

Análise de Sobrevivência

Iniciação Científica - PIBIC 2024/2025 (UFPA)

Breno Cauã Rodrigues da Silva

2024-11-12

Índice

Prefácio	3
1 Conceitos Básicos e Exemplos	4
1.1 Introdução	4
1.2 Tempo de Falha	4
1.3 Censura	5
1.4 Dados Truncados	6
1.5 Representação dos Dados de Sobrevivência	6
1.6 Especificando o Tempo de Sobrevivência	7
1.6.1 Função de Sobrevivência	7
1.6.2 Função de Taxa de Falha ou de Risco	7
1.6.3 Função de Taxa de Falha Acumulada	8
1.6.4 Tempo Médio e Vida Média Residual	8
1.7 Relações entre as Funções	9
2 Técnicas Não Paramétricas	10
2.1 Introdução	10
2.2 O Estimador de Kaplan-Meier	10
2.2.1 Propriedades do Estimador de Kaplan-Meier	12
2.2.2 Variância do Estimador de Kaplan-Meier	12
2.3 Outros Estimadores Não Paramétricos	15
2.3.1 Estimador de Nelson-Aalen	15
2.4 Testes de Hipóteses para Comparação de Curvas de Sobrevivência	15
3 Técnicas Paramétricas - Modelos Probabilísticos	16
3.1 Introdução	16
3.2 Distribuições do Tempo de Sobrevivência	16
3.2.1 Distribuição Exponencial	17
3.2.2 Distribuição Weibull	22
3.2.3 Distribuição lognormal	28
3.3 Estimação	33
3.3.1 Método de Máxima Verossimilhança	34

Prefácio

Este é um projeto desenvolvido...

1 Conceitos Básicos e Exemplos

1.1 Introdução

O primeiro capítulo do livro de Enrico Antônio Colosimo e Suely Ruiz Giolo tem como objetivo apresentar alguns *conceitos e fundamentos* de uma das áreas da Estatística e Análise de Dados que mais cresceram nas últimas duas décadas do século passado. Esse crescimento foi impulsionado pelo desenvolvimento e avanço de técnicas, juntamente com o progresso computacional.

Na Análise de Sobrevivência, a variável resposta é, geralmente, o tempo até a ocorrência de um determinado evento. De forma mais precisa, trata-se de uma técnica estatística utilizada para modelar e entender o tempo até que ocorra um evento de interesse, denominado **tempo de falha**. Para um entendimento inicial, Colosimo e Giolo dão os seguintes exemplos: tempo até a morte de um paciente, tempo até a cura ou até a recidiva de uma doença.

Uma questão que pode surgir é: por que não usar outras técnicas estatísticas? O uso de outras abordagens não é adequado para dados de sobrevivência devido à característica desses dados, que é a presença de **censura**. De forma simples, censura refere-se à observação parcial da resposta, o que ocorre quando o acompanhamento do paciente é interrompido por alguma razão. Sendo um conceito chave na análise de sobrevivência, podemos defini-la como a situação em que o tempo de falha real não é conhecido, apenas que ele excede certo ponto.

1.2 Tempo de Falha

Em Análise de Sobrevivência, é essencial definir alguns pontos fundamentais para o estudo. O primeiro deles é o tempo de início do estudo, que deve ser definido com precisão, garantindo que os indivíduos sejam comparáveis na origem do estudo, diferenciando-se apenas nas medidas das covariáveis. Existem muitas alternativas para definir o tempo inicial. Geralmente, esse tempo é o tempo real ou “de relógio”. Porém, em outras áreas, como a Engenharia, outras medidas podem ser utilizadas. Colosimo e Giolo fornecem exemplos como número de ciclos, quilometragem de um carro ou qualquer outra medida de carga.

Outro ponto importante relacionado ao Tempo de Falha é a definição do evento de interesse. Normalmente, esses eventos correspondem a situações indesejáveis, por isso são chamados de falhas. A definição da falha deve ser clara e precisa. Destaca-se um trecho do livro:

“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 inadequado para consumo quando atinge uma concentração específica de microrganismos por mm^2 de área.”

1.3 Censura

Frequentemente, estudos clínicos que assumem a resposta como uma variável temporal são prospectivos e de longa duração. Mesmo sendo longos, esses estudos costumam terminar antes que todos os indivíduos venham a falhar.

Uma característica comum a esses estudos é a presença de **censura**, ou seja, observações incompletas ou parciais. É importante ressaltar que, mesmo censuradas, essas observações fornecem informações valiosas sobre o tempo de vida dos pacientes. Colosimo e Giolo destacam a importância de manter os 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.”

São apresentados três tipos principais de censura:

- **Censura Tipo I:** O estudo é encerrado após um período de tempo pré-estabelecido.
- **Censura Tipo II:** O estudo termina quando o evento de interesse ocorre em um número específico de indivíduos.
- **Censura Aleatória:** Ocorre quando um paciente é retirado do estudo antes do evento.

No livro, a Figura 1.1 ilustra esses tipos de censura, todos conhecidos como censura à direita, pois o evento ocorre após o tempo registrado. Contudo, outros tipos de censura, como à esquerda e intervalar, também são possíveis.

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

“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, há censura à direita para crianças que não sabiam ler quando os dados foram coletados. Neste caso, os tempos de vida são considerados duplamente censurados (Turnbull, 1974).

De forma geral, a censura intervalar ocorre em estudos com visitas periódicas espaçadas, onde só se sabe que a falha ocorreu dentro de um intervalo de tempo. Quando o tempo de falha T é impreciso, é dito que ele pertence a um intervalo $T \in (L, U]$. Esses dados são conhecidos como sobrevivência intervalar ou dados de censura intervalar. Note que tempos exatos de falha, sejam censura à direita ou à esquerda, são casos especiais de sobrevivência intervalar com $L = U$. Em particular, $U = 0$ para censura à direita e $L = 0$ para censura à esquerda (Lindsey et al., 1998). veja a nota a seguir, que enfatiza um trecho que merece atenção no livro.

Nota: “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

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

1.5 Representação dos Dados de Sobrevivência

Seja uma amostra aleatória de tamanho n , o i -ésimo indivíduo no estudo é representado, em geral, 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 em análise de sobrevivência é representada por duas colunas no conjunto de dados.

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

Nota: A Seção 1.5 do livro apresenta exemplos de *Dados de Sobrevivência*.

1.6 Especificando o Tempo de Sobrevida

Seja T , uma variável aleatória (va) que, na maioria dos casos é contínua, representa o tempo de falha, assim, $T > 0$. Tal variável é geralmente pela sua *função risco* ou pela *função de taxa de falha* (ou risco). Tais funções, e outras relacionadas, são usados ao decorrer do processo de análise de dados de sobrevivência. A seguir, algumas definições.

