



Universidade Federal do Pará  
Instituto de Ciências Exatas e Naturais  
Faculdade de Estatística

Breno Cauã Rodrigues da Silva

Modelo Exponencial por Partes & Modelo  
Exponencial por Partes Potência

Belém/PA

2025

Breno Cauã Rodrigues da Silva

# Modelo Exponencial por Partes & Modelo Exponencial por Partes Potência

Trabalho de Conclusão de Curso, apresentado  
como requisito parcial para a obtenção do grau de  
Bacharel em Estatística, pela Universidade Federal  
do Pará.

Área de concentração: Análise de Sobrevida

Linha de pesquisa: Modelos semiparamétricos  
para dados de sobrevida

Orientador: Prof. Dr. Paulo Cerqueira dos Santos  
Júnior

Belém/PA  
2025

Breno Cauã Rodrigues da Silva

# Modelo Exponencial por Partes & Modelo Exponencial por Partes Potência

Trabalho de Conclusão de Curso, apresentado como requisito parcial para a obtenção do grau de Bacharel em Estatística, pela Universidade Federal do Pará.

Data de aprovação: DD/MM/AAAA

Conceito:  $E[\text{Conceito}] = \text{Excelente}$

---

Prof. Dr. Paulo Cerqueira dos Santos Júnior  
*Orientador - FAEST/ICEN/UFPa*

---

Professor (a)  
*Membro - FAEST/ICEN/UFPa*

---

Professor (a)  
*Membro - FAEST/ICEN/UFPa*

Belém/PA  
2025

# Agradecimentos

[...].

# Resumo

[...].

**Palavras-chave:** [...]; [...].

# Abstract

[...].

**Keywords:** [...]; [...].

## Lista de Tabelas

## Lista de Figuras

3.1	Curva de Sobrevida de Kaplan-Meier com IC de 95% . . . . .	17
3.2	Função Risco Acumulado e Função Sobrevida com IC de 95% segundo o Estimador de Nelson-Aalen. . . . .	18
3.3	Comparação Entre as Curvas de Sobrevida de Kaplan-Meier Nelson-Aalen. . . . .	19
3.4	Funções Densidade de Probabilidade, Sobrevida, Risco e Risco Acumulado segundo uma Distribuição Exponencial para diferentes valores do Parâmetro de Taxa. . . . .	21
3.5	Funções Densidade de Probabilidade, Sobrevida, Risco e Risco Acumulado segundo uma Distribuição Weibull para diferentes valores do Parâmetro de Forma e um valor fixo para o Parâmetro de Escala. . . . .	23
3.6	Funções Densidade de Probabilidade, Sobrevida, Risco e Risco Acumulado segundo uma Distribuição Log-normal para diferentes valores do Parâmetro de Localização e um valor fixo para o Parâmetro de Escala. . . . .	25

# Índice

<b>1</b>	<b>Introdução</b>	<b>8</b>
<b>2</b>	<b>Revisão Bibliográfica</b>	<b>9</b>
<b>3</b>	<b>Fundamentação Teórica</b>	<b>10</b>
3.1	Conceitos Básicos . . . . .	10
3.1.1	Tempo de Falha . . . . .	10
3.1.2	Censura . . . . .	10
3.1.3	Representação dos Dados de Sobrevida . . . . .	12
3.1.4	Especificando o Tempo de Sobrevida . . . . .	12
3.1.5	Relações entre as Funções . . . . .	14
3.2	Técnicas Não Paramétricas . . . . .	14
3.2.1	O Estimador de Kaplan-Meier . . . . .	14
3.2.2	Outros Estimadores Não Paramétricos . . . . .	17
3.3	Técnicas Paramétricas . . . . .	19
3.3.1	Distribuição Exponencial . . . . .	20
3.3.2	Distribuição Weibull . . . . .	22
3.3.3	Distribuição Log-normal . . . . .	24
3.3.4	Distribuição Exponencial por Partes . . . . .	25
3.3.5	Distribuição Exponencial por Partes Potência . . . . .	25
3.3.6	Estimação de Parâmetros - Método de Máxima Verossimilhança . . . . .	25
3.4	Modelos de Tempo de Vida Acelerados . . . . .	27
3.5	Censura Intervalar . . . . .	28
3.5.1	O Estimador de Turnbull . . . . .	28
3.5.2	Estimação de Parâmetros . . . . .	28
	<b>Referências</b>	<b>29</b>



# 1 Introdução

## 2 Revisão Bibliográfica

## 3 Fundamentação Teórica

### 3.1 Conceitos Básicos

A *Análise de Sobrevivência* é 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; tempo até a cura de uma doença ou até a recidiva de uma condição clínica.

É comum, surgir entre os pesquisadores que iniciam os estudos em análise de sobrevivência, a dúvida de: por que não utilizar outras técnicas estatísticas? Outros métodos convencionais acabam por se tornar inadequados 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.

#### 3.1.1 Tempo de Falha

Em 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.

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 inadequado para consumo quando atinge uma concentração específica de microrganismos por mm<sup>2</sup> de área.”*

#### 3.1.2 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.”*

Ao analisar dados de sobrevivência pode ocorrer a confusão entre os conceitos de **censura** e **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.

### 3.1.3 Representação dos Dados de Sobrevida

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

### 3.1.4 Especificando o Tempo de Sobrevida

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.

#### 3.1.4.1 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:

$$S(t) = P(T > t), \quad (3.1)$$

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,

$$F(t) = 1 - S(t). \quad (3.2)$$

#### 3.1.4.2 Função Taxa de Falha ou Função 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  $\Delta t$  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:

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

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.

#### 3.1.4.3 Função Taxa de Falha Acumulada ou Função Risco Acumulado

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:

$$\Lambda(t) = \int_0^t \lambda(u) du. \quad (3.4)$$

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.

