Análise de Sobrevivência

Iniciação Ciêntifica - PIBIC 2024/2025 (UFPA)

Breno Cauã Rodrigues da Silva

2024-01-25

Índice

Prefácio

Este é um projeto desenvolvido...

1 Conceitos Básicos e Exemplos

1.1 Introdução

O objetivo deste capítulo inicial é apresentar alguns *conceitos* e *fundamentos* de uma das áreas da Estatística e Análise de Dados que mais se desenvolveram nas últimas duas décadas do século XX. Esse avanço foi impulsionado pela evolução das técnicas estatísticas aliada ao progresso computacional.

Na Análise de Sobrevivência, a variável resposta é, em geral, o *tempo até a ocorrência de um evento de interesse*. Especificamente, essa área se concentra em modelar e compreender o tempo necessário para que um evento significativo ocorra, sendo este denominado **tempo de falha**. Como exemplo, Colosimo e Giolo (2006) mencionam casos como o tempo até a morte de um paciente, até a cura de uma doença ou até a recidiva de uma condição clínica.

Uma questão frequentemente levantada é: por que não utilizar outras técnicas estatísticas? Métodos tradicionais não são adequados para dados de sobrevivência devido a uma característica única: a **censura**. Esse conceito refere-se à observação parcial do tempo de falha, como ocorre quando o acompanhamento de um paciente é interrompido antes do evento de interesse. A censura, sendo um elemento essencial da Análise de Sobrevivência, caracteriza situações em que o tempo de falha real é desconhecido, sabendo-se apenas que ele excede determinado ponto.

1.2 Tempo de Falha

Na Análise de Sobrevivência, é fundamental estabelecer alguns pontos iniciais para o estudo. O primeiro deles é o **tempo inicial do estudo**, que deve ser claramente definido para garantir que os indivíduos sejam comparáveis no ponto de partida, diferenciando-se apenas pelas covariáveis medidas. Existem diversas maneiras de definir o tempo inicial, sendo o mais comum o **tempo cronológico**. Contudo, em áreas como Engenharia, outras métricas, como número de ciclos ou quilometragem, também podem ser utilizadas. Colosimo e Giolo (2006) apresentam exemplos práticos, como medidas de carga para equipamentos.

Outro aspecto essencial é a **definição do evento de interesse**, frequentemente associado a falhas ou situações indesejáveis. Para garantir resultados consistentes, a definição do evento deve ser clara e objetiva. Um exemplo elucidativo é fornecido por Colosimo e Giolo (2006):

> *"Em algumas situações, a definição de falha já é clara, como morte ou recidiva, mas em outras pode assumir termos ambíguos. Por exemplo, fabricantes de produtos alimentícios desejam saber o tempo de vida de seus produtos expostos em balcões frigoríficos de supermercados. O tempo de falha vai do momento de exposição (chegada ao supermercado) até o produto se tornar 'inapropriado para consumo'. Esse evento deve ser claramente definido antes do início do estudo. Por exemplo, o produto é considerado inapropriado para consumo quando atinge uma concentração específica de microrganismos por mm^2 de área."*

1.3 Censura

Estudos clínicos que tratam a resposta como uma variável temporal geralmente são prospectivos e de longa duração. No entanto, mesmo sendo extensos, esses estudos frequentemente se encerram antes que todos os indivíduos passem pelo evento de interesse.

Uma característica comum nesses estudos é a **censura**, que corresponde a observações incompletas ou parciais. Apesar disso, tais observações fornecem informações valiosas para a análise. Colosimo e Giolo (2006) destacam a relevância de incluir dados censurados na análise:

> *"Ressalta-se que, mesmo censurados, todos os resultados provenientes de um estudo de sobrevivência devem ser incluídos na análise estatística. Duas razões justificam esse procedimento: (i) mesmo sendo incompletas, as observações censuradas fornecem informações sobre o tempo de vida dos pacientes; (ii) a exclusão das censuras no cálculo das estatísticas pode levar a conclusões enviesadas."*

Existem três tipos principais de censura:

- **Censura Tipo I:** O estudo é encerrado após um período de tempo previamente definido.
- **Censura Tipo II:** O estudo termina quando um número específico de indivíduos passa pelo evento de interesse.
- **Censura Aleatória:** Ocorre quando um indivíduo é retirado do estudo antes do evento de interesse.

A censura mais comum é a **censura à direita**, em que o evento ocorre após o tempo registrado. Entretanto, outros tipos de censura, como **à esquerda** e **intervalar**, também são possíveis.

Censura à esquerda ocorre quando o evento já aconteceu antes do início da observação. Um exemplo é um estudo sobre a idade em que crianças aprendem a ler:

> *"Quando os pesquisadores começaram a pesquisa, algumas crianças já sabiam ler e não se lembravam com que idade isso ocorreu, caracterizando observações censuradas à esquerda."*

No mesmo estudo, observa-se censura à direita para crianças que ainda não sabiam ler no momento da coleta de dados. Nesse caso, os tempos de vida são classificados como **duplamente censurados** (Turnbull 1974).

A censura intervalar ocorre em estudos com visitas periódicas espaçadas, onde só se sabe que o evento ocorreu dentro de um intervalo de tempo. Quando o tempo de falha T é impreciso, considera-se que ele pertence a um intervalo $T \in (L, U]$, conhecido como **sobrevivência intervalar**. Casos especiais incluem tempos de falha exatos, em que $L = U$, sendo $U = 0$ para censura à direita e $L = 0$ para censura à esquerda (Lindsey e Ryan 1998). Destaca-se a seguinte observação de Colosimo e Giolo (2006):

> *"A presença de censura traz desafios para a análise estatística. A censura do Tipo II é, em princípio, mais tratável que os outros tipos, mas para situações simples, que raramente ocorrem em estudos clínicos (Lawless 1982). Na prática, utiliza-se resultados assintóticos para a análise dos dados de sobrevivência."*

1.4 Dados Truncados

O truncamento é uma característica de alguns estudos de sobrevivência que, muitas vezes, é confundida com a censura. Ele ocorre quando certos indivíduos são excluídos do estudo devido a uma condição específica. Nesse caso, os pacientes só são incluídos no acompanhamento após passarem por um determinado evento, em vez de serem acompanhados desde o início do processo.

1.5 Representação dos Dados de Sobrevivência

Considere uma amostra aleatória de tamanho n. O i-ésimo indivíduo no estudo é geralmente representado pelo par (t_i, δ_i), onde t_i é o tempo de falha ou censura, indicado pela variável binária δ_i, definida como:

$$\delta_i = \begin{cases} 1, & \text{se } t_i \text{ é um tempo de falha} \\ 0, & \text{se } t_i \text{ é um tempo de censura.} \end{cases}$$

Portanto, a variável resposta na análise de sobrevivência é representada por duas colunas no conjunto de dados.

Se o estudo também incluir covariáveis, os dados são representados por $(t_i, \delta_i, \mathbf{x}i)$. Caso a censura seja intervalar, a representação é $(li, u_i, \delta_i, \mathbf{x}_i)$.

Para exemplos de dados de sobrevivência, veja a Seção 1.5 do livro de Colosimo e Giolo (2006).

1.6 Especificando o Tempo de Sobrevivência

Seja T uma variável aleatória (v.a.), na maioria dos casos contínua, que representa o tempo de falha. Assim, o suporte de T é definido nos reais positivos $\mathbb{R}^+$. Tal variável é geralmente representada pela sua *função risco* ou pela *função de taxa de falha* (ou taxa de risco). Tais funções, e outras relacionadas, são usadas ao longo do processo de análise de dados de sobrevivência. A seguir, algumas dessas funções e as relações entre elas serão definidas.

1.6.1 Função de Sobrevivência

Esta é uma das principais funções probabilísticas usadas em análise de sobrevivência. A função sobrevivência é definida como a probabilidade de uma observação não falhar até certo ponto t, ou seja a probabilidade de uma observação sobreviver ao tempo t. Em probabilidade, isso pode ser escrito como:

$$S(t) = P(T > t), \tag{1.1}$$

uma conclusão a qual podemos chegar, é que a probabilidade de uma observação não sobreviver até o tempo t, é a acumulada até o ponto t, logo,

$$F(t) = 1 - S(t). \tag{1.2}$$

1.6.2 Função de Taxa de Falha ou de Risco

A probabilidade da falha ocorrer em um intervalo de tempo $[t_1, t_2)$ pode ser expressa em termos da função de sobrevivência como:

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

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

$$\frac{S(t_1) - S(t_2)}{(t_2 - t_1)S(t_1)}.$$

De forma geral, redefinindo o intervalo como $[t, t+\Delta t)$ a expressão assume a seguinte forma:

$$\lambda(t) = \frac{S(t) - S(t + \Delta_t)}{\Delta t \; S(t)}$$

Assumindo Δt bem pequeno, $\lambda(t)$ representa a taxa de falha instantânea no tempo t condicional à sobrevivência até o tempo t. Observe que as taxas de falha são números positivos, mas sem limite superior. A função de taxa de falha $\lambda(t)$ é bastante útil para descrever a distribuição do tempo de vida de pacientes. Ela descreve a forma em que a taxa instantânea de falha muda com o tempo. A função de taxa de falha de T é, então, definida como:

$$\lambda(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t} \tag{1.3}$$

A função de taxa de falha é mais informativa do que a função de sobrevivência. Diferentes funções de sobrevivência podem ter formas semelhantes, enquanto as respectivas funções de taxa de falha podem diferir drasticamente. Desta forma, a modelagem da função de taxa de falha é um importante método para dados de sobrevivência.

1.6.3 Função de Taxa de Falha Acumulada

Outra função útil em análise de dados de sobrevivência é a função taxa de falha acumulada. Esta função, como o próprio nome sugere, fornece a taxa de falha acumulada do indivíduo e é definida por:

$$\Lambda(t) = \int_0^t \lambda(u) du. \tag{1.4}$$

A função de taxa de falha acumulada, $\Lambda(t)$, não têm uma interpretação direta, mas pode ser útil na avaliação da função de maior interesse que é a função de taxa de falha, $\lambda(t)$. Isto acontece essencialmente na estimação não-paramétrica em que $\Lambda(t)$ apresenta um estimador com propriedades ótimas e $\lambda(t)$ é difícil de ser estimada.

1.6.4 Tempo Médio e Vida Média Residual

Outras duas quantidades de interesse em análise de sobrevivência são: o tempo médio de via e a vida média residual. A primeira é obtida pela área sob a função de sobrevivência. Isto é,

$$t_m = \int_0^\infty S(t) dt. \tag{1.5}$$

Já a vida média residual é definida condicional a um certo tempo de vida t. Ou seja, para indivíduos com idade t está quantidade mede o tempo médio restante de vida e é, então, a área sob a curva de sobrevivência à direita do tempo t dividida por $S(t)$. Isto é,

$$\text{vmr}(t) = \frac{\int_0^\infty (u-t) f(u) du}{S(t)} = \frac{\int_0^\infty S(u) du}{S(t)}, \tag{1.6}$$

sendo $f(\cdot)$ a função densidade de T. Observe que $\text{vmr}(0) = t_m$.

1.7 Relações entre as Funções

Para T uma variável aleatória contínua e não-negativa, tem-se, em termos das funções definidas anteriormente, algumas relações matemáticas importantes entre elas, a saber:

$$\lambda(t) = \frac{f(t)}{S(t)} = -\frac{d}{dt}\left[\log S(t)\right],$$

$$\Lambda(t) = \int_0^t \lambda(u) du = -\log S(t)$$

e

$$S(t) = \exp\left\{-\Lambda(t)\right\} = \exp\left\{-\int_0^t \lambda(u) du\right\}$$

Tais relações mostram que o conhecimento de uma das funções, por exemplo $S(t)$, implica no conhecimento das demais, isto é, $F(t)$, $f(t)$, $\lambda(t)$ e $\Lambda(t)$. Outras relações envolvendo estas funções são as seguintes:

$$S(t) = \frac{\text{vmr}(0)}{\text{vmr}(t)} \exp\left\{-\int_0^t \frac{du}{\text{vmr}(u)}\right\}$$

e

$$\lambda(t) = \left(\frac{d\left[\text{vmr}(t)\right]}{dt} + 1\right) / \text{vmr}(t).$$

2 Técnicas Não Paramétricas

2.1 Introdução

Este capítulo apresenta as técnicas não-paramétricas utilizadas para a análise de dados de sobrevivência. Essas técnicas são empregadas quando não se faz suposições sobre a forma específica da distribuição dos tempos de falha, sendo particularmente úteis para dados censurados.

2.2 O Estimador de Kaplan-Meier

Proposto por Kaplan e Meier (1958). É um estimador não-paramétrico utilizado para estimar a função de sobrevivência, $S(t)$. Tal estimador também é chamado de de *estimador limite-produto*. O Estimador de Kaplan-Meier é uma adaptação a $S(t)$ empiríca que, na ausência de censura nos dados, é definida como:

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

$\hat{S}(t)$ é uma função que tem uma formato gráfico de escada com degraus nos tempos observados de falha de tamanho $1/n$, onde n é o tamanho amostral.

O processo utilizado até se obter a estimativa de Kaplan-Meier é um processo passo a passo, em que o próximo passo depende do anterior. De forma suscetível, para qualquer t, $S(t)$ pode ser escrito em termos de probabilidades condicionais. Suponha que existam n pacientes no estudo e $k(\leq n)$ falhas distintas nos tempos $t_1 \leq t_2 \leq \cdots \leq t_k$. Considerando $S(t)$ uma função discreta com probabilidade maior que zero somente nos tempos de falha t_j, $j = 1, \cdots, k$, tem-se que:

$$S(t_j) = (1 - q_1)(1 - q_2)\cdots(1 - q_j), \tag{2.1}$$

em que q_j é a probabilidade de um indivíduo morrer no intervalo $[t_{j-1}, tj)$ sabendo que ele não morreu até t_{j-1} e considerando $t_0 = 0$. Ou seja, pode se escrever q_j como:

$$q_j = P(T \in [t_{j-1}, tj) | T \geq t_{j-1}), \tag{2.2}$$

para $j = 1, \cdots, k$.

A expressão geral do estimador de Kaplan-Meier pode ser apresentada após estas considerações preliminares, Formalmente, considere:

- $t_1 \leq t_2 \leq \cdots \leq t_k$, os k tempos distintos e ordenados de falha;
- d_j o número de falhas em t_j, $j = 1, \cdots, k$;
- n_j o número de indivíduos sob risco em t_j, ou seja, os indivíduos que não falharam e não foram censurados até o instante imediatamente anterior a t_j.