1.6.1 Função de Sobrevida

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:

$$S(t) = P(T > t), \quad (1.1)$$

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,

$$F(t) = 1 - S(t). \quad (1.2)$$

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:

$$\lambda(t) = \frac{S(t) - S(t + \Delta t)}{\Delta t S(t)} \quad (1.3)$$

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:

$$\lambda(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T \leq t + \Delta t | T \geq t)}{\Delta t} \quad (1.4)$$

A Figura 1.3, do livro, mostra três funções de taxa de falha. A função crescente indica que a taxa de falha do paciente aumenta com o transcorrer do tempo. Este comportamento mostra um efeito gradual de envelhecimento. A função constante indica que a taxa de falha não se altera com o passar do tempo. A função decrescente mostra que a taxa de falha diminui à medida que o tempo passa.

Sabe-se, ainda, que a taxa de falha para o tempo de vida de seres humanos é uma combinação das curvas apresentadas na Figura 1.3 em diferentes períodos de tempo. Ela é conhecida como *curva da banheira* e tem uma taxa de falha decrescente no período inicial, representando a mortalidade infantil, constante na faixa intermediária e crescente na porção final. Uma representação desta curva é mostrada na Figura 1.4, do livro.

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:

$$\Lambda(t) = \int_0^t \lambda(u) du. \quad (1.5)$$

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 vida e a vida média residual. A primeira é obtida pela área sob a função de sobrevivência. Isto é,

$$t_m = \int_0^{\infty} S(t)dt. \quad (1.6)$$

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,

$$\text{vmr}(t) = \frac{\int_0^{\infty} (u-t)f(u)du}{S(t)} = \frac{\int_0^{\infty} S(u)du}{S(t)}, \quad (1.7)$$

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:

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

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

e

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

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

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

2 Técnicas Não Paramétricas

2.1 Introdução

O segundo capítulo do livro que está sendo usado como um dos livros-base, 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 em 1958 por Edward L. Kaplan e Paul Meier. É um estimador não-paramétrico utilizado para estimar a função de sobrevivência, $S(t)$. Tal estimador também é chamado de *estimador limite-produto*. O Estimador de Kaplan-Meier é uma adaptação a $S(t)$ empírica 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 $t_1 \leq t_2 \leq \dots \leq t_k$. Considerando $S(t)$ uma função discreta com probabilidade maior que zero somente nos tempos de falha t_j , $j = 1, \dots, k$, tem-se que:

$$S(t_j) = (1 - q_1)(1 - q_2) \dots (1 - q_j), \quad (2.1)$$

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

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

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

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

- $t_1 \leq t_2 \leq \dots \leq t_k$, os k tempos distintos e ordenados de falha;
- d_j o número de falhas em t_j , $j = 1, \dots, k$;
- 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:

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

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* para $S(t)$. Seguindo certos passos, é possível provar que $\hat{S}_{KM}(t)$ é um estimador de máxima verossimilhança de $S(t)$. Supondo que d_j observações falham no tempo t_j , para $j = 1, \dots, k$, e m_j observações são censuradas no intervalo $[t_j, t_{j+1})$, nos tempos t_{j1}, \dots, t_{jm_j} . A probabilidade de falha no tempo t_j é, então,

$$S(t_j) - S(t_{j+}),$$

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

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

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 assintoticamente Normal. O intervalo de confiança com $100(1 - \alpha)\%$ de confiança para $\hat{S}_{KM}(t)$ é expresso por:

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

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, $\hat{U}(t) = \log[-\log(\hat{S}_{KM}(t))]$. Esta transformação foi sugerida por Kalbfleisch e Prentice (1980), tendo sua variância estimada por:

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

Veja uma aplicação do Estimador de Kaplan-Meier. Os dados dispostos no Apêndice (*) do livro de Análise de Sobrevida Aplicado, de Enrico e Giolo.

```
# -----
# [1] ATIVAÇÃO DE PACOTES
# -----

if (!require("survival")){
  install.packages("survival")
}
```

Carregando pacotes exigidos: survival

```
if (!require("ggplot2")) {
  install.packages("ggplot2")
}
```

Carregando pacotes exigidos: ggplot2

```
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] ESTIMAÇÃO
# -----

# -----
```

```

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

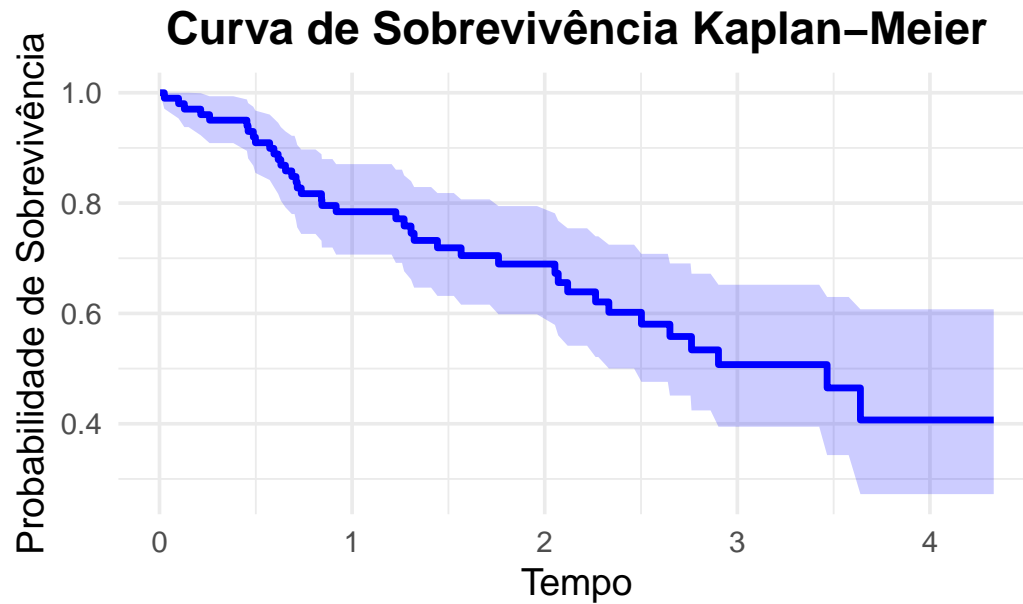
# -----
# [4] VISUALIZAÇÃO
# -----

# 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_step(color = "blue", size = 1.2) +
  geom_ribbon(aes(ymin = lower, ymax = upper), fill = "blue", alpha = 0.2) +
  labs(
    title = "Curva de Sobrevida Kaplan-Meier",
    x = "Tempo",
    y = "Probabilidade de Sobrevida",
    caption = "Fonte: Dados de leucemia"
  ) +
  theme_minimal(base_size = 14) +
  theme(
    plot.title = element_text(hjust = 0.5, face = "bold"),
    plot.caption = element_text(size = 10, hjust = 0.5)
  )

```

Warning: Using `size` aesthetic for lines was deprecated in ggplot2 3.4.0.
 i Please use `linewidth` instead.



