# Análise de sobrevivência e confiabilidade

## Introdução

**Prof. Paulo Cerqueira Jr**
**Faculdade de Estatística - FAEST**
**Instituto de Ciências Exatas e Naturais - ICEN**

https://github.com/paulocerqueirajr

# Representação do tempo de sobrevivência

# Representação do tempo de sobrevivência

- Uma representação simples do mecanismo de censura aleatória é feita usando duas variáveis aleatórias.
- Seja $Y$ uma variável aleatória representando o tempo de falha de um paciente;
- Seja $C$ uma variável aleatória independente de $T$ (censura não informativa), representando o tempo de censura associado a este paciente.
- O tempo observado é representado por

$$t = \min(Y, C) \quad e \quad \delta = \begin{cases} 1, & Y \leq C \\ 0, & Y > C \end{cases}$$

# Representação dos dados de sobrevivência

## Exemplos

Tabela: Dados de 32 pacientes com AIDS.

| | **3** | **18** | **29** | **54** | **60** | **84** | **110** | **112** | **116** | **123** |
|---|---|---|---|---|---|---|---|---|---|---|
| 145 | 151 | 151 | 158 | 173 | 194 | 214 | 329 | 331 | 371 | 408 |
| 490 | 514 | 541 | 555 | 688 | 780 | 801 | 858 | 887 | 998 | |

Tabela: Tempos de sobrevivência observados no estudo de hepatite

| **Grupo** | **Tempo de sobrevivência em semanas** |
|---|---|
| Controle | 1+, 2+, 3, 3, 3+, 5+, 5+, 16+, 16+, 16+, 16+, 16+, 16+ |
| Esteróide | 1, 1, 1, 1+, 4+, 5, 7, 8, 10, 10+, 12+, 16+, 16+, 16+ |

# Especificando o Tempo de Sobrevivência

## Função de sobrevivência

> Qual é a probabilidade de um paciente com aids sobreviver 365 dias ou mais? Isto é, qual a probabilidade de $T$ ser maior do que um determinado valor $t = 365$? Ou, mais formalmente, qual é $P(T > 365)$?

- Esta é uma das principais funções probabilísticas usadas para descrever estudos de sobrevivência.
- A função de sobrevivência é definida como a probabilidade de uma observação não falhar até um certo tempo $t$, ou seja, a probabilidade de uma observação sobreviver ao tempo $t$.
- Em termos probabilísticos, isto é escrito como

$$S(t) = P(T > t).$$

- Em consequência, a função de distribuição acumulada é dada por $F(t) = 1 - S(t)$.
- Definida como a probabilidade de uma observação não sobreviver ao tempo $t$.

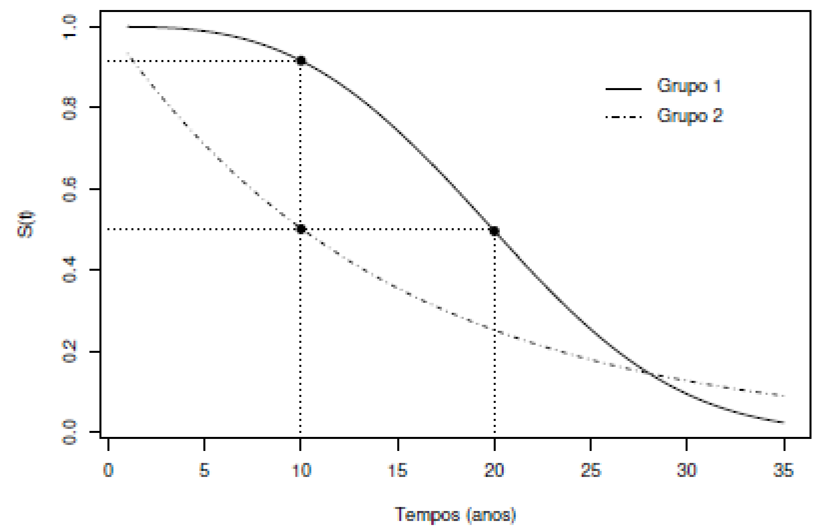
# Especificando o Tempo de Sobrevivência

## Função de sobrevivência

Estas curvas, supostas representarem as funções de sobrevivência de dois grupos diferentes de pacientes, o grupo 1 tratado com a droga A e o grupo 2 com a droga B.