Com isso, pode-se definir o estimador de Kaplan-Meier como:

$$\hat{S}_{KM}(t) = \prod_{j\,:\,t_j<t} \left(\frac{n_j - d_j}{n_j}\right) = \prod_{j\,:\,t_j<t} \left(1 - \frac{d_j}{n_j}\right) \tag{2.3}$$

De forma intuitiva, por assim dizer, a Equação 2.3 é proveniente da Equação 2.1, sendo está, uma decomposição de $S(t)$ em termos q_j's. Assim, a Equação 2.3 é justificada se os q_j's forem estimados por d_j/n_j, que em palavras está expresso na Equação 2.2. No artigo original de 1958, Kaplan e Meier provam que a Equação 2.3 é um *Estimador de Máxima Verossimilhança* (EMV) para $S(t)$. Seguindo certos passos, é possível provar que que $\hat{S}_{KM}(t)$ é EMV de $S(t)$. Supondo que d_j observações falham no tempo tempo t_j, para $j = 1, \cdots, k$, e m_j observações são censuradas no intervalo $[tj, t_{j+1})$, nos tempos $t_{j1}, \cdots, t_{jm_j}$. A probabilidade de falha no tempo t_j é, então,

$$S(t_j) - S(t_j+),$$

com $S(t_j+) = \lim_{\Delta t \to 0+} S(t_j + \Delta t)$, $j = 1, \cdots, k$. Por outro lado, a contribuição para a função de verossimilhança de um tempo de sobrevivência censurado em t_{jl} para $l = 1, \cdots, m_j$, é:

$$P(T > t_{jl}) = S(t_{jl}+).$$

A função de verossimilhança pode, então, ser escrita como:

$$L(S(\cdot)) = \prod_{j=0}^{k} \left\{ [S(t_j) - S(t_j+)]^{d_j} \prod_{l=1}^{m_j} S(t_{jl}+) \right\}.$$

Com isso, é possível provar que $S(t)$ que maximiza $L(S(\cdot))$ é exatamente a expressão dada pela Equação 2.3.

2.2.1 Propriedades do Estimador de Kaplan-Meier

Como um estimador de máxima verossimilhança, o estimador de Kaplan-Meier têm interessantes propriedades. As principais são:

- É não-viciado para grandes amostras;
- É fracamente consistente;
- Converge assintoticamente para um processo gaussiano.

A consistência e normalidade assintótica de $\hat{S}_{KM}(t)$ foram provadas sob certas condições de regularidade, por Breslow e Crowley (1974) e Meier (1975) e, no artigo original, Kaplan e Meier (1958) mostram que $\hat{S}_{KM}(t)$ é um EMV para $S(t)$, como já dito.

2.2.2 Variância do Estimador de Kaplan-Meier

Para que se possa construir intervalos de confiança e testar hipóteses para $S(t)$, se faz necessário ter conhecimento quanto variabilidade e precisão do estimador de Kaplan-Meier. Este estimador, assim como outros, está sujeito a variações que devem ser descritas em termos de estimações intervalares. A expressão assintótica do estimador de Kaplan-Meier é dada pela Equação 2.4.

$$\hat{Var}[\hat{S}_{KM}(t)] = [\hat{S}_{KM}(t)]^2 \sum_{j\,:\,t_j<t} \frac{d_j}{n_j(n_j - d_j)} \tag{2.4}$$

A expressão dada na Equação 2.4, é conhecida como fórmula de Greenwood e pode ser obtida a partir de propriedades do estimador de máxima verossimilhança. Os detalhes da obtenção da (Equação 2.4 estão disponíveis em Kalbfleisch e Prentice (1980, pag. 12-14).

Como $\hat{S}_{KM}(t)$, para um t fixo, tem distribuição assintóticamente Normal. O intervalo de confiança com $100(1-\alpha)\%$ de confiança para $\hat{S}_{KM}(t)$ é expresso por:

$$\hat{S}_{KM}(t) \pm z_{\alpha/2}\sqrt{\hat{Var}[\hat{S}_{KM}(t)]}.$$

Vale salientar que para valores extremos de t, este intervalo de confiança pode apresentar limites que não condizem com a teoria de probabilidades. Para solucionar tal problema, aplica-se uma transformação em $S(t)$ como, por exemplo, $\hat{U}(t) = \log\left[-\log\left(\hat{S}_{KM}(t)\right)\right]$. Esta transformação foi sugerida por Kalbfleisch e Prentice (1980), tendo sua variância estimada por:

$$\hat{Var}[\hat{U}(t)] = \frac{\sum_{j\,:\,t_j<t} \dfrac{d_j}{n_j(n_j-d_j)}}{\left[\sum_{j\,:\,t_j<t} \log\left(\dfrac{n_j-d_j}{n_j}\right)\right]^2} = \frac{\sum_{j\,:\,t_j<t} \dfrac{d_j}{n_j(n_j-d_j)}}{\left[\log \hat{S}_{KM}(t)\right]^2}$$

Logo, pode-se aproximar um intervalo com $100(1-\alpha)\%$ de confiança para $S(t)$ desta forma:

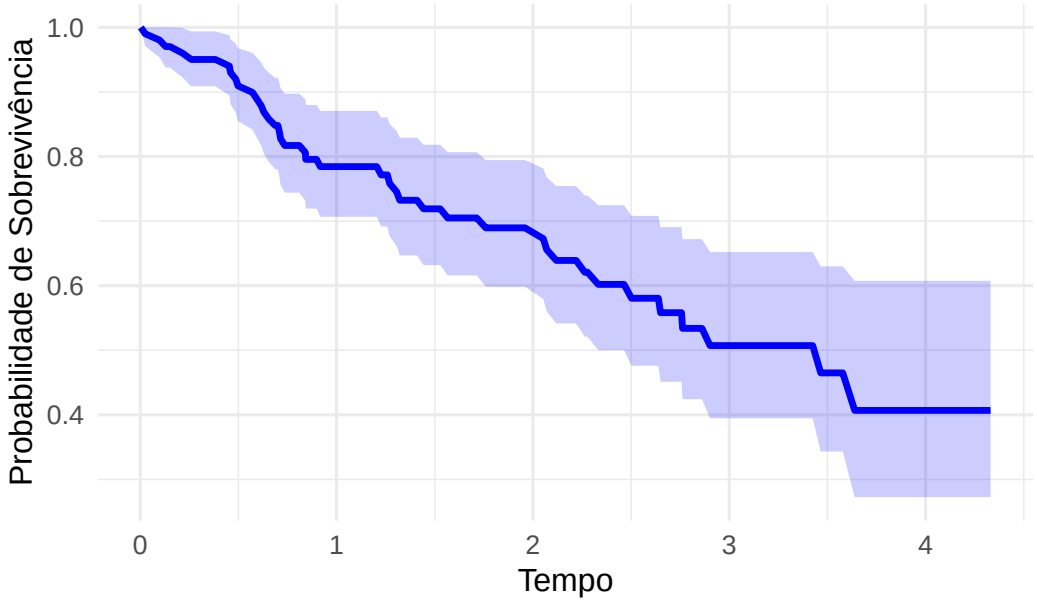
$$\left[\hat{S}(t)\right]^{\exp\left\{\pm z_{\alpha/2}\sqrt{\hat{Var}[\hat{U}(t)]}\right\}}.$$

Veja uma aplicação do Estimador de Kaplan-Meier. Os dados de Leucemia Pediátrica dispostos no Apêndice (A) do livro *Análise de Sobrevivência Aplicada* de Colosimo e Giolo (2006). De posse do conjunto de dados, pode-se estimar a curva de sobrevivência, tal curva foi ilustrada na Figura 2.1.

```r
# ----------------------
# [1] ATIVAÇÃO DE PACOTES
# ----------------------
library(survival)
library(ggplot2)

# --------------------------------
# [2] IMPORTAÇÃO E AJUSTE DOS DADOS
# --------------------------------

# Caminho URL para os dados
url <- "https://docs.ufpr.br/~giolo/asa/dados/leucemia.txt"

# Leitura dos dados
dados <- read.table(url, header = TRUE)

# ----------------------------
# [3] ESTIMADOR DE KAPLAN-MEIER
# ----------------------------
ekm <- survfit(Surv(tempos, cens) ~ 1, data = dados)

# -----------------
# [4] VISUALIZAÇÃO
# -----------------

# Preparando os dados para o ggplot2
ekm_data <- data.frame(time = ekm$time, survival = ekm$surv,
  lower = ekm$lower, upper = ekm$upper)

# Gráfico com ggplot2
ggplot(ekm_data, aes(x = time, y = survival)) +
  geom_line(color = "blue", lwd = 1.2) +
  geom_ribbon(aes(ymin = lower, ymax = upper), fill = "blue", alpha = 0.2) +
  labs(x = "Tempo", y = "Probabilidade de Sobrevivência",
    caption = "Fonte: https://docs.ufpr.br/~giolo/asa/dados/leucemia.txt") +
```

```
  theme_minimal(base_size = 12) +
  theme( plot.caption = element_text(hjust = 0.5, size = 12))
```

Figura 2.1: Curva de Sobrevivência de Kaplan-Meier com IC de 95%



Fonte: https://docs.ufpr.br/~giolo/asa/dados/leucemia.txt

2.3 Outros Estimadores Não Parâmetricos

O estimador de Kaplan-Meier é, indiscutivelmente, o mais utilizado para estimar $S(t)$ em análises de sobrevivência. Ele é amplamente disponibilizado em diversos pacotes estatísticos e abordado em inúmeros textos de estatística básica. Entretanto, outros dois estimadores de $S(t)$ também possuem relevância significativa na literatura especializada: o estimador de Nelson-Aalen e o estimador da tabela de vida.

O estimador de Nelson-Aalen, mais recente que o de Kaplan-Meier, apresenta propriedades similares às deste último. Já o estimador da tabela de vida possui importância histórica, tendo sido utilizado em informações derivadas de censos demográficos para estimar características associadas ao tempo de vida humano. Este estimador foi inicialmente proposto por demógrafos e atuários no final do século XIX, sendo empregado principalmente em grandes amostras.

Nesta seção será abordado apenas o estimador de Nelson-Aalen. Para conhecer mais sobre o estimador da Tabela de Vida ou Tabela Atuarial, consulte a Seção 2.4.2 do livro *Análise de Sobrevivência Aplicada** de Colosimo e Giolo (2006).

2.3.1 Estimador de Nelson-Aalen

Mais recente que o estimador de Kaplan-Meier, este estimador se baseia na função de sobrevivência expressa da seguinte forma:

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

em que $\Lambda(t)$ é a função de risco acumulado apresentada na Seção 1.6.3.

A estimativa para $\Lambda(t)$ foi inicialmente proposta por Nelson (1972) posteriormente retomada por Aalen (1978) que demonstrou suas propriedades assintóticas utilizando processos de contagem. Na literatura, esse estimador é amplamente conhecido como o estimador de Nelson-Aalen e é definido pela seguinte expressão:

$$\hat{\Lambda}(t) = \sum_{j:t_j<t} \left(\frac{d_j}{n_j}\right), \tag{2.5}$$

onde d_j e n_j são as mesmas definições usadas no estimador de Kaplan-Meier. A variância do estimador, conforme proposta por Aalen (1978), é dada por:

$$\hat{Var}(\hat{\Lambda}(t)) = \sum_{j:t_j<t} \left(\frac{d_j}{n_j^2}\right). \tag{2.6}$$

Uma alternativa para a estimativa da variância de $\hat{\Lambda}(t)$, proposta por Klein (1991), é:

$$\hat{Var}(\hat{\Lambda}(t)) = \sum_{j:t_j<t} \frac{(n_j - d_j)d_j}{n_j^3},$$

entretanto, o estimador da Equação 2.6 apresenta menor vício, tornando-o mais preferível que o proposto por Klein (1991).

Desta forma, podemos definir, com base no estimador de Nelson-Aalen, um estimador para a função de sobrevivência, podendo ser expressa por:

$$\hat{S}_{NA}(t) = \exp\left\{-\hat{\Lambda}(t)\right\}.$$

Deve-se, a variância deste estimador, a Aalen e Johansen (1978). Podendo ser mensurada pela expressão:

$$\hat{Var}(\hat{S}_{NA}(t)) = \left[\hat{S}_{NA}(t)\right]^2 \sum_{j:t_j<t} \left(\frac{d_j}{n_j^2}\right)$$

Vale destacar que o estimador de Nelson-Aalen aprenseta, na maioria dos casos, estimativas próximas ao estimador de Kaplan-Meier. Bohoris (1994) mostrou que $\hat{S}_{NA}(t) \geq \hat{S}_{KM}(t)$ para todo t, isto é, as estimativas obtidas pelo estimador de Nelson-Aalen são maiores ou iguais às estimativas obtidas pelo estimador de Kaplan-Meier.

2.4 Comparação de Curvas de Sobrevivência

Imagine um problema da área da saúde, onde se desejar comparar dois grupos, um grupo receberá tratamento através do uso de alguma droga e o outro será o grupo controle. Estatísticas mais comumente usadas podem ser vistas como generalizações para dados censurados, de conhecidos testes não-paramétricos. O teste *logrank* (Mantel 1966) é o mais usado em análise de sobrevivência. Gehan (1965) propôs uma generalização para a estatística de Wilcoxon. Outras generalizações foram propostas por Peto e Peto (1972) e Prentice (1978), entre outros. Latta (1981) fez uso de simulações de Monte Carlo para comparar vários testes não-paramétricos.

Neste texto, ênfase será dada ao teste *logrank*. Este teste é muito utilizado em análise de sobrevivência e é particularmente apropriado quando a razão das funções de risco dos grupos a serem comparados é aproximadamente constante. Isto é, as populações têm a propriedade de riscos proporcionais. A estatística deste teste é a diferença entre o número observado de falhas em cada grupo e uma quantidade que, para muitos propósitos, pode ser pensada como o correspondente número esperado de falhas sob a hipótese nula. A expressão do teste *logrank* é obtida de forma similar a do conhecido teste de Mantel e Haenszel (1959), para combinar tabelas de contigência. O teste *logrank* tem, também, a mesma expressão do teste de escore para o modelo de regressão de Cox que será apresentado no [...]. Outros testes também são apresentados nesta seção.

Considere, inicialmente, o teste de igualdade de duas funções de sobrevivência $S_1(t)$ e $S_2(t)$. Sejam $t_1 < t_2 < \cdots < t_k$ os tempos de falha distintos da amostra formada pela combinação das duas amostras individuais. Suponha que em um tempo t_j aconteçam d_j falhas e que n_j indivíduos estejam sob risco em um tempo imediatamente inferior a t_j na amostra combinada e, respectivamente, d_{ij} e n_{ij} na amostra i; $i = 1, 2$ e $j = 1, \cdots, k$. Em tempo de falha t_j, os dados podem ser dispostos em forma de uma tabela de contigência 2 x 2 com d_{ij} falhas e $n_{ij} - d_{ij}$ sobreviventes na coluna i. Isto é mostrado na Tabela 2.1.

Tabela 2.1: Tabela de contingência gerada no tempo t_j.