Fonte: Dados de leucemia

2.3 Outros Estimadores Não Parâmetros

Texto a ser preenchido...

2.3.1 Estimador de Nelson-Aalen

Texto a ser preenchido...

2.4 Testes de Hipóteses para Comparação de Curvas de Sobrevivência

3 Técnicas Paramétricas - Modelos Probabilísticos

3.1 Introdução

No Capítulo anterior, 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 covariáveis, 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, \dots, x_p , identifica-se o modelo de regressão como: *Linear*, *Poisson*, *Logístico*, 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 serem estimados.

3.2 Distribuições do Tempo de Sobrevivência

Seja T uma variável aleatória que representa “tempo de sobrevivência”, qual a distribuição de probabilidade poderia representá-la?

Como uma característica da variável aleatória T é contínua e não negativa, pode-se a partir dessa característica, 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 , obedece a seguinte relação:

$$\ln(T) = \mu + \sigma W, \quad (3.1)$$

onde μ é o parâmetro que representa a média de $\ln(T)$ e σ sua dispersão, mas usualmente denominados, respectivamente, 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 $T \sim \text{Exp}(\alpha)$, a sua função densidade de probabilidade é expressa da seguinte forma:

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

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

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

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

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

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:

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

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 :

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

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, color = expression(alpha)) +
scale_color_manual(values = color_values,
                    labels = lapply(alphas, function(a) bquote(alpha == .(a)))) +
theme_minimal()
}

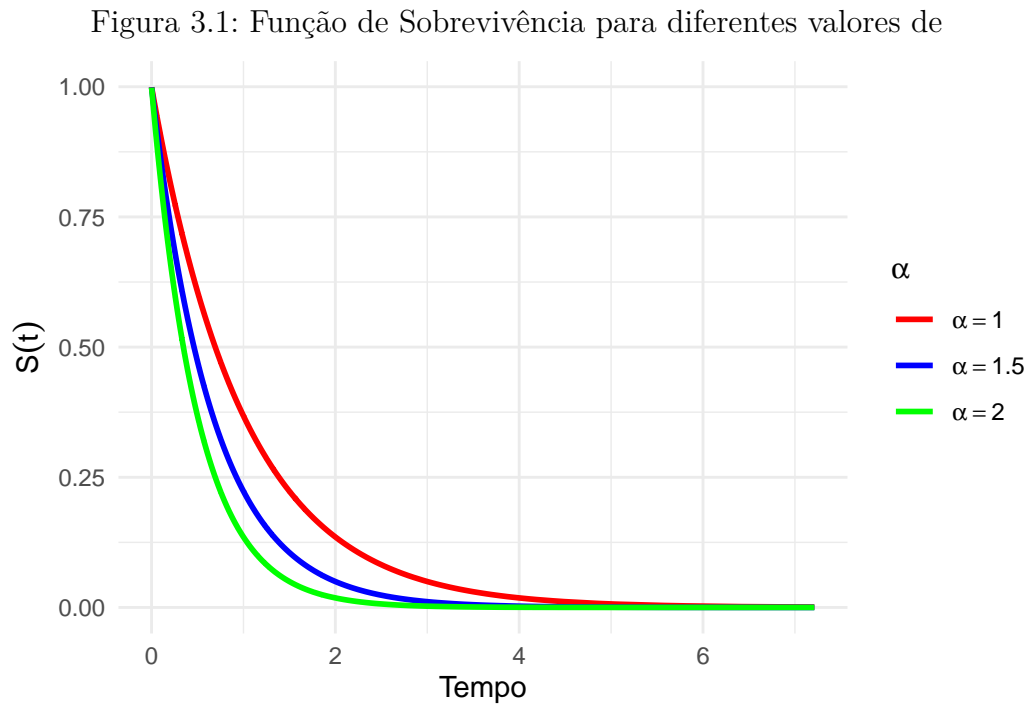
```

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)))

```



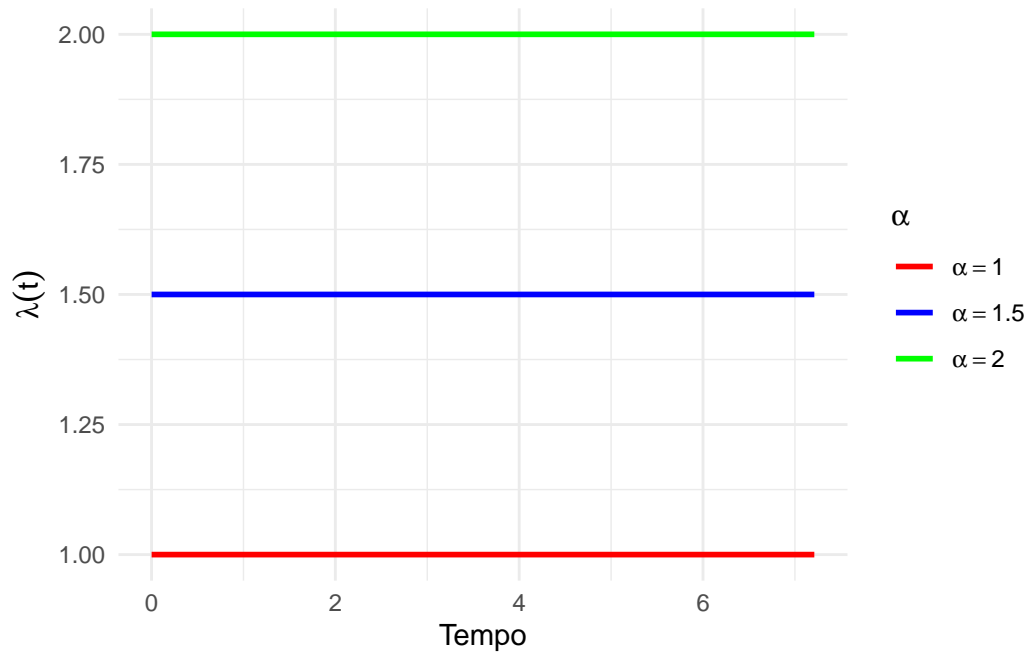
Veja a seguir a Figura 3.2. Tal Figura mostra a função de risco para diferentes valores do parâmetro α .

```

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