Note, por exemplo, que o tempo de vida dos pacientes do grupo 1 é superior ao dos pacientes do grupo 2.

Para os pacientes do grupo 1, o tempo para o qual cerca de 50% (tempo mediano) deles estarão mortos é de 20 anos, enquanto que para os pacientes do grupo 2, este tempo é menor (10 anos).



# Representação dos dados de sobrevivência

- Os dados de sobrevivência para o indivíduo $i(i = 1, \ldots, n)$ sob estudo, são representados, em geral, pelo par $(t_i, \delta_i)$,
- $t_i$: o tempo de evento ou de censura e $\delta_i$ a variável indicadora de falha ou censura, ou seja,

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

- Na presença de covariáveis medidas no i-ésimo indivíduo tais como, dentre outras, $x_i = (\text{sexo}, \text{idade}, \text{tratamento recebido})$,

$$(y_i, \delta_i, x_i).$$

- Na presença de censura intervalar temos $(l_i, u_i, \delta_i, x_i)$, em que $l_i$ e $u_i$ são os limites inferiores e superiores observados, respectivamente.

# Especificando o Tempo de Sobrevivência

## Função Risco ou taxa de falha

- Qual é o risco de um paciente com aids vir a óbito após sobreviver 365 dias?
- Esse risco de morrer aumenta ou diminui com o tempo?

$h(t)$: probabilidade instantânea de um indivíduo sofrer o evento em um intervalo de tempo $t$ e $(t + \epsilon)$ dado que ele sobreviveu até o tempo $t$.

Sendo $\epsilon$ infinitamente pequeno, $\lambda(t)$ expressa o risco instantâneo de ocorrência de um evento, dado que até então o evento não tenha ocorrido.

# Especificando o Tempo de Sobrevivência

## Função Risco ou taxa de falha

$$h(t) = \lim_{\epsilon \to 0} \frac{P(t < T < t + \epsilon | T \geq t)}{\epsilon}$$

- $h(t)$ também é denominada:
  - função ou taxa de incidência,
  - força de infecção,
  - taxa de falha,
  - força de mortalidade,
  - força de mortalidade condicional.
- Apesar do nome risco, $h(t)$ é uma taxa (tempo$^{-1}$).

Pode assumir qualquer valor positivo (não é probabilidade)

# Especificando o Tempo de Sobrevivência

## Estimando a função taxa de falha

Na ausência de censuras, temos que a fórmula para estimar a função de taxa de falha é dada por

$$\hat{h}_x(t) = \frac{N_x(t)}{R_x(t) \times \Delta_x}$$

$N_x(t)$: Número de eventos observados no intervalo de classe $x$.

$\Delta_x$: amplitude de $x$.

Uma maneira alternativa de estimar $\lambda(t)$ é utilizar as relações entre $S(t)$, $f(t)$ e $h(t)$.

Comum nas tábuas de vida - demografia.

# Especificando o Tempo de Sobrevivência

## Estimando função de sobrevivência

| **Intervalo** | $R_x(t)$ | $\hat{S}_x(t)$ |
|---|---|---|
| (0,3] | 32 | 1.00000 |
| (3,18] | 31 | 0.96875 |
| (18,29] | 30 | 0.93750 |
| (29,54] | 29 | 0.90625 |
| (54,60] | 28 | 0.87500 |
| (60,84] | 27 | 0.84375 |
| (84,110] | 26 | 0.81250 |
| (110,112] | 25 | 0.78125 |
| (112,116] | 24 | 0.75000 |
| (116,123] | 23 | 0.71875 |
| (123,134] | 22 | 0.68750 |
| (134,145] | 21 | 0.65625 |

| **Intervalo** | $R_x(t)$ | $\hat{S}_x(t)$ |
|---|---|---|
| (145,151] | 20 | 0.62500 |
| (151,158] | 18 | 0.56250 |
| (158,173] | 17 | 0.53125 |

Tabela: Estimativa da sobrevivência para os 32 pacientes com AIDS.