	Grupo 1	Grupo 2	
Falha	d_{1j}	d_{2j}	d_j
Não Falha	$n_{1j} - d_{1j}$	$n_{2j} - d_{2j}$	$n_j - d_j$
	n_{1j}	n_{2j}	n_j

Condicional à experiência de falha e censura até o tempo t_j (fixando as marginais de coluna) e ao número de falhas no tempo t_j (fixando as marginais de linha), a distribuição de d_{2j} é, então, uma hipergeométrica:

$$\frac{\binom{n_{1j}}{d_{1j}}\binom{n_{2j}}{d_{2j}}}{\binom{n_j}{d_j}},$$

A média de d_{2j} é $w_{2j} = n_{2j}d_j n_j^{-1}$, o que equivale a dizer que, se não houver diferença entre as duas populações no tempo t_j, o número total de falhas (d_j) pode ser dividido entre as duas amostras de acordo com a razão entre o número de indivíduos sob risco em cada amostra e o número total sob risco. A variância de d_{2j} obtida a partir da distribuição hipergeométrica é:

$$(V_j)_2 = n_{2j}(n_j - n_{2j})d_j(n_j - d_j)n_j^{-2}(n_j - 1)^{-1}.$$

Então, a estatística $d_{2j} - w_{2j}$ tem média zero e variância $(V_j)_2$. Se as k tabelas de contingência forem independentes, um teste aproximado para a igualdade das duas funções de sobrevivência pode ser baseado na estatística:

$$T = \frac{\left[\sum_{j=1}^{k}(d_{2j} - w_{2j})\right]^2}{\sum_{j=1}^{k}(V_j)_2}, \tag{2.7}$$

que, sob a hipótese nula $H_0 : S_1(t) = S_2(t)$ para todo t no período de acompanhamento, tem uma distribuição qui-quadrado com 1 grau de liberdade para grandes amostras.

Com o intuito de exemplificar o teste de *logrank* com dados reais aplicou-se o mesmo aos dados de Leucemia Pediátrica dispostos no Apêndice (A) do livro *Análise de Sobrevivência Aplicada* de Colosimo e Giolo (2006), os mesmo dados usados para gerar a Figura 2.1. No teste executado o obejtivo é testar se a curva de sobrevivência das categorias da covariável `r6` são iguais, sob as hipóteses:

$$\begin{cases} H_0 : \text{As curvas de sobrevivência dos grupos são iguais ao longo do tempo.} \\ H_1 : \text{As curvas de sobrevivência dos grupos são diferentes ao longo do tempo.} \end{cases}$$

```
# ----------------------
# [1] ATIVAÇÃO DE PACOTES
# ----------------------
library(survival)
library(ggplot2)
library(dplyr)

# --------------------------------
```

```r
# [2] IMPORTAÇÃO E AJUSTES DOS DADOS
# ----------------------------------

# Caminho URL para os dados
url <- "https://docs.ufpr.br/~giolo/asa/dados/leucemia.txt"

# Leitura dos dados
df <- read.table(url, header = TRUE)

# Decodificando a coluna r_6
df <- df %>%
  mutate(grupo = ifelse(r6 == 0, "Category Zero", "Category One"))

# ----------------------
# [3] TESTE DE LOGRANK
# ----------------------

# Aplicando o Teste de Logrank
TestLogrank <- survdiff(Surv(tempos, cens)~grupo, data = df, rho = 0)
print(TestLogrank )
```

```
Call:
survdiff(formula = Surv(tempos, cens) ~ grupo, data = df, rho = 0)

                     N Observed Expected (O-E)^2/E (O-E)^2/V
grupo=Category One  95       34    37.16     0.269      5.73
grupo=Category Zero  8        5     1.84     5.429      5.73

 Chisq= 5.7  on 1 degrees of freedom, p= 0.02
```

Se for fixado o nível de significância em 5%, ou seja, $\alpha = 0,05$, rejeitamos a hipótese nula. Chega-se a essa conclusão olhando para o $p-valor$ (probabilidade de significância) do teste, mensurado em $p-valor = 0.02$. Como o $p-valor < \alpha$, rejeita-se H_0, logo, as curvas de sobrevivência dos grupos são diferentes ao longo do tempo, ao nível de significância de 5%.

A generalização do teste *logrank* para a igualdade de $r > 2$ funções de sobrevivência $S_1(t), S_2(t), ..., S_r(t)$ não é complicada. Considere a mesma notação anterior, com o índice i variando, agora, entre 1 e r. Desta forma, os dados podem ser arranjados em forma de uma tabela de contingência 2 x r falhas d_{ij} e $n_{ij} - d_{ij}$ sobreviventes na coluna i. Ou seja, a Tabela 2.1 passaria a ter r colunas em vez de simplesmente duas.

3 Técnicas Paramétricas - Modelos Probabilísticos

3.1 Introdução

No Capítulo anteriror, foi vista uma abordagem não paramétrica, onde, a estimação é feita sem se referir a uma distribuição de probabilidade específica para o tempo de sobrevivência.

Obtendo os estimadores não paramétricos diretamente do conjunto de dados. Supondo que o mecanismo gerador dessas informações opere de forma distinta em diferentes momentos no tempo. Funcionando de forma quase que independente, desta forma, conclui-se que a estimação não paramétrica têm tantos parâmetros quanto intervalos no tempo. Ao incluir covariáveis, o modelo de *Kaplan-Meier* não permite estimar o "efeito" das cováriaveis, mas apenas comparar e testar a igualdade entre duas curvas de sobrevivência.

De acordo com a distribuição de probabilidade que acredita-se descrever a variável resposta Y, e de acordo com a função escolhida para a relação de Y com as covariáveis $x_1, x_2, \cdots, x_p$, identifica-se o modelo de regressão como: *Linear*, *Poisson*, *Logísitico*, entre outros. Aplica-se a ideia em análise de sobrevivência, de forma que o tempo de ocorrência até um evento de interesse é a variável resposta.

Nesse contexto, neste Capítulo é vista uma abordagem paramétrica para estimar as funções básicas de sobrevivência. Onde se assume como conhecida a distribuição de probabilidade do tempo de evento e, desta forma, os parâmetros seram estimados.

3.2 Distribuições do Tempo de Sobrevivência

Seja T uma variável aleatória que representa "tempo de sobrevivênvia", qual a distribuição de probabilidade poderia representa-lá?

Como uma característica da variável aleatória T é contínua e não negativa, pode-se a partir dessa caracater´sitica, remover algumas distribuições da lista de possíveis distribuições de probabilidades de T. Desta forma, a distribuição normal já não se torna adequada, pois tal distribuição permite valores negativos. Além disso, o tempo de sobrevivência contém, frequentemente, uma forte assimetria à direita.

Entre os modelos paramétricos, utiliza-se muito a classe **tempo de vida acelerado**. Em tal classe, o tempo de sobrevivência T, obdesce a seguinte relação:

$$\ln(T) = \mu + \sigma W, \tag{3.1}$$

onde μ é o parâmetro que representa a média de $\ln(T)$ e σ sua dispersão, mas usualmente denominados, respectivamnete, parâmetros de locação e escala. W é uma variável aleatória que possa representar $\ln(T)$ a partir de uma distribuição de probabilidade.

3.2.1 Distribuição Exponencial

Se $T \sim Exp(\alpha)$, a sua função densidade de probabilidade é expressa da seguinte forma:

$$f(t) = \alpha e^{-\alpha t},\ t \geq 0 \text{ e } \alpha > 0. \tag{3.2}$$

Desta forma, para obtermos, por exemplo, a função de sobrevivência basta integrarmos a função densidade de probabilidade, veja:

$$\begin{aligned} S(t) = P(T \geq t) &= \int_t^\infty \alpha e^{-\alpha t}\ dt \\ &= \alpha \int_t^\infty e^{-\alpha t}\ dt = \alpha \int_t^\infty e^u\ \frac{du}{-\alpha} \\ &= \frac{\alpha}{-\alpha} \int_t^\infty e^u\ du = -\left[e^u\right]_t^\infty = -\left[e^{-\alpha t}\right]_t^\infty \\ &= -[e^{-\alpha\infty} - e^{-\alpha t}] = -[0 - e^{-\alpha t}] = e^{-\alpha t}. \end{aligned}$$

Assim, foramalmente, a função de sobrevivência é expressa por:

$$S(t) = e^{-\alpha t}. \tag{3.3}$$

Note que o parâmetro α é a velocidade de queda da função sobrevivência. Através das relações entre as funções em análise de sobrevivência, temos a função risco ou taxa de falha. Obtida pela razão entre a densidade de probabilidade e a função de sobrevivência:

$$\lambda(t) = \frac{f(t)}{S(t)} = \frac{\alpha e^{-\alpha t}}{e^{-\alpha t}} = \alpha = \text{constante}. \tag{3.4}$$

Sendo a função risco constante para todo tempo observado t, o risco acumulado é função linear no tempo com uma inclinação na reta dado por α:

$$\Lambda = -\ln[S(t)] = -\ln[e^{-\alpha t}] = -(-\alpha t) = \alpha t \tag{3.5}$$

Afim de ilustrar para visualização e melhor aprendizado do autor e de quem for ler tal material foram simuladas as funções de sobrevivência, risco e risco acumulado variando o parâmetro α.

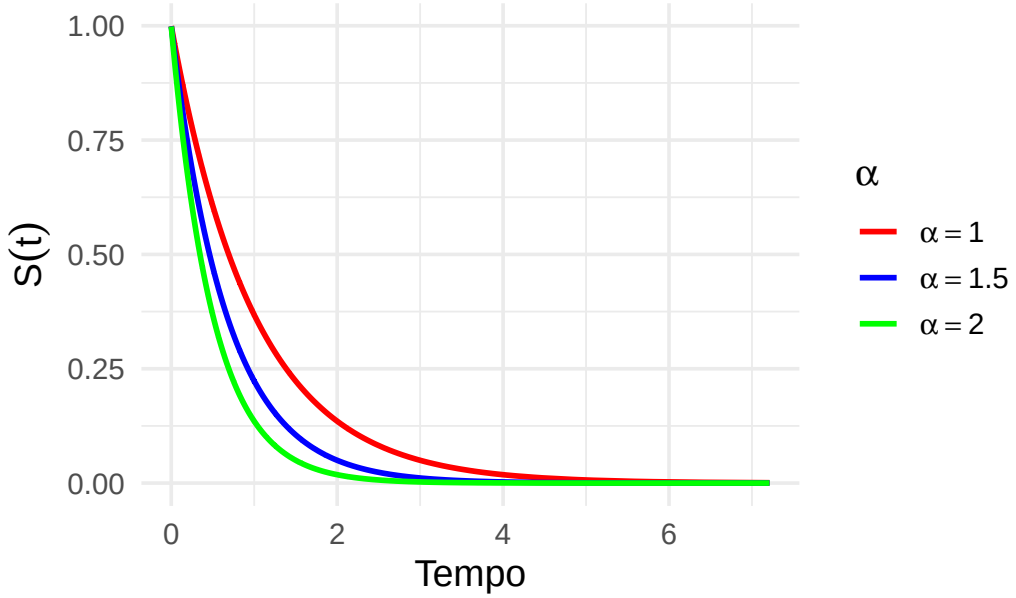
```
library(ggplot2)
set.seed(123)
n <- 1000

# --------------------------
# [1] DISTRIBUIÇÃO EXPONENCIAL
# --------------------------

# -------------
# [1.1] FUNÇÕES
# -------------
# As funções de sobrevivência, risco e risco acumulado são simplificadas
Stexp <- function(t, alpha) exp(-alpha * t)
htexp <- function(alpha) rep(alpha, length(t))
Ltexp <- function(t, alpha) alpha * t

# ---------------------------------------
# [1.2] SIMULAÇÃO E VARIAÇÃO DE PARÂMETROS
# ---------------------------------------
tempo <- rexp(n, rate = 1) # Simulando dados de uma exponencial
alphas <- c(1, 1.5, 2)     # Valores de alpha a serem avaliados

# Criando um Data Frame com valores das funções
dados <- do.call(rbind, lapply(alphas, function(alpha) {
  data.frame(
    tempo = tempo,
    St = Stexp(tempo, alpha),
    ht = htexp(alpha),
    Lt = Ltexp(tempo, alpha),
    alpha = factor(alpha)
  )
}))

# -------------
# [1.3] GRÁFICOS
# -------------

# Criando uma função para gerar gráficos
plot_func <- function(data, y_var, y_label, color_values, y_expression) {
  ggplot(data, aes(x = tempo, y = !!sym(y_var), color = alpha)) +
    geom_line(stat = "summary", fun = mean, size = 1) +
```

```
    labs(x = "Tempo", y = y_expression,
         caption = "Fonte: Elaborado pelo autor.",
         color = expression(alpha)) +
    scale_color_manual(values = color_values,
                       labels = lapply(alphas, function(a) bquote(alpha == .(a)))) +
    theme_minimal(base_size = 14) +
  theme(
    plot.caption = element_text(hjust = 0.5, size = 10)
  )
}
```

Veja a seguir a Figura 3.1. Tal Figura mostra as curvas de sobrevivência para diferentes valores do parâmetro α.

```
# Função de Sobrevivência
plot_func(dados, "St", "S(t)", c("red", "blue", "green"), expression(S(t)))
```

Figura 3.1: Função de Sobrevivência segundo uma Distribuição Exponencial para diferentes valores do parâmetro de taxa.

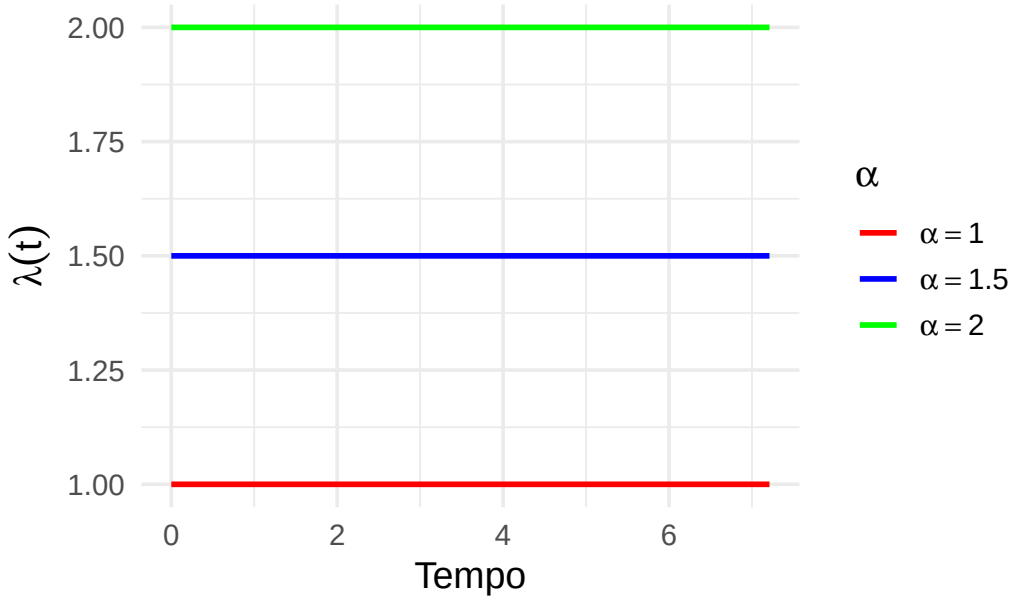


Fonte: Elaborado pelo autor.

Veja a seguir a Figura 3.2. Tal Figura mostra a função de risco para diferentes valores do parâmetro α.

```
# Função de Risco
plot_func(dados, "ht", expression(lambda(t)), c("red", "blue", "green"), expression(lamb
```

Figura 3.2: Função de Risco segundo uma Distribuição Exponencial para diferentes valores do parâmetro de taxa.