```

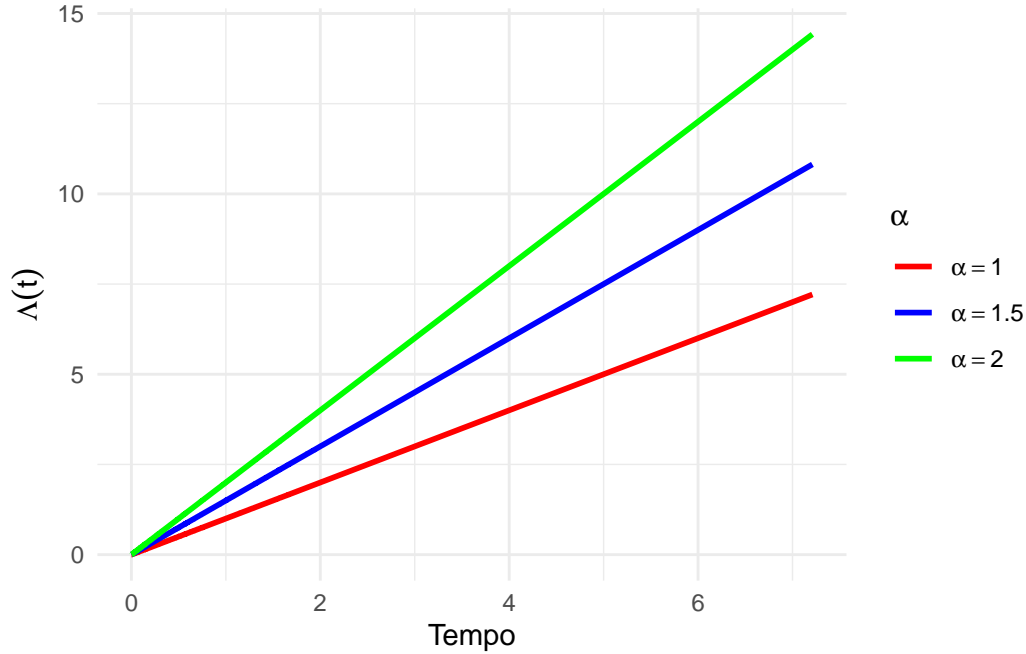
Figura 3.2: Função de Risco para diferentes valores de



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(Lambda(t)))
```

Figura 3.3: Função de Risco Acumulado para diferentes valores de



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

Quando $\alpha = 1$, diz-se que a **distribuição exponencial padrão**. A média ($E[t]$) e variância ($Var[T]$) do tempo de sobrevivência, quando este seguir uma distribuição em Exponencial, 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 parâmetros utilizados 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:

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

Em suma, o modelo exponencial se torna adequado quando o período do experimento é curto para que a suposição de risco constante possa ser atendida.

3.2.2 Distribuição Weibull

Na maioria dos 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:

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

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

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

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:

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

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

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

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
# -----
```

```

# 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, color = expression(gamma)) +
    scale_color_manual(values = color_values,
                      labels = lapply(gammas, function(g) bquote(gamma == .(g)))) +
    theme_minimal()
}

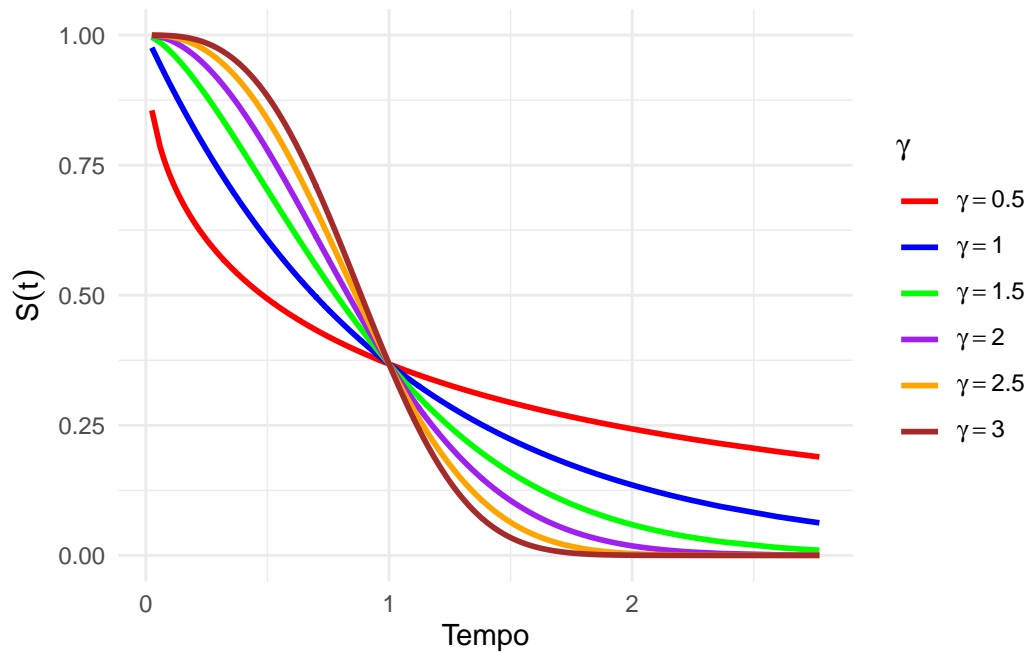
# Paleta de cores
color_values <- c("red", "blue", "green", "purple", "orange", "brown")

```

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)))
```

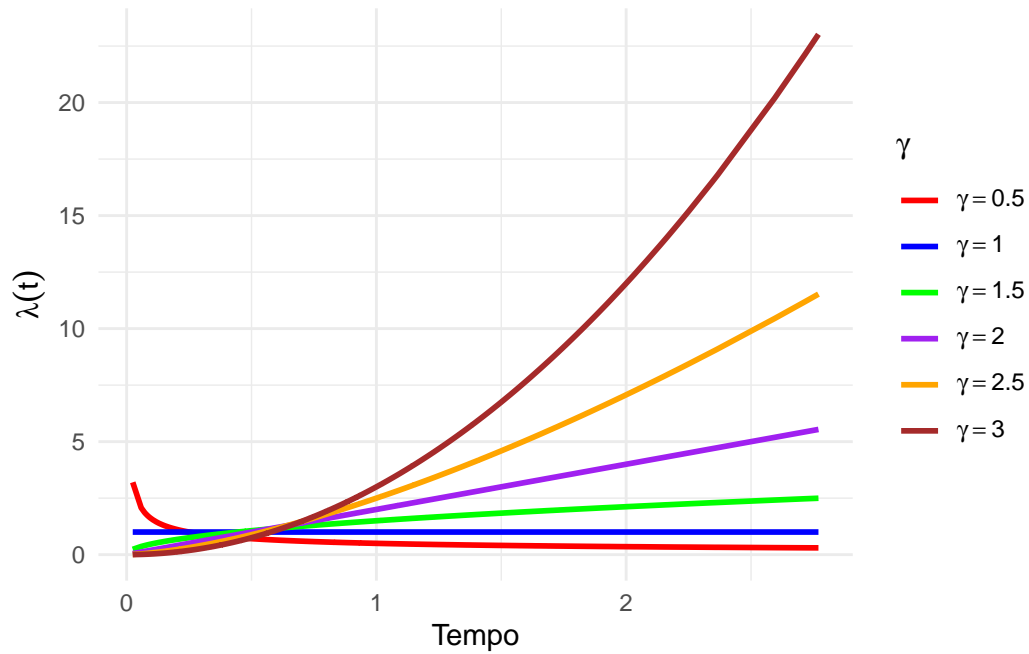
Figura 3.4: Função de Sobrevivência para diferentes valores de



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)))
```

