delta_i)\frac{\partial^2 lnS_0(z_i)}{\partial z_i^2}\right] \tag{40}$$

$$\begin{aligned}\frac{\partial^2 logL(u,b)}{\partial b^2} &= \frac{r}{b^2}\sum_{i=1}^{n}\left[\delta_i z_i\frac{\partial lnf_0(z_i)}{\partial z_i} + (1-\delta_i)z_i\frac{\partial lnS_0(z_i)}{\partial z_i}\right] + \\ &\quad \frac{1}{b^2}\sum_{i=1}^{n}\left[\delta_i z_i^2\frac{\partial^2 lnf_0(z_i)}{\partial z_i^2} + (1-\delta_i)z_i^2\frac{\partial^2 lnS_0(z_i)}{\partial z_i^2}\right]\end{aligned} \tag{41}$$

$$\frac{\partial^2 lnL(u,b)}{\partial u \partial b} = \frac{1}{b^2}\sum_{i=1}^{n}\left[\delta_i \frac{\partial ln f_0(z_i)}{\partial z_i} + (1-\delta_i)\frac{\partial lnS_0(z_i)}{\partial z_i}\right] + \frac{1}{b^2}\sum_{i=1}^{n}\left[\delta_i z_i \frac{\partial^2 ln f_0(z_i)}{\partial z_i^2} + (1-\delta_i) z_i \frac{\partial^2 lnS_0(z_i)}{\partial z_i^2}\right] \quad (42)$$

A matriz de informação observada é dada por

$$I(u,b) = \begin{pmatrix} -\frac{\partial^2 lnL(u,b)}{\partial u^2} & -\frac{\partial^2 lnL(u,b)}{\partial u \partial b} \\ -\frac{\partial^2 lnL(u,b)}{\partial u \partial b} & -\frac{\partial^2 lnL(u,b)}{\partial b^2} \end{pmatrix}. \quad (43)$$

Com a matriz acima, obtém-se através de simulação os estimadores com base na dimensão da amostra e padrão de censura. A distribuição conjunta dos estimadores de máxima verossimilhança u e b é aproximadamente uma distribuição normal bivariada para grandes amostras com vetor de média (u, b) e matriz de covariância $I(\hat{u}, \hat{b})^{-1}$. Mais detalhes sobre a estimação dos parâmetros para o modelo log-normal, pode ser obtido em Lawlees (2003).

De forma geral, na presença de covariáveis pode-se inferir sobre os parâmetros do modelo de regressão com a função de verossimilhança

$$\begin{aligned} L(\boldsymbol{\beta}) &= \prod_{i=1}^{n}[f(t_i;\boldsymbol{\beta}|\boldsymbol{x_i})]^{\delta_i}[S(t_i;\boldsymbol{\beta}|\boldsymbol{x_i})]^{1-\delta_i} \\ &= \prod_{i=1}^{n}[f(t_i;\boldsymbol{\beta}|\boldsymbol{x_i})]^{\delta_i}\frac{S(t_i;\boldsymbol{\beta}|\boldsymbol{x_i})}{[S(t_i;\boldsymbol{\beta}|\boldsymbol{x_i})]^{\delta_i}} \\ &= \prod_{i=1}^{n}[h(t_i;\boldsymbol{\beta}|\boldsymbol{x_i})]^{\delta_i}S(t_i;\boldsymbol{\beta}|\boldsymbol{x_i}) \end{aligned} \quad (44)$$

em que $f(t_i;\boldsymbol{\beta}|\boldsymbol{x_i})$ representa a função densidade, $S(t_i;\boldsymbol{\beta}|\boldsymbol{x_i})$ a função de sobrevivência, δ_i a censura ou falha no tratamento e $h(t_i;\boldsymbol{\beta}|\boldsymbol{x_i})$ a função de risco e $\boldsymbol{x_i}$ representa as covariáveis. A estimação do parâmetro β está detalhada em Lawlees (2003).

2.6 Teste Log-Rank

Ao analisar as curvas de sobrevivência ajustadas para diferentes populações, por exemplo, comparar o tempo de vida de dois produtos diferentes, com-

parar um processo novo com um antigo, há a possibilidade de agrupar variáveis que apresentam comportamentos semelhantes, ou fazer comparações entre duas ou mais curvas da sobrevivência dos dados. Para tais comparações existe o teste Log-rank que foi proposto por Mantel (1966).

O teste Log-rank é bastante utilizado em análise de sobrevivência, principalmente quando, nos dados ocorre a proporcionalidade na função de risco, porém tal proporcionalidade não é condição necessária para uso do teste. Com base em Andreozzi et al. (2011), o teste é a diferença entre o número observado de falhas em cada grupo, assim como uma quantidade que corresponde ao número esperado de falhas sobre uma hipótese nula de que é possível reduzir o número de curvas de sobrevivências existentes no trabalho. O teste Log-rank é verificado através de um $p - valor$ com base em um nível de significância pré-estabelecido.

2.7 Critérios para seleção de modelos

Um dos problemas com que o pesquisador se depara ao modelar um conjunto de dados é sobre qual modelo utilizar. O que se procura é um modelo de menor ordem possível, que consiga se adequar satisfatoriamente aos dados. Porém, dados reais têm uma grande chance de nunca se adequarem perfeitamente a algum modelo, ou seja, escolher um modelo que consiga captar todas as características dos dados a serem modelados.

Uma solução é aumentar a ordem do modelo, ou o tamanho da amostra, permitindo, assim, que o modelo capte características mais complexas dos dados. A questão se torna, então, até onde é razoável aumentar a ordem do modelo para conseguir uma melhor adequação aos dados.

Com isso surgem alguns critérios para selecionar, de maneira mais precisa, um modelo para se adequar, da melhor forma possível, ao conjunto de dados. Segundo Burnhan & Anderson (2004) é importante usar critérios que sejam baseados em princípios científicos e então tem-se dois critérios mais conhecidos e usados que são, Critério de Informação de Akaike (Akaike's Information Criterion) conhecido

como (AIC) e Critério de Informação Bayesiano (Bayesian Information Criterion) conhecido como (BIC).

2.8 Critério de Informação de Akaike

Akaike (1974) sugeriu um critério conhecido como (AIC) Bozdogan (1987) que admite a existência de um modelo "real"desconhecido que descreve os dados e escolhe, dentre um grupo de modelos avaliados, o que minimiza a divergência de Kullback-Leibler (K−L). O valor de K−L para um modelo com parâmetros em relação ao modelo "real", representado por f é $l(f,g) = \int f(x) ln\left(\frac{f(x)}{g(x|\theta)}\right)$, onde $f(x)$ é modelo real e $g(x|\theta)$ representa um modelo com melhor aproximação para o modelo real $f(x)$.

O AIC, embora largamente aceito e utilizado, tem limitações. Ele foi desenvolvido sob o conceito de que, assintoticamente (quando o tamanho da amostra tende a infinito), ele converge para o valor exato da divergência de Kullback−Leibler. Com isto, por vezes o AIC não só falha em escolher um modelo mais parcimonioso, como por vezes escolhe o modelo de maior ordem entre todos os modelos comparados.

Diante desta situação, alguns métodos foram sugeridos para conseguir trabalhar satisfatoriamente com pequenas amostras, como o AICc (AIC corrigido), KIC (Kullback Information Criterion), KICc (KIC corrigido), AKICc (Aproximated KICc) e AICF (AIC Finite Sample). A diferença entre os métodos citados é o termo da penalização.

O Critério de Informação de Akaike (AIC) é definido como: $AIC = -2L(\theta)+2.[(p+1)+1]$, em que $L(\theta)$ corresponde à função de máxima verossimilhança do modelo e p o número de variáveis explicativas consideradas no modelo.

2.9 Critério de Informação Bayesiano

O critério de Informação Bayesiano BIC proposto por Schwarz (1978) usa a probabilidade a posteriori, ou seja, a probabilidade condicional a uma hipótese

válida. Assim como o AIC, tem como pressuposto a existência de um "modelo verdadeiro"que descreve a relação entre a variável dependente e as diversas variáveis explanatórias entre os diversos modelos sob seleção. Assim, o critério é definido como a estatística que maximiza a probabilidade de se identificar o verdadeiro modelo dentre os avaliados. O modelo com menor BIC é considerado o de melhor ajuste.

Todos os métodos, à exceção do BIC, são métodos assintoticamente eficientes, ou seja, à medida que o número de amostras tende ao infinito, eles tendem a escolher o modelo que diminui o erro. Porém, o BIC é um método consistente, que escolhe o modelo de ordem correto com probabilidade 1 à medida que o número de amostras tende ao infinito, desde que o modelo correto esteja no conjunto de modelos a ser testado.

"O Critério de Informação Bayesiano BIC, leva em consideração o tamanho da amostra e é definido como $BIC = -2.L(\theta)+[(p+1)+1].log(n)$ em que $L(\theta)$ também corresponde a função de máxima verossimilhança do modelo e p o número de variáveis explicativas consideradas no modelo"(Schwarz, 1978). Assim como o AIC, o BIC aumenta conforme soma dos quadrados dos erros SQE aumenta. Como modelos com mais variáveis tendem a produzir menor SQE mas usam mais parâmetros, a melhor escolha é balancear o ajuste com relação a quantidade de variáveis usadas, além disso, ambos critérios penalizam modelos com muitas variáveis. Para a definição do modelo a ser usado, escolhe o modelo que apresentar menores valores de AIC e BIC.

Os critérios apresentados apesar de conceitualmente diferentes acerca dos modelos em avaliação, utilizam o mesmo critério estatístico, o máximo da função de verossimilhança como medida do ajustamento, entretanto, definem valores críticos diferentes. 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, ou o modelo verdadeiro e então, tenta−se escolher o modelo que minimize a divergência (K−L). Com o BIC, está implícito que existe o modelo que descreve a relação entre as variáveis envolvidas e o critério tenta maximizar a

probabilidade de escolha do verdadeiro modelo (André & Regazzi, 2014).

2.10 Teste da razão de Verossimilhança

No intuito de diminuir a subjetividade na escolha de um modelo, além dos critérios de AIC e BIC existe, por exemplo, teste da razão de verossimilhanças. "Essa técnica utiliza a máxima verossimilhança, a qual é determinada em modelos em que um é o caso particular do outro. Como exemplo, temos a gama generalizada que apresenta os modelos Exponencial, Weibull, Log-normal e Gama, como caso particular"(Pascoa, 2012).

Segundo Colosimo & Giolo (2006), o teste da razão de verossimilhança compara valores do logaritmo da função de verossimilhança maximizada $logL(\hat{\theta})$ e que não tem restrição sob a hipótese nula $logL(\hat{\theta_0})$.

Segundo o mesmo autor acima a estatística para o teste é dada por

$$TRV = -2ln\left[\frac{L\left(\hat{\theta}_0\right)}{L\left(\hat{\theta}\right)}\right] = 2\left[lnL\left(\hat{\theta}\right) - lnL\left(\hat{\theta}_0\right)\right], \tag{45}$$

e sob a $H_0 : \theta = \theta_0$ segue aproximadamente uma distribuição qui-quadrado com p graus de liberdade.

"Ainda quando é utilizado modelos encaixados, o teste da razão de verossimilhança testa a hipótese nula H_0 de que o modelo de interesse é adequado versus uma hipótese alternativa de que o modelo não é adequado"(Cox & Hinkley, 1974). Por exemplo, como o modelo Weibull é uma generalização do modelo exponencial, se $H_0 : \gamma = \gamma_0 = 1$ temos a exponencial.

2.11 Modelo de Riscos Proporcionais

Quando se trabalha com pesquisas clínicas, um dos modelos de regressão em análise de sobrevivência mais utilizado é o modelo de riscos proporcionais de Cox. Um dos motivos é que o modelo de Cox possui a presença de componentes não paramétricos. "Esse modelo proporciona uma análise estatística cuja resposta

é o tempo até a ocorrência de um evento de interesse que pode ser ajustado por covariáveis"(Colosimo & Giolo, 2006).

A função risco do modelo proporcional de Cox é

$$h(t) = h_0(t)g(\mathbf{x}^{'}\boldsymbol{\beta}) \tag{46}$$

em que h(.) representa a função de risco, g(.) é uma função especificada, $g(0) = 1$, $h_0(t)$ é uma função com componente não paramétrica também chamada de função base, $\boldsymbol{\beta}$ sendo um vetor de coeficientes das covariáveis e $\mathbf{x}$ é um vetor das covariáveis de regressão e $t > 0$ representa o tempo de estudo (Cox, 1972).

O coeficiente $\boldsymbol{\beta}$ no modelo de Cox representa o efeito das covariáveis sobre a função da taxa de falha, além disso outra suposição é que as funções de risco com diferentes níveis de covariáveis não dependam do tempo t.

Considere um estudo que consiste na comparação de tempos de falha ou risco de dois grupos em que pacientes, por exemplo, recebem tratamento padrão (grupo 0) ou outro tratamento (grupo 1). Representando as funções de risco do primeiro e segundo grupos por $h_0(t)$ e $h_1(t)$ e assumindo proporcionalidade entre as funções, tem-se

$$\frac{h_1(t)}{h_0(t)} = K \tag{47}$$

em que K é a razão de risco, constante para todo tempo t de acompanhamento do estudo. Representando $K = e^{\boldsymbol{x}^{'}\boldsymbol{\beta}}$ tem-se

$$\frac{h_1(t)}{h_0(t)} = e^{\boldsymbol{x}^{'}\boldsymbol{\beta}} \tag{48}$$

A partir da equação (48), observa-se que o quociente das taxas de falha entre dois indivíduos é constante no tempo t. Assim, por exemplo, se um grupo tem, no início de um estudo um risco de morte que é igual a duas vezes mais que o risco de outro grupo, supõe-se que a razão de risco será a mesma em todo o período de tratamento.