#### 3.1.4.4 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,

$$t_m = \int_0^\infty S(t) dt. \quad (3.5)$$

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,

$$\text{vmr}(t) = \frac{\int_0^\infty (u - t) f(u) du}{S(t)} = \frac{\int_0^\infty S(u) du}{S(t)}, \quad (3.6)$$

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

### 3.1.5 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:

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

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

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

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

## 3.2 Técnicas Não Paramétricas

As técnicas não paramétricas desempenham um importante papel na análise de sobrevivência, pois permitem a estimativa da função de sobrevivência sem a necessidade de pressupor uma distribuição específica dos tempos até a ocorrência do evento de interesse. Essa abordagem é especialmente útil em estudos onde a distribuição dos tempos de falha é desconhecida ou onde se deseja evitar suposições rígidas sobre sua forma.

Diferentemente dos métodos paramétricos, que assumem distribuições predefinidas (como Weibull ou exponencial), as técnicas não paramétricas operam apenas com a ordenação dos eventos observados, tornando-se mais flexíveis e robustas e particularmente úteis na presença de dados censurados.

### 3.2.1 O Estimador de Kaplan-Meier

Proposto por Kaplan e Meier (1958). É um estimador não paramétrico utilizado para estimar a função de sobrevivência. 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:

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

$\hat{S}(t)$  é uma função que tem um 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:

$$S(t_j) = (1 - q_1)(1 - q_2) \dots (1 - q_j), \quad (3.7)$$

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:

$$q_j = P(T \in [t_{j-1}, t_j) | T \geq t_{j-1}), \quad (3.8)$$

para  $j = 1, \dots, k$ . A expressão geral do estimador de Kaplan-Meier pode ser apresentada após estas considerações preliminares. 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  $t_j$ , com  $j = 1, \dots, k$ ;
- $n_j$  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:

$$\hat{S}_{KM}(t) = \prod_{j: t_j < t} \left( \frac{n_j - d_j}{n_j} \right) = \prod_{j: t_j < t} \left( 1 - \frac{d_j}{n_j} \right) \quad (3.9)$$

De forma intuitiva, por assim dizer, a Equação 3.9 é proveniente da Equação 3.7, sendo está, uma decomposição de  $S(t)$  em termos  $q_j$ 's. Assim, a Equação 3.9 é justificada se os  $q_j$ 's forem estimados por  $d_j/n_j$ , expresso em termos de probabilidade na Equação 3.8. No artigo original de 1958, Kaplan e Meier provam que a Equação 3.9 é um *Estimador de Máxima Verossimilhança* (EMV) para  $S(t)$ . Seguindo certos passos, é possível provar que  $\hat{S}_{KM}(t)$  é EMV 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 3.9.



### 3.2.1.1 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).

### 3.2.1.2 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 da variância assintótica do estimador de Kaplan-Meier é dada pela Equação 3.10.

$$\widehat{Var} [\hat{S}_{KM}(t)] = [\hat{S}_{KM}(t)]^2 \sum_{j: t_j < t} \frac{d_j}{n_j(n_j - d_j)} \quad (3.10)$$

A expressão dada na Equação 3.10, 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 3.10 estão disponíveis em Kalbfleisch e Prentice (1980).

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  $\hat{S}_{KM}(t)$  como, por exemplo,  $\hat{U}(t) = \ln \left\{ -\ln \left\{ \hat{S}_{KM}(t) \right\} \right\}$ . Esta transformação foi sugerida por Kalbfleisch e Prentice (1980), tendo sua variância estimada por:

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

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

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

Veja uma aplicação do estimador de Kaplan-Meier para os dados de *Leucemia Pediátrica* dispostos no Apêndice (A) do livro *Análise de Sobrevida 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 3.1.

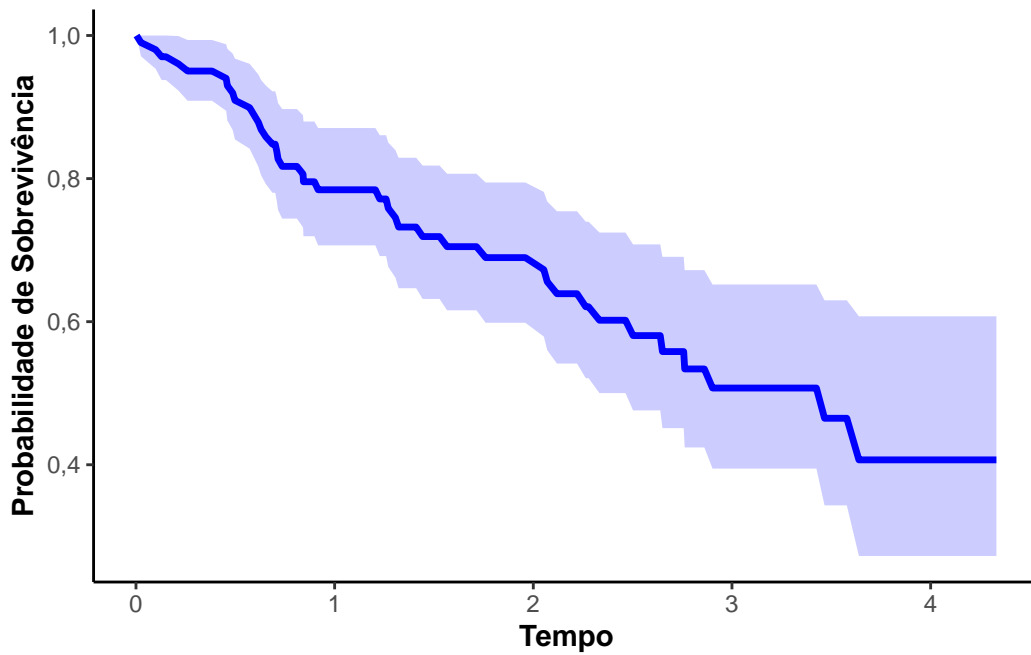


Figura 3.1: Curva de Sobrevivência de Kaplan-Meier com IC de 95%

### 3.2.2 Outros Estimadores Não Parâmetros

O estimador de Kaplan-Meier é amplamente utilizado para estimar a função de sobrevivência  $S(t)$ . Ele está disponível em diversos pacotes estatísticos e é frequentemente abordado em materiais introdutórios de estatística. No entanto, dois outros estimadores também possuem relevância na literatura: o estimador de Nelson-Aalen e o estimador de Tabela de Vida.