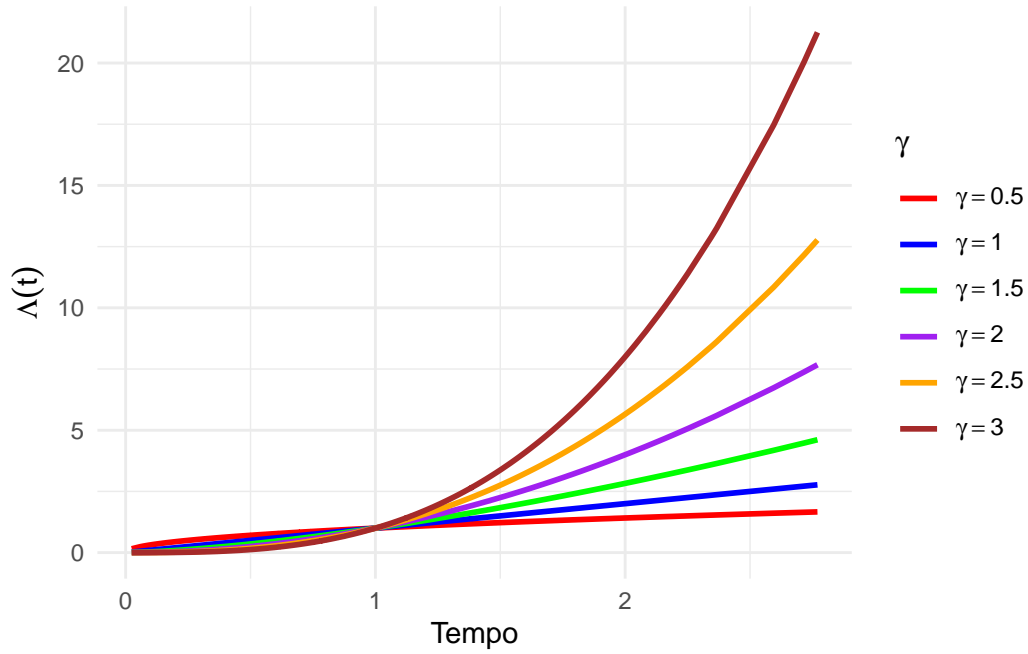
Figura 3.5: Função de Risco para diferentes valores de



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 para diferentes valores de



3.2.2.1 Algumas Considerações

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

$$E[T] = \alpha \Gamma[1 + (1/\gamma)]$$

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

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:

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

Assim, quando o tempo de sobrevivência segue uma distribuição log-normal, sua função de sobrevivência e as demais não tem uma forma analítica 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:

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

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

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
```

```

# -----
# [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, color = expression(mu)) +
    scale_color_manual(values = color_values,
                      labels = lapply(mus, function(m) bquote(mu == .(m)))) +
    theme_minimal()
}

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

```

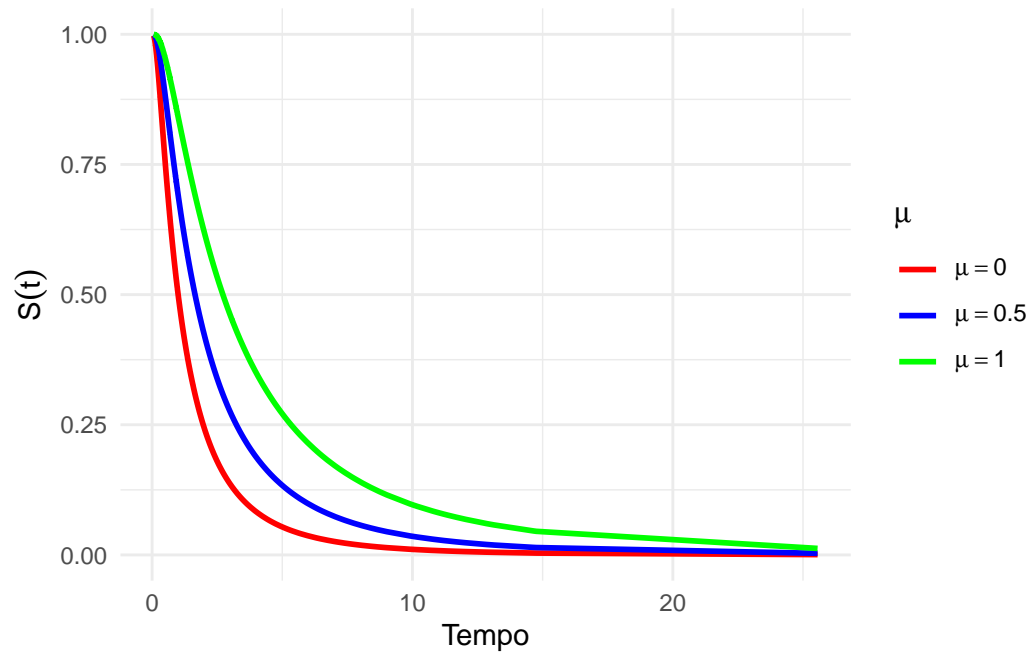
Veja a seguir a Figura 3.7. Tal Figura mostra as curvas de sobrevivência para diferentes valores do parâmetro μ .