Generalizando para p covariáveis com $\boldsymbol{x} = (x_1, ..., x_p)'$, tem-se para a razão o modelo de regressão de Cox.

$$h(t) = h_0(t)g(\boldsymbol{x}^{'}\boldsymbol{\beta}) \tag{49}$$

em que h_0 é não paramétrico e $g(\boldsymbol{x'\beta})$ é paramétrico.

$$g(\boldsymbol{x'\beta}) = \exp(\boldsymbol{x'\beta}) = \exp(\beta_1 x_1 + ... + \beta_p x_p) \tag{50}$$

neste caso $\boldsymbol{\beta}$ é o vetor de parâmetros associados às covariáveis.

A razão das funções de taxa de falha para indivíduos i e j é

$$\frac{h_i(t)}{h_j(t)} = \frac{h_0(t)e^{\boldsymbol{x_i'\beta}}}{h_0(t)e^{\boldsymbol{x_j'\beta}}} = e^{\boldsymbol{x_i'\beta - x_j'\beta}} \tag{51}$$

em que tal quociente não depende do tempo t.

"Para o ajuste do modelo de Cox é necessário fazer um condicionamento na função de verossimilhança devido ao componente não paramétrico $h_0(t)$. Esse método é conhecido como máxima verossimilhança parcial de Cox"(Cox, 1975). Mais detalhes pode ser visto em Lawlees (2003), Colosimo & Giolo (2006) e Andreozzi et al. (2011).

2.12 Tempos Medianos e Percentis

Tão importante quanto usar corretamente os métodos para escolher um modelo que se ajuste bem aos dados, é fazer a interpretação adequada dos resultados obtidos. Em análise de sobrevivência dependendo do modelo com que foi feito o ajuste, a interpretação do resultado não é imediata.

Segundo Colosimo & Giolo (2006), um ajuste feito, por exemplo, com o modelo Weibull ou log-normal sofre uma transformação logarítmica na escala da resposta e tal transformação impede que a interpretação de coeficientes estimados seja feita de maneira direta como é feito, por exemplo, em uma regressão linear.

Segundo o mesmo autor, assim como em uma regressão linear se fixado os outros termos do modelo e feito a estimação de um parâmetro, por exemplo, $\hat{\beta} = 2$, significa que para um aumento de uma unidade na covariável correspondente a média do logaritmo do tempo, fica aumentada em duas unidades. Porém ao ajustar aos dados um modelo que sofre uma transformação logarítmica na sua escala, a

interpretação como dada antes para a regressão linear fica errada, pois sabe-se que a esperança do logaritmo de uma variável é diferente do logaritmo da esperança da mesma variável.

Com essa situação, uma possibilidade sugerida por Colosimo & Giolo (2006) é usar a razão de tempos medianos para a análise dos modelos ajustados em análise de sobrevivência. Segundo ele, para uma covariável binária a razão do tempo mediano, ou de qualquer percentil, é

$$\frac{t_p\left(x=1,\hat{\boldsymbol{\beta}}\right)}{t_p\left(x=0,\hat{\boldsymbol{\beta}}\right)} = exp\left\{\hat{\boldsymbol{\beta}}\right\} \tag{52}$$

em que $\hat{\boldsymbol{\beta}}$ é o parâmetro estimado do modelo e $x=0$ e $x=1$ é a covariável binária e p é o percentil de interesse.

Por exemplo, se T tem uma distribuição de Weibull com parâmetros $exp\left\{\beta_0+\beta_1 x_1+\beta_2 x_2\right\}$ e γ tem-se que

$$t_p\left(\boldsymbol{x},\hat{\boldsymbol{\beta}}\right) = (-logp)^{\hat{\gamma}} exp\left\{\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2\right\} \tag{53}$$

que é o mesmo que

$$\begin{aligned}\frac{t_p\left(x_1=1,x_2=0,\hat{\boldsymbol{\beta}}\right)}{t_p\left(x_1=0,x_2=0,\hat{\boldsymbol{\beta}}\right)} &= \frac{(-logp)^{\hat{\gamma}}\, exp\left\{\hat{\beta}_0+\hat{\beta}_1 x_1+\hat{\beta}_2 x_2\right\}}{(-logp)^{\hat{\gamma}}\, exp\left\{\hat{\beta}_0\right\}} \\ &= exp\left\{\hat{\beta}_1\right\} \end{aligned} \tag{54}$$

Esta interpretação também pode ser usada para covariáveis categóricas.

A seguir, apresenta-se a análise de resíduo que possibilita avaliar se o modelo ajustado é adequado aos dados.

2.13 Análise de Resíduo

Em toda análise estatística, é de grande importância fazer uma análise de resíduo afim de confirmar a adequação do modelo que foi ajustado com base em técnicas anteriores. Porém, segundo Buuren & Miranda (2001), é bom ressaltar que

fazer uma análise de resíduo significa uma maneira de rejeitar modelos que não se ajustam bem ao conjunto de dados do trabalho, mas não demonstra que um modelo particular está totalmente bem ajustado.

Segundo Andreozzi et al. (2011), "a definição de uma medida de resíduo quando se trabalha com a análise de sobrevivência não é tão clara e direta se comparado com uma abordagem de modelos lineares". No ajuste de modelos paramétricos em análise de sobrevivência, dois resíduos são bastante utilizados para analisar o ajuste global do modelo, são eles: os resíduos padronizados e os resíduos de Cox-Snell. Em ambos os resíduos, se o modelo utilizado estiver bem ajustado, o gráfico dos resíduos versus os riscos acumulados apresentará um comportamento próximo a uma reta, ou seja, quanto mais linear estiver o gráfico dos resíduos, melhor é o ajuste do respectivo modelo.

2.14 Resíduo de Cox-Snell

Os resíduos de Cox-Snell (1968) são determinados por $\hat{e}_i = \hat{H}\left(t_i|\boldsymbol{x_i}\right)$, em que $\hat{H}(.)$ corresponde a função de risco acumulada obtida do modelo ajustado. Para os modelos exponencial, Weibull e log-normal, os resíduo de Cox-Snell são dados respectivamente por

$$\hat{e}_i = \left[t_i exp\left\{-\boldsymbol{x_i'}\hat{\boldsymbol{\beta}}\right\}\right] \tag{55}$$

$$\hat{e}_i = \left[t_i exp\left\{-\boldsymbol{x_i'}\hat{\boldsymbol{\beta}}\right\}\right]^{\hat{\gamma}} \tag{56}$$

$$\hat{e}_i = -log\left[1 - \Phi\left(\frac{log(t_i) - \boldsymbol{x_i'}\hat{\boldsymbol{\beta}}}{\hat{\sigma}}\right)\right] \tag{57}$$

Segundo Colosimo e Giolo (2006) e Lawlees (1982), os resíduos $\hat{e}_i$ referem-se a uma população homogênea e devem seguir uma distribuição exponencial padrão caso o modelo ajustado seja adequado. O gráfico de $\hat{e}_i$ versus o $\hat{H}(\hat{e}_i)$ deve ser "aproximadamente"uma reta com inclinação 1, quando o modelo for o adequado para o ajuste.

Outro fato que ajuda na verificação do ajuste do modelo utilizado, é o gráfico das curvas de sobrevivência dos resíduos, obtidas pelo estimador de Kaplan-Meier, e pelo modelo exponencial padrão. Quantos mais próximos eles estiverem, melhor estará o ajuste do modelo.

2.15 Resíduos Padronizados

Outro resíduo que, assim como o de Cox-Snell, são estimativas de erros oriundos de um população homogênea, é o resíduo padronizado. Seu cálculo é dado por

$$\hat{v}_i = \frac{y_i - \boldsymbol{x}_{\boldsymbol{i}}^{'}\hat{\boldsymbol{\beta}}}{\hat{\sigma}}, \tag{58}$$

em que $y_i = log(t_i)$, para $i = 1, ..., n$ e $\hat{\sigma}$ é o estimador do parâmetro de escala de uma distribuição de valor extremo.

A partir da equação (58), as probabilidades de sobrevivência obtidas para os resíduos padronizados, $\hat{S}(\hat{v}_i)$ estimadas por Kaplan-Meier, devem ser "aproximadamente"uma reta.

3 MATERIAL E MÉTODO

Os dados utilizados neste trabalho, foram obtidos por uma empresa que atua no ramo petroquímico e que utiliza águas do rio Jaguari em seu processo de produção.

Para verificar a qualidade da água devolvida ao rio Jaguari, são realizadas amostras nos locais antes e após o uso pela empresa e, para efeito de comparação (controle), são coletadas amostras do rio Atibaia assim como do local onde as planárias são criadas. Os locais de onde as amostras de água são extraídas são classificados como:

Local 0 Local com condições ambientais controladas e onde são criadas as planárias;

Local 1 Local onde ocorre a captação da água do rio Jaguari para uso na empresa;

Local 3 Local onde é devolvida a água tratada ao rio Jaguari após o uso na empresa;

Local 4 Nascente do rio Atibaia;

Local 8 Jusante do rio Atibaia, localizada a 500 metros da nascente.

A água do rio Atibaia não é usada pela empresa, porém serve de parâmetro de qualidade para o tratamento da água que é usada na empresa, tratada e devolvida ao rio Jaguari.

4 RESULTADOS E DISCUSSÃO

Neste capítulo é descrito o ajuste dos tempos de vida das planárias através de modelos de sobrevivência. Como foi mencionado no capítulo anterior os dados são referentes à observações feitas no polo petroquímico na cidade de Paulínia no estado de São Paulo no mês de março de 2011, sobre o tempo de sobrevivência de planárias com variáveis tempo, censura e locais de coleta.

Os tempos de vida das planárias foram medidos em dias, fixado por um período de 30 dias e as observações realizadas de dois em dois dias, sendo considerado como censurado os tempos de vida das planárias que não obtiveram o evento de interesse (óbito) até o período fixado. Na Tabela (1) abaixo, constam os locais e o número de planárias vivas até determinado momento.

Tabela 1: Número de planárias vivas durante os 30 dias das observações

Dias	Local 0				Local 1				Local 3				Local 4				Local 8			
0	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10	10
2	10	10	10	10	10	10	10	10	9	9	9	10	10	10	10	10	10	10	10	10
4	10	10	10	10	10	10	10	10	9	8	9	7	9	10	10	9	10	10	10	10
6	10	10	10	10	10	9	10	10	9	8	9	7	9	10	10	8	10	10	10	10
8	10	10	10	10	10	9	10	10	9	8	9	7	9	10	10	8	10	10	10	10
10	10	10	9	10	10	9	10	10	9	8	9	7	9	10	10	8	10	10	10	10
12	10	10	9	10	10	9	10	10	9	7	9	6	9	9	10	8	10	10	10	10
14	10	10	9	10	10	9	10	10	9	7	8	7	9	9	10	8	9	10	10	10
16	10	10	9	10	10	9	10	10	9	7	8	7	9	9	10	8	9	10	10	10
18	10	10	9	10	10	9	10	10	9	7	8	7	9	9	9	8	9	10	10	10
20	10	10	9	10	10	8	10	8	9	7	8	6	9	9	9	8	9	10	10	10
22	10	10	8	10	10	8	8	8	9	6	8	6	9	9	9	6	9	10	10	10
24	10	10	8	10	9	8	7	9	9	6	8	6	9	9	9	6	9	10	10	10
26	10	10	8	10	9	8	7	9	9	6	8	6	9	9	9	6	9	9	9	9
28	10	10	8	10	9	8	7	9	9	6	8	6	8	8	8	6	9	9	9	9
30	10	10	8	10	9	8	7	9	9	6	8	5	8	8	8	6	9	9	9	9

Fonte: Replan-Paulínia SP.

A partir dessas informações, realizou-se a modelagem através da análise de sobrevivência e está descrita a seguir.

4.1 Número de Empates e Censura Intervalar

Como mencionado antes e também expresso na tabela anterior, as observações do tempo de vida das planárias ocorreu de dois em dois dias. Sendo assim, sabe-se o intervalo da morte da planária mas não o instante exato em que ocorreu a falha caracterizando censura intervalar.

Como as inspeções foram realizadas de dois em dois dias e o período total das observações foi de 30 dias, houve um alto número de empates. Chalita et al. (2002), sugere, levando em consideração a proporção de empates existentes, uma forma para decidir qual tipo de modelo deve ser ajustado aos dados, ou seja, usar um modelo contínuo ou discreto, com base em simulações de Monte Carlo. O cálculo é dado pela equação

$$pe = \frac{d-k}{n}, \tag{59}$$

em que pe corresponde à proporção de empates, d ao número total de falhas dos dados, k o número de falhas distintas e n ao tamanho do conjunto de dados. Se o valor de pe for menor que 20 deve ser usado modelo contínuo com aproximações para a função de verossimilhança parcial, se pe estiver entre 20 e 25, pode ser usado modelo contínuo ou um modelo discreto com aproximação para a função de verossimilhança parcial, agora se o valor de pe for maior que 25 deve ser usado um modelo discreto.

No estudo presente temos que a proporção de empates foi $pe = \frac{32-15}{200} = 0.085$, pois existem um total de 32 falhas, o número de falhas distintas é 15 e o tamanho da amostra corresponde a $n = 200$. Como pe é menor que 20, neste trabalho a abordagem foi realizada por um modelo contínuo com aproximações para a função de verossimilhança parcial.