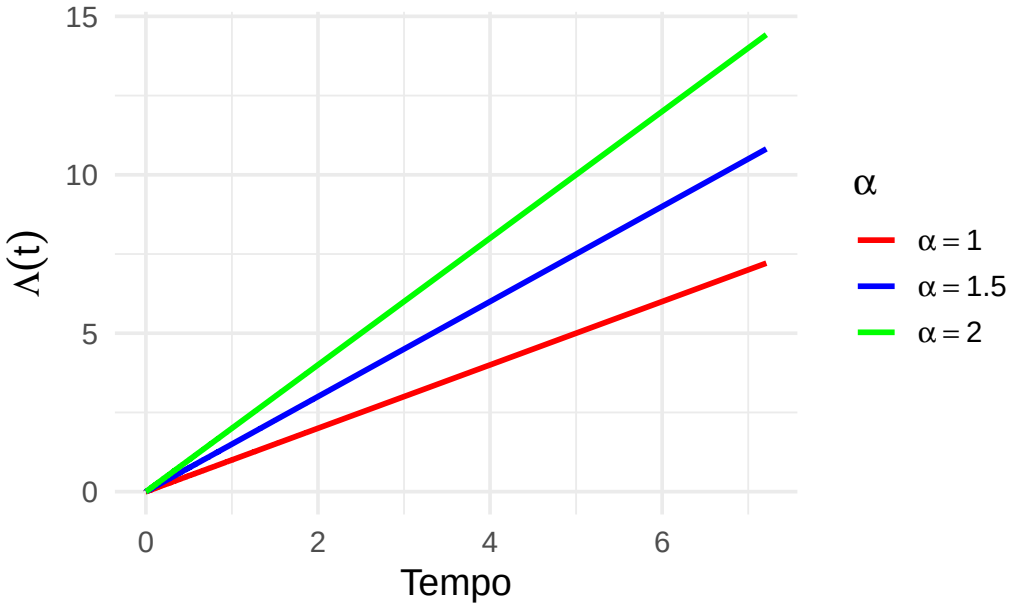


Fonte: Elaborado pelo autor.

Veja a seguir a Figura 3.3. Tal Figura mostra a função de risco acumulado para diferentes valores do parâmetro α.

```
# Função de Risco Acumulado
plot_func(dados, "Lt", expression(Lambda(t)), c("red", "blue", "green"), expression(Lamb
```

Figura 3.3: Função de Risco segundo uma Distribuição Exponencial para diferentes valores do parâmetro de taxa.



Fonte: Elaborado pelo autor.

3.2.1.1 Algumas Considerações

Note que quanto maior o valor de α (risco), mas abruptamente a função de sobrevivência $S(t)$ decresce e maior é a inclinação que representa o risco acumulado.

Como a distribuição exponencial possui um único parâmetro se torna matematicamente simples além de possuir um formato assimétrico.

O seu uso em análise de sobrevivência tem certa analogia com a presuposição de normalidade em outras técnicas e áreas da estatísticas. Porém, seu presuposto de risco constante é uma afirmação muuito forte a se fazer a respeito do risco. Por exemplo, está sendo realizado um estudo sobre o cancêr, o tempo de evento de tal experimento é o tempo até que ocorra a morte ou cura do paciente. Para aplicar a distribuição exponecial para modelar esse problema seria necessário pensar que o tempo desde o diagnóstico da doênça não afeta o tempo de ocorrência do evento. O que é delicado de se aceitar, tendo em vista que o prórprio passar do tempo afeta a probabilidade de sobrevivência, o risco, risco acumulado, etc. Tendo em vista que, isso pode ser simplismente por causas naturais como aumento da idade ao passar do tempo (envelhecimento), por exemplo. Desta forma, está consequência da distribuição exponencial, isso é denominado de **falta de memória da distribuição normal**.

Quando $\alpha = 1$, diz-se que a **distribuição exponecial padrão**. A média ($E[t]$) e variância ($Var[T]$) do tempo de sobrevivência, quando este seguir uma distribuição em Exponecial, são obtidas a partir da inversa do risco (α). Quanto maior o risco, menor o tempo médio de sobrevivência e menor variabilidade deste em torno na média.

$$E[T] = \frac{1}{\alpha}$$

$$Var[T] = \frac{1}{\alpha^2}$$

Ao afirmar que o tempo de sobrevivência T segue uma distribuição de exponencial equivale a dizer que na Equação 3.1, W segue uma distribuição valor extremo padrão, $\sigma = 1$. Assim para cada tempo t, a Equação 3.1 é escrita da seguinte forma: $\ln(T) = \mu + w$, como a $E[T] = \frac{1}{\alpha}$ na distribuição exponencial, tem-se que:

$$\mu = -\ln(\alpha).$$

> Essa é uma forma de parametrização dos modelos parâmetricos utilizada no R, isto é, o parâmetro $\alpha = e^{-\mu}$.

Como a distribuição de T é assimétrica, se torna mais usual utilizar o *tempo mediano de sobrevivência* ao invés de tempo médio. Pode-se obter o tempo mediano de sobrevivência a partir de um tempo t, tal que, $S(t) = 0,5$, logo

$$\begin{aligned} S(t) = 0,5 \Leftrightarrow e^{-\alpha t} = 0,5 \Leftrightarrow -\alpha t = \ln(1/2) \\ \alpha t = -\ln(2^{-1}) \Leftrightarrow \alpha t = \ln((2^{-1})^{-1}) \\ \alpha t = -\ln(2). \end{aligned}$$

Desta forma, o tempo mediano de sobrevivência é definido como:

$$T_{mediano} = \frac{\ln(2)}{\alpha}.$$

Em suma, o modelo exponecial se torna adequado quando o período do experimento é curto para que a supocisão de risco constante possa ser antendida.

3.2.2 Distribuição Weibull

Na maioria do casos de análise de sobrevivência na área da saúde, se torna mais lógico supor que o risco não é constante ao longo do tempo.

Atualmente, a *Distribuição Weibull* é mais utilizada, pois permite a variação do risco ao longo do tempo. Será possível ver que a distribuição exponencial é um caso particular da distribuição weibull.

Se o tempo de sobrevivência T segue uma distribuição Weibull, isto é, $T \sim Weibull(\alpha, \gamma)$, sua função densidade de probabilidade é expressa por:

$$f(t) = \gamma\alpha^{\gamma}t^{\gamma-1}\exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\}. \tag{3.6}$$

A partir da Equação 3.6 é possível chegar a função de sobrevivência da distribuição Weibull sendo está função definida como:

$$S(t) = \exp\left\{-\left(\frac{t}{\alpha}\right)^{\gamma}\right\}, \tag{3.7}$$

onde $t \geq 0$, α o parâmetro taxa e γ parâmetro de forma. Ambos os parâmetros sempre positivos.

A função de risco, $\lambda(t)$, depende do tempo de sobrevivência. Apresentando variação no tempo conforme a expressão:

$$\lambda(t) = \gamma\alpha^{\gamma}t^{\gamma-1} \tag{3.8}$$

e a função de risco acumulado da distribuição Weibull é dada por:

$$\Lambda(t) = -\ln S(t) = -\left(\frac{t}{\alpha}\right)^{\gamma}. \tag{3.9}$$

Note que, o parâmetro γ determina a forma função de risco da seguinte maneira:

- $\gamma < 1 \rightarrow$ função de risco decresce;
- $\gamma > 1 \rightarrow$ função de risco cresce;
- $\gamma > 1 \rightarrow$ a função de risco se torna constante, caindo no caso particular da distribuição exponencial.

Afim de ilustrar para visualização e melhor aprendizado do autor e de quem for ler tal material foram simuladas as funções de sobrevivência, risco e risco acumulado variando o parâmetro γ.

```
library(ggplot2)
set.seed(123)
n <- 1000

# ------------------------
# [2] DISTRIBUIÇÃO WEIBULL
# ------------------------

# -------------
# [2.1] FUNÇÕES
# -------------
```

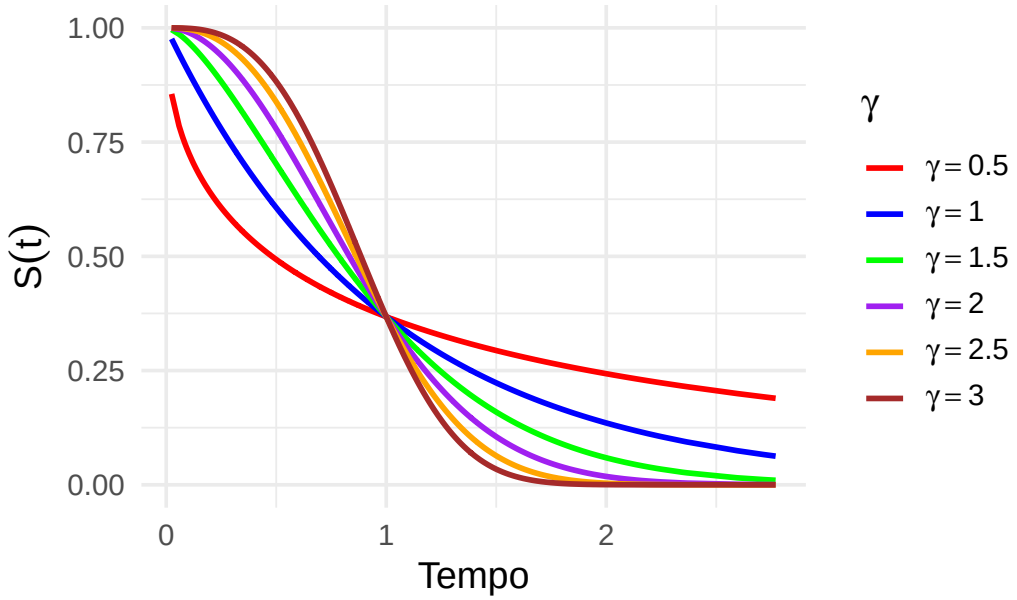
```r
# Funções para Weibull
StWei <- function(t, alpha, gamma) exp(-(t/alpha)^gamma)
htWei <- function(t, alpha, gamma) gamma * (alpha^gamma) * t^(gamma - 1)
LtWei <- function(t, alpha, gamma) (t/alpha)^gamma

# ----------------------------------------
# [2.2] SIMULAÇÃO E VARIAÇÃO DE PARÂMETROS
# ----------------------------------------

# Simulando dados de uma Weibull
tempo <- rweibull(n, shape = 2, scale = 1)
alpha <- 1   # Fixo para simplificar
gammas <- c(0.5, 1.0, 1.5, 2.0, 2.5, 3.0) # Valores de gamma

# Criando um Data Frame com valores das funções
dados <- do.call(rbind, lapply(gammas, function(gamma) {
  data.frame(
    tempo = tempo,
    St = StWei(tempo, alpha, gamma),
    ht = htWei(tempo, alpha, gamma),
    Lt = LtWei(tempo, alpha, gamma),
    gamma = factor(gamma)
  )
}))

# --------------
# [2.3] GRÁFICOS
# --------------

# Função genérica para gráficos
plot_func <- function(data, y_var, y_label, color_values, y_expression) {
  ggplot(data, aes(x = tempo, y = !!sym(y_var), color = gamma)) +
    geom_line(stat = "summary", fun = mean, size = 1) +
    labs(x = "Tempo", y = y_expression,
         caption = "Fonte: Elaborado pelo autor.", color = expression(gamma)) +
    scale_color_manual(values = color_values,
                       labels = lapply(gammas, function(g) bquote(gamma == .(g)))) +
    theme_minimal(base_size = 14) +
  theme(
    plot.caption = element_text(hjust = 0.5, size = 10)
  )
}

# Paleta de cores
```

```
color_values <- c("red", "blue", "green", "purple", "orange", "brown")
```

Veja a seguir a Figura 3.4. Tal Figura mostra as curvas de sobrevivência para diferentes valores do parâmetro γ.

```
# Função de Sobrevivência
plot_func(dados, "St", expression(S(t)), color_values, expression(S(t)))
```

Figura 3.4: Função de Sobrevivência segundo uma Distribuição Weibull para diferentes valores do parâmetro de forma.

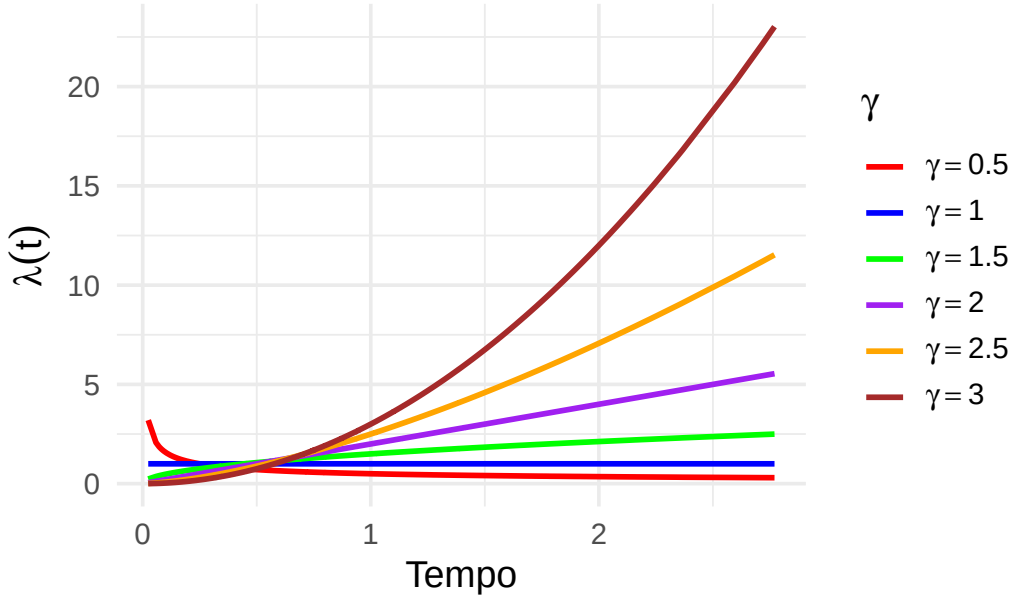


Fonte: Elaborado pelo autor.

Veja a seguir a Figura 3.5. Tal Figura mostra a função de risco para diferentes valores do parâmetro γ.

```
# Função de Risco
plot_func(dados, "ht", expression(lambda(t)), color_values, expression(lambda(t)))
```

Figura 3.5: Função de Risco segundo uma Distribuição Weibull para diferentes valores do parâmetro de forma.