O estimador de Nelson-Aalen, desenvolvido posteriormente, apresenta similaridades com Kaplan-Meier em termos de propriedades, mas adota uma abordagem diferente ao focar na função risco acumulado  $\Lambda(t)$ .

Já o estimador da Tabela de Vida, também chamado de tabela atuarial, tem um importante valor histórico, sendo amplamente utilizado por demógrafos e atuários desde o século XIX sendo empregado principalmente em grandes amostras. Seu uso é especialmente relevante em contextos demográficos e atuariais, como estudos de expectativa de vida e análise de dados censitários.

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).

#### 3.2.2.1 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 3.1.4.3.

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:

$$\hat{\Lambda}(t) = \sum_{j:t_j < t} \left( \frac{d_j}{n_j} \right), \quad (3.11)$$

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:

$$\widehat{Var} [\hat{\Lambda}(t)] = \sum_{j:t_j < t} \left( \frac{d_j}{n_j^2} \right). \quad (3.12)$$

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

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

entretanto, o estimador da Equação 3.12 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:

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

Uma aplicação do estimador de Nelson-Aalen foi desenhada na Figura 3.2 em dois subgráficos. O primeiro apresenta a função de risco acumulado  $\Lambda(t)$  estimada conforme a Equação 3.11. O segundo mostra a curva de sobrevivência de Nelson-Aalen através das relações entre as funções de análise de sobrevivência.

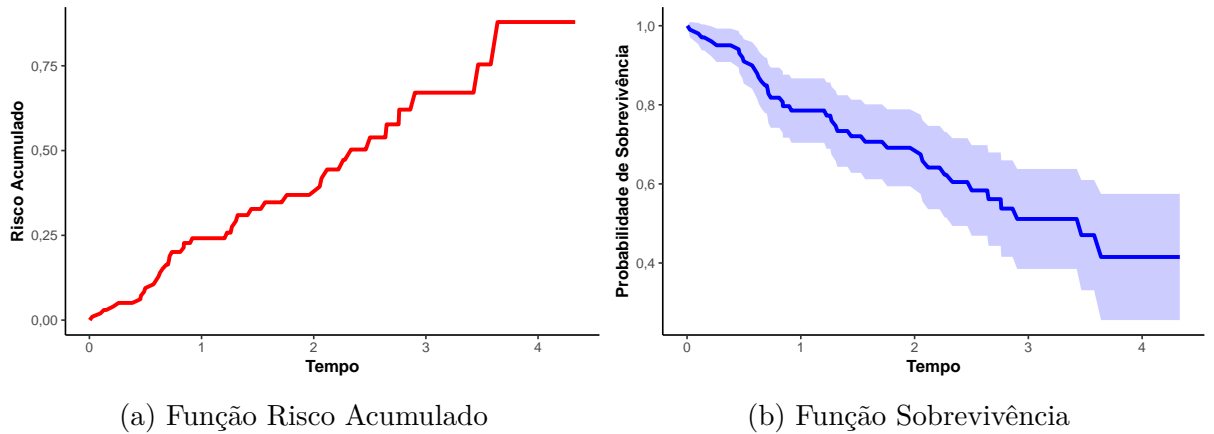


Figura 3.2: Função Risco Acumulado e Função Sobrevivência com IC de 95% segundo o Estimador de Nelson-Aalen.

Vale destacar que o estimador de Nelson-Aalen apresenta, 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.

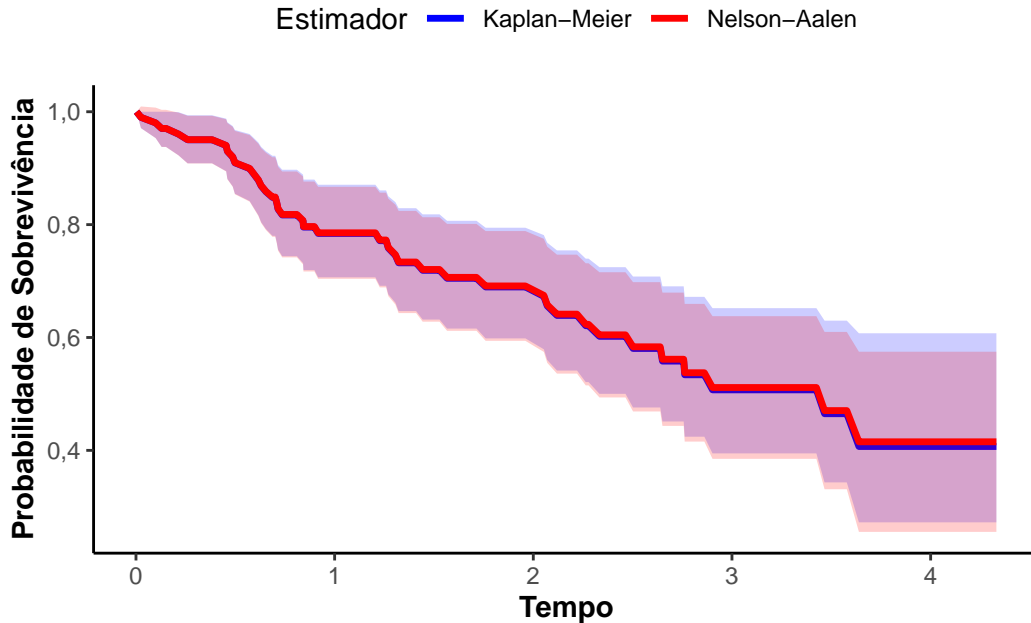


Figura 3.3: Comparação Entre as Curvas de Sobrevivência de Kaplan-Meier Nelson-Aalen.