O tempo de vida das planárias considerado para a análise foi o limite inferior do intervalo correspondente, por exemplo, na Tabela 1 do oitavo para o décimo dia na terceira amostra do local 0, houve a morte de uma planária, nesse caso o tempo considerado para a ocorrência do evento, foi o oitavo dia, ou seja, foi utilizado o menor tempo de vida. Também foram verificados os resultados de ajustes realizado com o tempo médio de quando

ocorreu a falha, assim como um ajuste do modelo semi-paramétrico de Cox. Para o tempo médio das falhas o modelo paramétrico que melhor se ajustou continuou sendo o log-normal, já o modelo semi-paramétrico de Cox não apresentou bons resultados. Os resultados de tais ajustes estão no apêndice deste trabalho.

4.2 Ajuste e comparação dos modelos para o tempo de Sobrevivência

Em uma análise inicial dos tempos de falha foi possível observar a porcentagem de observações censuradas para cada local de amostra de água. Como a planária Dugesia Tigrina possui um longo ciclo de vida (Alvarado & Newmark, 1998) e o tempo das mesmas foram acompanhadas por apenas 30 dias, sendo as observações realizadas no intervalo de dois em dois dias, era esperado que a porcentagem de dados censurados fosse alta.

Para o local 0 (criadouro) foram apresentadas apenas 2 falhas e, sendo assim, totalizado 95% de dados censurados a direita. Os locais 1 e 3, relativos ao rio Jaguari, apresentaram 82,5% e 70% de censura a direita, respectivamente. Os locais 4 e 8, relativos ao rio Atibaia, apresentaram 75% e 90% de censura a direita, respectivamente.

Com o objetivo de verificar se existe diferença entre tempos de vida das planárias da espécie Dugesia Tigrina para os diferentes locais de coleta de água, foram calculadas as estimativas não paramétricas de sobrevivência (KM), proposta por Kaplan e Meier (1958), considerando para cada nível o fator "*local*", apresentadas na Figura (7).

Os softwares usados na análise estatística para o ajuste dos modelos, foi o software R: (2013) e SAS (2011).

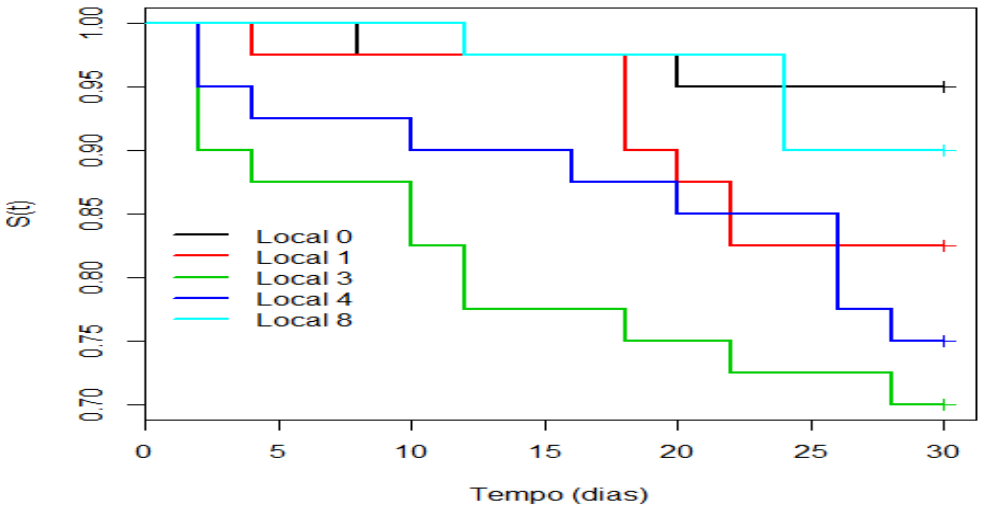

Figura 7: Curvas de sobrevivência estimada para os diferentes locais.

Após o cálculo dos tempos de sobrevivência empíricos, foi realizado o teste de *log-rank*, para verificar a existência de diferença significativa entre as curvas de sobrevivência para os diferentes locais. O $valor - p$ encontrado foi de $p = 0,014$, considerando o nível de significância como $\alpha = 0,05$, logo a hipótese nula de que não existe diferença significativa entre os níveis da covariável *local* foi rejeitada.

Como foi observado a existência de diferença significativa nos tempos de vida das planárias, considerando os diferentes locais, foi necessário realizar uma análise para detectar e agrupar as curvas que não apresentaram diferença. Com a Figura 7 é possível observar que a curva da função de sobrevivência formada pelo local 0 é similar a curva formada pelo local 8, ou seja, as planárias inseridas no local 0, controle, e as planárias inseridas no local 8, jusante do rio Atibaia, possuem o tempo de duração de vida muito próximo ao longo dos 30 dias do estudo, o mesmo ocorreu com os locais 1 e 4.

Utilizando o teste de log-rank para a confirmação com maior rigor dos comportamentos parecidos dos tempos de vida entre os locais, o valor-p encontrado quando comparado às curvas de sobrevivência dos locais 0 $versus$ 8 foi de $p = 0,423$ e o valor-p obtido quando comparado aos locais e 1 $versus$ 4 foi de $p = 0,416$. Sendo assim, podemos

concluir que os animais inseridos no local 0 e 8, assim como os animais inseridos no local 1 e 4 possuem a curva de sobrevivência com distribuição semelhante ao nível de 5% de significância e assim os mesmos podem ser agrupados.

A ideia de agrupar os locais tem como objetivo aumentar a precisão das estimativas, assim como diminuir o número de parâmetros do modelo de regressão, já que as estimativas dos mesmos serão próximas e estatisticamente iguais. Sendo assim, foram formados dois novos grupos dados por: **grupo 1** composto pelo local 0 e local 8; **grupo 2** composto pelo local 1 e local 4. Com os grupos formados, tem-se a nova variável *grupos* composta pelos níveis *grupo* 1, *grupo* 2 e *local* 3, substituindo a variável *local*.

Após o uso de técnicas não paramétricas, foram ajustados, com base na literatura, os modelos paramétricos de regressão exponencial, Weibull e log-normal. Para realizar a seleção dos modelos, foi verificado o critério de informação de Akaike (AIC) assim como o critério de informação bayesiano (BIC), que foram obtidos com o uso do *software* estatístico (SAS), em anexo, e cujos valores dos critérios são apresentados pela Tabela (2).

Tabela 2: Critério de informação, logaritmo da função de verossimilhança e resultado do TRV.

Modelo	AIC	BIC	$\log(\mathrm{L}(\hat{\theta}))$	TRV	valor-p
Gama generalizado	251,22	267,71	-120,61	-	-
Exponencial	245,07	260,96	-122,53	3,848	0,145
Weibull	245,01	266,20	-122,50	3,789	0,050
Log-Normal	242,50	263,69	-121,25	1,278	0,258

Analisando os valores da Tabela (2), vemos que o modelo log-normal apresentou o menor valor para o AIC e o modelo exponencial apresentou o menor valor para o BIC. Para realizar a comparação entre esses dois modelos, foi necessário realizar o ajuste do modelo gama generalizado que contém os modelos exponencial, log-normal e também a Weibull como modelos encaixados. Com o modelo gama generalizado ajustado, foram realizados os testes da razão de verossimilhanças (TRV) com o intuito de selecionar um de seus modelos particulares.

Com os resultados dos TRV, apresentados pela Tabela (2), é possível observar que o modelo log-normal apresenta menor diferença significativa se comparado com o modelo exponencial, ambos em relação ao modelo gama generalizado, pois para o modelo log-normal o TRV é $2[-120.61 - (-121.25)] = 1.28$.

4.3 Análise de Resíduo dos Modelos de Sobrevivência

Após utilizar os critérios de seleção de modelos AIC e BIC e o TRV e verificar que a distribuição log-normal melhor se ajusta ao conjunto de dados, é apresentado uma análise de qualidade de ajuste, afim de confirmar os resultados obtidos nos testes para a escolha do modelo.

Desta maneira foi verificado o comportamento de retas que são construídas pela função de sobrevivência estimadas pelo Kaplan-Meier e pela função de sobrevivência estimada para cada distribuição pelo estimador de máxima verossimilhança. Segundo Papa (2007), o modelo mais adequado será aquele em que a reta da função de sobrevivência estimada por cada modelo versus a função de sobrevivência estimada pelo Kaplan-Meier, mais se aproximar de reta $y = x$

Sendo assim foi realizada na Figura (8), a comparação das estimativas de sobrevivência obtidas pelo método de Kaplan-Meier e pelos modelos paramétricos Weibull, exponencial e log-normal. Assim como nos testes anteriores, novamente o modelo log-normal se apresenta como o mais indicado para representar os tempos de vida das planárias da espécie Dugesia Tigrina.

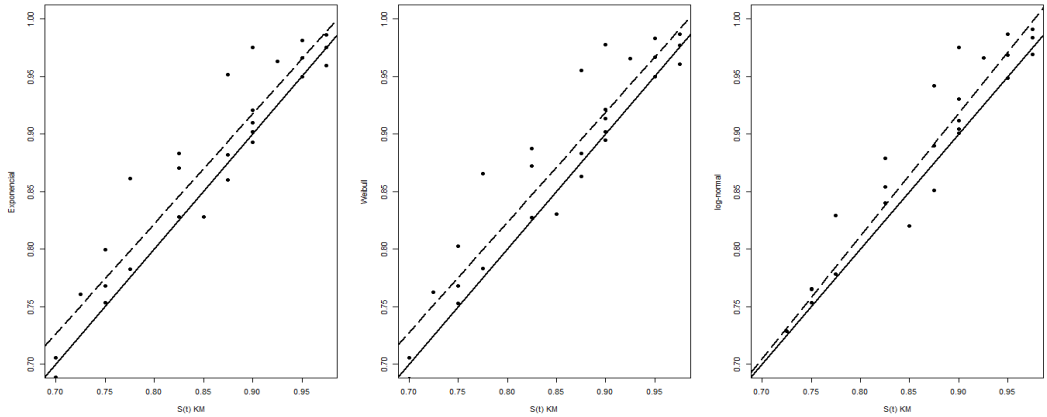

Figura 8: Gráfico da sobrevivência estimada por Kaplan-Meier versus a sobrevivência estimada pelos modelos exponencial, weibull, log-normal (retas tracejadas).

O gráfico das probabilidades de sobrevivência dos resíduos padronizado estimados por Kaplan-Meier e pelo modelo log-normal, assim como os gráficos das curvas de sobrevivência estimadas, encontram-se na Figura (9) a seguir. Vemos que o gráfico da sobrevivência dos resíduos estimada pelo método de Kaplan-Meier e pelo modelo log-normal, não apresentam um afastamento marcante de uma reta e que as respectivas curvas de sobrevivência estimadas, gráfico a direita na Figura (9), mostra um bom ajuste entre as mesmas.

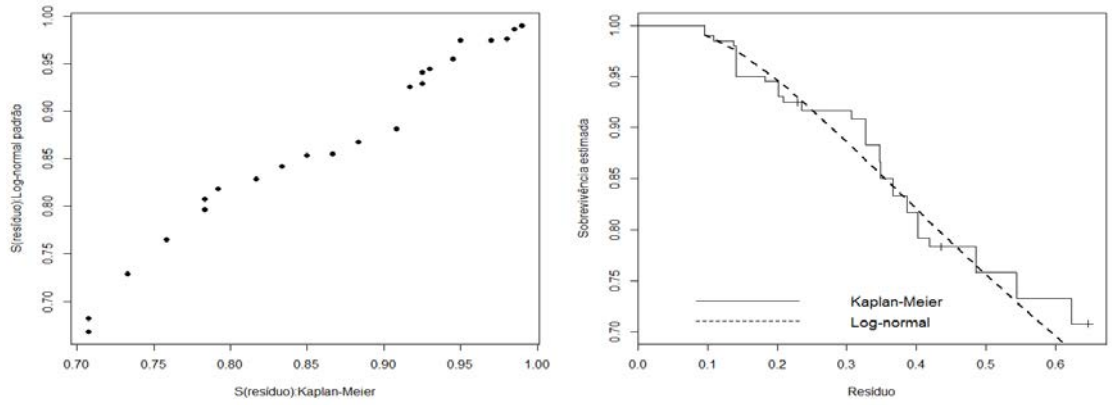

Figura 9: Sobrevivências dos resíduos estimada pelo método de Kaplan-Meier e pelo modelo log-normal (gráfico a esquerda) e respectivas curvas de sobrevivência estimadas (gráfico a direita).

Outra maneira de analisar que um modelo de regressão seja considerado adequado é que os resíduos de Cox-Snell devem seguir uma distribuição exponencial padrão, o que é verificado para o modelo log-normal na Figura (10). Percebe-se que o gráfico da sobrevivência dos resíduos de Cox-Snell estimada pelo método de Kaplan-Meier e pelo modelo log-normal também não apresenta um afastamento marcante de uma reta e as respectivas curvas de sobrevivência estimadas, gráfico a direita na Figura (10), mostra um bom ajuste entre as mesmas. Logo, também se confirma pelas análises de resíduos, que o modelo paramétrico log-normal ajusta-se bem aos conjunto de dados.