# Especificando o Tempo de Sobrevivência

## Estimando a função taxa de falha

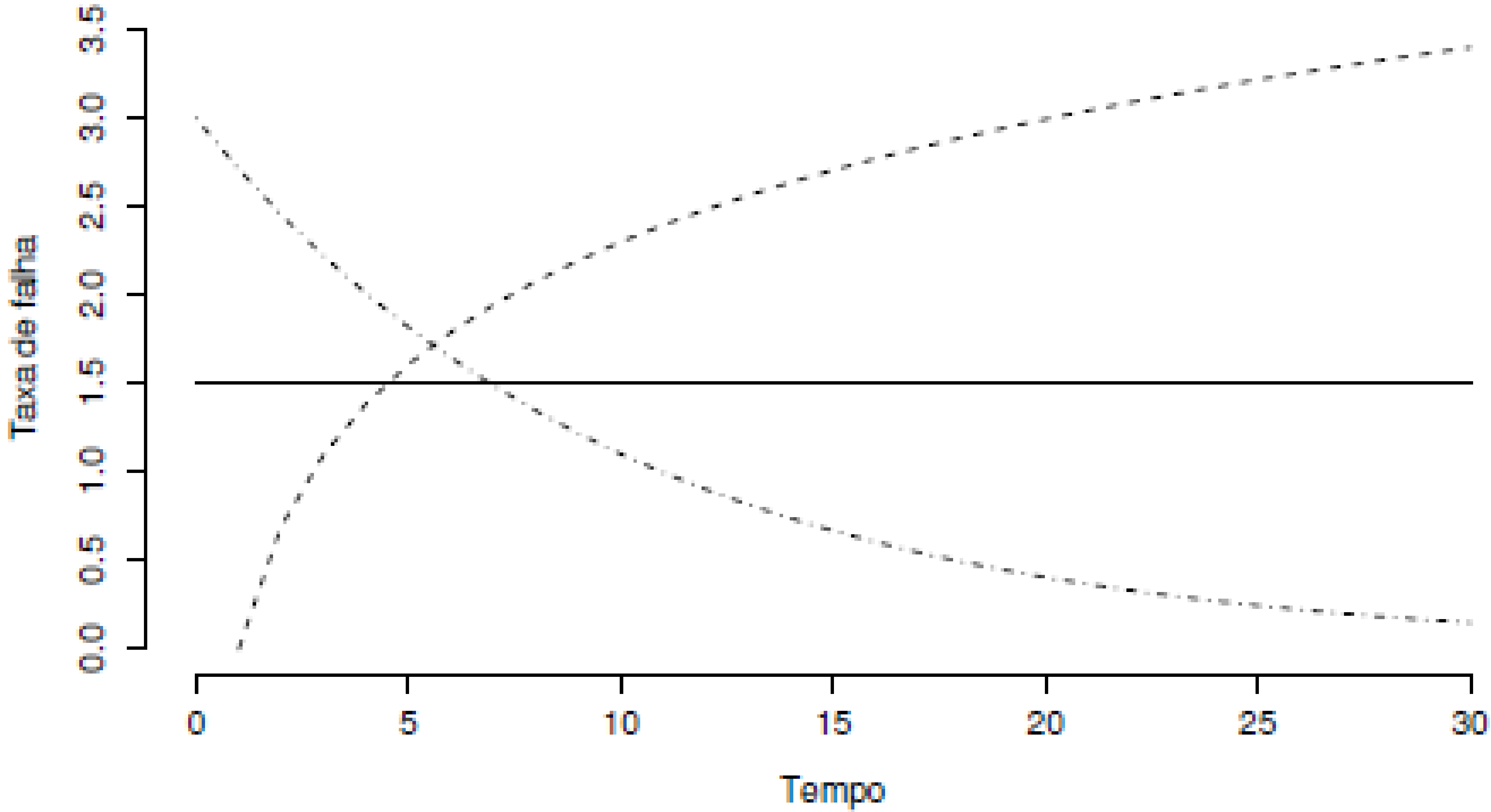
| Intervalo | $R_x(t)$ | $N_x(t)$ | $\Delta_x$ | $\hat{h}_x(t)$ |
|---|---|---|---|---|
| (0,3] | 32 | 1 | 3 | 0.01042 |
| (3,18] | 31 | 1 | 15 | 0.00215 |
| (18,29] | 30 | 1 | 11 | 0.00303 |
| (29,54] | 29 | 1 | 25 | 0.00138 |
| (54,60] | 28 | 1 | 6 | 0.00595 |
| (60,84] | 27 | 1 | 24 | 0.00154 |
| (84,110] | 26 | 1 | 26 | 0.00148 |
| (110,112] | 25 | 1 | 2 | 0.02000 |
| (112,116] | 24 | 1 | 4 | 0.01042 |
| (116,123] | 23 | 1 | 7 | 0.00621 |
| (123,134] | 22 | 1 | 11 | 0.00413 |
| (134,145] | 21 | 1 | 11 | 0.00433 |