### 3.3 Técnicas Paramétricas

Na análise de sobrevivência, métodos não paramétricos estimam funções sem assumir uma distribuição prévia para o tempo de falha. O estimador de Kaplan-Meier, por exemplo, calcula probabilidades diretamente dos dados, sendo útil para comparar curvas de sobrevivência entre grupos. No entanto, essa abordagem não permite avaliar diretamente o impacto de covariáveis, como idade ou tipo de tratamento, sobre a sobrevivência.

Os modelos paramétricos, por outro lado, assumem uma distribuição específica para o tempo de ocorrência do evento, permitindo estimativas mais estruturadas e eficientes. Assim como ocorre em modelos de regressão (linear, Poisson, logístico), esses métodos possibilitam relacionar covariáveis diretamente ao tempo de sobrevivência, garantindo uma análise mais detalhada e interpretável.

Entretanto, quais distribuições de probabilidade são adequadas para representar o tempo até a ocorrência do evento de interesse? Como o tempo de sobrevivência  $T$  é uma variável contínua e não negativa, algumas distribuições comuns — como a normal — não são apropriadas, pois permitem valores negativos. Além disso, dados de sobrevivência frequentemente apresentam assimetria à direita, indicando que poucos indivíduos sobrevivem por períodos longos enquanto a maioria tem eventos precoces. Reforçando a inadequação de algumas distribuições para representação probabilística do tempo de sobrevivência.

### 3.3.1 Distribuição Exponencial

Se  $T \sim \text{Exponencial}(\alpha)$ . Sua função densidade de probabilidade é expressa da seguinte forma:

$$f(t) = \alpha \exp \{-\alpha t\}. \quad (3.13)$$

Desta forma, podemos obter a função sobrevivência com base no completar da distribuição acumulada de  $T$ :

$$\begin{aligned} S(t) &= P(T > t) = 1 - P(T \leq t) = 1 - F(t) \\ &= 1 - [1 - \exp \{-\alpha t\}]. \end{aligned}$$

Assim definimos, formalmente, a função sobrevivência como:

$$S(t) = \exp \{-\alpha t\}. \quad (3.14)$$

Note que o parâmetro  $\alpha$  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 da função densidade de probabilidade e a função sobrevivência:

$$\lambda(t) = \frac{f(t)}{S(t)} = \frac{\alpha \exp \{-\alpha t\}}{\exp \{-\alpha t\}} = \alpha. \quad (3.15)$$

Sendo a função risco constante para todo tempo observado  $t$ , o risco acumulado é função linear no tempo com inclinação da reta dada por  $\alpha$ :

$$\Lambda(t) = -\ln \{S(t)\} = -\ln \{\exp \{-\alpha t\}\} = -(-\alpha t) = \alpha t \quad (3.16)$$

Veja, a seguir, a Figura 3.4 que mostra as curvas de densidade de probabilidade, de sobrevivência, risco e risco acumulado para diferentes valores do parâmetro  $\alpha$ .

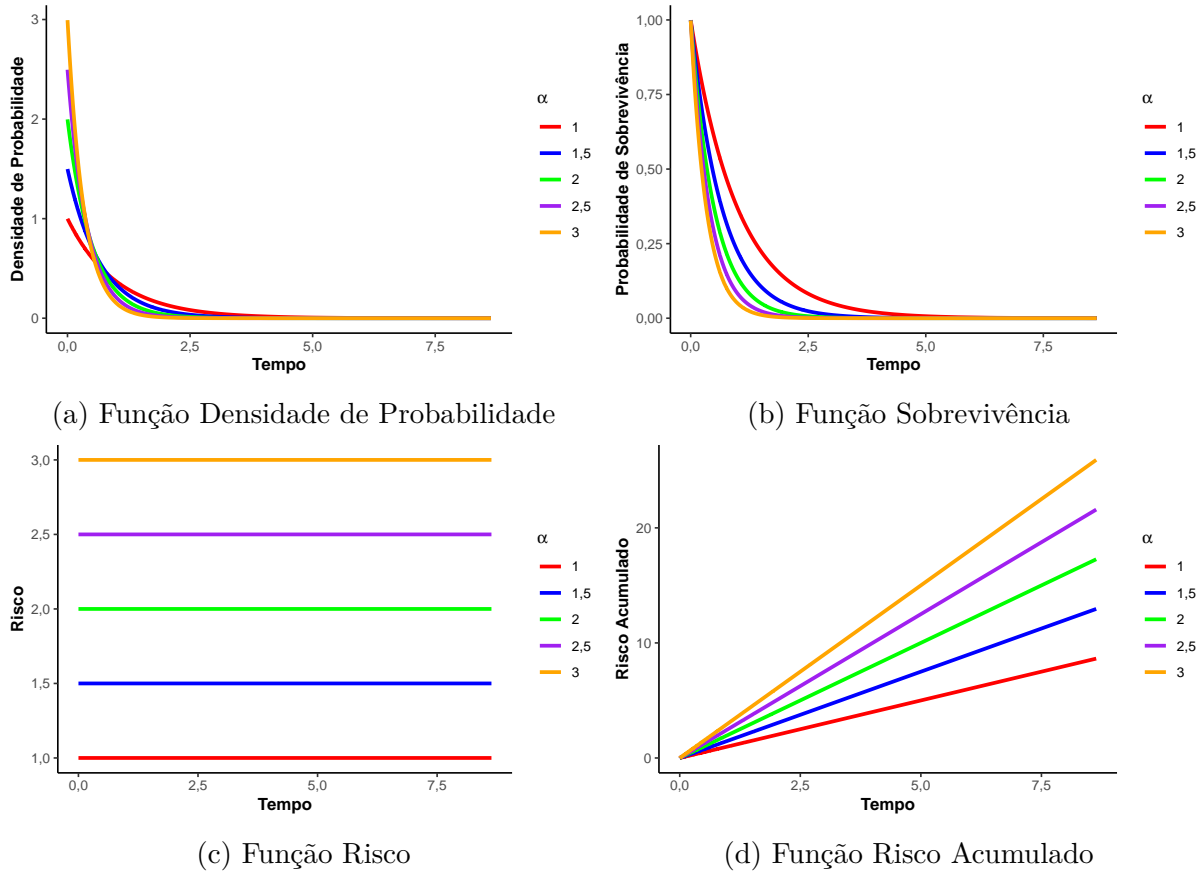


Figura 3.4: Funções Densidade de Probabilidade, Sobrevivência, Risco e Risco Acumulado segundo uma Distribuição Exponencial para diferentes valores do Parâmetro de Taxa.