```

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

```

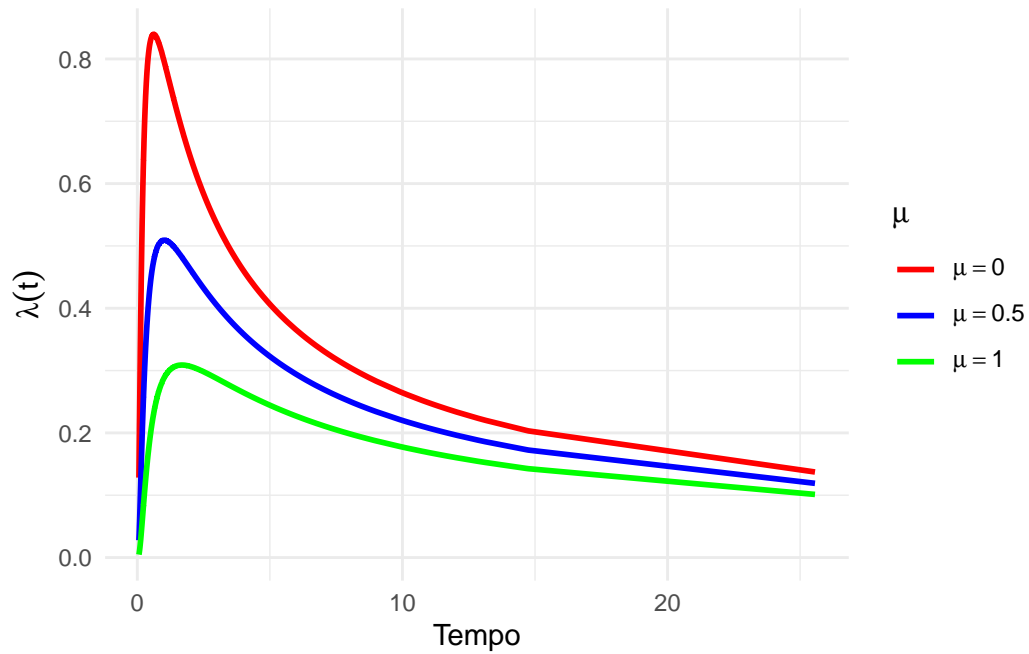
Figura 3.7: Função de Sobrevivência para diferentes valores de



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)))
```

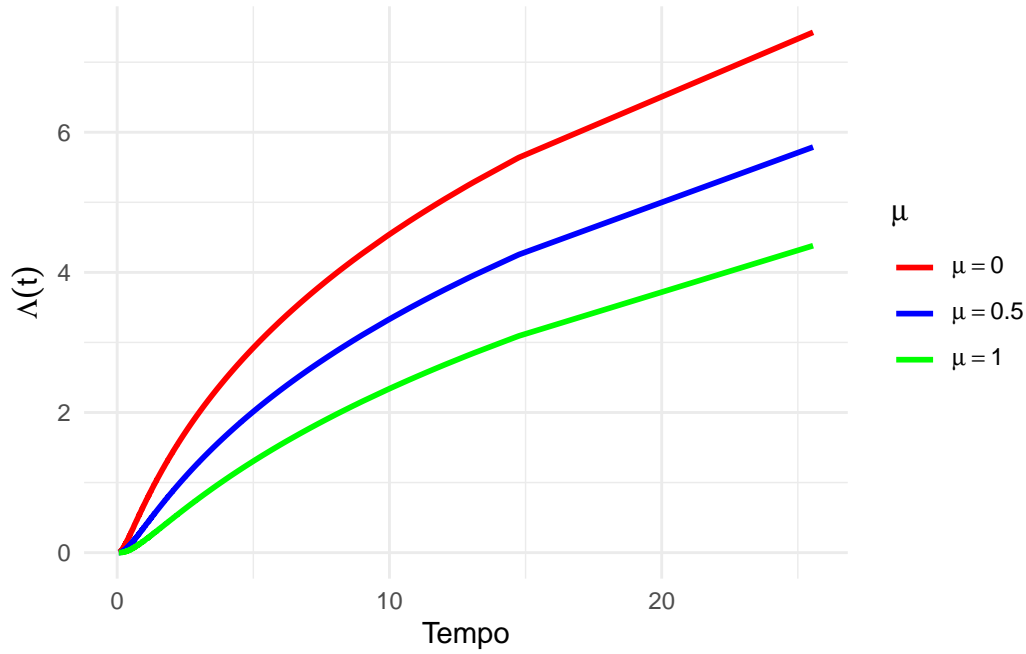
Figura 3.8: Função de Risco para diferentes valores de



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 para diferentes valores de



3.2.3.1 Algumas Considerações

A média de T é dada por:

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

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 , $i = 1, 2, \dots, n$. 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 contribuição 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:

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

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 Exponencial

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 $\ln L(\alpha)$:

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

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

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

Ao igualar a derivada a zero e isolando α em um dos lados da igualdade e assumindo que a forma analítica 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 exponencial 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(knitr)

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

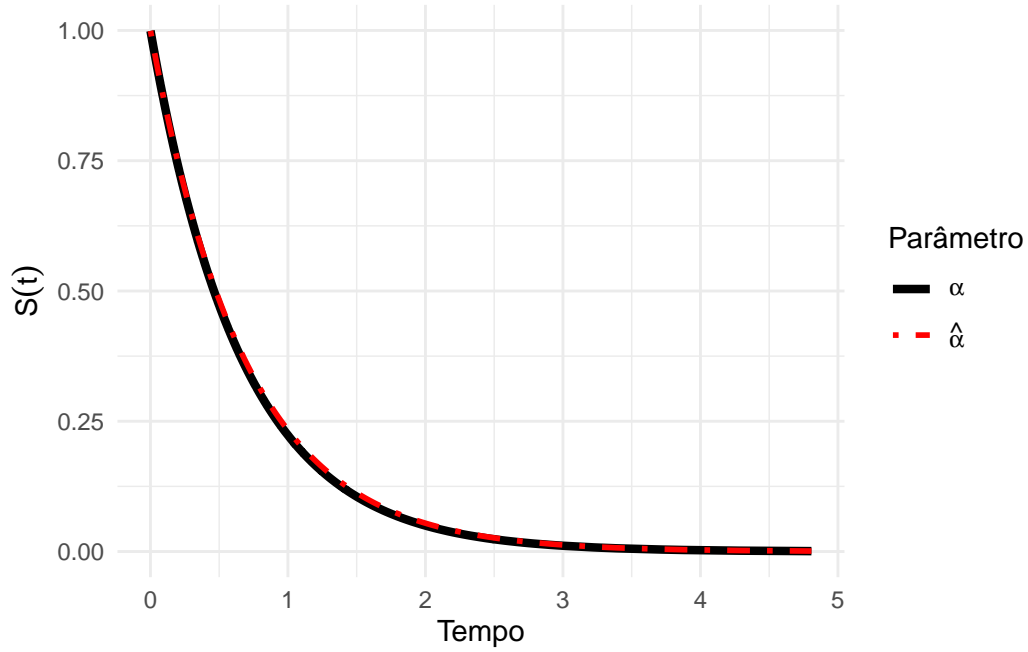
Tabela 3.1: Real e Estimado para as Funções de Sobrevivência

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.

```
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 analítica 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 aprensado 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:

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

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
```



```

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.