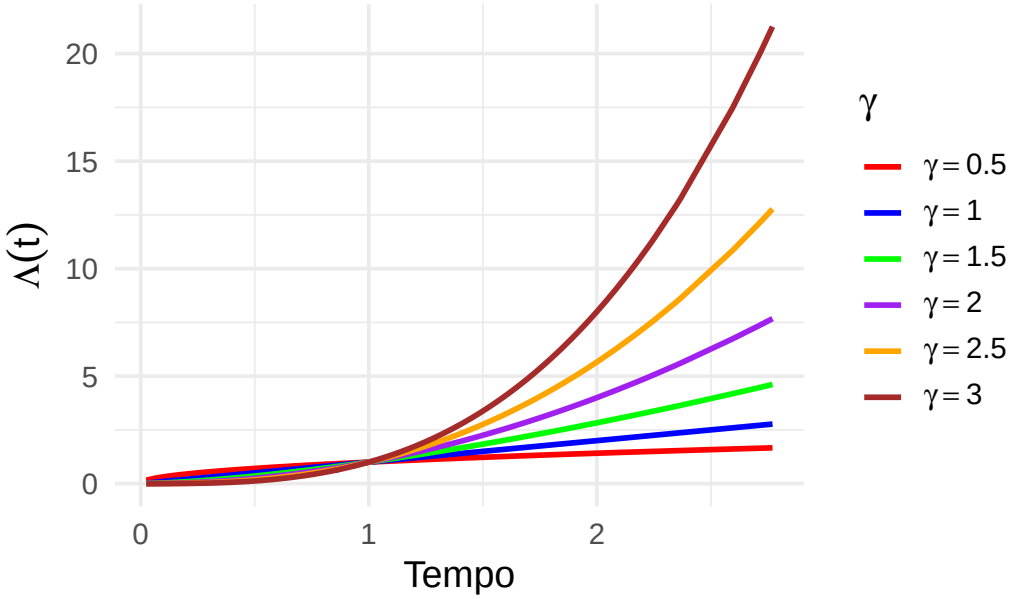


Fonte: Elaborado pelo autor.

Veja a seguir a Figura 3.6. Tal Figura mostra a função de risco acumulado para diferentes valores do parâmetro γ.

```
# Função de Risco Acumulado
plot_func(dados, "Lt", expression(Lambda(t)), color_values, expression(Lambda(t)))
```

Figura 3.6: Função de Risco Acumulado segundo uma Distribuição Weibull para diferentes valores do parâmetro de forma.



Fonte: Elaborado pelo autor.

3.2.2.1 Algumas Considerações

É incluso a função gama na média e variância da distribuição Weibull, assim,

$$E[T] = \alpha\Gamma[1 + (1/\gamma)]$$
$$Var[T] = a^2\left[\Gamma[1 + (2/\gamma)] - \Gamma[1 + (1/\gamma)]^2\right]$$

sendo a função gama $\Gamma[k]$, expressa por $\Gamma[k] = \int_0^\infty x^{k-1}e^{-x}dx$.

Afim de se obter o tempo mediano de sobrevivência, igualamos a probabilidade de sobrevivência a 0,5. Desta forma:

$$S(t) = 0,5 \Leftrightarrow e^{(-\alpha t)^\gamma} = 1/2$$
$$-(\alpha t)^\gamma = -\ln(2^{-1}) \Leftrightarrow \alpha t = \ln(2)$$
$$(\alpha t)^\gamma = \ln(2).$$

Logo,

$$T_{mediano} = \frac{\ln(2)^{1/\gamma}}{\alpha}.$$

3.2.2.2 Distribuição do valor extremo ou de Gambel

Um ponto que deve ser chamada atenção é a relação da distribuição Weibull com outra distribuição. Esta outra distribuição é chamada de *distribuição do valor extremo* ou de *Gambel*. Tal distribuição surge ao se tomar o logaritmo de uma variável T com distribuição de Weibull com $f(t)$ dada por Equação 3.6, desta forma, $Y = \ln(T)$ tem distribuição do valor extremo com densidade da forma:

3.2.3 Distribuição lognormal

Uma outra possibilidade para modelar o tempo de sobrevivência é a *distribuição Log-normal*. Dizer que $T \sim Normal(\mu, \sigma^2)$ implica em dizer que $\ln(T) \sim log - Normal(\mu, \sigma^2)$ em que μ é a média do logaritmo do tempo de falha e σ^2 sua variância. Pode-se fazer uso desta relação para modelar o tempo de sobrevivência conforme uma distribuição normal, desde que, se aplique o logaritmo aos dados observados. A função densidade para tal distribuição é dada por:

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\left(\frac{\ln(t) - \mu}{\sigma}\right)^2\right\} \tag{3.10}$$

Assim, quando o tempo de sobrevivência segue uma distribuição log-normal, sua função de sobrevivência e as demais não tem uma forma análitica explícita, desde modo, deve-se fazer uso das relações entre as funções para se obter a função taxa de falha e taxa de falha acumulada. Desta forma, essas funções são expressas, respectivamente, por:

$$S(t) = \Phi\left(\frac{-\ln(t) + \mu}{\sigma}\right) \tag{3.11}$$

$$\lambda(t) = \frac{f(t)}{S(t)} \quad \text{e} \quad \Lambda(t) = -\ln S(t)$$

em que $\Phi(\cdot)$ é a função de distribuição acumulada da normal padrão.

Afim de ilustrar para visualização e melhor aprendizado do autor e de quem for ler tal material foram simuladas as funções de sobrevivência, risco e risco acumulado variando o parâmetro μ.

```
library(ggplot2)
library(dplyr)

set.seed(123)
n <- 1000
```

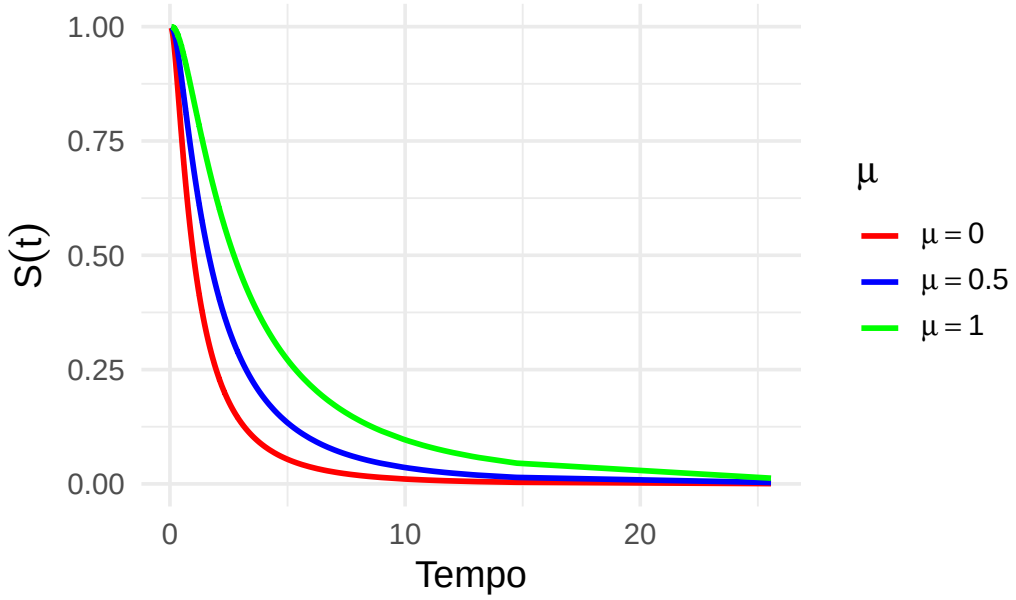
```r
# -------------------------
# [4] DISTRIBUIÇÃO LOG-NORMAL
# -------------------------

# -------------
# [4.1] FUNÇÕES
# -------------

# Função densidade (f)
ftLogNormal <- function(t, mu, sigma) {
  (1 / (t * sigma * sqrt(2 * pi))) * exp(-0.5 * ((log(t) - mu) / sigma)^2)
}

# Função de Sobrevivência (S)
StLogNormal <- function(t, mu, sigma) {
  pnorm(-(log(t) - mu) / sigma, lower.tail = TRUE)
}

# Função de Risco (h)
htLogNormal <- function(t, mu, sigma) {
  ftLogNormal(t, mu, sigma) / StLogNormal(t, mu, sigma)
}

# Função de Risco Acumulado (Lambda)
LtLogNormal <- function(t, mu, sigma) {
  -log(StLogNormal(t, mu, sigma))
}

# ---------------------------------------
# [4.2] SIMULAÇÃO E VARIAÇÃO DE PARÂMETROS
# ---------------------------------------

# Simulando dados da distribuição log-normal
tempo <- rlnorm(n, meanlog = 0, sdlog = 1)
mus <- c(0, 0.5, 1) # Valores de mu
sigma <- 1          # Valor fixo de sigma

# Criando um Data Frame com valores das funções
dados <- do.call(rbind, lapply(mus, function(mu) {
  data.frame(
    tempo = tempo,
    ft = ftLogNormal(tempo, mu, sigma),
    St = StLogNormal(tempo, mu, sigma),
    ht = htLogNormal(tempo, mu, sigma),
```

```
    Lt = LtLogNormal(tempo, mu, sigma),
    mu = factor(mu)
  )
}))

# -------------
# [4.3] GRÁFICOS
# -------------

# Função genérica para gráficos
plot_func <- function(data, y_var, y_label, color_values, y_expression) {
  ggplot(data, aes(x = tempo, y = !!sym(y_var), color = mu)) +
    geom_line(stat = "summary", fun = mean, size = 1) +
    labs(x = "Tempo", y = y_expression,
         caption = "Fonte: Elaborado pelo autor", color = expression(mu)) +
    scale_color_manual(values = color_values,
                       labels = lapply(mus, function(m) bquote(mu == .(m)))) +
    theme_minimal(base_size = 14) +
  theme(
    plot.caption = element_text(hjust = 0.5, size = 10)
  )
}

# Paleta de cores
color_values <- c("red", "blue", "green")
```

Veja a seguir a Figura 3.7. Tal Figura mostra as curvas de sobrevivência para diferentes valores do parâmetro μ.

```
# Função de Sobrevivência
plot_func(dados, "St", expression(S(t)), color_values, expression(S(t)))
```

Figura 3.7: Função de Sobrevivência segundo uma Distribuição Log-Normal para diferentes valores do parâmetro de locação.

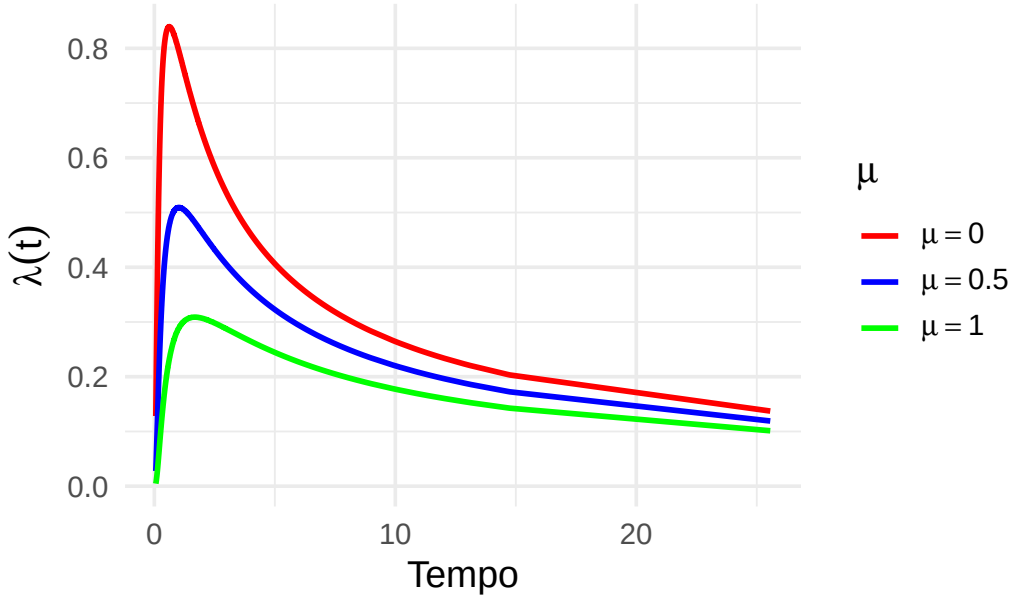


Fonte: Elaborado pelo autor

Veja a seguir a Figura 3.8. Tal Figura mostra a função de risco para diferentes valores do parâmetro μ.

```
# Função de Risco
plot_func(dados, "ht", expression(lambda(t)), color_values, expression(lambda(t)))
```

Figura 3.8: Função de Risco segundo uma Distribuição Log-Normal para diferentes valores do parâmetro de locação.

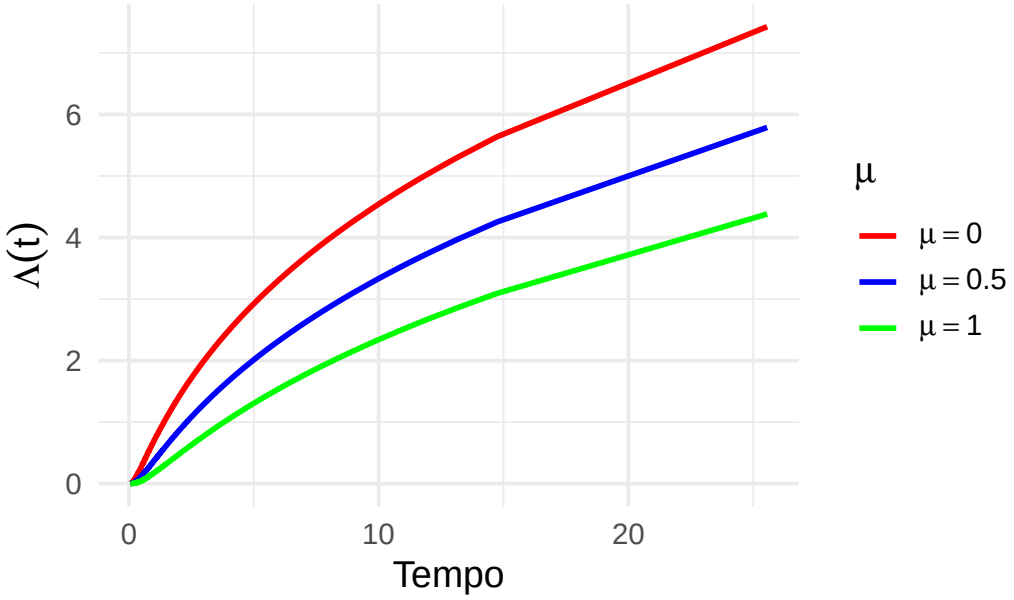


Fonte: Elaborado pelo autor

Veja a seguir a Figura 3.9. Tal Figura mostra a função de risco acumulado para diferentes valores do parâmetro μ.

```
# Função de Risco Acumulado
plot_func(dados, "Lt", expression(Lambda(t)), color_values, expression(Lambda(t)))
```

Figura 3.9: Função de Risco Acumulado segundo uma Distribuição Log-Normal para diferentes valores do parâmetro de locação.



Fonte: Elaborado pelo autor

3.2.3.1 Algumas Considerações

A média de T é dada por:

$$E[T] = \exp\{\mu + \frac{\sigma^2}{2}\},$$

e a variância de T definida como:

$$Var[T] = \exp\{2\mu + \sigma^2\}(\exp\{\sigma^2\} - 1).$$

3.3 Estimação

Foi mostrado alguns modelos probabilísticos. Porém, tais modelos apresentam quantidades desconhecidas denominadas de parâmetros ou parâmetro quando o modelo de probabilidade depende apenas de uma quantidade desconhecida, por exemplo, a distribuição exponencial.