Note que, o parâmetro  $\alpha$  deve ser sempre positivo e quanto maior o valor de  $\alpha$  (taxa), mais abruptamente a função sobrevivência  $S(t)$  decresce, e maior é a inclinação da função de risco acumulado. Quando  $\alpha = 1$ , a distribuição é denominada exponencial padrão.

A distribuição exponencial, por possuir um único parâmetro, é matematicamente simples e apresenta um formato assimétrico. Seu uso em análise de sobrevivência tem uma analogia com a suposição de normalidade em outras técnicas e áreas da estatística. Entretanto, a suposição de risco constante associada a essa distribuição é bastante restritiva e, em muitos casos, pode não ser realista. Essa característica da distribuição exponencial é conhecida como falta de memória, o que significa que o risco futuro é independente do tempo já decorrido.

A média e a variância do tempo de sobrevivência, para uma variável que segue a distribuição exponencial, são expressas como funções inversas do parâmetro de taxa ( $\alpha$ ). Assim, quanto maior o risco, menor o tempo médio de sobrevivência e menor a variabilidade em torno da média. As expressões são dadas por:

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

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

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,

$$S(t) = 0,5 \Leftrightarrow \exp\{-\alpha t\} = 0,5 \Leftrightarrow -\alpha t = \ln\{2^{-1}\}$$

$$\alpha t = -(-\ln\{2\}) \Leftrightarrow \alpha t = \ln\{2\}.$$

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

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

Em resumo, o modelo exponencial é apropriado para situações em que o período do experimento é curto o suficiente para que a suposição de risco constante seja plausível.

### 3.3.2 Distribuição Weibull

Na maioria dos casos de análise de sobrevivência - principalmente na área da saúde - é mais razoável supor que o risco varia ao longo do tempo, em vez de permanecer constante. Atualmente, a Distribuição Weibull é amplamente utilizada, pois permite modelar essa variação do risco ao longo do tempo. Como será demonstrado, a distribuição exponencial é um caso particular da distribuição Weibull.

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

$$f(t) = \frac{\gamma}{\alpha^\gamma} t^{\gamma-1} \exp\left\{-\left(\frac{t}{\alpha}\right)^\gamma\right\}. \quad (3.17)$$

A partir da Equação 3.17 é possível chegar a função sobrevivência da distribuição Weibull sendo esta função definida como:

$$S(t) = \exp\left\{-\left(\frac{t}{\alpha}\right)^\gamma\right\}, \quad (3.18)$$

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

$$\lambda(t) = \frac{f(t)}{S(t)} = \frac{\gamma}{\alpha^\gamma} t^{\gamma-1} \quad (3.19)$$

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

$$\Lambda(t) = -\ln\{S(t)\} = -\ln\left\{\exp\left\{-\left(\frac{t}{\alpha}\right)^\gamma\right\}\right\} = \left(\frac{t}{\alpha}\right)^\gamma. \quad (3.20)$$

Note que, o parâmetro  $\gamma$  determina a forma função risco da seguinte maneira:

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

Veja, a seguir, a Figura 3.5 que mostra as curvas de densidade, sobrevivência, risco e risco acumulado para diferentes valores do parâmetro de forma  $\gamma$  e o de escala  $\alpha = 1$ .