```
# -----
# 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.

```
library(knitr)

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
0.7315788	0.7883051	0.7921747
1.4183502	0.4089769	0.4128247
0.5290778	0.8830174	0.8858229

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

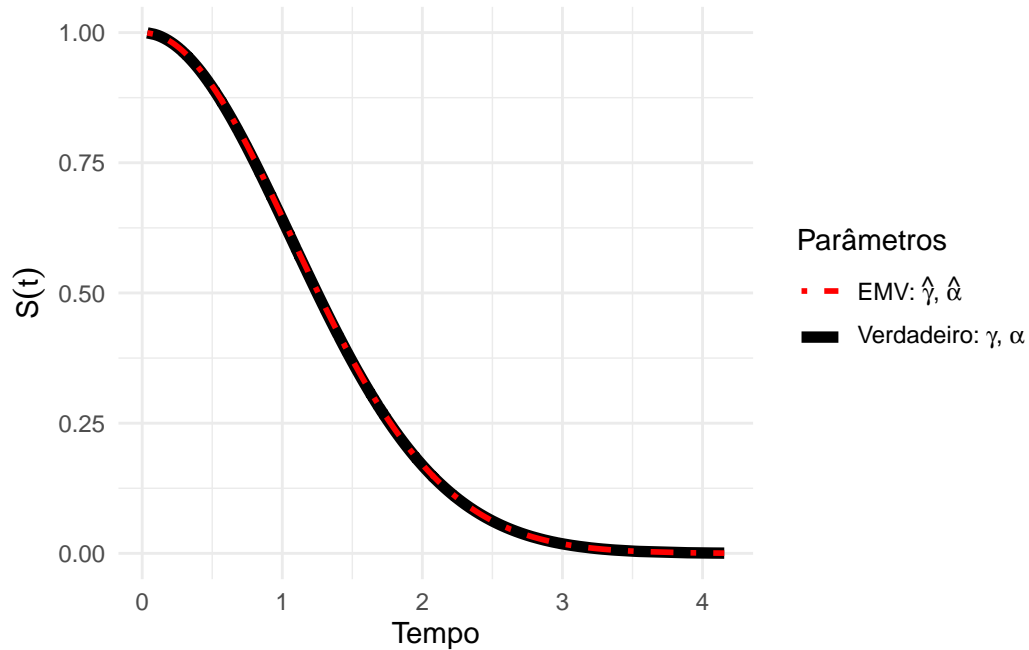
Tempo	S(t)	S(t) EMV
0.3716200	0.9404673	0.9422483
2.6362482	0.0455565	0.0456849

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 γ e α 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
# -----

if (!require("survival")) install.packages("survival")
```

Carregando pacotes exigidos: survival

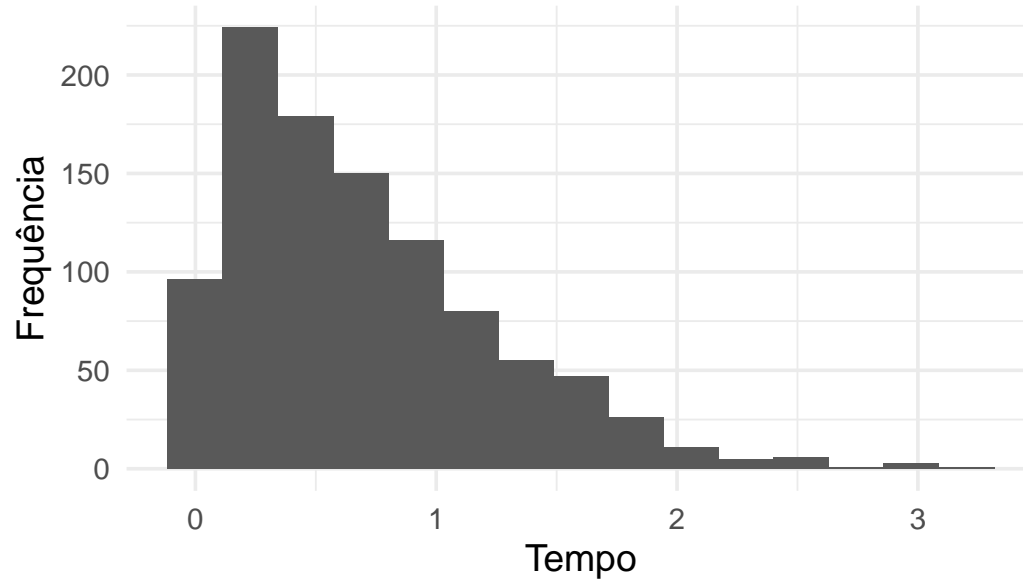
```
if (!require("ggplot2")) install.packages("ggplot2")

library(survival)
library(ggplot2)
```

2. Simulamos os dados com censura:

```
# -----  
# [2] Simulação dos Dados  
# -----  
  
# Definindo Semente  
set.seed(123)  
  
# Tamanho amostral  
n <- 1000  
  
# Parâmetros - Distribuição Weibull  
gamma <- 2  
alpha <- 1.5  
  
# Geração de tempos de falha e censura  
Tfalha <- rweibull(n, shape = gamma, scale = alpha)  
Tcensu <- rexp(n, rate = 1)  
  
# Determinação do tempo observado e indicador de censura  
Tobservado <- pmin(Tfalha, Tcensu)  
indCensura <- as.numeric(Tfalha <= Tcensu)  
  
# Dados simulados  
dados <- data.frame(Tempo = Tobservado, Censura = indCensura)  
  
# Histograma  
ggplot(data = dados, aes(x = Tempo)) +  
  geom_histogram(bins = 15) +  
  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"))
```

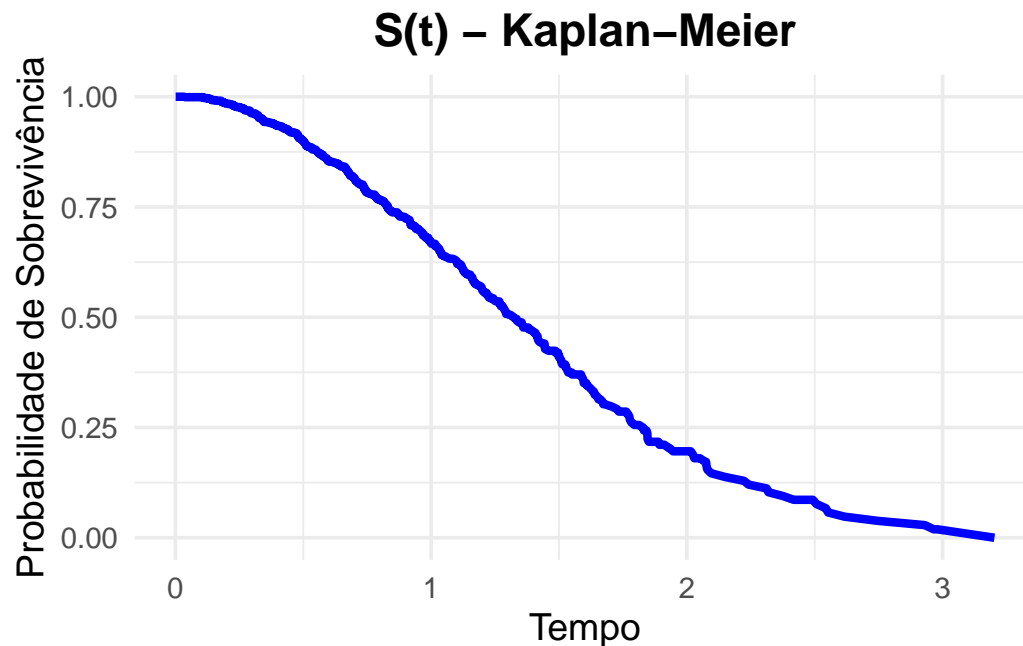
Histograma do Tempo de Sobrevivência – Simulação