| **Intervalo** | $R_x(t)$ | $N_x(t)$ | $\Delta_x$ | $\hat{h}_x(t)$ |
|---|---|---|---|---|
| (145,151] | 20 | 2 | 6 | 0.01667 |
| (151,158] | 18 | 1 | 7 | 0.00794 |
| (158,173] | 17 | 1 | 15 | 0.00392 |

Tabela: Estimativa da função de taxa de falha para os 32 pacientes com AIDS.

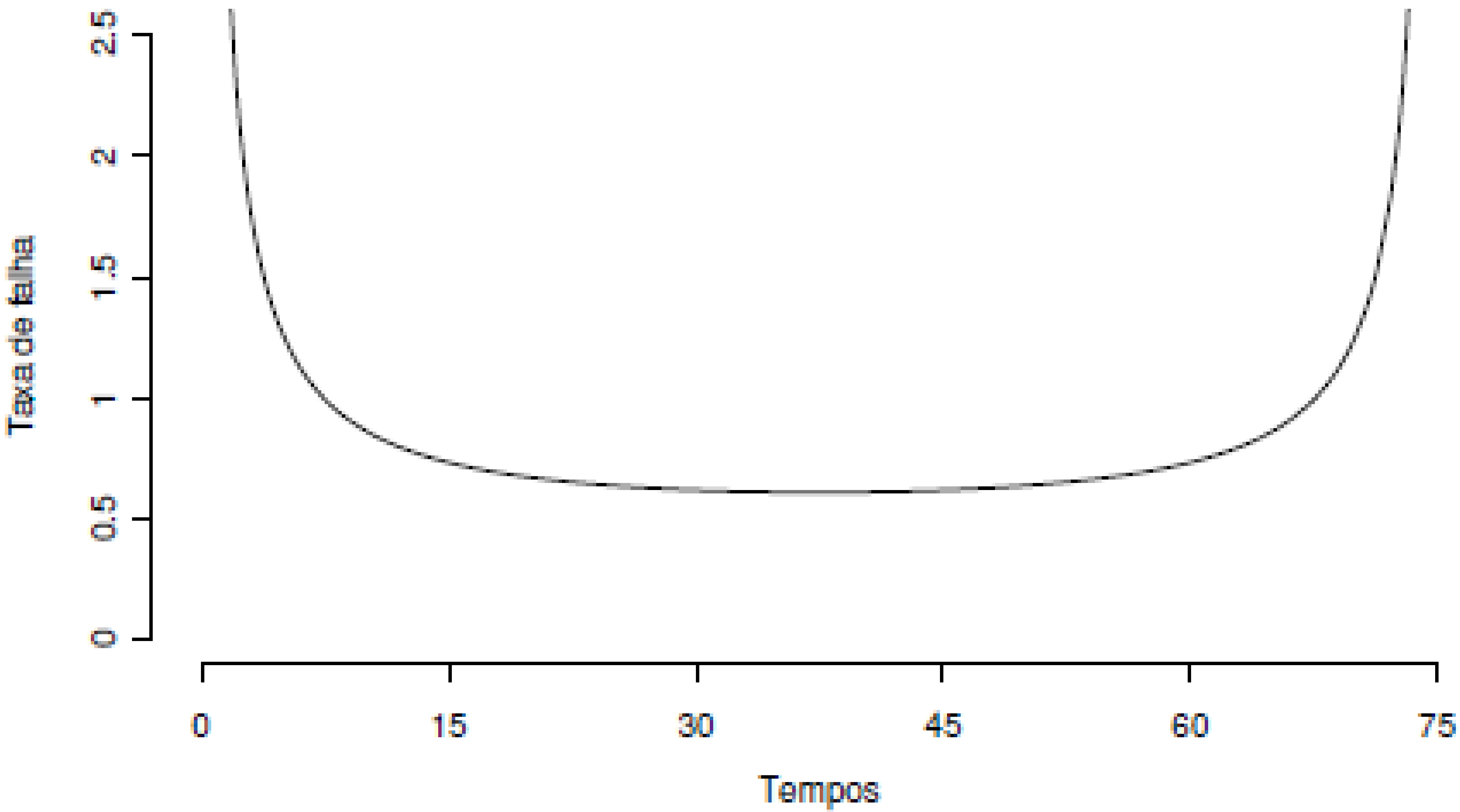
# Especificando o Tempo de Sobrevivência