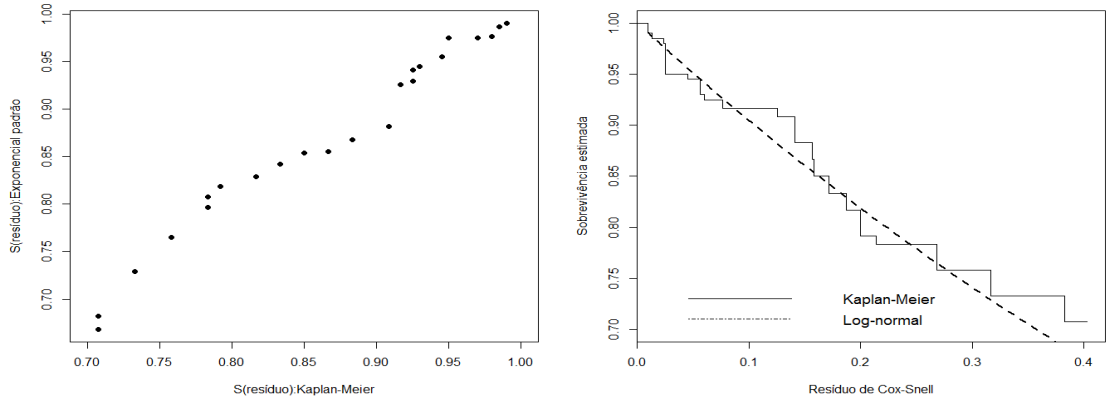

Figura 10: Sobrevivências dos resíduos de Cox-Snell estimada pelo método de Kaplan-Meier e pelo modelo log-normal (gráfico a esquerda) e respectivas curvas de sobrevivência estimadas (gráfico a direita).

Na sequência encontram-se os resíduos padronizados e de Cox-Snell para os modelos exponencial e Weibull.

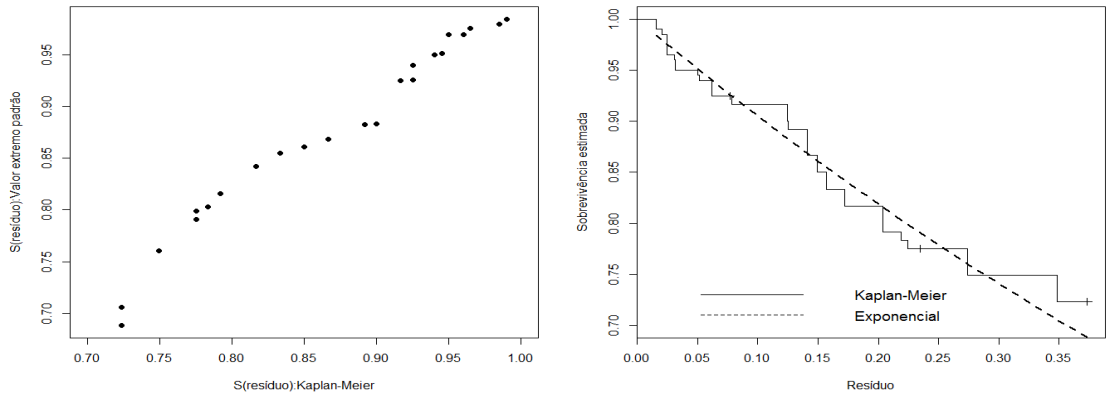

Figura 11: Sobrevivências dos resíduos padronizados estimada pelo método de Kaplan-Meier e pelo modelo exponencial (gráfico a esquerda) e respectivas curvas de sobrevivência estimadas (gráfico a direita).

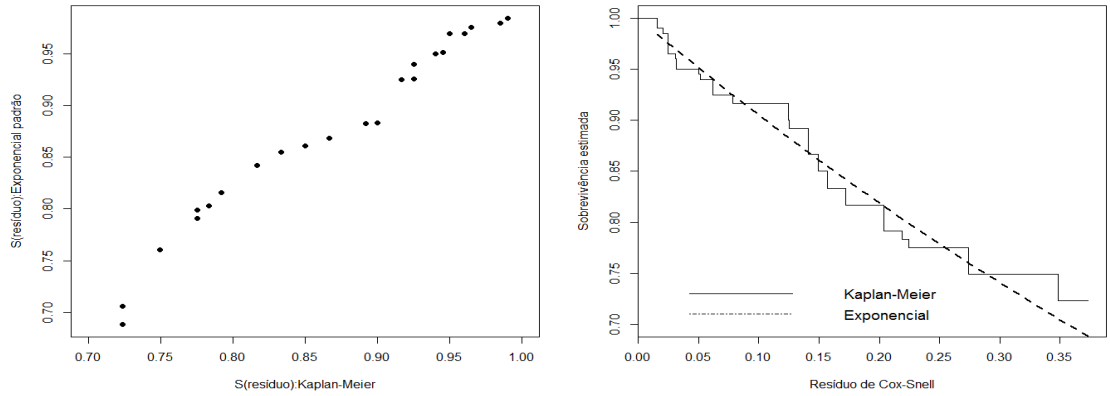

Figura 12: Sobrevivências dos resíduos de Cox-Snell estimada pelo método de Kaplan-Meier e pelo modelo exponencial (gráfico a esquerda) e respectivas curvas de sobrevivência estimadas (gráfico a direita).

Na sequência encontram-se os resíduos padronizado e de Cox-Snell para o modelo Weibull.

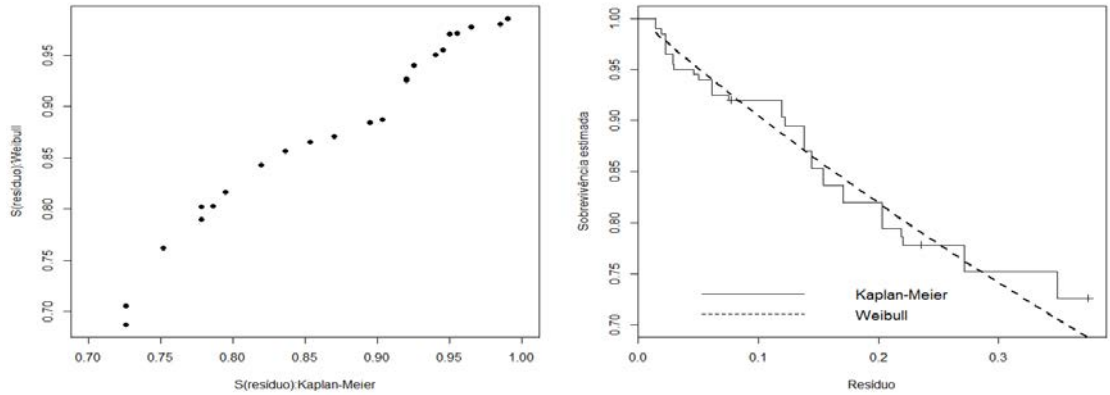

Figura 13: Sobrevivências dos resíduos padronizados estimada pelo método de Kaplan-Meier e pelo modelo Weibull (gráfico a esquerda) e respectivas curvas de sobrevivência estimadas (gráfico a direita).

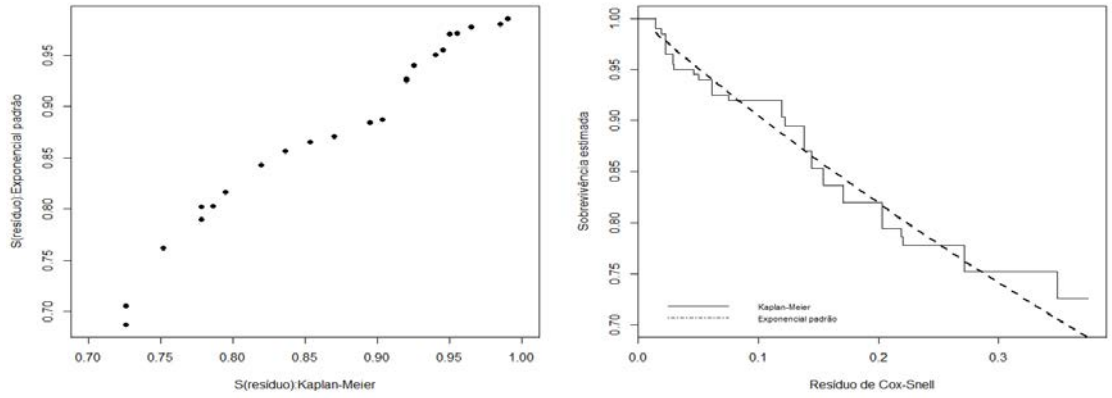

Figura 14: Sobrevivências dos resíduos de Cox-Snell estimada pelo método de Kaplan-Meier e pelo modelo Weibull (gráfico a esquerda) e respectivas curvas de sobrevivência estimadas (gráfico a direita).

Com os gráficos, vemos que os resíduos padronizados e de Cox-Snell, tanto para o modelo exponencial (Figuras (11) e (12)) e Weibull (Figuras (13) e (14)) se apresentaram adequado. Desta forma, com base nos resultados dos testes de seleção de modelos AIC, BIC e TRV, assim como na análise de resíduo, foi escolhido o modelo log-normal para o ajuste do tempo de vida das planárias.

Outro fato que levou à escolha do modelo log-normal, foi devido a além de todos os métodos e análises já indicados no trabalho, sua melhor adequação em relação a curva de sobrevivência do local 3, pois esse é o local onde a água coletada é aquela que foi usada e tratada pela empresa e que foi devolvida ao rio novamente.

4.4 Interpretação do ajuste do modelo log-normal

Após a análise de resíduos, ajustou-se o modelo log-normal considerando a covariável grupo. A função de sobrevivência do modelo de regressão log-normal é dada por

$$S(t) = \Phi\left(\frac{-log(t) + \beta_0 + \beta_1 x_1 + \beta_2 x_2}{\sigma}\right), \tag{60}$$

em que $\Phi(.)$ representa a função de distribuição acumulada de uma normal padrão e x_i representa a variável indicadora de grupo, indicada na Tabela (3). As estimativas dos

parâmetros do modelo de regressão log-normal são apresentadas pela Tabela (3) e as curvas de sobrevivência ajustadas na Figura (15).

Tabela 3: Estimativas, erro padrão e intervalo de confiança dos parâmetros do modelo log-normal.

Parâmetro	Variável explanatória	Estimativa	Erro Padrão	LI	LS
β_0	Intercepto	4,173	0,397	3,394	4,952
β_1	grupo 1 (X_1)	1,847	0,536	0,796	2,898
β_2	grupo 2 (X_2)	0,702	0,439	-0,157	1,563
σ	-	1,776	0,249	1,349	2,340

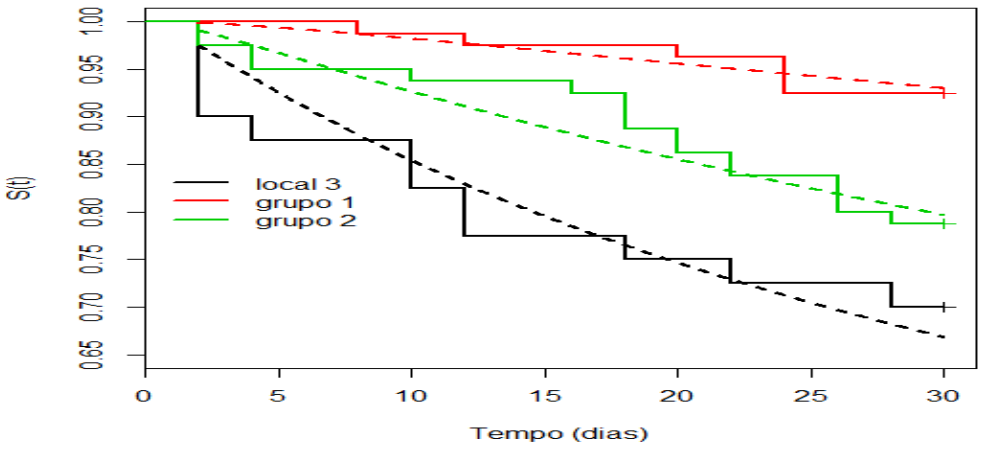

Figura 15: Curvas de sobrevivência empírica e estimada pelo modelo log-normal.

Analisando os valores das estimativas dos parâmetros do modelo log-normal apresentados na Tabela (3), observa-se que a estimativa do parâmetro β_1, referente ao grupo 1, apresenta valor positivo, ou seja, os tempos de vida das planárias que foram mantidas nas águas dos locais 0 e 8 apresentam expectativa de vida maior do que as mesmas adicionados nas águas do local 3. Já a estimativa do parâmetro β_2, referente ao grupo 2, o valor zero

pertence ao intervalo de confiança, assim a expectativa de vida das planárias que foram mantidas nas águas dos locais 1 e 4, podem apresentar expectativas de vida semelhante em relação às planárias adicionadas nas águas do local 3.

Foi realizado, com base em Colosimo e Giolo (2006), a comparação entre os tempos medianos de vida ajustados pelo modelo log-normal. Pelo resultado da Tabela 3, temos que $\hat{\beta}_1 = 1.847$ e assim a razão entre o tempo mediano com o modelo log-normal para o grupo 1 e local 3 é

$$\frac{t_p\left(x_1 = 1, x_2 = 0, \hat{\boldsymbol{\beta}}\right)}{t_p\left(x_1 = 0, x_2 = 0, \hat{\boldsymbol{\beta}}\right)} = \frac{\exp\{\hat{\sigma}\, z_p\}\exp\{\hat{\beta}_0 + \hat{\beta}_1\}}{\exp\{\hat{\sigma}\, z_p\}\exp\{\hat{\beta}_0\}} = \exp\{\hat{\beta}_1\} = 6,34. \tag{61}$$

Logo, o tempo mediano de vida das planárias inseridas na água coletada no ambiente de controle (local 0) e na jusante do rio Atibaia (local 8) é 6,34 vezes maior que o tempo de vida mediano das planárias inseridas na água utilizada pela empresa (local 3).