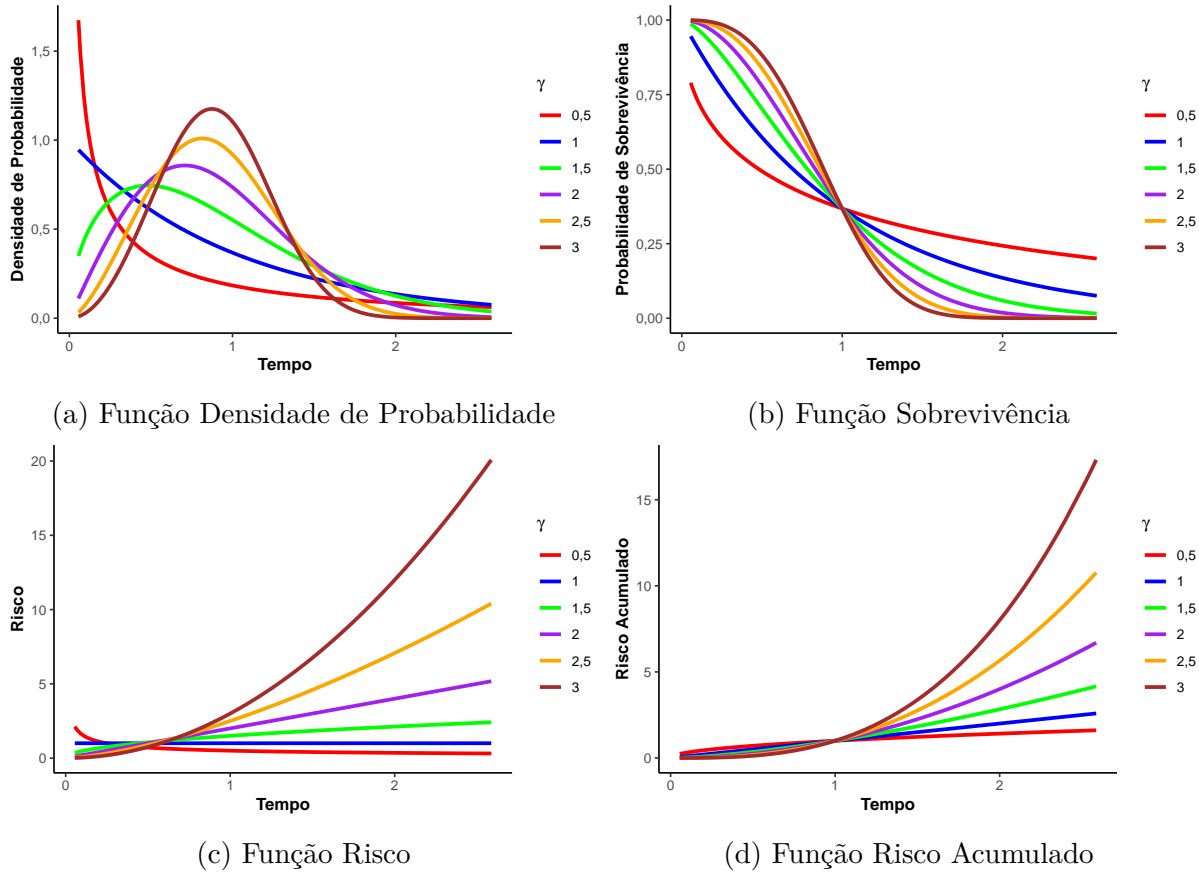


Figura 3.5: Funções Densidade de Probabilidade, Sobrevivência, Risco e Risco Acumulado segundo uma Distribuição Weibull para diferentes valores do Parâmetro de Forma e um valor fixo para o Parâmetro de Escala.

Perceba que  $\alpha$  - parâmetro escala - e  $\gamma$  - parâmetro de forma - são definidos dentro dos  $\mathbb{R}^+$ . É incluso a função Gama na média e variância da distribuição Weibull, assim,

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

e

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

sendo a função Gama  $\Gamma[k]$  expressa por  $\Gamma[k] = \int_0^\infty t^{k-1} \exp\{-t\} dt$ . Afim de se obter o tempo mediano de sobrevivência, igualamos a probabilidade de sobrevivência a 0,5. Desta forma:

$$\begin{aligned} S(t) = 0,5 &\Leftrightarrow \exp \left\{ - \left( \frac{t}{\alpha} \right)^\gamma \right\} = 0,5 \\ - \left( \frac{t}{\alpha} \right)^\gamma &= \ln \{2^{-1}\} \Leftrightarrow \left( \frac{t}{\alpha} \right)^\gamma = \ln \{2\} \\ \frac{t}{\alpha} &= [\ln \{2\}]^{1/\gamma}. \end{aligned}$$

Logo, definimos o tempo mediano de sobrevivência da distribuição Weibull como:

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



### 3.3.3 Distribuição Log-normal

Uma alternativa para modelar o tempo de sobrevivência é a distribuição log-normal. Dizemos que uma variável aleatória  $T$  tem distribuição log-normal com parâmetros  $\mu$  e  $\sigma^2$ , denotado por  $T \sim \text{Log-normal}(\mu, \sigma^2)$ , quando o logaritmo natural de  $T$ , ou seja,  $T^* = \ln \{T\}$ , segue uma distribuição normal, isto é,  $T^* \sim \text{Normal}(\mu, \sigma^2)$ . Nesse caso,  $\mu$  e  $\sigma^2$  correspondem à média e variância de  $\ln \{T\}$ . De forma equivalente, pode-se dizer que se  $T^* \sim \text{Normal}(\mu, \sigma^2)$ , então  $T = \exp \{T^*\} \sim \text{Log-normal}(\mu, \sigma^2)$ . A função densidade de probabilidade da distribuição log-normal é dada por:

$$f(t) = \frac{1}{t\sigma\sqrt{2\pi}} \exp \left\{ -\frac{1}{2} \left( \frac{\ln(t) - \mu}{\sigma} \right)^2 \right\}. \quad (3.21)$$

Quando o tempo de sobrevivência segue essa distribuição, a função sobrevivência  $S(t)$  é expressa por meio da função de distribuição da normal padrão:

$$S(t) = \Phi \left( \frac{-\ln \{t\} + \mu}{\sigma} \right). \quad (3.22)$$

Já as funções risco e risco acumulado não têm formas analíticas simples. Porém, podem ser obtidas por meio das relações entre as funções de análise de sobrevivência. Com isso, definimos, respectivamente, a função risco e a função risco acumulado pela expressão:

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

A Figura 3.6 ilustra as curvas usadas na análise de sobrevivência segundo uma distribuição log-normal, variando o parâmetro de locação  $\mu$  e fixando o parâmetro de escala  $\sigma = 1$ .