3.3.1 Método de Máxima Verossimilhança

O *Método de Máxima Verossimilhança* se baseia na ideia de que, a partir de uma amostra aleatória, a estimativa para o parâmetro de interesse maximiza a probabilidade de tal amostra aleatória ser obtida.

Em termos simples, o método de máxima verossimilhança condensa toda informação contida, através da função de verossimilhança, na amostra. Afim de encontrar o(s) parâmetro(s) da distribuição que melhor expliquem essa amostra é realizado o produtório da densidade $(f(t))$ para cada observação amostral t_i, $i = 1, 2, \cdots, n$. Em livros de estatística básica, a seguinte definição da função verossimilhança é adotada, para um parâmetro (ou conjunto de parâmetros) θ qualquer:

$$L(\theta) = \prod_{i=1}^{n} f(t_i, \theta).$$

Perceba que L é função de θ, sendo este um único parâmetro ou um conjunto de parâmetros, como na distribuição log-normal, onde $\theta = (\mu, \sigma^2)$. Entretanto, nota-se que para qualquer observação não censurada, sua contribuição para $L(\theta)$ é a sua densidade, o que na análise de sobrevivência não é o suficiente, já que os dados apresentam censura, implicando no tempo de falha ser na verdade superior ao tempo de censura observado.

Desta forma, faz-se uso da variável indicadora δ_i, apresentada na Seção 1.5, que nos diz se o i-ésimo tempo é tempo de falha ou de censura. Logo, são feitos alguns ajustes na função de verossimilhança. Tais ajustes fazem com que para $\delta_i = 1$, o i-ésimo tempo é tempo de falha e a contribução para L é a própria função densidade de probabilidade, em contraste a isso, se $\delta_i = 0$, o i-ésimo tempo é tempo de censura e a contribuição dessa observação é a função de sobrevivência. Assim, a função de verossimilhança para um parâmetro ou um conjunto de parâmetros θ pode ser escrita como:

$$L(\theta) = \prod_{i=1}^{n} [f(t_i, \theta)]^{\delta_i} \ [S(t_i)]^{1-\delta_i} \tag{3.12}$$

A partir da deriva do log da verossimilhança igualada a zero,

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

é possível encontrar um valor para θ que maximize $\ln L(\theta)$, portanto, maximize $L(\theta)$.

3.3.1.1 Aplicações

Será mostrado nessa seção como encontrar o estimador ou estimadores de máxima verossimilhança para os parâmetros das disitribuições citadas.

3.3.1.1.1 Distribuição Exponecial

Para uma distribuição exponencial conforme descrita na Seção 3.2.1. O *Estimador de Máxima Verossimilhança* do parâmetro α pode ser obtido de acordo com os seguintes passos:

1. Determinar a função verossimilhança $L(\alpha)$:

$$\begin{aligned} L(\alpha) &= \prod_{i=1}^{n} [\alpha \exp\{-\alpha t_i\}]^{\delta_i} [\exp\{-\alpha t_i\}]^{1-\delta_i} \\ &= \prod_{i=1}^{n} \alpha^{\delta_i} \exp\{-\alpha t_i\}. \end{aligned}$$

2. Tomar o logaritmo da função verossimilhança $\ln L(\alpha)$:

$$\begin{aligned} \ln L(\alpha) &= \sum_{i=1}^{n} \ln \left[\alpha^{\delta_i} \exp\{-\alpha t_i\}\right] = \sum_{i=1}^{n} \ln \left[\alpha^{\delta_i}\right] + \sum_{i=1}^{n} \ln \left[\exp\{-\alpha t_i\}\right] \\ &= \sum_{i=1}^{n} \delta_i \ln \alpha + \sum_{i=1}^{n} -\alpha t_i = \ln \alpha \sum_{i=1}^{n} \delta_i - \alpha \sum_{i=1}^{n} t_i. \end{aligned}$$

3. Derivar a função do log da verossimilhança $\dfrac{\partial \ln L(\theta)}{\partial \theta} = 0$:

$$\frac{\partial \ln L(\theta)}{\partial \theta} = \frac{1}{\alpha} \sum_{i=1}^{n} \delta_i - \sum_{i=1}^{n} t_i.$$

Ao igualar a derivada a zero e isolando α em um dos lados da igualdade e assuminado que a forma análitica de α obtida é um estimador de máxima verossimilhança temos:

$$\begin{aligned} \frac{\partial \ln L(\theta)}{\partial \theta} &= 0 \\ \frac{1}{\hat{\alpha}} \sum_{i=1}^{n} \delta_i - \sum_{i=1}^{n} t_i &= 0 \\ \hat{\alpha} &= \frac{\sum_{i=1}^{n} \delta_i}{\sum_{i=1}^{n} t_i} \end{aligned}$$

Note que, para o caso em que não se tem censura o numerador, $\sum_{i=1}^{n} \delta_i$, equivale ao tamanho da amostra n.

A seguir, temos um exemplo computacional. Simulou-se uma amostra proveniente de uma distribuição exponecial e a partir de tal amostra se obteve a estimativa de máxima verossimilhança do parâmetro α de tal amostra.

```
set.seed(123)
n <- 1000

# -------------
# [2.1] FUNÇÕES
# -------------
# Função de Sobrevivência
Stexp <- function(t, alpha) exp(-alpha * t)

# ---------------------------------------
# [2.2] SIMULAÇÃO E VARIAÇÃO DE PARÂMETROS
# ---------------------------------------

alpha <- 1.5 # Parâmetro de taxa (escala) populacional
tempo <- rexp(n, rate = alpha) # Simulando dados de uma exponencial

emvExp <- n / sum(tempo) # EMV

# Criando um Data Frame com valores das funções
dados <- data.frame(
  Tempo = tempo,
  St = Stexp(tempo, alpha),
  emvSt = Stexp(tempo, emvExp)
)
```

O valor de verdadeiro do parâmetro $\alpha = 1.5$. A estimativa de máxima verossimilhança obtida foi $\hat{\alpha} = 1.46$.

Veja a Tabela 3.1 que mostra as dez primeiras observações e suas respectivas funções de sobrevivência, sobrevivência real e sobrevivência estimada.

```
library(kableExtra)
```

```
Anexando pacote: 'kableExtra'
```

```
O seguinte objeto é mascarado por 'package:dplyr':
```

```
    group_rows
```

```r
knitr::kable(
  head(dados),
  col.names = c("Tempo", "S(t)", "S(t) EMV"),
  escape = FALSE,
  align = 'c',
  booktabs = TRUE
)
```

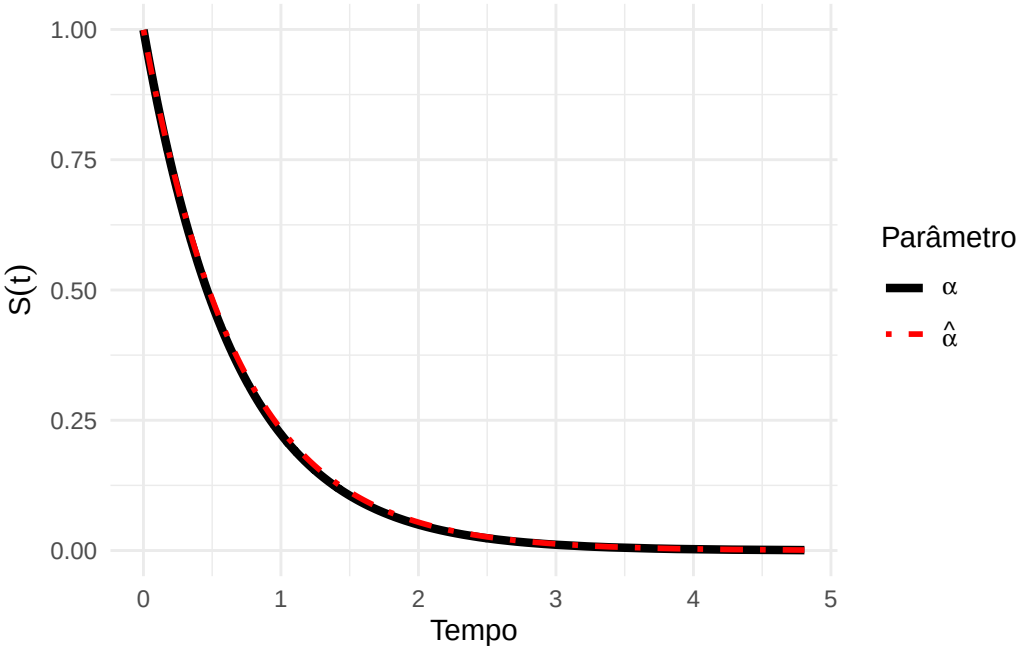
Tabela 3.1: Real e Estimado para as Funções de Sobrevivência

Tempo	S(t)	S(t) EMV
0.5623048	0.4302206	0.4409133
0.3844068	0.5617995	0.5713079
0.8860366	0.2647273	0.2751688
0.0210516	0.9689160	0.9698070
0.0374740	0.9453397	0.9468876
0.2110008	0.7286941	0.7354381

Temos também a comparação dessas duas curvas de sobrevivência, ilustradas na Figura 3.10.

```r
ggplot(dados, aes(x = Tempo)) +
  geom_line(aes(y = St, color = "alpha"), lwd = 1.5) +
  geom_line(aes(y = emvSt, color = "emvAlpha"), lwd = 1, lty = 4) +
  scale_color_manual(
    values = c("alpha" = "black", "emvAlpha" = "red"),
    labels = c(expression(alpha), expression(hat(alpha)))) +
  labs(
    x = "Tempo",
    y = expression(S(t)),
    color = "Parâmetro") +
  theme_minimal()
```

Figura 3.10: Comparação do verdadeiro valor do parâmetro com sua estimativa de máxima verossimilhança.



3.3.1.1.2 Distribuição Weibull

Para uma distribuição Weibull, descrita na Seção 3.2.2 não há uma forma análitica para γ e α. Logo, para obter a sua estimativa de máxima verossimilhança se usa um método de aproximação numérica, será introduzido aqui o *Método Iterativo de Newton-Raphson*.

O **Método de Newton-Raphson** é uma abordagem iterativa eficiente para resolver equações não lineares, sendo amplamente utilizado na estimação de parâmetros de distribuições estatísticas. Quando aplicado ao ajuste de distribuições, como a **Weibull** no contexto de análise de sobrevivência, o método busca maximizar a função de verossimilhança, resolvendo o sistema de equações derivado das condições de otimalidade (gradiente nulo).

A fórmula iterativa é:

$$\theta_{n+1} = \theta_n - \mathbf{H}^{-1}(\theta_n)\nabla L(\theta_n),$$

onde:

- θ_n é o vetor de parâmetros estimados na iteração n;
- $L(\theta)$ é a função log-verossimilhança;
- $\nabla L(\theta)$ é o vetor gradiente (derivadas parciais de $L(\theta)$);
- $\mathbf{H}(\theta)$ é a matriz Hessiana (segunda derivada de $L(\theta)$).

Vantagens no ajuste de distribuições:

- **Eficiência**: O método converge rapidamente quando o ponto inicial θ_0 está próximo dos valores reais dos parâmetros.
- **Flexibilidade**: Adequa-se a diferentes modelos, como a distribuição Weibull, usada para modelar tempos de vida ou sobrevivência.

Cuidados na aplicação:

- **Convergência**: Garantida apenas se o ponto inicial estiver próximo da solução e as condições de regularidade forem atendidas.
- **Cálculo da Hessiana**: Pode ser computacionalmente intensivo para distribuições complexas.

No caso da distribuição Weibull, a aplicação do método Newton-Raphson envolve derivadas em relação aos parâmetros de forma (γ) e escala (α), permitindo ajustar o modelo aos dados observados de tempos de sobrevivência de forma precisa e eficiente.

Para utilizar o Método Iterativo de Newton-Raphson, pode-se escrever o algoritmo passo a passo. Outra forma, é usar a função do R `optim`. Será aprensatado as duas formas e seus detalhes serão comentados.

Começando pela construção do algoritmo passo a passo, precisamos definir algumas funções. A primeira é a função de verossimilhança da distribuição Weibull, que pode ser obtida a partir da Equação 3.12 ao substituir respectivamente a função densidade e sobrevivência da distribuição Weibull respectivamente. Assim:

$$\begin{aligned} L(\gamma,\alpha) &= \prod_{i=1}^{n}\left[\gamma\alpha^{\gamma}t_i^{\gamma-1}\exp\left\{-\left(\frac{t_i}{\alpha}\right)^{\gamma}\right\}\right]^{\delta_i}\left[\exp\left\{-\left(\frac{t_i}{\alpha}\right)^{\gamma}\right\}\right]^{1-\delta_i} \\ &= \prod_{i=1}^{n}\left[\frac{\gamma}{\alpha^{\gamma}}t^{\gamma-1}\right]^{\delta_i}\exp\left\{-\left(\frac{t_i}{\alpha}\right)^{\gamma}\right\} \end{aligned}$$

Toma-se o logaritmo de $L(\gamma,\alpha)$, logo:

$$\begin{aligned} \ln L(\gamma,\alpha) &= \sum_{i=1}^{n}\delta_i\ln\{\gamma\} - \sum_{i=1}^{n}\delta_i\gamma\ln\{\alpha\} + \sum_{i=1}^{n}\delta_i(\gamma-1)\ln\{t_i\} + \sum_{i=1}^{n}-(\alpha^{-1}t_i)^{\gamma} \\ &= \ln\{\gamma\}\sum_{i=1}^{n}\delta_i - \gamma\ln\{\alpha\}\sum_{i=1}^{n}\delta_i + (\gamma-1)\sum_{i=1}^{n}\delta_i\ln\{t_i\} + \sum_{i=1}^{n}-(\alpha^{-1}t_i)^{\gamma} \end{aligned}$$

Agora, aplica-se as derivadas de primeira ordem em relação a γ e α.

$$\frac{\partial\ln L(\gamma,\alpha)}{\partial\gamma} = \frac{\sum_{i=1}^{n}\delta_i}{\gamma} - \ln\{\alpha\}\sum_{i=1}^{n}\delta_i + \sum_{i=1}^{n}\delta_i\ln\{t_i\} - \sum_{i=1}^{n}(\alpha^{-1}t_i)^{\gamma}\ln\{(\alpha^{-1}t_i)^{\gamma}\}$$

$$\frac{\partial \ln L(\gamma,\alpha)}{\partial \alpha} = -\frac{\gamma \sum_{i=1}^{n} \delta_i}{\alpha} + \gamma \alpha^{-\gamma-1} \sum_{i=1}^{n} t_i^{\gamma}$$

Toma-se agora as derivadas de segunda ordem.