Observa-se, analogamente à comparação anterior, que o tempo mediano de vida das planárias inseridas nas águas dos locais 1 e 4, relativos ao grupo 2, é aproximadamente o dobro do tempo mediano de vida das mesmas inseridas nas águas do (local 3), pois

$$\frac{t_p(x_1 = 0, x_2 = 1, \hat{\boldsymbol{\beta}})}{t_p(x_1 = 0, x_2 = 0, \hat{\boldsymbol{\beta}})} = \frac{\exp\{\hat{\sigma}\, z_p\}\exp\{\hat{\beta}_0 + \hat{\beta}_2\}}{\exp\{\hat{\sigma}\, z_p\}\exp\{\hat{\beta}_0\}} = \exp\{\hat{\beta}_2\} = \exp\{0,702\} = 2,02. \tag{62}$$

Por fim, foi possível observar que o tempo mediano de vida das planárias inseridas nas águas do grupo 1 foi maior que o tempo mediano de vida do grupo 2, ou seja,

$$\frac{exp\left\{\hat{\beta}_1\right\}}{exp\left\{\hat{\beta}_2\right\}} = \frac{exp\left\{1,847\right\}}{exp\left\{0,702\right\}} = \frac{6,34}{2,02} = 3,14 \tag{63}$$

assim, as planárias colocados na água do local de controle e na jusante do rio Atibaia, tem um tempo de vida mediano 3,14 vezes maior se comparado com o tempo de vida das planárias colocados na água que adentra a empresa e na nascente do rio Atibaia.

5 CONCLUSÃO

Depois de realizadas as etapas mencionadas neste trabalho, foi possível obter algumas conclusões, a começar pela importância da estatística, neste caso a análise de sobrevivência, uma vez que torna-se mais preciso, tomar atitudes quando usa-se a estatística para fazer inferência sobre determinada situação. Por exemplo, neste trabalho as conclusões obtidas foi com base no ajuste do modelo paramétrico log-normal. Esse modelo foi comparado com o modelo exponencial e weibull através dos métodos de seleção e análise de resíduo, ou seja, a teoria estatística serve como alicerce e fundamenta as conclusões e afirmações realizadas com base nos resultados obtidos.

Sobre os dados do trabalho e com os resultados da análise realizada, verificou-se que a qualidade da água devolvida ao rio Jaguari, após o uso pela empresa, foi inferior à qualidade da mesma quando captada. Desta forma, faz-se necessário que a empresa realize melhorias nos tratamentos da água utilizada, assim como o acompanhamento do controle de qualidade dos tratamentos a serem utilizados.

Outro resultado importante obtido, foi que a empresa não deve usar como comparação (controle) as águas do rio Atibaia, pois percebe-se claramente com a Figura (7) que a nascente sofre de fatores não controlados, ou não observados, influenciando na qualidade da mesma.

Outra situação que é sabida por todos, mas que é reforçada neste trabalho, é a importância da preservação da natureza, no caso deste trabalho dos rios, pois o descaso com o cuidado dos mesmos acarreta danos muitas vezes irreversíveis ao meio ambiente e por consequência aos seres humanos, uma indicação disto é a preocupação que as empresas tem e que as levam a cuidar da qualidade da água por elas usada e que são lançadas ao meio ambiente.

APÊNDICE

Ajustes modelos paramétricos com tempo médio de falhas

A porcentagem de censura com a análise do tempo médio de vida das planárias foram as mesmas porcentagens do ajuste com o tempo a inferior. A curva de sobrevivência estimada por Kaplan-Meier segue na Figura (16) a seguir.

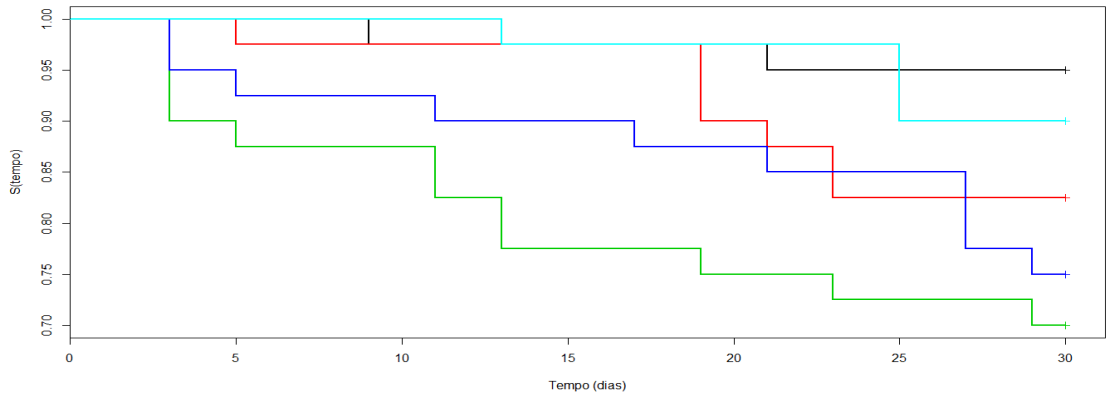

Figura 16: Curvas de sobrevivência empírica para os diferentes locais.

As retas tracejadas na Figura (17), representam as estimativas de

ajuste dos modelos exponencial, Weibull e log-normal e a contínua é uma reta $y = x$. Quanto mais próximo a reta tracejada estiver da reta contínua, melhor será o ajuste do modelo.

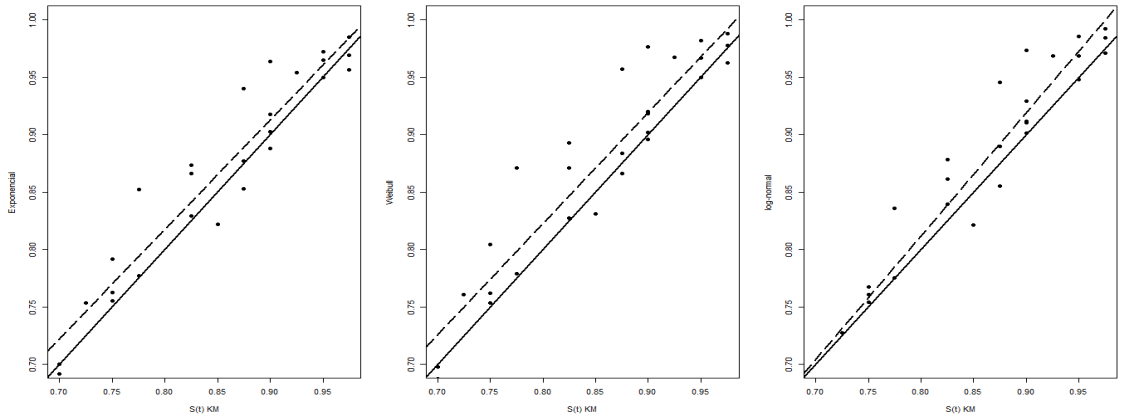

Figura 17: Gráfico da sobrevivência estimada por Kaplan-Meier (reta contínua) versus a sobrevivência estimada pelos modelos exponencial, Weibull, log-normal (retas tracejadas).

Percebe-se que o modelo log-normal se apresenta como o mais indicado para representar os tempos de vida das planárias, pois como vemos na Figura (17) a reta tracejada do modelo log-normal é a que mais se aproxima da reta contínua.

Os resíduos padronizados e de Cox-Snell para o modelo log-normal, estão representados nas Figuras (18) e (19).

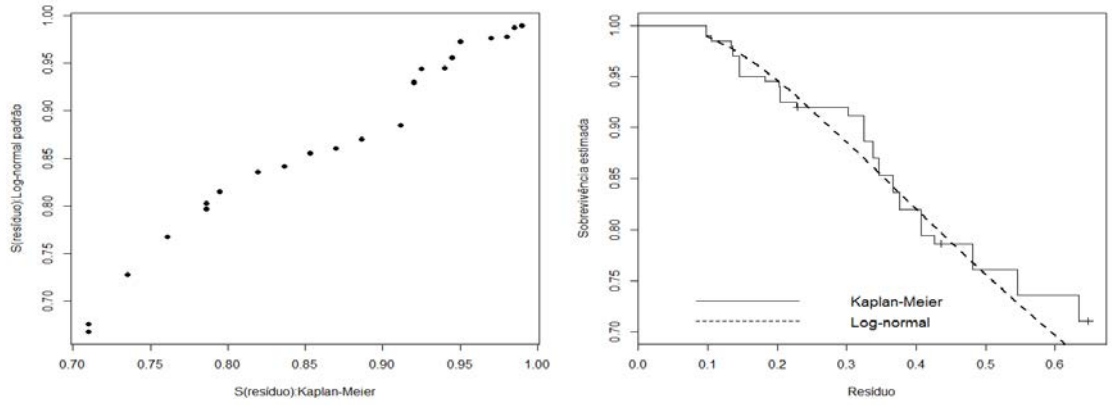

Figura 18: Sobrevivências dos resíduos estimadas pelo método de Kaplan-Meier e pelo modelo log-normal (gráfico a esquerda) e respectivas curvas de sobrevivência estimadas (gráfico a direita).

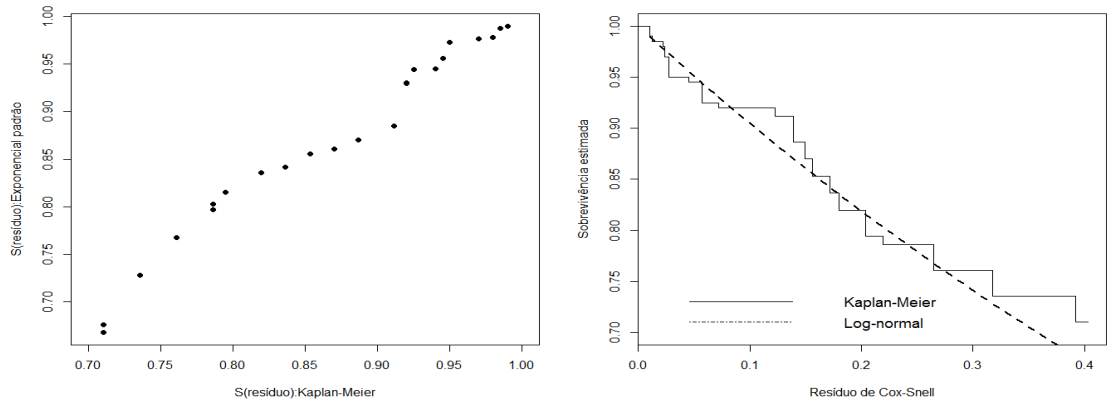

Figura 19: Sobrevivências dos resíduos de Cox-Snell estimada pelo método de Kaplan-Meier e pelo modelo log-normal (gráfico a esquerda) e respectivas curvas de sobrevivência estimadas (gráfico a direita).

Nas Figuras (20) e (21) estão os resíduos padronizados e de Cox-Snell para o modelo exponencial com tempo médio de vida das planárias em cada observação em que ocorreu a falha.

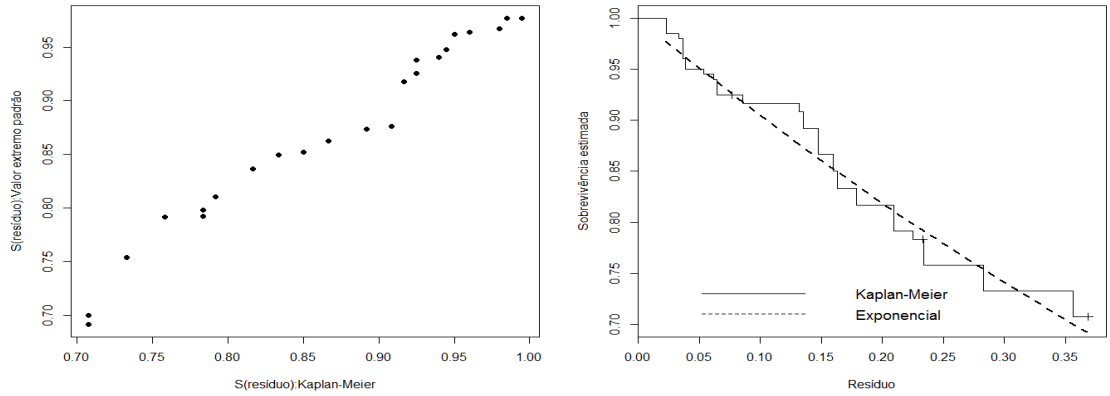

Figura 20: Sobrevivências dos resíduos estimada pelo método de Kaplan-Meier e pelo modelo exponencial (gráfico a esquerda) e respectivas curvas de sobrevivência estimadas (gráfico a direita).

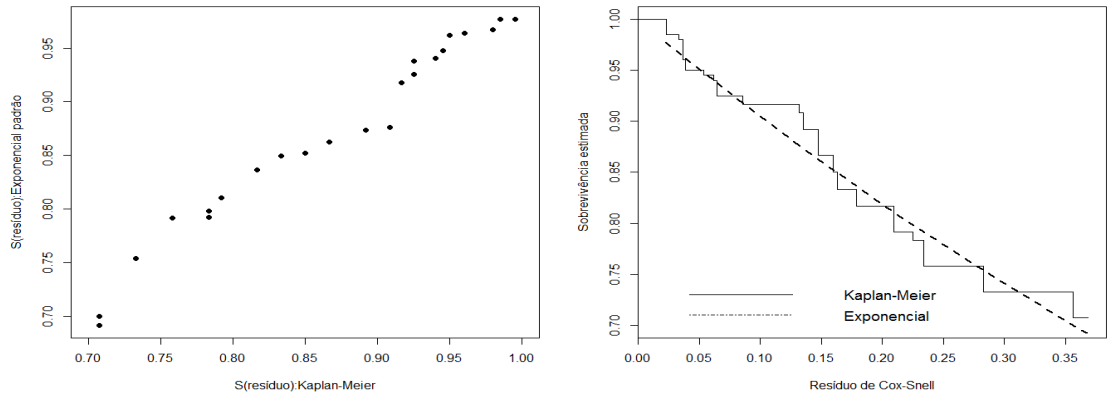

Figura 21: Sobrevivências dos resíduos de Cox-Snell estimada pelo método de Kaplan-Meier e pelo modelo exponencial (gráfico a esquerda) e respectivas curvas de sobrevivência estimadas (gráfico a direita).