3. Começamos o processo de estimação.

4. 1. Estimação Não Paramétrica.

```
# -----  
# [3.1] Estimação Não Paramétrica  
# -----  
# -----  
# [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"))
```



3. 2. Estimação Não Paramétrica.
4. 2. 1. Distribuição Exponencial.

```
# -----
# [3.2.1] Distribuição Exponencial
# -----

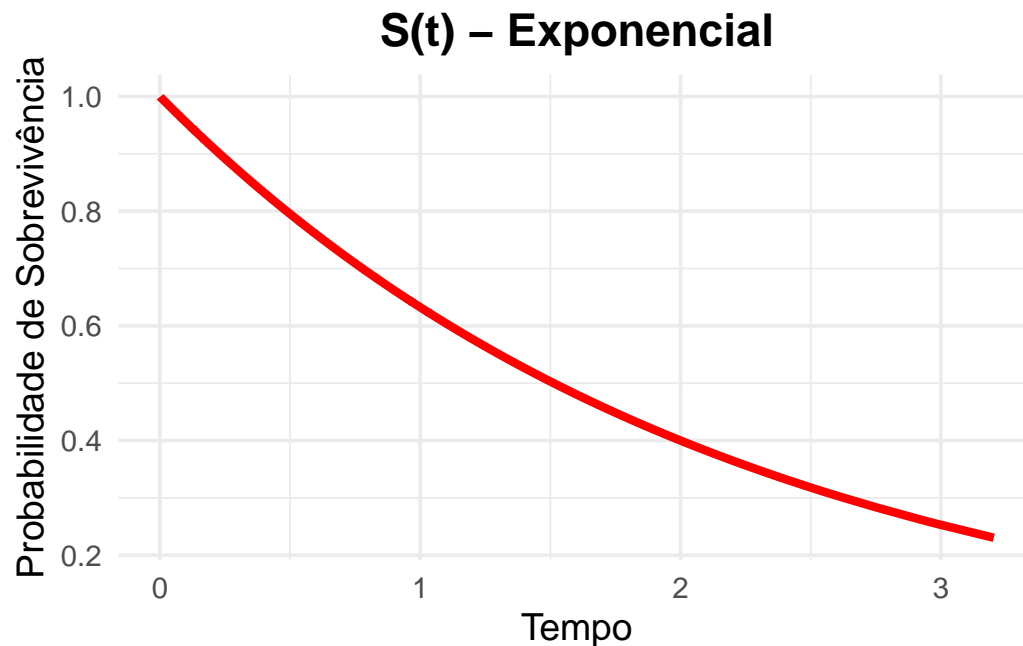
# Função de Sobrevivência
Stexp <- function(t, alpha) exp(-alpha * t)

# 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"))
```



3. 2. 2. Distribuição Weibull.

::: {.cell}

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

# Função de Sobrevida
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(alpha): NaNs produzidos
```

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

```
:::
```

```
estimate
```

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

```
$par
```

```
[1] 2.021559 1.561521
```

```
$value
```

```
[1] 455.8422
```

```
$counts
```

```
function gradient
      30      11
```

```
$convergence
```

```
[1] 0
```

```
$message
```

```
NULL
```

```
$hessian
```

```
      [,1]      [,2]
[1,] 169.43753  71.93369
[2,]  71.93369 536.32537
```

```
:::
```

```
# EMV da Sobrevida
```

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

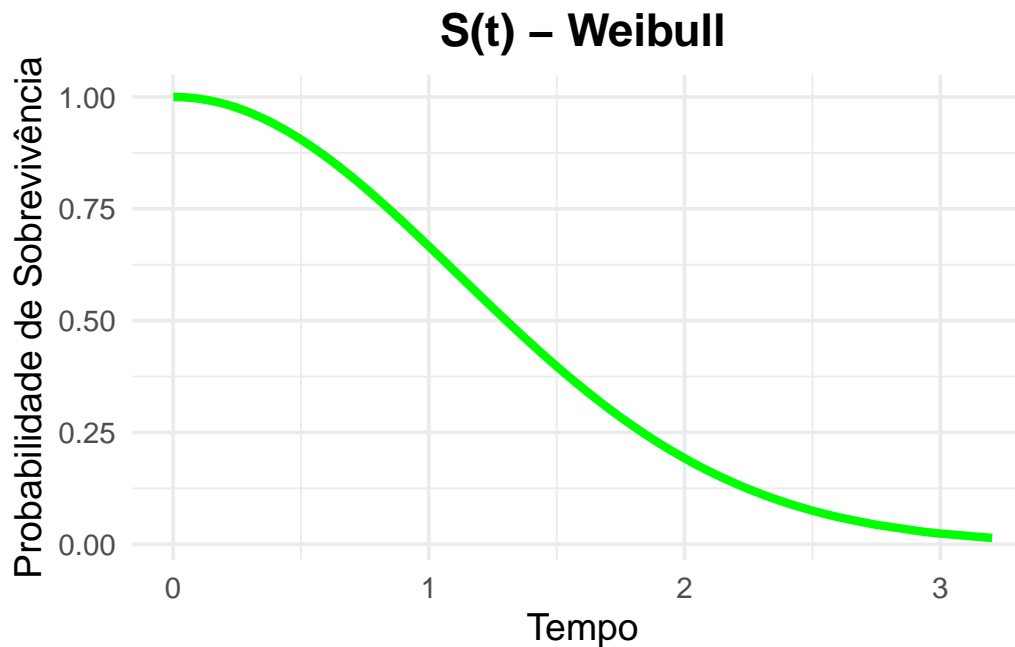
```
# Formatando como DataFrame
```

```
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 Sobrevida") +
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 Sobrevida") +
  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")
```

$\hat{S}(t)$: Kaplan–Meier, Weibull e Exponencial