## Comportamentos da função de taxa de falha



# Especificando o Tempo de Sobrevivência

## Comportamentos da função de taxa de falha



# Especificando o Tempo de Sobrevivência

## Função taxa de falha acumulada

- Qual o risco de um paciente com aids vir a óbito no primeiro ano após o diagnóstico?
- Qual é o risco dele vir a óbito nos primeiros 2 anos?

$H(t)$: função de taxa de falha (risco) acumulada.

Mede o risco de ocorrência do evento até o tempo $t$.

É a soma (integral) de todos os riscos em todos os tempos até o tempo $t$

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

Também é uma taxa, logo não está restrita ao intervalo $[0; 1]$.

# Especificando o Tempo de Sobrevivência

## Estimando $H(t)$

$$\hat{H}_x(t) = \sum_{i=1}^{x-1} \hat{h}_x(t) \times \Delta_x .$$

- O risco acumulado até o tempo $t$ é igual a:
  - o risco acumulado até o tempo $t-1$ mais
  - o risco instantâneo do período anterior vezes o intervalo de tempo até $t$.

# Especificando o Tempo de Sobrevivência

## Estimando $H(t)$

| Intervalo | $R_x(t)$ | $N_x(t)$ | $\Delta_x$ | $\hat{h}_x(t)$ | $\hat{H}_x(t)$ |
|---|---|---|---|---|---|
| (0,3] | 32 | 1 | 3 | 0.01042 | 0.00000 |
| (3,18] | 31 | 1 | 15 | 0.00215 | 0.03125 |
| (18,29] | 30 | 1 | 11 | 0.00303 | 0.06351 |
| (29,54] | 29 | 1 | 25 | 0.00138 | 0.09684 |
| (54,60] | 28 | 1 | 6 | 0.00595 | 0.13132 |
| (60,84] | 27 | 1 | 24 | 0.00154 | 0.16704 |
| (84,110] | 26 | 1 | 26 | 0.00148 | 0.20408 |
| (110,112] | 25 | 1 | 2 | 0.02000 | 0.24254 |
| (112,116] | 24 | 1 | 4 | 0.01042 | 0.28254 |
| (116,123] | 23 | 1 | 7 | 0.00621 | 0.32420 |
| (123,134] | 22 | 1 | 11 | 0.00413 | 0.36768 |
| (134,145] | 21 | 1 | 11 | 0.00433 | 0.41314 |

| **Intervalo** | $R_x(t)$ | $N_x(t)$ | $\Delta_x$ | $\hat{h}_x(t)$ | $\hat{H}_x(t)$ |
|---|---|---|---|---|---|
| (145,151] | 20 | 2 | 6 | 0.01667 | 0.46076 |
| (151,158] | 18 | 1 | 7 | 0.00794 | 0.56076 |
| (158,173] | 17 | 1 | 15 | 0.00392 | 0.61631 |

Tabela: Estimativa da função de taxa acumulada para os 32 pacientes com AIDS.

# Especificando o Tempo de Sobrevivência

## Tempo médio e médio 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,

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

Já a vida média residual é definida condicional a um certo tempo de vida $t$. Ou seja, para indivíduos com idade $t$ esta quantidade mede o tempo médio restante de vida. Isto é,

$$vmr(t) = \frac{\int_t^\infty S(u)du}{S(t)}.$$

# Especificando o Tempo de Sobrevivência

- A variável aleatória não-negativa $T$, que representa o tempo de evento, é usualmente especificada em análise de sobrevivência pela:
  - sua função de sobrevivência.
  - ou pela função de taxa de falha (risco).
- Estas duas funções, e funções relacionadas, que são extensivamente usadas na análise de dados de sobrevivência são apresentadas a seguir.