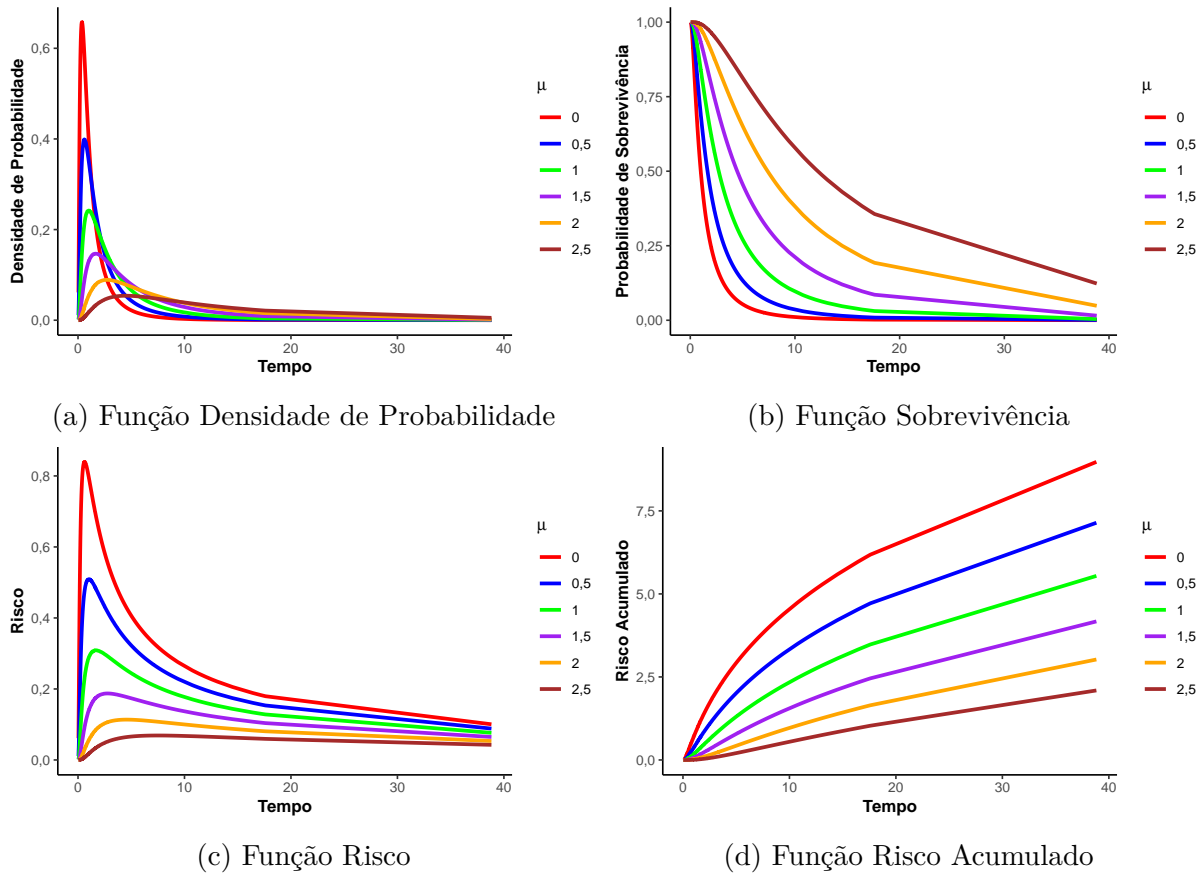


Figura 3.6: Funções Densidade de Probabilidade, Sobrevivência, Risco e Risco Acumulado segundo uma Distribuição Log-normal para diferentes valores do Parâmetro de Locação e um valor fixo para o Parâmetro de Escala.

O valor esperado e a variância da distribuição log-normal podem ser expressas em termos da distribuição normal. Desta forma, o valor esperado de  $T$  é expresso por:

$$E[T] = \exp \{ \mu + \sigma^2/2 \}.$$

Já a variância de  $T$  é dada por:

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

### 3.3.4 Distribuição Exponencial por Partes

### 3.3.5 Distribuição Exponencial por Partes Potência

### 3.3.6 Estimação de Parâmetros - Método de Máxima Verossimilhança

Foram apresentados alguns modelos probabilísticos. Esses modelos possuem quantidades desconhecidas, denominadas **parâmetros**, ou **parâmetro**, quando o modelo depende de uma única quantidade desconhecida, como no caso da distribuição exponencial.

O *Método de Máxima Verossimilhança* baseia-se no princípio de que, a partir de uma amostra aleatória, a melhor estimativa para o parâmetro de interesse é aquela que maximiza a probabilidade daquela amostra ter sido observada (Bussab e Morettin 2010), tornando a amostra mais verossímil.

De forma simples, o método de máxima verossimilhança condensa toda a informação contida na amostra, por meio da *função de verossimilhança*, para encontrar o(s) parâmetro(s) da distribuição que melhor expliquem os dados. Essa abordagem utiliza o produtório das densidades  $f(t)$  para cada observação  $t_i$ ,  $i = 1, 2, \dots, n$ . Em livros introdutórios de estatística, a função de verossimilhança é definida da seguinte maneira, para um parâmetro ou vetor de parâmetros  $\theta$ :

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

Observe que  $L$  é uma função de  $\theta$ , que pode ser um único parâmetro ou um vetor de parâmetros, como ocorre na distribuição log-normal, onde  $\theta = (\mu, \sigma^2)$ . No entanto, em análise de sobrevivência, essa definição tradicional de função de verossimilhança é insuficiente, pois os dados frequentemente apresentam **censura**, o que implica que o tempo de evento pode ser apenas parcialmente observado.

Para lidar com essa característica, utiliza-se a variável indicadora  $\delta_i$ , apresentada na Seção 3.1.3, que identifica se o  $i$ -ésimo tempo é um tempo de evento ou de censura. Com base nessa informação, a função de verossimilhança é ajustada da seguinte forma:

- Para  $\delta_i = 1$ , o  $i$ -ésimo tempo é um tempo de evento, e sua contribuição para  $L(\theta)$  é a densidade de probabilidade  $f(t_i|\theta)$ ;
- Para  $\delta_i = 0$ , o  $i$ -ésimo tempo é um tempo censurado, e sua contribuição para  $L(\theta)$  é a função de sobrevivência  $S(t_i|\theta)$ .

Assim, a função de verossimilhança ajustada, que incorpora dados censurados, é expressa como:

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

Para encontrar o valor de  $\theta$  que maximiza  $L(\theta)$ , utiliza-se a derivada do logaritmo de base neperiana da verossimilhança igualada a zero:

$$\frac{\partial \ln[L(\theta)]}{\partial \theta} = 0.$$

A solução dessa equação fornece o valor de  $\theta$  que maximiza  $\ln[L(\theta)]$ , e consequentemente,  $L(\theta)$ .

### 3.3.6.1 Método Iterativo de Newton-Raphson

Para algumas distribuições, apresentadas na seção anterior, e outras denifidas na literatura, não há forma analítica para as estimativas de máxima verossimilhança. Assim, as estimativas de tais parâmetros depende de métodos numéricos, sendo o **Método Iterativo de Newton-Raphson** uma abordagem amplamente utilizada.