$$\frac{\partial^2 \ln L(\gamma,\alpha)}{\partial \gamma^2} = \frac{\sum_{i=1}^{n} \delta_i}{\gamma^2} - \sum_{i=1}^{n} (\alpha^{-1} t_i)^{\gamma} [\ln\{(\alpha^{-1} t_i)^{\gamma}\}]^2$$

$$\frac{\partial^2 \ln L(\gamma,\alpha)}{\partial \alpha^2} = \frac{\gamma \sum_{i=1}^{n} \delta_i}{\alpha^2} - \gamma(\gamma+1)\alpha^{-\gamma-2} \sum_{i=1}^{n} t_i^{\gamma}$$

$$\frac{\partial^2 \ln L(\gamma,\alpha)}{\partial \gamma \partial \alpha} = -\frac{\sum_{i=1}^{n} \delta_i}{\alpha} + \sum_{i=1}^{n} \frac{t_i^{\gamma}}{\alpha^{\gamma+1}} \left(\gamma \ln \left\{ \frac{t_i}{\alpha} \right\} + 1 \right)$$

Com todas as derivadas definidas, é possível definirmos algumas funções e variáveis que iremos precisar para utilizar o algoritmo de Newton-Raphson.

```
cat("Era para ser um bloco de código. Porém, tal código está em manutenção")
```

```
Era para ser um bloco de código. Porém, tal código está em manutenção
```

O bloco de código abaixo contém o algoritmo de Newton-Raphson.

```
cat("Era para ser um bloco de código. Porém, tal código está em manutenção")
```

```
Era para ser um bloco de código. Porém, tal código está em manutenção
```

O bloco abaixo mostra o uso da função `optim` para otimização.

```
# ----------
# Otimização
# ----------

# Semente
set.seed(123)

# --------------------------------
# Simulação e Visualização do Dados
# --------------------------------
# Tamanho da amostra
```

```r
n <- 1000

# Parâmetros da distribuição
wShape <- 2
wScale <- 1.5

# Simulação
dadosWeibull <- rweibull(n, shape = wShape, scale = wScale)

# --------------------------
# Função Log-verossimilhança
# --------------------------

logWeibull <- function(theta, dados){
  gamma <- theta[1] # Parâmetro de forma
  alpha <- theta[2] # Parâmetro de escala
  n <- length(dados)
  t <- dados

logverossimil <- (n * log(gamma)) - (gamma * log(alpha) * n) + (gamma - 1) * sum(log(t))
return(-logverossimil)
}

# ------------------------
# Aplicando a função optim
# ------------------------

theta0 <- c(1.5, 1) # Chute inicial
estimate <- optim(par = theta0, fn = logWeibull, gr = NULL , method = "BFGS" ,
             hessian = TRUE, dados=dadosWeibull)
estimate
```

```
$par
[1] 2.015515 1.507207

$value
[1] 999.1172

$counts
function gradient
      52        9

$convergence
[1] 0
```

```
$message
NULL

$hessian
          [,1]     [,2]
[1,]  451.7492 -281.952
[2,] -281.9520 1788.248
```

Assim como na distribuição exponencial, será feita uma comparação entre o real e estimado.

```r
# ---------------------
# AJUSTES DE FORMATAÇÃO
# ---------------------

# Função de Sobrevivência
StWeibull <- function(t, gamma, alpha) exp(-(t/alpha)^gamma)

# Data Frame
dfWeibull <- data.frame(
  Tempo = dadosWeibull,
  St = StWeibull(dadosWeibull, wShape, wScale),
  EMVSt = StWeibull(dadosWeibull, estimate$par[1], estimate$par[2])
)
```

Veja a Tabela 3.2 que mostra as dez primeiras observações e suas respectivas funções de sobrevivência, sobrevivência real e sobrevivência estimada.

```r
library(knitr)
library(kableExtra)

knitr::kable(
  head(dfWeibull),
  col.names = c("Tempo", "S(t)", "S(t) EMV"),
  escape = FALSE,
  align = 'c',
  booktabs = TRUE
)
```

Tabela 3.2: Real e Estimado para as Funções de Sobrevivência da Distribuição Weibull

Tempo	S(t)	S(t) EMV
1.6745421	0.2875775	0.2904304

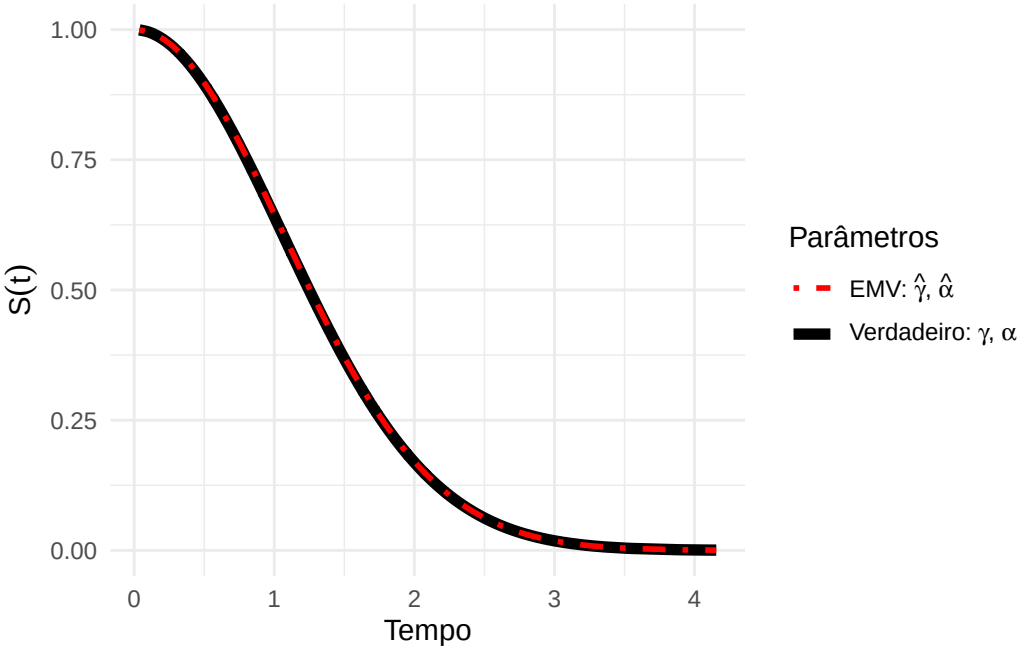
Tabela 3.2: Real e Estimado para as Funções de Sobrevivência da Distribuição Weibull

Tempo	S(t)	S(t) EMV
0.7315788	0.7883051	0.7921747
1.4183502	0.4089769	0.4128247
0.5290778	0.8830174	0.8858229
0.3716200	0.9404673	0.9422483
2.6362482	0.0455565	0.0456849

Temos também a comparação dessas duas curvas de sobrevivência, ilustradas na Figura 3.11.

```
library(ggplot2)

ggplot(dfWeibull, aes(x = Tempo)) +
  geom_line(aes(y = St, color = "Verdadeiro"), lwd = 2) +
  geom_line(aes(y = EMVSt, color = "EMV"), lwd = 1, lty = 4) +
  scale_color_manual(
    values = c("Verdadeiro" = "black", "EMV" = "red"),
    labels = c(
      "Verdadeiro" = expression(paste("Verdadeiro: ", gamma, ", ", alpha)),
      "EMV" = expression(paste("EMV: ", hat(gamma), ", ", hat(alpha)))
    )) +
  labs(
    x = "Tempo",
    y = expression(S(t)),
    color = "Parâmetros") +
  theme_minimal()
```

Figura 3.11: Comparação do verdadeiro valor dos parâmetros e com suas estimativas de máxima verossimilhança.



3.3.1.1.3 Distribuição Log-Normal

3.3.1.2 Aplicações com Censura

Podemos utilizar os conhecimentos até aqui obtidos para modelar e estimar parâmetros, e consequentemente, curvas de sobrevivência onde exista censura. Vejamos os blocos de códigos a seguir.

1. Ativamos os pacotes necessários:

```
# ----------------------
# [1] Ativação de Pacotes
# ----------------------

library(survival)
library(ggplot2)
```

2. Simulamos os dados com censura:

```r
# ----------------------
# [2] Simulação dos Dados
# ----------------------

# Definindo Semente
set.seed(123)

# Parâmetros
n <- 1000       # Número total de observações
gamma <- 2      # Parâmetro shape da Weibull
alpha <- 1.5    # Parâmetro scale da Weibull
TaxaExp <- 1    # Taxa da distribuição exponencial
propCens <- 1/3 # Proporção desejada de censuras

# Vetores para armazenar os resultados
Tobservado <- numeric(n)
indCensura <- numeric(n)

# Contadores
nFalhas <- 0
nCensuras <- 0

# Loop para gerar os tempos
for (i in 1:n) {
  # Gerar um tempo de falha e um tempo de censura
  Tfalha <- rweibull(1, shape = gamma, scale = alpha)
  Tcensu <- rexp(1, rate = TaxaExp)

  # Verificar qual é o menor tempo
  if (Tfalha <= Tcensu) {
    Tobservado[i] <- Tfalha
    indCensura[i] <- 1 # Falha
    nFalhas <- nFalhas + 1
  } else {
    if (nCensuras < propCens * n) {
      Tobservado[i] <- Tcensu
      indCensura[i] <- 0 # Censura
      nCensuras <- nCensuras + 1
    } else {
      Tobservado[i] <- Tfalha
      indCensura[i] <- 1 # Falha
      nFalhas <- nFalhas + 1
    }
  }
```

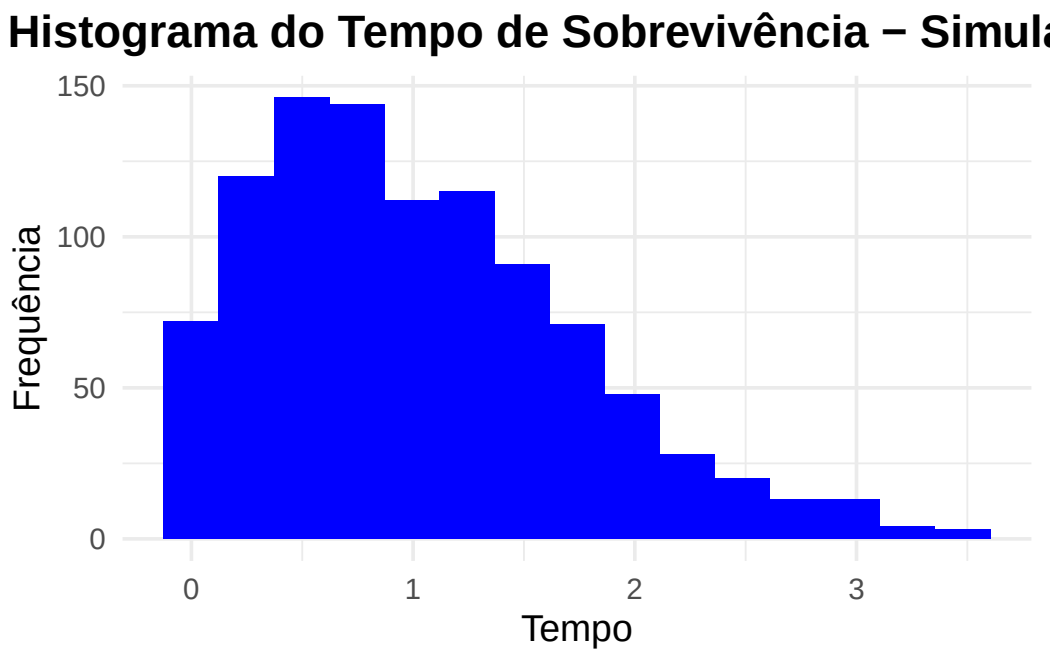
```
}

# Verificar as proporções
cat("Proporção de falhas:", nFalhas / n, "\n")
```

```
Proporção de falhas: 0.666
```

```
cat("Proporção de censuras:", nCensuras / n, "\n")
```

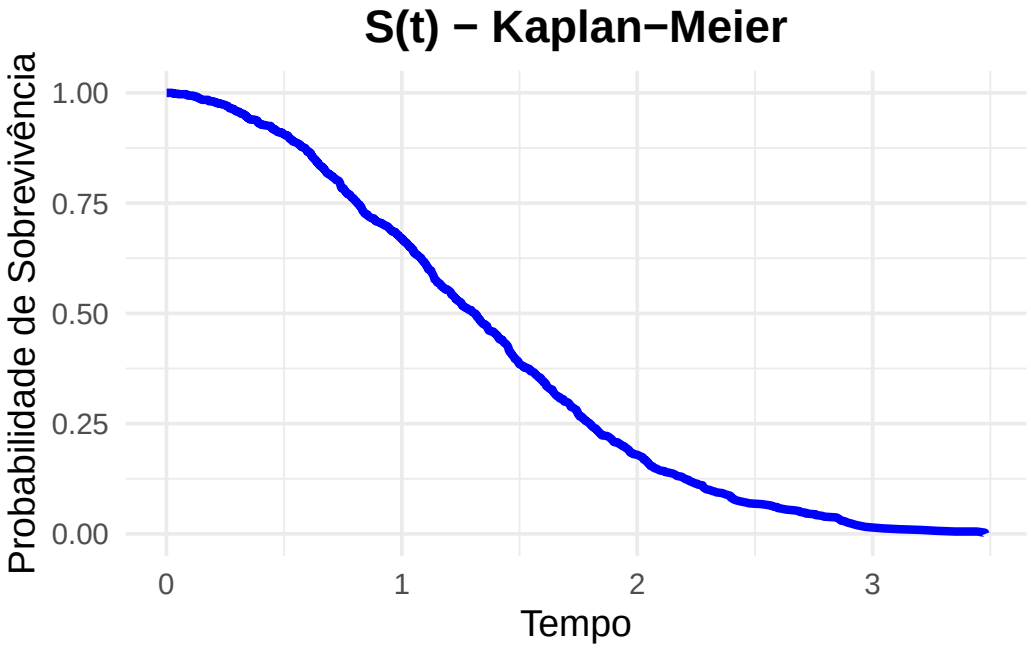
```
Proporção de censuras: 0.334
```

```
# Dados simulados
dados <- data.frame(Tempo = Tobservado, Censura = indCensura)

# Histograma
ggplot(data = dados, aes(x = Tempo)) +
  geom_histogram(bins = 15, fill = "blue") +
  labs(title = "Histograma do Tempo de Sobrevivência - Simulação",
       x = "Tempo", y = "Frequência") +
  theme_minimal(base_size = 14) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
```