Nas Figuras (22) e (23) estão os resíduos padronizados e de Cox-Snell para o modelo Weibull com tempo médio de vida das planárias em cada observação em que ocorreu a falha.

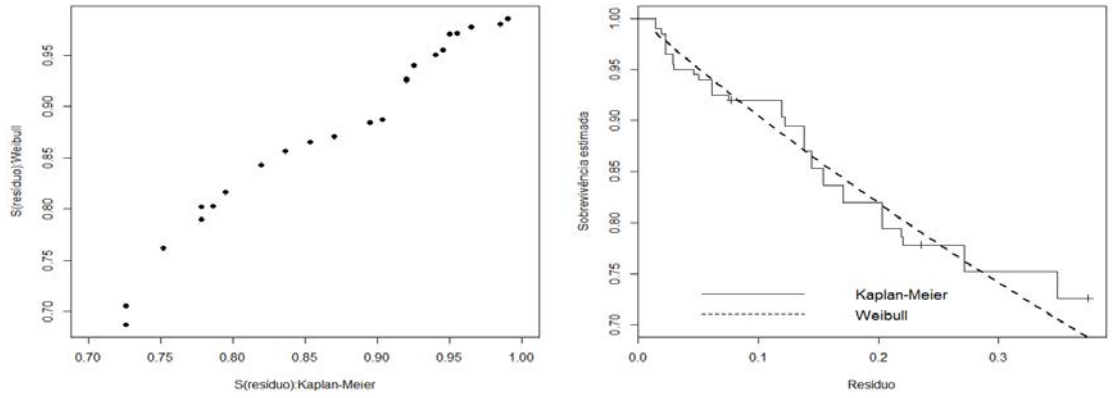

Figura 22: Sobrevivências dos resíduos estimada pelo método de Kaplan-Meier e pelo modelo Weibull (gráfico a esquerda) e respectivas curvas de sobrevivência estimadas (gráfico a direita).

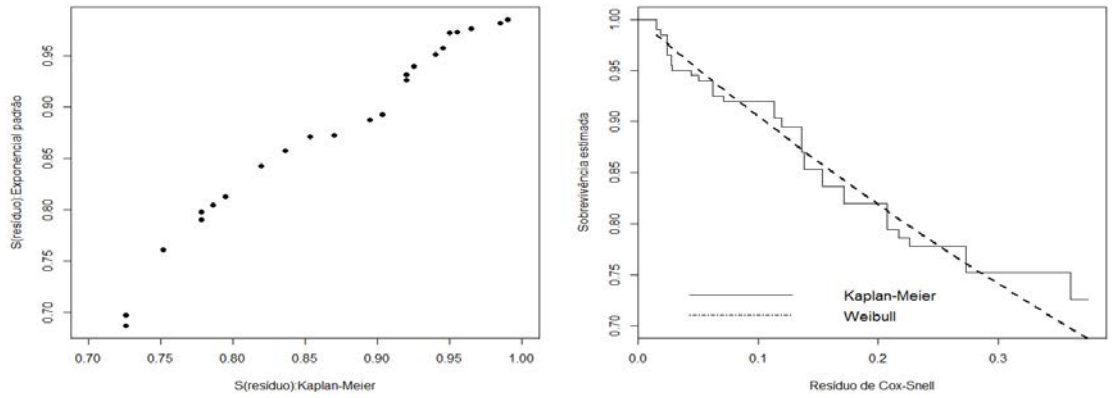

Figura 23: Sobrevivências dos resíduos de Cox-Snell estimada pelo método de Kaplan-Meier e pelo modelo Weibull (gráfico a esquerda) e respectivas curvas de sobrevivência estimadas (gráfico a direita).

Percebe-se, assim como com os tempos de vida a esquerda nas medições, tempos utilizados no trabalho, que ambos os modelos apresentam, com base na literatura, um resíduo adequado.

A Figura (24) a seguir traz a representação do modelo log-normal ajustado.

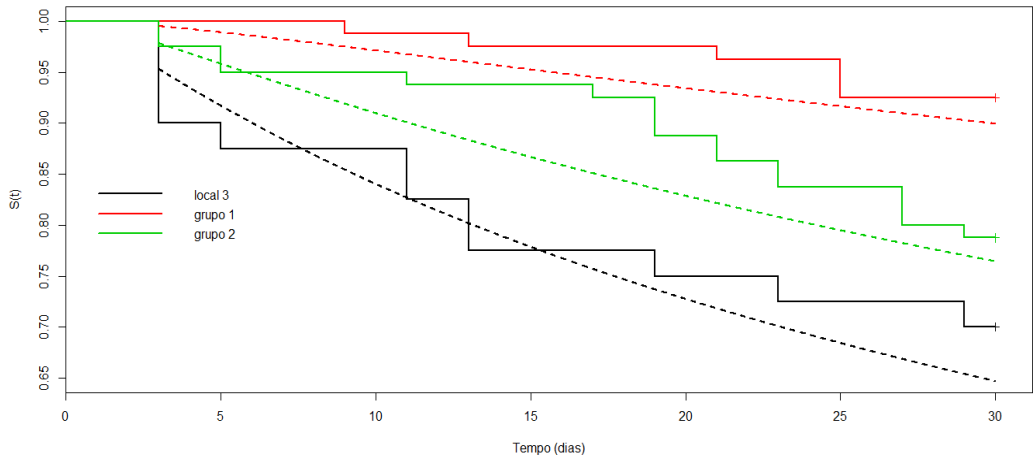

Figura 24: Curvas de sobrevivência empírica e estimada pelo modelo log-normal.

Ajuste do modelo semi-paramétrico de Cox

Assim como foi verificado o ajuste de modelos paramétricos com o tempo médio de ocorrência da falha, nesta seção foi verificada a suposição de riscos proporcionais com o ajuste do modelo semi-paramétrico de Cox, através de métodos gráficos indicados na literatura, por exemplo, por Andreozzi et al. (2011) e Colosimo & Giolo (2006).

Na Tabela (4) estão os resultados para o teste da proporcionalidade dos riscos no modelo ajustado e na Figura (25) tem-se a suposição de riscos proporcionais para as variáveis grupo 1 e grupo 2, fazendo uso do resíduo padronizado de Schoenfeld. Para que a suposição de riscos proporcionais seja adequada, segundo Colosimo

& Giolo (2006) é preciso que as estimativas do coeficiente de correlação de Pearson (ρ) sejam próximos de zero e que não haja tendências nas curvas relacionadas pelo tempo e variáveis.

Tabela 4: Teste da proporcionalidade dos riscos no modelo ajustado

Covárivel	rho (ρ)	χ^2	valor p
grupo 1	0.296	3.01	0.0828
grupo 2	0.304	3.19	0.0740
Global	-	4.30	0.1165

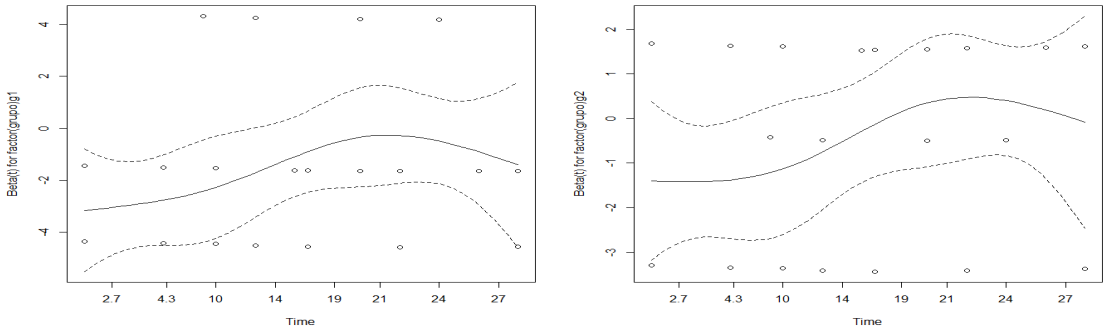

Figura 25: Suposição de riscos proporcionais para as variáveis grupo 1 e grupo 2, fazendo uso do resíduo padronizado de Schoenfeld.

Percebe-se com a Tabela (4) que os valores do coeficiente de correlação de Pearson ρ, não estão próximo ao valor zero, assim como no gráfico da Figura (25) as curvas apresentam uma certa tendência, principalmente a curva da direita que é relacionada ao grupo 2, o que implica uma possível violação da suposição da proporcionalidade nos riscos e assim o ajuste de um modelo semi-paramétrico pode não ser o mais indicado.

"Outra maneira de verificar o ajuste do modelo semi-paramétrico é analisar os resíduos martingale e deviance"(Andreozzi et al., 2011). O gráfico des-

tes resíduos precisam ter comportamentos aleatórios em torno de zero para que o ajuste do modelo semi-paramétrico seja adequado. A Figura (26) traz os resíduos martingale e deviance para o ajuste.

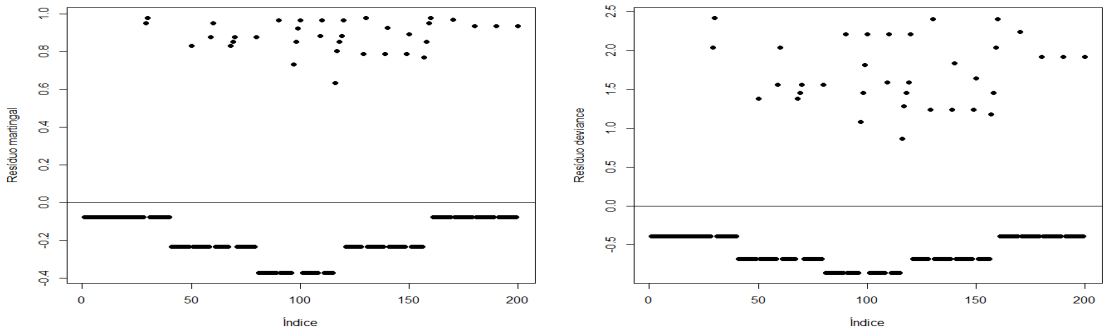

Figura 26: Resíduos martingal e deviance do modelo de Cox ajustado.

Através da Figura (26), também percebe-se que os resíduos apresentam comportamentos diferentes dos indicados pela literatura para um bom ajuste do modelo de Cox, pois percebe-se com o gráfico da Figura (26), que tais resíduos não apresentam comportamento aleatório em torno de zero.

Anexos

Comandos Software R

```
#Análise
require(survival)
dados<-read.table("plancompleta.txt",h=T)

#Adicionando uma nova variavel local como fator
dados$local2<-as.factor(dados$local)

attach(dados)

#Estimativas de kaplam e Meier
ekm<-survfit(Surv(tempo,censura)~local2)

#Verificando graficamente se existe diferença
# entre locais

dados<-read.table("plancompleta.txt",h=T)
```

```
plot(ekm,lty=c(1),lwd=2,xlab="Tempo (dias)",
ylab="S(tempo)",col=c(1:5),ylim=c(0.7,1))
legend(0,0.85,c("Local 0","Local 1","Local 3",
"Local 4","Local 8"),
       lty=c(1),col=c(1:5),bty="n",lwd=2)
dev.off()

#Verificando por teste de logrank
survdiff(Surv(tempo,censura)~local2)

###porcentagens de censura
tbl<-table(local,censura)
tbl[,1]#censura
tbl[,2]#falha
tbl[,1]+tbl[,2]#total
tbl[,1]/(tbl[,1]+tbl[,2])#% de censura

#Realizando desdobramento do teste Log-rank
length(tempo)
dados
tlocal<-list()
tlocal[[1]]<-tempo[1:40]
tlocal[[2]]<-tempo[41:80]
tlocal[[3]]<-tempo[81:120]
tlocal[[4]]<-tempo[121:160]
tlocal[[5]]<-tempo[161:200]

dados[81:160,]
cen<-list()
```

```
cen[[1]]<-censura[1:40]
cen[[2]]<-censura[41:80]
cen[[3]]<-censura[81:120]
cen[[4]]<-censura[121:160]
cen[[5]]<-censura[161:200]

library(survival)
t66<-list()
t666<-list()
hipot66<-matrix(0,ncol=5,nrow=5)
for(i in 1:5){
  for (j in 1:5){
    t66[[j]]<-c(tlocal[[j]], tlocal[[i]])
    c66<-c(cen[[j]], cen[[i]])
    oi<-Surv(t66[[j]],c66)
    trr<-c(rep('local1',length(tlocal[[i]])),
           rep('local2',length(tlocal[[j]])) )
    chiquad<-survdiff(oi ~ trr)
    hipot66[i,j]<-round(pchisq(q=chiquad$chisq,
    df=1,lower.tail=F),3)
  }
  t666[[i]]<-t66
}

t666 # combinações dos tempos de cada par de local
hipot66 # Matriz de teste de hipótese! testes 1 a 1
igual survdiff
hipot66[upper.tri(hipot66)]<- NA
```

```
# Agrupando
(dados2<-dados)
dados2$grupo<-(dados2$local)
ekm<-survfit(Surv(tempo,censura)~grupo, data=dados2)
plot(ekm,lty=c(1),xlab="Tempo (dias)",ylab="S(t)",
col=c(1:5),
ylim=c(0.7,1),conf.int=FALSE)

dados2$grupo[dados2$grupo=="1" | dados2$grupo=="4"]<-"g2"
ekm<-survfit(Surv(tempo,censura)~grupo, data=dados2)
plot(ekm,lty=c(1:5),xlab="Tempo (dias)",ylab="S(t)",
col=c(1:5),
ylim=c(0.7,1),conf.int=FALSE)

dados2$grupo[dados2$grupo=="0" | dados2$grupo=="8"]<-"g1"
ekm<-survfit(Surv(tempo,censura)~grupo, data=dados2)
plot(ekm,lty=c(1:5),xlab="Tempo (dias)",ylab="S(t)",
col=c(1:5),
ylim=c(0.7,1),conf.int=FALSE)
```