O Método de Newton-Raphson é um procedimento iterativo eficiente para resolver equações não lineares, muito empregado na estimação de parâmetros de modelos estatísticos. No ajuste de distribuições o método busca maximizar a função de

verossimilhança resolvendo o sistema de equações derivado das condições de otimalidade (gradiente nulo). A fórmula iterativa é:

$$\theta_{n+1} = \theta_n - \mathbf{H}^{-1}(\theta_n) \nabla \ln[L(\theta_n)], \quad (3.24)$$

onde:

- $\theta_n$  é o vetor de parâmetros estimados na iteração  $n$ ;
- $\ln[L(\theta_n)]$  é o vetor gradiente, contendo as derivadas parciais de  $\ln[L(\theta_n)]$  em relação as coordenadas do vetor  $\theta$  (parâmetros);
- $\mathbf{H}(\theta)$  é a matriz Hessiana, composta pelas segundas derivadas de  $\ln[L(\theta_n)]$ .

O método apresenta vantagens convenientes no ajuste de parâmetros de modelos estatísticos. Uma das vantagens é a *eficiência* do método, que apresenta convergência rápida quando o ponto inicial  $\theta_0$  está próximo dos valores reais dos parâmetros. Outra vantagem, é *flexibilidade*, pois pode ser aplicado a diversos modelos probabilísticos, como o modelo Weibull, que é amplamente utilizada para modelar tempos de vida e dados de sobrevivência.

Entretanto, deve-se, também, atentar-se aos cuidados na aplicação do método. Pois, a *convergência* do método não é garantida caso o ponto inicial esteja muito distante da solução ou se as condições de regularidade do modelo não forem atendidas. Outro ponto que merece atenção é o cálculo da *matriz Hessiana*, que pode ser computacionalmente custoso, especialmente em modelos com maior complexidade.

Para um melhor entendimento do Método Iterativo de Newton-Raphson veja o Apêndice (D) do livro *Análise de Sobrevivência Aplicada* de Colosimo e Giolo (2006).

### 3.4 Modelos de Tempo de Vida Acelerados

Na seção anterior, foram apresentados modelos paramétricos para dados de sobrevivência. Entretanto, esses modelos não contemplam a inclusão de covariáveis na análise do tempo de sobrevivência. Neste capítulo, exploraremos esse método.

No modelo de regressão linear clássico, a relação entre a variável resposta  $Y$  e as covariáveis  $\mathbf{x}^\top$  é aditiva, ou seja, mudanças nas covariáveis alteram  $Y$  de maneira linear. O modelo de regressão linear clássico é expresso como:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_p X_p + \varepsilon, \quad (3.25)$$

onde  $\varepsilon$  é a parte estocástica (erro) que segue uma distribuição Normal(0;  $\sigma^2$ ).

No entanto, em análise de sobrevivência, essa suposição não se sustenta, pois o efeito das covariáveis geralmente acelera ou retarda o tempo de falha, tornando necessária uma abordagem multiplicativa. Este modelo de regressão é chamado de Modelo de *Tempo de Vida Acelerado* (Accelerated Failure Time - AFT).

No modelo AFT, assume-se que o tempo de falha  $T$  é afetado por um fator de aceleração exponencial das covariáveis. Esse fator multiplicativo indica se o tempo até o evento será prolongado ou encurtado. Assim, o modelo é definido como:

$$T = \exp\{\mathbf{x}^\top \beta\} \varepsilon = \exp\{\beta_0 + \beta_1 X_1 + \beta_2 X_2 \dots + \beta_p X_p\} \varepsilon, \quad (3.26)$$

onde  $\varepsilon$  é um termo de erro multiplicativo que captura a variabilidade não explicada pelas covariáveis. Aplicando a transformação logarítmica em  $T$  obtém-se a forma linearizável da Equação 3.26 que aproxima-se da Equação 3.25, de forma que

$$\ln[T] = \beta_0 + \beta_1 X_1 + \beta_2 X_2 \dots + \beta_p X_p + v,$$

onde  $v = \ln[\varepsilon]$  segue uma distribuição de valor extremo. Essa escolha para a distribuição dos erros decorre do fato de que os tempos de sobrevivência frequentemente apresentam forte assimetria à direita. Portanto, os erros não podem ser adequadamente representados por uma distribuição normal, sendo mais apropriado assumir distribuições como Log-normal, Weibull ou Exponencial.

Nos modelos AFT, a função de sobrevivência sofre um ajuste devido ao efeito das covariáveis, que podem acelerar ou retardar o tempo de falha. Assim, a função de sobrevivência condicional às covariáveis é expressa como:

$$S(t|\mathbf{x}) = P(T > t / \exp\{\mathbf{x}^\top \beta\}). \quad (3.27)$$

Como o tempo de falha é ajustado pelo fator de aceleração, a função de risco também precisa ser reformulada para incorporar o efeito das covariáveis. A forma geral da função de risco em modelos AFT é dada por:

$$\lambda(t|\mathbf{x}) = \lambda_0(t)g(\mathbf{x}). \quad (3.28)$$

Nesta expressão,  $\lambda_0(t)$ , representa a função de risco basal, isto é, representa o risco no tempo  $t$  quando todas as covariáveis são iguais a zero, ou seja, na ausência de efeitos das covariáveis. Já o termo  $g(\mathbf{x}) = \exp\{-\mathbf{x}^\top \beta\}$  age como um fator de ajuste, mensurando o impacto das covariáveis na taxa de falha.

## 3.5 Censura Intervalar

### 3.5.1 O Estimador de Turnbull

### 3.5.2 Estimação de Parâmetros

## Referências

- 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>.
- Bussab, Wilton de Oliveira, e Pedro Alberto Morettin. 2010. *Estatística Básica*. 6<sup>a</sup> ed. São Paulo: Saraiva.
- Colosimo, Enrico Antonio, e Suely Ruiz Giolo. 2006. *Análise de Sobrevivência Aplicada*. 1<sup>o</sup> ed. São Paulo, Brasil: Blucher.
- 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>.
- 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](https://doi.org/10.1002/(SICI)1097-0258(19980130)17:2%3C219::AID-SIM735%3E3.0.CO;2-D).
- 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>.
- 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>.