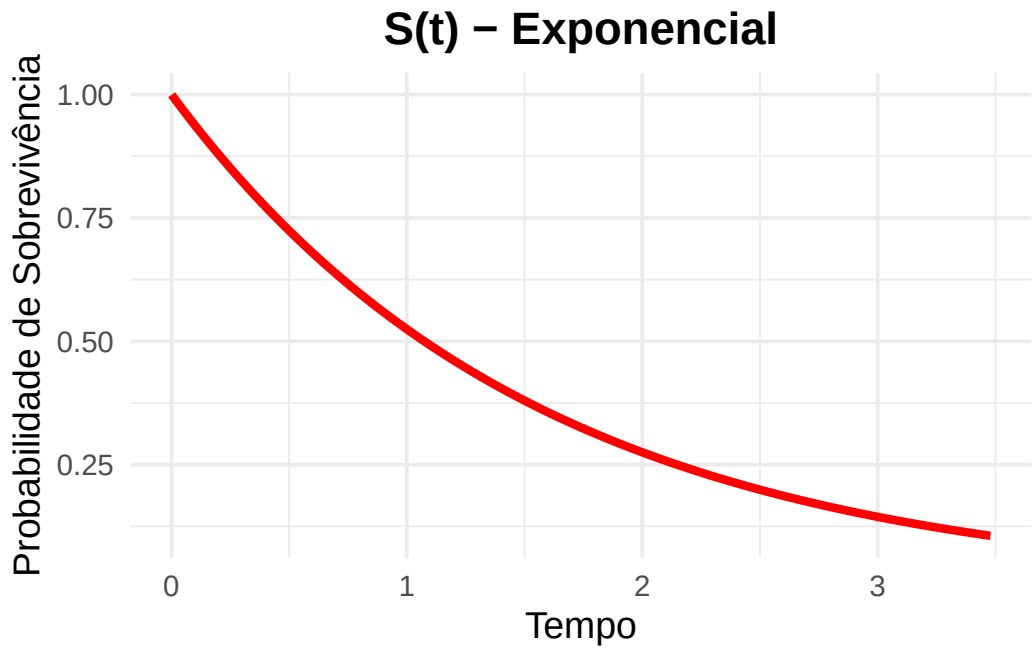
3. Estimação Não Paramétrica.

```r
# -------------
# [3] Estimação
# -------------

# -------------------------------
# [3.1] Estimação Não Paramátrica
# -------------------------------
# --------------------------------
# [3.1.1] Estimador de Kaplan-Meier
# --------------------------------

# Modelo de Kaplan-Meier
ekm <- survfit(Surv(Tempo, Censura) ~ 1, data = dados)

# Formatando como DataFrame
DataEKM <- data.frame(Time = ekm$time, Survival = ekm$surv, Type = "Kaplan-Meier")

# Visualização da curva
ggplot(data = DataEKM, aes(x = Time, y = Survival)) +
  geom_line(color = "blue", lwd = 1.5) +
  labs(title = "S(t) - Kaplan-Meier",
       x = "Tempo", y = "Probabilidade de Sobrevivência") +
  theme_minimal(base_size = 14) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
```



4. Estimação Não Paramétrica - Distribuição Exponencial.

```r
# --------------------------
# [3.2] Estimação Paramátrica
# --------------------------
# -------------------------------
# [3.2.1] Distribuição Exponencial
# -------------------------------

# Função de Sobrevivência
Stexp <- function(t, alpha) exp(-alpha * t)

# EMV de
emvExp <- sum(dados$Censura) / sum(dados$Tempo)

# EMV da Sobrevivência
EMVSurvExp <- Stexp(dados$Tempo, emvExp)

# Formatando como DataFrame
DataExp <- data.frame(Time = dados$Tempo, Survival = EMVSurvExp, Type = "Exponencial")

# Visualização da curva
ggplot(data = DataExp, aes(x = Time, y = Survival)) +
  geom_line(color = "red", lwd = 1.5) +
  labs(title = "S(t) - Exponencial",
       x = "Tempo", y = "Probabilidade de Sobrevivência") +
  theme_minimal(base_size = 14) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
```



5. 4. Estimação Não Paramétrica - Distribuição Weibull.

::: {.cell}

```
# ----------------------------
# [3.2.1] Distribuição Weibull
# ----------------------------

# Função de Sobrevivência
StWeibull <- function(t, gamma, alpha) exp(-(t / alpha)^gamma)

# EMV de   e
# 1. Função Log-verossimilhança
logWeibull <- function(theta, dados){
  gamma <- theta[1] # Parâmetro de forma
  alpha <- theta[2] # Parâmetro de escala

  t <- dados$Tempo   # Tempo de falha
  c <- dados$Censura # Variável indicadora

  logv <- (sum(c) * log(gamma)) - (gamma * log(alpha) * sum(c)) +
(gamma - 1) * sum(c * log(t)) - sum((t / alpha)^gamma)
  return(-logv)
}

# 2. Otimizando
Theta0 <- c(1.5, 1)
estimate <- optim(par = Theta0, fn = logWeibull,
                  gr = NULL, method = "BFGS", hessian = TRUE, dados = dados)
```

:::

::: {.cell-output .cell-output-stderr}

```
Warning in log(gamma): NaNs produzidos
Warning in log(gamma): NaNs produzidos
Warning in log(gamma): NaNs produzidos
```

:::

::: {.cell-output .cell-output-stderr}

```
Warning in log(alpha): NaNs produzidos
Warning in log(alpha): NaNs produzidos
Warning in log(alpha): NaNs produzidos
Warning in log(alpha): NaNs produzidos
Warning in log(alpha): NaNs produzidos
Warning in log(alpha): NaNs produzidos
```

:::

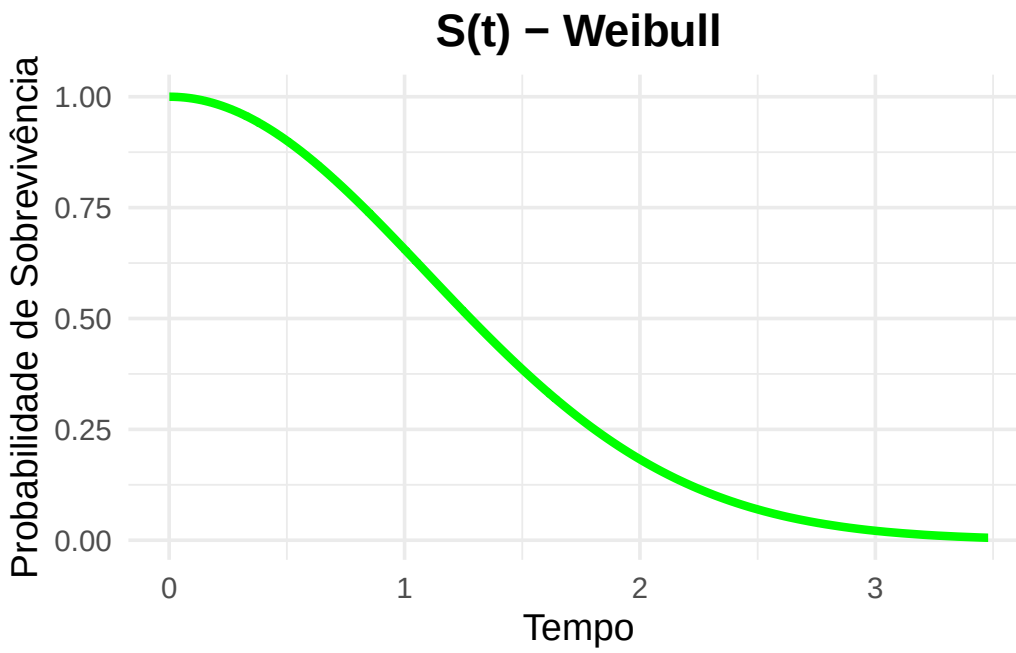
```
estimate
```

::: {.cell-output .cell-output-stdout}

```
$par
[1] 2.013314 1.535915

$value
[1] 745.9756

$counts
function gradient
      31        9

$convergence
[1] 0

$message
NULL

$hessian
          [,1]       [,2]
[1,] 304.76432  -96.91392
[2,] -96.91392 1144.24624
```

:::

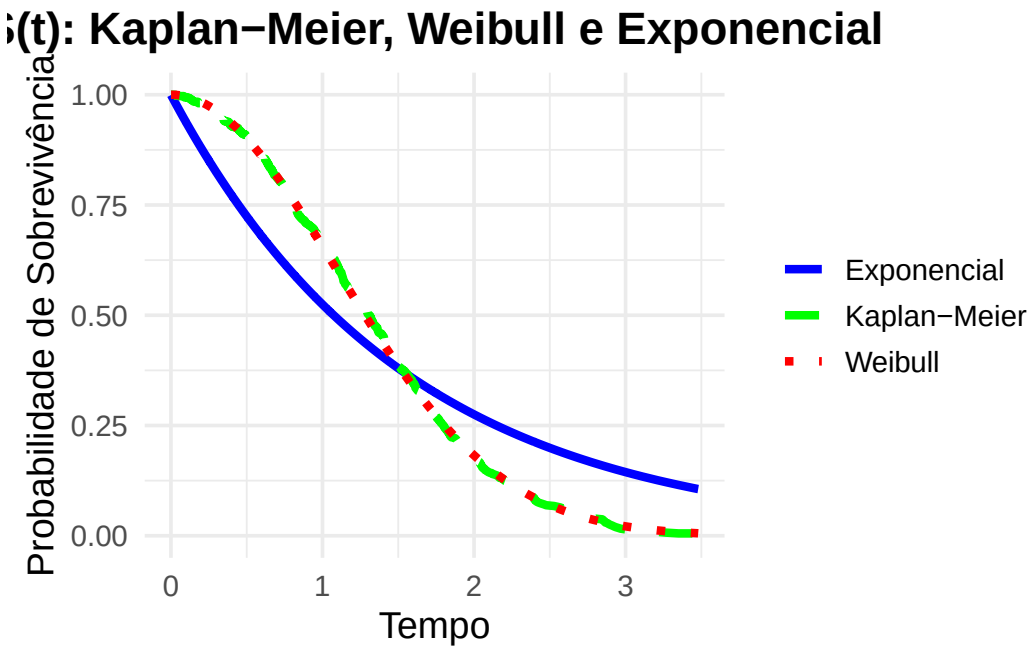
```
# EMV da Sobrevivência
EMVSurvWeib <- StWeibull(dados$Tempo, estimate$par[1], estimate$par[2])

# Formatando como DataFrame
DataWeib <- data.frame(Time = dados$Tempo, Survival = EMVSurvWeib, Type = "Weibull"

# Visualização da curva
ggplot(data = DataWeib, aes(x = Time, y = Survival)) +
  geom_line(color = "green", lwd = 1.5) +
  labs(title = "S(t) - Weibull",
   x = "Tempo", y = "Probabilidade de Sobrevivência") +
  theme_minimal(base_size = 14) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"))
```



::: {.cell-output-display}
::: :::

Por fim, com o objetivo de fazer uma comparação, visual, temos a Figura abaixo.

```
# --------------------
# [4] Análise Conjunta
# --------------------

# Unindo os dados para visualização
AllData <- rbind(DataEKM, DataExp, DataWeib)

# Gráfico com ggplot2
ggplot(AllData, aes(x = Time, y = Survival)) +
  geom_line(aes(color = Type, linetype = Type), lwd = 1.5) +
  labs(title = "S(t): Kaplan-Meier, Weibull e Exponencial",
       x = "Tempo", y = "Probabilidade de Sobrevivência") +
  theme_minimal(base_size = 14) +
  theme(plot.title = element_text(hjust = 0.5, face = "bold"),
        legend.title = element_blank()) +
  scale_color_manual(values = c("blue", "green", "red")) +
  scale_linetype_manual(values = c(1, 2, 3)) # c("solid", "dashed", "dotted")
```


Ŝ(t): Kaplan−Meier, Weibull e Exponencial
Probabilidade de Sobrevivência
1.00
0.75
0.50
0.25
0.00
0
1
2
3
Tempo
Exponencial
Kaplan−Meier
Weibull

Referêcias

Aalen, Odd O. 1978. «Nonparametric Inference for a Family of Counting Processes». *Annals of Statistics* 6 (4): 701–26. https://doi.org/10.1214/aos/1176344247.

Aalen, Odd O., e Søren Johansen. 1978. «An Empirical Transition Matrix for Non-Homogeneous Markov Chains Based on Censored Observations». *Scandinavian Journal of Statistics* 5 (3): 141–50.

Bohoris, G. A. 1994. «Comparison of the Cumulative-Hazard and Kaplan-Meier Estimators of the Survivor Function». *IEEE Transactions on Reliability* 43 (2): 230–32. https://doi.org/10.1109/24.293488.

Breslow, Norman, e John Crowley. 1974. «A Large Sample Study of the Life Table and Product Limit Estimates under Random Censorship». *The Annals of Statistics* 2 (3): 437–53. https://doi.org/10.1214/aos/1176342705.

Colosimo, Enrico Antonio, e Suely Ruiz Giolo. 2006. *Análise de Sobrevivência Aplicada*. 1.ª ed. São Paulo, Brasil: Blucher.

Gehan, Edmund A. 1965. «A Generalized Wilcoxon Test for Comparing Arbitrarily Singly-Censored Samples». *Biometrika* 52 (1-2): 203–24. https://doi.org/10.2307/2333825.

Kalbfleisch, John D., e Ross L. Prentice. 1980. *The Statistical Analysis of Failure Time Data*. Wiley Series em Probability e Mathematical Statistics. New York: Wiley.

Kaplan, Edward L., e Paul Meier. 1958. «Nonparametric Estimation from Incomplete Observations». *Journal of the American Statistical Association* 53 (282): 457–81. https://doi.org/10.1080/01621459.1958.10501452.

Klein, John P. 1991. «Small Sample Moments of Some Estimators of the Variance of the Kaplan-Meier and Nelson-Aalen Estimators». *Scandinavian Journal of Statistics* 18 (4): 333–40. https://doi.org/10.2307/4616203.

Latta, Robert B. 1981. «A Monte Carlo Study of Some Two-Sample Rank Tests with Censored Data». *Journal of the American Statistical Association* 76 (375): 713–19. https://doi.org/10.2307/2287572.

Lawless, J. F. 1982. *Statistical Models and Methods for Lifetime Data*. Wiley Series em Probability e Statistics. New York: John Wiley & Sons.

Lindsey, Jane C., e Louise M. Ryan. 1998. «Methods for Interval-Censored Data». *Statistics in Medicine* 17 (2): 219–38. https://doi.org/10.1002/(SICI)1097-0258(19980130)17:2%3C219::AID-SIM735%3E3.0.CO;2-D.

Mantel, Nathan. 1966. «Evaluation of Survival Data and Two New Rank Order Statistics Arising in Its Consideration». *Cancer Chemotherapy Reports* 50 (3): 163–70.

Mantel, Nathan, e William Haenszel. 1959. «Statistical Aspects of the Analysis of Data from Retrospective Studies of Disease». *Journal of the National Cancer Institute* 22 (4): 719–48.

Meier, Paul. 1975. «Estimation of a Survival Curve from Incomplete Data». *Journal of the American Statistical Association* 70 (351): 607–10. https://doi.org/10.1080/01621459.1975.10479872.

Nelson, Wayne. 1972. «Theory and Applications of Hazard Plotting for Censored Failure Data». *Technometrics* 14 (4): 945–66. https://doi.org/10.1080/00401706.1972.10488981.

Peto, Richard, e Julian Peto. 1972. «Asymptotically Efficient Rank Invariant Test Procedures». *Journal of the Royal Statistical Society: Series A (General)* 135 (2): 185–98. https://doi.org/10.2307/2344317.

Prentice, Ross L. 1978. «Linear Rank Tests with Right Censored Data». *Biometrika* 65 (1): 167–79. https://doi.org/10.2307/2335206.

Turnbull, Bruce W. 1974. «Nonparametric Estimation of a Survivorship Function with Doubly Censored Data». *Journal of the American Statistical Association* 69 (345): 169–73. https://doi.org/10.1080/01621459.1974.10480146.