Ajuste Modelo exponencial, Weibull e log-normal

```
names(survreg.distributions) #distribuições

#Exponencial
exp.fit <-survreg(Surv(tempo, censura)~1,
```

```
dist='exponential', data=dados)
(scal<-exp.fit$scale)
(b0<-exp.fit$coef[1])

#Weibull
weibull.fit <-survreg(Surv(tempo, censura)~1,
dist='weibull', data=dados)
(scal2<-weibull.fit$scale)
(b02<-weibull.fit$coef[1])

#Log-normal
lnorm.fit <-survreg(Surv(tempo, censura)~1,
dist='lognormal', data=dados)
(scal3<-lnorm.fit$scale)
(b03<-lnorm.fit$coef[1])

ekm<-survfit(Surv(tempo,censura)~1)
plot(ekm,lty=c(1:5),xlab="Tempo
(dias)",ylab="S(t)",col=c(1:5),
ylim=c(0.75,1),conf.int=FALSE)

aa<-seq(min(tempo),max(tempo)-0.01,0.01)
lines(aa,1-psurvreg(aa,(b0),scal,dist="exponential"),
col=1,lwd=1,lty=1)
lines(aa,1-psurvreg(aa,(b02),scal2,dist="weibull"),
col=2,lwd=1,lty=1)
lines(aa,1-psurvreg(aa,(b03),scal3,dist="lognormal"),
col=3,lwd=1,lty=1)
```

```
legend(5,0.9,c("exp","weibull","lognormal"),
lty=c(1),col=c(1,2,3),seg.len=2,bty="n",lwd=2)
```

Ajuste dos modelo para os locais e teste razão de verossimilhança

```
# Distribuição exponencial
aa<-seq(min(tempo),max(tempo)-0.01,0.01)
exp.fit <-survreg(Surv(tempo, censura)~grupo,
dist='exponential', data=dados2)
summary(exp.fit)
(scal<-exp.fit$scale)
exp.fit$coef
(b0<-exp.fit$coef[1])
(g1<-exp.fit$coef[2])
(g2<-exp.fit$coef[3])

ekm<-survfit(Surv(tempo,censura)~grupo,data=dados2)
plot(ekm,lty=c(1),xlab="Tempo (dias)",ylab="S(t)",
col=c(1:3),ylim=c(0.7,1),lwd=2)
legend(0,0.85,c("local 3","grupo 1","grupo 2"),
       lty=c(1),col=c(1:3),seg.len=2,bty="n",lwd=2)

lines(aa,1-psurvreg(aa,(b0),scal,dist="exponential"),
col=1,lwd=2,lty=1)
```

```
lines(aa,1-psurvreg(aa,(b0+g1),scal,dist="exponential"),
col=2,lwd=2,lty=1)
lines(aa,1-psurvreg(aa,(b0+g2),scal,dist="exponential"),
col=3,lwd=2,lty=1)

# Distribuição weibull
weibull.fit <-survreg(Surv(tempo, censura)~grupo,
dist='weibull', data=dados2)
summary(weibull.fit)
(scal<-weibull.fit$scale)
(b0<-weibull.fit$coef[1])
(g1<-weibull.fit$coef[2])
(g2<-weibull.fit$coef[3])

#ajuste Weibull e Kaplan-Meier
lines(aa,1-psurvreg(aa,(b0),scal,dist="weibull"),
col=1,lwd=2,lty=3)
lines(aa,1-psurvreg(aa,(b0+g1),scal,dist="weibull"),
col=2,lwd=2,lty=3)
lines(aa,1-psurvreg(aa,(b0+g2),scal,dist="weibull"),
col=3,lwd=2,lty=3)

# Distribuição log-normal
lnorm.fit <-survreg(Surv(tempo, censura)~grupo,
dist='lognormal', data=dados2)
summary(lnorm.fit)
(scal<-1.776)
lnorm.fit$coef[1]
lnorm.fit$coef[2]
```

```
lnorm.fit$coef[3]
lines(aa,1-psurvreg(aa,(b0),scal,dist="lognormal"),
col=1,lwd=2,lty=2)
lines(aa,1-psurvreg(aa,(b0+g1),scal,dist="lognormal"),
col=2,lwd=2,lty=2)
lines(aa,1-psurvreg(aa,(b0+g2),scal,dist="lognormal"),
col=3,lwd=2,lty=2)

legend(10,0.75,c("exp","lognorm","Weibull"),
       lty=c(1,2,3),col=c(1),seg.len=2,bty="n",lwd=2)

dev.off()

# Teste da razão de verossimilhaças

l1<- -122.5368
l2<- -122.5070
l3<- -121.2517
l4<- -120.6124

(trv1<-2*(-l1+l4)) # exp      x  Gama G
(trv2<-2*(-l2+l4)) # weibull  x  Gama G
(trv3<-2*(-l3+l4)) # lognorm  x  Gama G

pchisq(trv1,2,lower.tail=FALSE)
pchisq(trv2,1,lower.tail=FALSE)
pchisq(trv3,1,lower.tail=FALSE)

#Ajuste final distribuição log-normal
```

```
ekm<-survfit(Surv(tempo,censura)~grupo,data=dados2)
plot(ekm,lty=c(1),xlab="Tempo (dias)",ylab="S(t)",
col=c(1:3),
ylim=c(0.65,1),lwd=2)
legend(0,0.85,c("local 3","grupo 1","grupo 2"),
       lty=c(1),col=c(1:3),seg.len=2,bty="n",lwd=2)
lines(aa,1-psurvreg(aa,(b0),scal,dist="lognormal"),
col=1,lwd=2,lty=2)
lines(aa,1-psurvreg(aa,(b0+g1),scal,dist="lognormal"),
col=2,lwd=2,lty=2)
lines(aa,1-psurvreg(aa,(b0+g2),scal,dist="lognormal"),
col=3,lwd=2,lty=2)

dev.off()
```

Análise de Resíduos

```
ekm<-survfit(Surv(tempo,censura)~local2)
ekm$strata
time<-ekm$time
st<-ekm$surv #estimativas KM

# Distribuição exponencial
exp.fit <-survreg(Surv(tempo, censura)~local2,
```

```
dist='exponential', data=dados)
(scal<-exp.fit$scale)
(b0<-exp.fit$coef[1])
(lc1<-exp.fit$coef[2])
(lc3<-exp.fit$coef[3])
(lc4<-exp.fit$coef[4])
(lc8<-exp.fit$coef[5])

te<-NULL
ste<-NULL
ekm$strata # tamanho de cada grupo
(te<-1-psurvreg(ekm$time[1:3],(b0),scal,
dist="exponential"))
(ste<-c(ste,te))
te<-1-psurvreg(ekm$time[4:8],(b0+lc1),scal,
dist="exponential")
ste<-c(ste,te)
te<-1-psurvreg(ekm$time[9:16],(b0+lc3),scal,
dist="exponential")
ste<-c(ste,te)
te<-1-psurvreg(ekm$time[17:24],(b0+lc4),scal,
dist="exponential")
ste<-c(ste,te)
te<-1-psurvreg(ekm$time[25:27],(b0+lc8),scal,
dist="exponential")
(ste<-c(ste,te))

#Distribuição Weibull
weibull.fit <-survreg(Surv(tempo, censura)~local2,
```

```
dist='weibull', data=dados)
summary(weibull.fit)
(scal<-weibull.fit$scale)
(b0<-weibull.fit$coef[1])
(lc1<-weibull.fit$coef[2])
(lc3<-weibull.fit$coef[3])
(lc4<-weibull.fit$coef[4])
(lc8<-weibull.fit$coef[5])

te<-NULL
stw<-NULL
(te<-1-psurvreg(ekm$time[1:3],(b0),scal,
dist="weibull"))
(stw<-c(stw,te))
te<-1-psurvreg(ekm$time[4:8],(b0+lc1),scal,
dist="weibull")
stw<-c(stw,te)
te<-1-psurvreg(ekm$time[9:16],(b0+lc3),scal,
dist="weibull")
stw<-c(stw,te)
te<-1-psurvreg(ekm$time[17:24],(b0+lc4),scal,
dist="weibull")
stw<-c(stw,te)
te<-1-psurvreg(ekm$time[25:27],(b0+lc8),scal,
dist="weibull")
(stw<-c(stw,te))

lnorm.fit <-survreg(Surv(tempo, censura)~local2,
dist='lognormal', data=dados)
```

```
(scal<-lnorm.fit$scale)
(b0<-lnorm.fit$coef[1])
(lc1<-lnorm.fit$coef[2])
(lc3<-lnorm.fit$coef[3])
(lc4<-lnorm.fit$coef[4])
(lc8<-lnorm.fit$coef[5])

te<-NULL
stl<-NULL
(te<-1-psurvreg(ekm$time[1:3],(b0),scal,
dist="lognormal"))
(stl<-c(stl,te))
te<-1-psurvreg(ekm$time[4:8],(b0+lc1),scal,
dist="lognormal")
stl<-c(stl,te)
te<-1-psurvreg(ekm$time[9:16],(b0+lc3),scal,
dist="lognormal")
stl<-c(stl,te)
te<-1-psurvreg(ekm$time[17:24],(b0+lc4),scal,
dist="lognormal")
stl<-c(stl,te)
te<-1-psurvreg(ekm$time[25:27],(b0+lc8),scal,
dist="lognormal")
(stl<-c(stl,te))

library(MASS)

par(mfrow=c(1,3))
plot(st,ste,xlab="S(t) KM",ylab="Exponencial",
```

```
pch=16,ylim=c(0.7,1))
r1<-lm(ste~st) # reta ajustada
lines(c(0,1),c(0,1),type="l",lwd=2)
abline(r1,col="black",lwd=2,lty=5)

plot(st,stw,xlab="S(t) KM",ylab="Weibull",pch=16,
ylim=c(0.7,1))
r2<-lm(stw~st)
lines(c(0,1),c(0,1),type="l",lwd=2)
abline(r2,col="black",lwd=2,lty=5)

plot(st,stl,xlab="S(t) KM",ylab="log-normal",pch=16,
ylim=c(0.7,1))
r3<-lm(stl~st)
lines(c(0,1),c(0,1),type="l",lwd=2)
abline(r3,col="black",lwd=2,lty=5)

dev.off()

##Resíduos padronizados para lognormal

xb<-lnorm.fit$coefficients[1]+lnorm.fit$coefficients[2]*v1+
lnorm.fit$coefficients[3]*v2
sigma<-lnorm.fit$scale
res<-(log(tempo)-(xb))/sigma#resíduo padronizado
resid<-exp(res)
ekm<-survfit(Surv(resid,censura)~1)
resid<-ekm$time
sln<-pnorm(-log(resid))
```

```
par(mfrow=c(1,2))
plot(ekm$surv,sln,xlab="S(resíduo):Kaplan-Meier",
ylab="S(resíduo):
Log-normal padr?o",pch=16)
plot(ekm,conf.int=F,ylim=c(0.7,1), xlab="Resíduo",
ylab="Sobrevivência estimada",pch=16)
lines(resid,sln,lwd=2,lty=2)
legend(0.01,0.75,lty=c(1,2),c("Kaplan-Meier","Log-normal"),
cex=0.9,bty="n")

##resíduo Cox-snell  para log normal#

ei<- -log(1-pnorm(res))#resíduo de Cox-snell
ekm1<-survfit(Surv(ei,censura)~1)
t<-ekm1$time
st<-ekm1$surv
sexp<-exp(-t)
par(mfrow=c(1,2))
plot(st,sexp,xlab="S(resíduo):Kaplan-Meier",xlim=c(0.7,1),
ylab="S(resíduo):Exponencial padrão",pch=16)
plot(ekm1,conf.int=F,mark.time=F,ylim=c(0.7,1),
xlab="Resíduo de Cox-Snell",
ylab="Sobrevivência estimada")
lines(t,sexp,lwd=2,lty=2)
legend(0.00,0.73,lty=c(1,4),c("Kaplan-Meier","Log-normal"),
cex=0.9,bty="n")

##resíduos padronizados para exponencial#
```

```
xbexp<-exp.fit$coefficients[1]+exp.fit$coefficients[2]*v1+
exp.fit$coefficients[3]*v2
sigma1<-exp.fit$scale
resexp<-(log(tempo)-(xbexp))/sigma1
residexp<-exp(resexp)
ekmexp<-survfit(Surv(residexp,censura)~1)
residexp<-ekmexp$time
slnexp<-exp(-(residexp))
par(mfrow=c(1,2))
plot(ekmexp$surv,slnexp,xlab="S(resíduo):Kaplan-Meier",
xlim=c(0.7,1),ylab="S(resíduo):Valor extremo padrão",pch=16)
plot(ekmexp,conf.int=F,ylim=c(0.7,1), xlab="Resíduo",
ylab="Sobrevivência estimada",pch=16)
lines(residexp,slnexp,lwd=2,lty=2)
legend(0.01,0.75,lty=c(1,2),c("Kaplan-Meier","Exponencial"),
cex=1.2,bty="n")

##resíduo Cox-snell  para exponencial

xb1<-exp.fit$coefficients[1]+exp.fit$coefficients[2]*v1+
exp.fit$coefficients[3]*v2
ei1<- tempo*exp(-(xb1))#resíduo de Cox-snell
ekm2<-survfit(Surv(ei1,censura)~1)
t2<-ekm2$time
st2<-ekm2$surv
sexp1<-exp(-t2)
par(mfrow=c(1,2))
plot(st2,sexp1,xlab="S(resíduo):Kaplan-Meier",
xlim=c(0.7,1),
```

```
ylab="S(resíduo):Exponencial padrão",pch=16)
plot(ekm2,conf.int=F,mark.time=F,ylim=c(0.7,1),
xlab="Resíduo de Cox-Snell",
ylab="Sobrevivência estimada")
lines(t2,sexp1,lwd=2,lty=2)
legend(0.00,0.75,lty=c(1,4),c("Kaplan-Meier","Exponencial"),
cex=1.2,bty="n")

##resíduos padronizados para weibull#

xbw<-weibull.fit$coefficients[1]+
weibull.fit$coefficients[2]*v1+
weibull.fit$coefficients[3]*v2
sigmaw<-weibull.fit$scale
resw<-(log(tempo)-(xbw))/sigmaw
residw<-exp(resw)
ekmw<-survfit(Surv(residw,censura)~1)
residw<-ekmw$time
slnw<-exp(-(residw))
par(mfrow=c(1,2))
plot(ekmw$surv,slnw,xlab="S(resíduo):Kaplan-Meier",xlim=c(0.7,1),
ylab="S(resíduo):Weibull",pch=16)
plot(ekmw,conf.int=F,ylim=c(0.7,1), xlab="Resíduo",
ylab="Sobrevivência estimada",pch=16)
lines(residw,slnw,lwd=2,lty=2)
legend(0.01,0.75,lty=c(1,2),c("Kaplan-Meier","Weibull"),
cex=1.2,bty="n")

##resíduo Cox-snell  para weibull#
```

```
xb2<-weibull.fit$coefficients[1]+
weibull.fit$coefficients[2]*v1+
weibull.fit$coefficients[3]*v2
ei2<- ((tempo*exp(-(xb2)))^(1/weibull.fit$scale))
ekm3<-survfit(Surv(ei2,censura)~1)
t3<-ekm3$time
st3<-ekm3$surv
sexp2<-exp(-t3)
par(mfrow=c(1,2))
plot(st3,sexp2,xlab="S(resíduo):Kaplan-Meier",xlim=c(0.7,1),
ylab="S(resíduo):Exponencial padrão",pch=16)
plot(ekm3,conf.int=F,mark.time=F,ylim=c(0.7,1),xlab="Resíduo de Cox-Snell",
ylab="Sobrevivência estimada")
lines(t3,sexp2,lwd=2,lty=2)
legend(0.00,0.75,lty=c(1,4),c("Kaplan-Meier","Weibull"),cex=1.2,bty="n")
```