# Especificando o Tempo de Sobrevivência

## Relações entre funções

$$S(t) = 1 - F(t)$$

$$h(t) = -\frac{d \ln S(t)}{dt}$$

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

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

$$H(t) = -\ln(S(t))$$

$$S(t) = \exp(-H(t))$$

# Especificando o Tempo de Sobrevivência

## No R

- O pacote mais usual de análise de sobrevivência no `R` é o `survival`.
- Ele contém uma grande quantidade de métodos para o assunto.

A função `Surv` serve para definir um objeto do tipo de sobrevivência, para duas notações:

`Surv(tempo, status)` `Surv(inicio, final, status)`

Vamos ver um exemplo no `R`.

# Especificando o Tempo de Sobrevivência

## No R

- O pacote mais usual de análise de sobrevivência no R é o *survival*.
- Ele contém uma grande quantidade de métodos para o assunto.

A função `Surv` serve para definir um objeto do tipo de sobrevivência, para duas notações:

- **`Surv(tempo, status)`**
- **`Surv(inicio, final, status)`**

Vamos ver um exemplo no R.

# Estimação não-paramétrica

# Introdução

- Introdução
- Kaplan-Meier
- Nelson-Aalen
- Intervalos de confiança
- Tempo Mediano de sobrevivência
- Kaplan-Meier com estratificação
- Teste de Log-Rank
- Teste de Peto

# Introdução

Duas formas não paramétricas de estimação das funções de sobrevivência:

- **Kaplan-Meier** – $S(t)$
- **Nelson-Aalen** – $H(t)$

Ambas técnicas envolvem censura.

Sem suposições sobre a distribuição do tempo.

# Estimação não-paramétrica

## Estimador de Kaplan-Meier

- O estimador não-paramétrico de Kaplan-Meier, proposto por Kaplan e Meier (1958), para estimar a função de sobrevivência.
- A probabilidade de sobrevida até o tempo $t$ é estimada considerando que a sobrevivência até cada tempo é independente da sobrevivência até outros tempos.
- A probabilidade de chegar até o tempo $t$ é o produto da probabilidade de chegar até cada um dos tempos anteriores.
- Estimador produto (ou estimador limite produto)

# Estimação não-paramétrica

## Estimador de Kaplan-Meier

- Sejam $t_1 < t_2 < \cdots < t_m$ os $m$ tempos onde ocorreram os eventos;
- $R(t_j)$ é o total de pessoas a risco no tempo $t_j$.
- $\Delta N(t_j)$ é o número de eventos ocorridos precisamente em $t_j$.
- Para os $m$ tempos $t_j$ em que ocorre um evento, a probabilidade de sobrevivência será estimada pelo número dos que sobreviveram até aquele tempo $(R(t_j) - \Delta N(t_j))$ sobre os que estavam em risco naquele tempo $(R(t_j))$.
- Como os eventos são independentes, $S(t)$ é o produto das probabilidades de sobrevivência a cada tempo $t_j \leq t$.

# Estimação não-paramétrica

## Estimador de Kaplan-Meier

Dessa forma, temos a seguinte expressão:

$$\hat{S}_{KM}(t) = \frac{R(t_1) - \Delta N(t_1)}{R(t_1)} \times \frac{R(t_2) - \Delta N(t_2)}{R(t_2)} \times \cdots \times \frac{R(t_m) - \Delta N(t_m)}{R(t_m)}$$