Teste Razão Verossimilhança no SAS

```
data dados1;
input local$ tempo  censura;
y=log(tempo);
if local=0 then grupo=1;
if local=8 then grupo=1;
if local=1 then grupo=2;
if local=4 then grupo=2;
```

```
 if local=3 then grupo=3;

proc lifereg data = dados1; /*exponential */
class grupo;
 model tempo*censura(0) = grupo / dist=exponential;
run;

proc lifereg data = dados1; /*weibull */
class grupo;
 model tempo*censura(0) = grupo / dist=weibull;
run;


proc lifereg data = dados1; /*lognormal */
class grupo;
 model tempo*censura(0) = grupo / dist=lognorm;
run;

proc lifereg data = dados1; /*Gamma */
class grupo;
 model tempo*censura(0) = grupo / dist=gamma MAXITER=550;

run;
```

Modelo de Cox dados agrupados

```
m<-coxph(Surv(tempo,censura)~factor(grupo),data=dados2)
summary(m)

#ANOVA da o TRV no modelo semiparamétrico
anova(m)

#modelo saturado
mod.saturado<-coxph(Surv(tempo,censura)~factor(id),data=dados2)
mod.saturado
summary(mod.saturado)

#resíduo de Schoenfeld
res.sch<-cox.zph(m)
plot(res.sch)

dev.off()

res.sch<-cox.zph(mod.saturado)
plot(res.sch)

#resíduo scaledsch (sup risco proporcional para g1 e g2 pelo
# res de schoenfeld)

resid(m,type="scaledsch")
cox.zph(m,transform="identity")
par(mfrow=c(2,3))
```

```
plot(cox.zph(m))

#res martingal e deviance
par(mfrow=c(1,2))
rm<-resid(m,type="martingale")#res martingale
rd<-resid(m,type="deviance") #res deviance
pl<-(m$linear.predictors)
plot(rm,xlab="índice",ylab="Resíduo martingal",pch=16)
abline(h=0)
plot(rd,xlab="índice",ylab="Resíduo deviance",pch=16)
abline(h=0)

m.sch<-cox.zph(m)
m.sch

dev.off()
```

REFERÊNCIAS BIBLIOGRÁFICAS

AKAIKE, H. A new look at the statistical model identification. **IEEE Transactions on Automatic Control, Boston**, v.19, n.6, p.716–723, 1974.

ALVARADO, A. S.; NEWMARK, P. **The use of planarians to dissect the molecular basis of metazoan regeneration**. Rio Grande do Sul: Wound Repair Regen, 1998. 413p.

ANDRÉ, C. M. G.; REGAZZI, A. J. Critérios para seleção de modelos baseados na razão de verossimilhança. UFV, 2014.

ANDREOZZI, V. L.; BARBOSA, M. T. S.; CAMPOS, D. P.; CODEÇO, C. T.; SHIMAKURA, S. E.; CARVALHO, M. S. **ANÁLISE DE SOBREVIVÊNCIA Teoria e aplicações em saúde**. Rio de Janeiro: Editora Fiocruz, 2011. 434p.

BARROS, R. T. V.; CHERNICHARO, C. A. L.; HELLER, L.; SPERLING, M. V. **Manual de Saneamento e Proteção Ambiental para apoio aos Municípios**. Belo Horizonte: Escola de Engenharia da UFMG, 1995. 221p.

BOZDOGAN, H. **Model selection and Akaike's Information Criterion (AIC): the general theory and its analytical extensions**. New York: Psychometrika, 1987. 345p.

BRESLOW, N.; CROWLEY, J. A large Sample Study of the Life Table and Product Limit Estimates Under Random Censorship. **Annals of Statistics**, , n.2, p.437–453, 1974.

BRITO, A. C.; SOUZA, M. L. Proposta de Melhoria de um Método de Estimação da Taxa de Falhas em Interconexões de Segundo Nível de Componentes Eletrônicos. **INPE**, , n.2, p.1–15, 2010.

BRUSCA, R. C.; BRUSCA, G. J. **Invertebrados**. Rio de Janeiro: Guanabara Koogan, 2007. 968p.

BURNHAN, K. P.; ANDERSON, D. R. Multimodel inference: understanding aic and bic in model selection. **Sociological Methods and Research. Beverly Hills**, v.33, n.2, p.261–304, 2004.

BUUREN, V. S.; MIRANDA, F. Worm plot: a simple diagnostic device for modelling growth reference curves. **STATISTICS IN MEDICINE**, v.20, p.1259–1277, 2001.

CHALITA, L. V. A. S.; COLOSIMO, E. A.; DEMÉTRIO, C. B. G. Likelihood Approximations and Discrete Models for Tied Survival Data, Communications in Statistics - Theory and Methods. **STATISTICS IN MEDICINE**, v.31, p.1215–1229, 2002.

COLLET, E. **Modelling Survival Data in Medical Research**. London: Chapman and Hall, 1994.

COLOSIMO, E. A. **Análise de Sobrevivência Aplicada**. São Paulo: Egard Blucher LTDA, 2001. 145p.

COLOSIMO, E. A.; GIOLO, S. R. **Análise de Sobrevivência Aplicada**. São Paulo: Egard Blucher LTDA, 2006. 317p.

COX, D. R. Regression Models and Life-Tables. **Journal of the Royal Statistical Society. Series B (Methodological)**, v.34, n.2, p.187–220, 1972.

COX, D. R. Partial Likelihood. **Biometrika**, v.62, n.2, p.269–276, 1975.

COX, D. R.; HINKLEY, D. V. Theoretical Statistic. **Chapman and Hall, London**, 1974.

GAVA, A. J. **Princípios de Tecnologia de Alimentos**. São Paulo: Editora Nobel, 1984. 248p.

GUJARATI, D. **Econometria Básica**. Rio de Janeiro: Seção 3, ed.Elsevier, 2006. 791p.

JUNIOR, P. L.; SILVEIRA, F. L.; OSTERMANN, F. Análise de sobrevivência aplicada ao estudo do fluxo escolar nos cursos de graduação em física: um exemplo de uma universidade brasileira. **Arcos Design 6**, v.34, n.1, p.1403;1–1403;10, 2011.

KAPLAN, E. L.; MEIER, P. Nonparametric estimation from incomplete observations. **Journal of the American Statistical Association**, p.53, 1958.

KLEIN, M.; KLEINBAUM, D. G. **Survival Analysis: A Self-Learning**. New York: Text. 2. ed. New York Springer, 2005. 590p.

LAU, A. H. Testes de Genotoxicidade em Planárias: Análise de Aberrações Cromossômicas e Teste Cometa. Porto Alegre, 1998. 83p. Dissertação (Mestrado) - Universidade Federal do Rio Grande do Sul.

LAU, A. H. Avaliação Múltipla do Potencial Genotóxico da Poluição Urbana de Porto Alegre - RS. Porto Alegre, 2002. 118p. Tese (Doutorado) - Universidade Federal do Rio Grande do Sul.

LAWLEES, J. F. **Statistical Models and Methods for Lifetime**. New York: John Wiley and Sons, 1982.

LAWLEES, J. F. **Statistical Models and Methods for Lifetime**. New York: John Wiley and Sons, 2003.

LORA, E. E. S. **Prevenção e controle da poluição nos setores energético, industrial e de transporte**. Rio de Janeiro: Interciência, 2002. 221p.

MANTOVANI, A.; FRANCO, M. A. P. Estudo da Distribuição Assintótica dos Estimadores dos Parâmetros da Distribuição Weibull na Presença de Dados Sujeitos

a Censura Aleatória. **Revista Matemática e Estatística, São Paulo**, v.22, n.3, p.7–20, 2004.

MOORE, J. **Uma Introdução aos Invertebrados**. São Paulo: Santos, 2003. 356p.

MOTA, T. S.; SILVEIRA, L. V. A.; ANTUNES, A. A. Modelo de Cox para eventos cardiovasculares recorrentes em pacientes sob diálise com covariáveis medidas no tempo. **Revista Brasileira de Biometria, São Paulo**, v.30, n.1, p.150–159, 2012.

NELSON, W. B. **Accelerated Testing: Statistical Models, Test Plans, and Data Analyses**. New York: John Wiley Sons, 1990.

PAPA, M. C. O. Estudo do Efeito das Incertezas na Variável de Estresse em Ensaios Acelerados. Piracicaba, 2007. 130p. Dissertação (Mestrado) - Universidade Metodista de Piracicaba.

PARREIRA, D. R. M. Um Modelo de Risco Proporcional Dependente do Tempo. São Carlos, 2007. 57p. Tese (Doutorado) - Universidade Federal de São Carlos.

PASCOA, A. R. M. Extensões da Distribuição Gama Generalizada: Propriedades e Aplicações. Piracicaba, 2012. Tese (Doutorado) - Escola Supeior de Agricultura Luiz de Queiroz, Universidade de São Paulo.

PINHEIRO, J. C.; BATES, D. M. Statistics and computting: Mixed-effects models en S and S-Plus. **New York:Springer**, v.63, n.3, p.394–396, 2001.

R:, D. C. T. A language and envinroment for Statistical compunting, 2013.

RAMOS, P. L.; ACHCAR, J. A.; RAMOS, E. Método Eficiente para Calcular os Estimadores de Máxima Verossimilhança da Distribuição Gama Generalizada. **Revista Brasileira de Biometria, São Paulo**, v.32, n.2, p.267–281, 2014.

RIBEIRO, A. R. Potencial do uso de planárias na avaliação de contaminantes ambientais. Campinas, 2012. Dissertação (Mestrado) - Universidade de Campinas.

SABESP. Operação Sistema Cantareira. **São paulo**, 1989.

SAS, I. SAS/STAT user's guide. v.2, n.6, 2011.

SCHWARZ, G. **Estimating the dimension of a model**. Philadelphia: Annals of Statistics, 1978. 461p.

SEAP/PR. Manual de procedimentos para implantação de estabelecimentos industrial de pescado: produtos frescos e congelados. **Ministério da Agricultura, Pecuária e Abastecimento; Secretaria Especial de Aquicultura e Pesca. Brasília: MAPA: SEAP/PR**, p.116, 2007.

SÁFADI, R. S. Emprego de planárias de água doce Giradia tigrina (Girard, 1850)(Platyhelmintes, Tricadida, Paludicola)na avaliação de toxicidade de compostos metálicos. São Paulo, 1993. 203p. Dissertação (Mestrado) - Universidade de São Paulo.

STRAPASSON, E. Comparação de modelos com censura intervalar em análise de sobrevivência. Piracicaba, 2007. 135p. Tese (Doutorado) - Escola Supeior de Agricultura Luiz de Queiroz, Universidade de São Paulo.

SUZI, M. M.; ANAMARIA, M.; JORGE, B.; PAOLO, P. C. Método estatístico "Análise de Sobrevivência"aplicado a avaliação de produtos. **Arcos Design 6**, p.58–75, 2010.