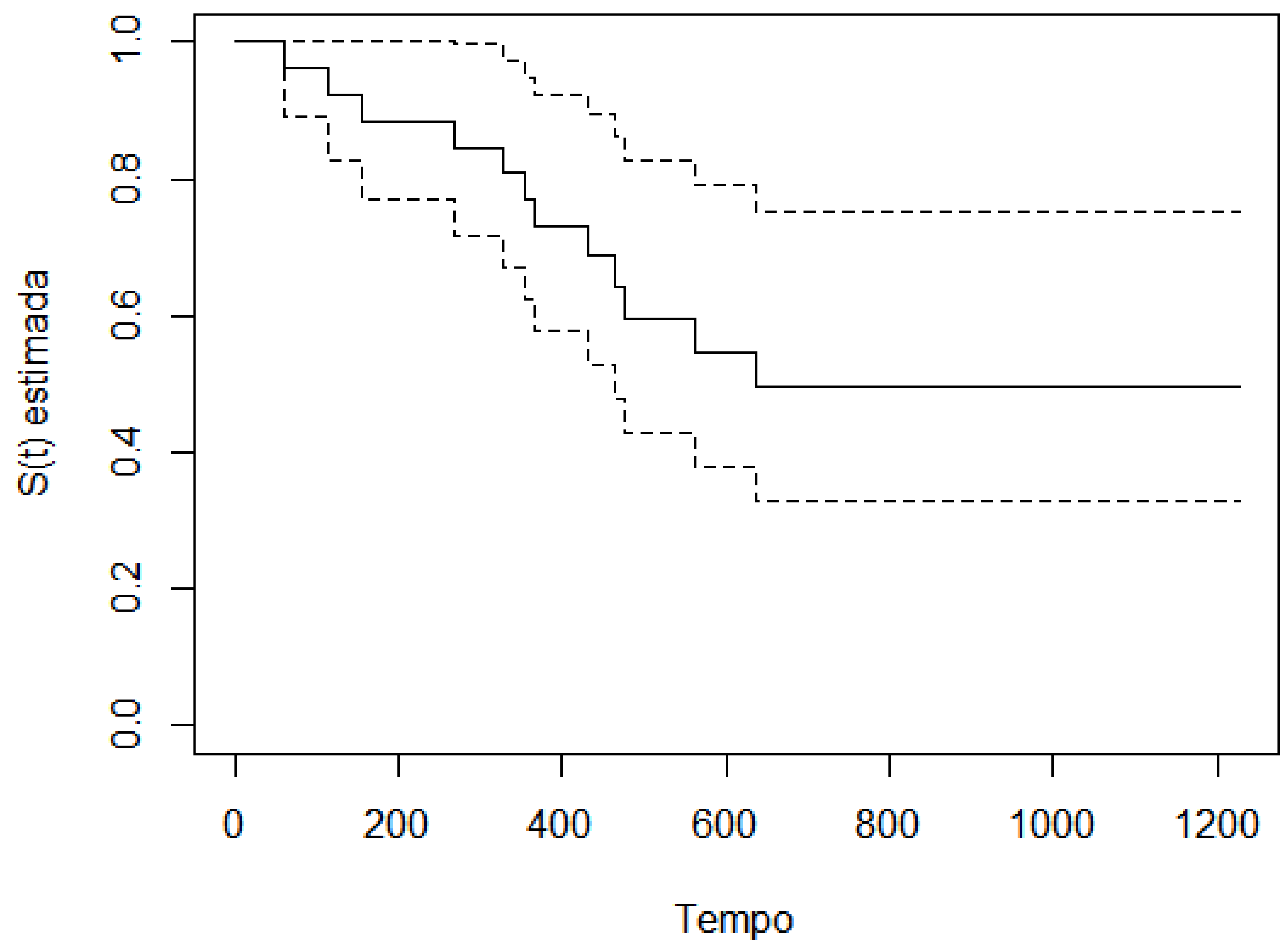
ou simplesmente

$$\hat{S}_{KM}(t) = \prod_{j:t_j \leq t} \frac{R(t_j) - \Delta N(t_j)}{R(t_j)}$$

# Estimação não-paramétrica

## Estimador de Kaplan-Meier

No R, o estimador de Kaplan-Meier está implementado na função `survfit`, que deve ser aplicada em um objeto `Surv`.



# Estimação não-paramétrica

## Estimador de Kaplan-Meier

Função de taxa de falha Acumulada

$$\hat{H}_{KM}(t) = \ln \hat{S}_{KM}(t)$$

Logo, pode-se estimar qualquer das funções.

No `R`, basta usar o argumento `cumhaz=TRUE` na funcao `plot`.